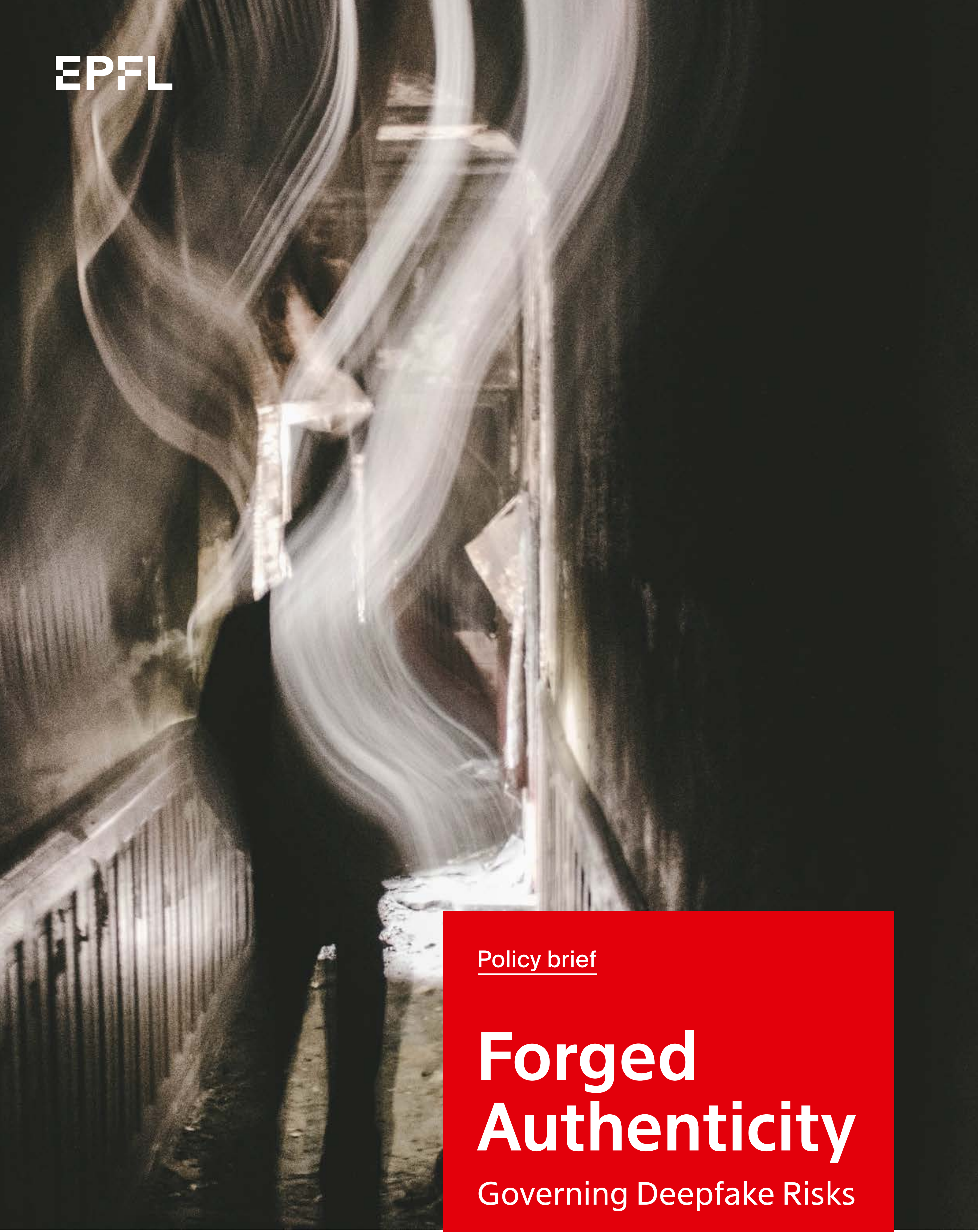


Policy brief

Forged Authenticity

Governing Deepfake Risks

International Risk
Governance Center

Authorisation to reproduce material in this publication is granted, subject to full acknowledgement. It should be cited as: Collins, Aengus. (2019). *Forged Authenticity: Governing Deepfake Risks*. Lausanne: EPFL International Risk Governance Center.
DOI: 10.5075/epfl-irgc-273296

This publication can be downloaded from irgc.epfl.ch.
For any questions, please contact irgc@epfl.ch.

Photos from Unsplash: cover by Echo Grid, page 04 by Simon King, page 08 by Alfaz Sayed, page 12 by Faris Mohammed, and page 14 by Valentina Conde.

Forged Authenticity

Governing Deepfake Risks

Executive summary

The phenomenon of deepfake digital content—fabricated images, video, audio or text created using machine-learning tools—is advancing rapidly. For video in particular the production of increasingly sophisticated fabricated content is becoming exponentially quicker, easier and cheaper. What is true of many emerging technologies is particularly true of deepfakes: the pace of change is running ahead of our ability to understand the implications and respond to them.

In September 2019, the EPFL International Risk Governance Center (IRGC) convened a group of experts to consider the risk-governance challenges posed by deepfakes. The group comprised individuals drawn from a diverse range of fields: policy, law, technology, academia, the media and business. This report builds on the proceedings of that workshop to highlight the potential risks posed by deepfakes and to suggest a range of potential responses.

What are deepfakes?

There are four main technological factors that contribute to the deepfake phenomenon. Three of these relate to the process of using machine learning to create a deepfake: (i) machine-learning algorithms; (ii) the computing power needed to execute the algorithms; and (iii) the datasets needed to train the algorithms. Internet platform technologies represent the fourth factor—they are central to the dissemination of much deepfake content and so are central to deepfake governance.

The manipulation of digital content is not a new phenomenon, but the application of machine learning to the creation of deepfakes has had three important effects. It has radically altered the quality of output that is routinely achievable. It has slashed the resources required to produce hyper-realistic fake artefacts, enabling their production at previously unimaginable scales. And it has “democratized” the process through the distribution of user-friendly software tools and paid-for services.

What are the main risks?

As you will see in the summary table on page 11, we highlight three key impacts of deepfakes: reputational damage, financial damage, and manipulation of decision-making processes. And we plot these against three different levels at which these impacts can be felt: individual, organizational and societal. More work is needed to build an evidence base demonstrating which areas of life are most vulnerable to disruption by deepfakes.

In order to prioritize among various deepfake risks—or instance, to decide where governance responses are most needed—it is necessary to consider the following three dimensions: severity (the level of harm caused by the deepfake), scale (how widespread the harm is) and resilience (the ability of the “target” to withstand the impact). We suggest that there is a prima facie case for prioritizing responses to deepfakes that cause intense harm to individuals or that contribute to systemic societal risks such as the erosion of trust and truth.

What can be done?

We present a total of fifteen potential responses to the deepfake phenomenon, which are summarized below. They are offered not as a definitive response to the deepfake phenomenon, but with a view to stimulating further research and debate in this area.

Risk management

Granular assessments	More detailed work to assess the potential impact of deepfakes in specific domains is needed
Incident recording	We suggest a two-stage process that would build on reporting systems that are already in place for other purposes

Technology

Detection	Continued research into technologies to distinguish between authentic and fabricated digital content
Provenance	Techniques designed to verify the origin and integrity of digital artefacts, such as trusted-hardware schemes or ways of preserving metadata
Image rights and control	Greater control for individuals over digital content that relates to them, including potential "takedown" rights
Digital corroboration	The use of multiple independent data sources, analogous to the familiar process of corroborating eye-witness testimony
Secure digital processes	A greater focus on authentication and verification to make digital communication less vulnerable to deepfakes
Platform nudges	Interventions to influence the way people – and algorithms – share digital content

Law and regulation

Awareness-raising	More should be done to build an understanding of deepfakes throughout the legal system
Legal guidance	Clarification of the ways in which existing legal frameworks – such as the EU's GDPR for example – apply to deepfakes
Hard law	There is a strong case for legal restrictions where harm can be clearly delineated, even if identifying and prosecuting culprits may be difficult
Penalties	The persistent nature of some harms involving digital content may require changes in the way they are penalized
Soft law	Various soft-law measures may be easier to agree than new hard law, but they suffer from limited transparency, accountability and effectiveness

Society

Education	Education is not a panacea, but a stronger focus on digital responsibility (among both consumers and developers) would be welcome
Digital governance	Deepfakes prompt wider questions about internet governance, including the role of prevailing incentive structures and business models

Table of Contents

Introduction

Emerging and converging technologies play an increasingly pivotal role in all aspects of modern life, but they challenge many traditional tools of policy and governance. In part this reflects the sheer pace of technological evolution, but it also reflects numerous structural changes in the way the world is organized, such as the deep globalization of economic activity, the increasing importance of intangible cross-border flows (notably of data), and changes in the balance between public and private sectors. Against this backdrop, dealing with the risks associated with new technologies has become increasingly complicated.

At IRGC we focus on the broader concept of risk governance rather than risk management, because it is better at capturing the breadth of the societal challenges involved in dealing with risk in a world characterized by complexity, uncertainty and ambiguity. Governance is not the same as government or regulation. We define it as the totality of actors, rules, traditions and institutions by which authoritative decisions are taken. It is precisely these broad questions related to taking legitimate and authoritative decisions about potential threats that come to the fore in relation to emerging technologies.

In this report we focus on one technology that has emerged particularly rapidly in recent years: the use of machine learning to produce "deepfakes"—increasingly realistic fabricated digital content. The deepfake phenomenon encapsulates in microcosm many of the wider governance difficulties related to new technologies: pace of evolution, declining cost and difficulty of creation, reliance on general-purpose underlying technologies, embeddedness in a deeply interconnected online information ecosystem, regulatory ambiguity, and uncertainty about both the short-term and long-term impacts (both positive and negative).

Each year we hold an interdisciplinary workshop at the Swiss Re Center for Global Dialogue near Zurich. In recent years, the focus of these workshops has been on various aspects of the risks associated with the global digital transformation. In 2019, we focused on deepfakes. We convened a group of interdisciplinary experts (see page 28 for a list of participants) for two days of debate, aimed at clarifying the key issues and plotting some possible routes forward.

This report summarizes and elaborates upon proceedings at the workshop. It does not pretend to be the last word on deepfake risk governance. On the contrary, it is an effort to open up a number of important areas of research and deliberation at an early stage in the maturation of this emerging use of machine learning. After an initial brief discussion of the technology underpinning deepfakes, the report addresses two broad questions: what potential risks do they pose, and what responses (legal, technological, societal) are open to us?

We hope that our suggested answers to these questions will be taken up and pursued by researchers, practitioners and policymakers. We hope also that this work on the governance of deepfake risks might generate useful insights relating to the risk governance of emerging technologies more generally.

BOILER HOUSE

Chapter 1

What are deepfakes?

In this report, we define deepfakes as the fabrication or manipulation of digital content using machine learning technology. The most familiar examples of deepfakes in current circulation are videos, but it is important to note that deepfakes can also take the form of manipulated text, static images and audio files (see the box below).

The manipulation of digital content is not a new phenomenon. But the application of machine learning algorithms to the creation of deepfakes marks an important development, for a number of reasons. First, it has radically altered the quality of output that is routinely achievable. Second, allowing computers rather than individual creative skills to produce hyper-realistic fake artefacts has slashed the resources required, allowing for production and dissemination at previously unimaginable scales. Third, this whole process is being "democratized" through the development of user-friendly software tools and paid-for services that lower or remove the technological barriers to deepfake-creation.

> DEEPFAKE TECHNOLOGIES HAVE RADICALLY ALTERED THE QUALITY OF FABRICATED OUTPUT THAT IS ROUTINELY ACHIEVABLE

There is a temptation to conflate deepfakes with concerns about "fake news", but the two are distinct. Although deepfakes have the potential to play further havoc with reliable reporting of what has taken place, the direct risks they pose potentially run wider and deeper than the disruption of the media landscape. We suggest in Chapter 2 that numerous domains are potentially affected, including individual wellbeing, national security, business and finance, and the judicial system. More fundamental concerns have already been voiced, relating to the health of democracy, society and the international system (Agarwal et al., 2019). Herb Lin has gone as far as drawing dystopian parallels between deepfakes, climate change and nuclear war (Lin, 2019).

Deepfakes in the wild

The first deepfake videos surfaced in 2017, created using algorithms trained on pornographic movies and celebrity images. Pornography remains a key use case for deepfake technologies. One study published in 2019 counted 14,678 deepfake videos on the internet, of which 96% were pornographic (Ajder, Cavalli, Patrini, Giorgio, & Cullen, 2019).

Although the highest profile deepfakes have been videos, other categories have been reported. In one striking example which highlights the increasing sophistication of audio deepfakes, the synthesized voice of a company's chief executive was used to convince finance departments to execute unauthorized cash transactions (BBC, 2019). Meanwhile advances on the machine-learning generation of written content is such that OpenAI, a research lab, delayed by months the release of the full version of their synthetic text generator system, GPT-2, owing to concerns that the system would be misused (OpenAI, 2019).

The most widely circulated deepfakes tend to be parodic in character, involving high-profile individuals, such as Donald Trump, Barack Obama, Mark Zuckerberg and entertainment-sector celebrities. One of the most frequently viewed deepfakes in 2019 saw Arnold Schwarzenegger's face swapped onto the actor Bill Hader when the latter was impersonating the former during a chat-show interview (Ctrl Shift Face, 2019). It is remarkably seamless. This prevalence of parodic deepfakes highlights the fact that not all deepfakes are created with malicious intent.

The entertainment industry provides perhaps the best example of potentially beneficial uses of deepfake technology. In most cases, high-end digital effects work is still done using meticulous, time-consuming modelling techniques. However, 2019 saw the reported use of deepfake techniques in a Hollywood film for the first time (Bradshaw, 2019). A range of other potential positive uses have been suggested, including: voice-synthesis for medical purposes (Chesney & Citron, 2018), digital forensics in criminal investigations (Rothman, 2018), therapeutic applications, and 'reanimation' services whereby the audio or video likeness of a deceased person would be created (and perhaps integrated with a digital assistant).

1. The underlying technology

There are four main technological factors that contribute to the deepfake phenomenon. Three of these relate to the process of using machine learning to create a deepfake: (i) machine-learning algorithms; (ii) the computing power needed to execute the algorithms; and (iii) the datasets needed to train the algorithms. Internet platform technologies such as video services, social media networks, and so on, are the fourth factor. These platforms do not provide the only means of distributing deepfake content – consider, for example, the submission of a fraudulent insurance claim backed by fabricated evidence. However, online platforms are central to the dissemination of most deepfake content – as well as to the technologies that create deepfakes – and so they are at the heart of debates about deepfake governance.

> *THE PROCESS OF CREATING DEEPFAKES IS BEING "DEMOCRATIZED" WITH THE DEVELOPMENT OF USER-FRIENDLY TOOLS AND SERVICES*

Machine-learning algorithms

A key category of algorithms for generating deepfakes is the generative adversarial network, or GAN (Goodfellow, et al., 2014). GANs employ two machine learning models in tandem. First, a deep neural network uses a sample dataset to learn the characteristics of the target phenomenon (i.e., the thing being faked) and generate "fake" samples. The GAN's second model assesses the quality of these fakes by comparing them to the samples in the original dataset. This feedback loop – the generator creating new samples and the discriminator assessing whether they are fake or authentic – is the innovation that makes GANs so efficient in producing convincing deepfakes. The decision of the discriminator is fed back to the generator and, similarly, the generator tells the discriminator whether the submitted samples were actually real or fake. The two models use this reciprocal feedback to improve their performance – the generator tries to produce more realistic samples, and the discriminator tries to get better at identifying fake from authentic samples. It is the pitching of the two learning models as adversaries competing to outdo each other that helps the GAN reach an optimum where the generator produces hyper-realistic fakes that the discriminator can only classify randomly as it cannot distinguish between the fake and real anymore.

Computing power

In addition to the development of powerful new algorithms, the spread of deepfakes also reflects the increasing availability and affordability of computing resources, whether conventional CPUs (central processing units), specialized GPUs (graphics processing units) or high-performance supercomputers. The cost of accessing high-performance computing resources has fallen rapidly, with the advent of cloud computing services having a particularly dramatic impact. This can be illustrated by looking at the resources required to train ResNet-50 (He, Zhang, Ren, & Sun, 2015), a 50-layer deep convolutional neural network, on 1 million images from the ImageNet database, classified into 1000 object categories, such as animals, pencil, keyboard, mouse (Deng, et al., 2009). While it would require an investment of around US$120,000 to purchase a supercomputer capable of training ResNet-50 in a few hours, cloud-based supercomputers can be used to train ResNet-50 within eight minutes at a cost of less than US$50 (Nvidia, 2019) (Google, 2019a). These estimates are based on classification tasks, but the underlying workloads are representative of all machine learning tasks including the deep generative models used in the production of deepfakes.

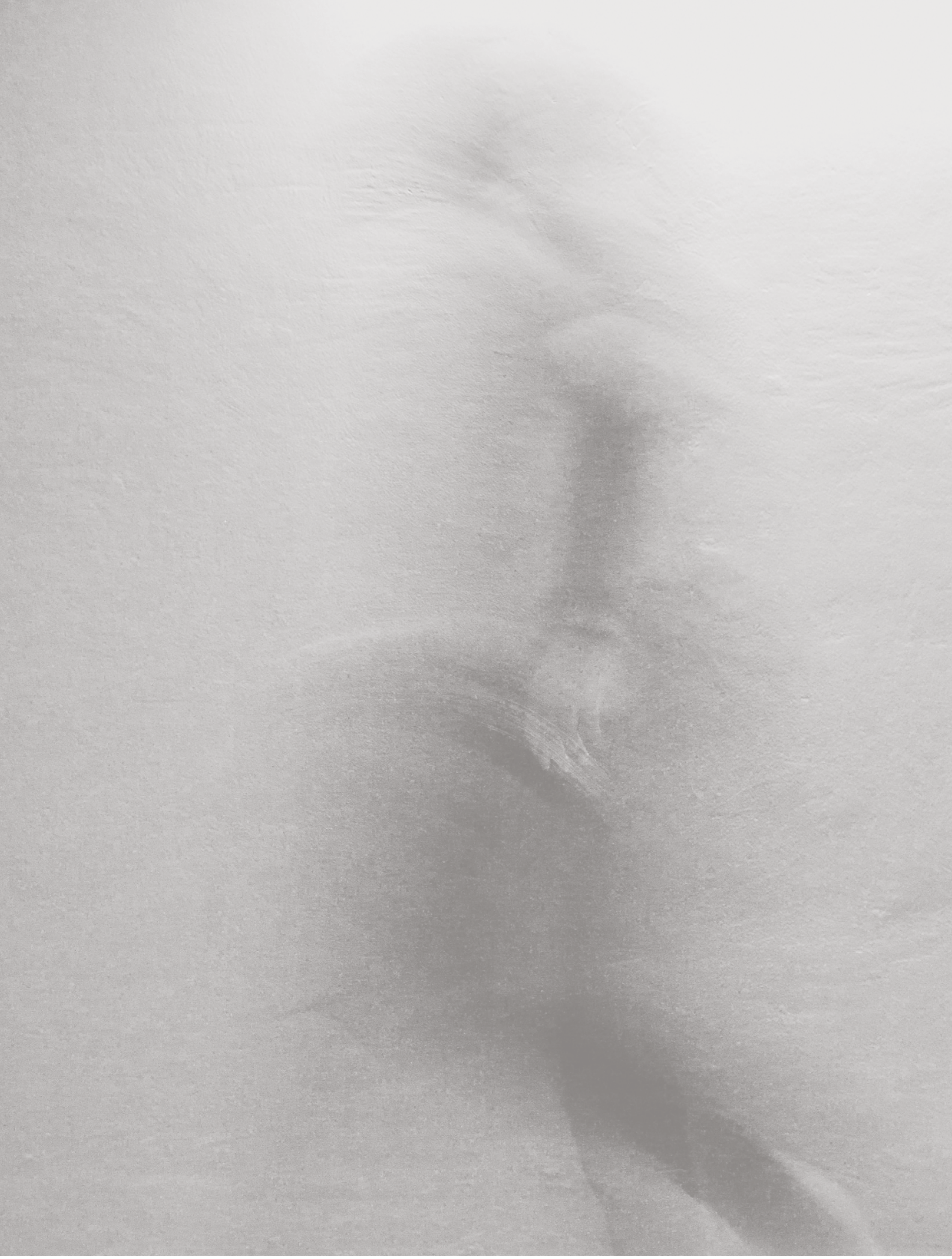
Data requirements

The third technical requirement for generating deepfakes is the training dataset. Conventionally, large volumes of data have been required to learn the features of the target object to generate convincing fakes: the more training data is available, the better the quality of the generated fake. This is one reason that public figures have been among the most prominent targets of deepfakes – the size of their digital footprint (in terms of publicly available images, videos and speech) provides ample training data. However, the proliferation of user-posted images and other online content has greatly increased the digital footprints of non-public figures, while recent advances in machine learning have drastically reduced the volume of data required to create a deepfake of a specific target (Finn, Abbeel, & Levine, 2017). In 2019 it was demonstrated that only a handful of images of the target are needed in order to create deepfake talking-head videos of people (Zakharov, Shysheya, Burkov, & Lempitsky, 2019).

Internet platforms

The risks posed by deepfakes are exacerbated by the power of the internet – and of social media platforms in particular – to disseminate digital content rapidly and widely. As noted above, this isn't to say that all deepfake risks involve the sharing of content via online platforms. However, in many cases the potential harms caused by deepfakes will be proportional to their rapid online proliferation. The digital ecosystem that has evolved over the past decades thrives on such proliferation and the tendency for deepfakes to spread rapidly is heightened by the fact that information with a high novelty or surprise quotient is more likely to be shared (Vosoughi, Roy, & Aral, 2018). The role of the major internet platform companies in this sharing ecosystem is contentious. Their business models incentivize the rapid sharing of digital content, but because the content is generated and propagated by users the companies do not face the same responsibilities as entities deemed to be the publishers of content, such as broadcasters. However, in at least some jurisdictions there is an increasing appetite to regulate how the platforms deal with harmful content (Mullin, 2017).There are also increasing signs that the platforms recognize the tensions and challenges in this area, but as yet there is no consistent approach to false content including deepfakes (see Soft law, page 21).

> *IN MANY CASES THE POTENTIAL HARM CAUSED BY DEEPFAKES WILL BE PROPORTIONAL TO THE EXTENT OF THEIR ONLINE PROLIFERATION*



Chapter 2

What are the main risks?

The relative novelty of the deepfake phenomenon, coupled with the pace at which the underlying technologies are developing, mean that in addition to managing risks that have already begun to crystallize, one of the pressing governance challenges is to anticipate where new risks might emerge. (Arguably this uncertainty in the face of rapid change is a defining feature of risk governance related to emerging technologies in general.[1]) It is notable that some of the individuals who have made significant contributions to the technologies underpinning deepfakes — including Ian Goodfellow, who was instrumental in developing GANs (Giles, 2018) — are also prominent among the voices warning of risks in this area (Knight, 2019). As with other technologies, a key trade-off that arises in the governance of deepfakes is between risk mitigation on the one hand, and the maintenance of incentives for beneficial technological innovation on the other hand.

> *UNCERTAINTY IN THE FACE OF RAPID CHANGE IS A DEFINING FEATURE OF RISK GOVERNANCE RELATED TO EMERGING TECHNOLOGIES IN GENERAL*

[1] For a wider discussion of the malicious use of artificial intelligence, see (Brundage, et al., 2018).

1.

Which areas could be affected?

The first deepfakes to circulate widely were pornographic in character (Cole, 2017) and as noted in the box on page 6, pornographic videos continue to account for the vast majority of documented deepfakes. These typically involve a woman's face being swapped into a pornographic video, raising the prospect of deepfakes as potential instruments of individual intimidation, coercion or defamation. This phenomenon is increasingly referred to as image-based sexual abuse, but still frequently dubbed "revenge porn". In July 2019, the state of Virginia in the US amended its laws to include deepfakes in its prohibition of harassment via the sharing of sexual images (Robertson, 2019).

Deepfake technologies have emerged against a backdrop of worries about "fake news" and disinformation, so a second early concern has been that deepfakes might mark an important intensification of the erosion of norms related to truth and trust. Of particular importance here is the idea of the liar's dividend: the fact that deepfakes exist will allow dishonest actors to claim that authentic digital content that happens to be inconvenient or incriminating is in fact fake. This in turn risks worsening the developing crisis surrounding the integrity and trustworthiness of the information ecosystem and of society more broadly. It also has potential national security implications, and in mid-2019 the Intelligence committee of the US House of Representatives held an open hearing on this issue (US House of Representatives Permanent Select Committee on Intelligence, 2019).[2]

Between these two poles of targeted individual harm and damage on a systemic or societal scale, there is a potentially very large range of harms that could be caused by deepfakes in specific sectors or domains. On the face of it, for example, any decision-making process that relies on the trustworthiness of documentary evidence is potentially vulnerable. In this context, questions have been raised about the potential use of deepfakes to undermine audio and video evidence in court cases (Maras & Alexandrou, 2019); to facilitate fraudulent insurance claims (McMahon, 2018); and to manipulate financial markets (BBC, 2019). At our workshop, numerous other potential risks were cited including:

- Personal reputational damage
- Corporate brand reputational damage
- Fraud and identity theft
- Identity "creation" – deepfakes used to bypass checks designed to prevent the creation of fake online identities or bots
- Media verification of the authenticity of video and audio
- Public opinion manipulation, particularly in polarized or conflictual societies
- Influence operations and other national security threats
- Sowing of uncertainty as to whether or not human rights breaches have taken place
- Undermining the historical record – a digital-age equivalent of having people erased from historical photographs
- A plagiarism-like challenge in education, with papers being created using machine-learning text-synthesis tools.

This list is not intended to be comprehensive and it is not restricted to areas where the use of deepfakes has already been documented. Its purpose is to illustrate the potential range of harmful applications for deepfake technologies, and to prompt further discussion and debate as to where potential vulnerabilities may exist. As has been noted already, a key element of the risk-governance challenge related to a phenomenon as new and fluid as deepfakes lies not in finding solutions to known problems, but in imagining potential instances of those problems that have not yet crystallized, and weighing their significance.

> *ADVERSE SOCIETAL IMPACTS INCLUDE EROSION OF TRUST, THE "LIAR'S DIVIDEND", DIMINISHING SOCIAL COHESION AND THE UNDERMINING OF SHARED STANDARDS OF TRUTH*

[2] For a discussion of the role of social media in state and non-state "influence operations", see (Bonfanti, 2018).

2.

Motivations and impacts

One way of thinking about potential deepfake risks is to consider their likely motivations and impacts, as plotted in the table below. We have suggested three key potential deepfake impacts:

- reputational damage
- financial fraud or extortion
- manipulation of decision-making processes

The boundaries between these categories are not rigid. For example, reputational damage might be caused or threatened not for its own sake, but as a means to fraud or extortion. Similarly, an attempt to manipulate decision-making processes might have unforeseen and unintended financial spillovers.
It is also worth noting that not all deployments of deepfake technologies will necessarily be deliberate attempts to cause targeted harm. There is also an anarchist strand of internet (counter)culture in which transgression is an end in itself, resulting in patterns of “abusive pranksterism” (MacDougald, 2017).

As the table also indicates, adverse impacts from deepfakes can occur on three levels:

- individual
- organizational/institutional (including both public and private sectors)
- national/societal/systemic

One assumption here is that the last of these three categories is somewhat different. Adverse impacts on individuals and organizations will often be direct and intended. While this kind of purposeful impact can occur at a society-wide level—consider a state’s use of deepfakes to disrupt a geopolitical rival—our suggestion is that some of the most corrosive societal impacts from deepfakes are likely to be an unintended result of their growing prevalence. Impacts of this type might include a deepening erosion of trust, the liar’s dividend mentioned above, a diminishing of social cohesion, and the undermining of shared standards of truth (with important consequences for democratic discourse and decision-making).

Issues of trust and truth recur repeatedly in discussion of deepfakes. In part this is because concerns about these values were already elevated when deepfakes first emerged. These worries reflect the fact that interpersonal trust and the idea of a broadly shared and knowable truth are closely woven into the way democratic societies operate. For centuries, one of the justifications for openness as a governing principle for society has been its relationship to truth. If that relationship is called into question, it therefore strikes a democratic nerve. Milton asked the question: “who ever knew Truth put to the worse, in a free and open encounter?” One answer from the 21st century might be: “anyone who has spent time on social media.”

Table 1: Impacts of harmful deepfakes

	Impact		
	Reputational damage	Financial	Manipulation of decision-making
Individual level	• Intimidation / abuse • Defamation	• Identity theft • Phishing-type scams • Extortion	• Attacks on politicians
Organizational level	• Brand damage • Undermining of trust in the organization	• Stock-price manipulation • Insurance fraud	• Fabricated court evidence • Media manipulation • Faked education papers • Attacks on political parties, advocacy groups, etc.
Societal level	• Damage to societal cohesion, norms of trust and truth, etc. • Domestic or foreign electoral manipulation • Deliberate stoking of tension / panic / conflict		

> THERE IS A PRIMA FACIE CASE FOR PRIORITIZING ACTION ON DEEPFAKES THAT CAUSE INTENSE HARM TO INDIVIDUALS OR THAT CONTRIBUTE TO SYSTEMIC SOCIETAL RISKS

3.

Risk-governance priorities

Table 1 suggests how we might categorize the potential impacts associated with deepfakes, but more work is needed to identify and better understand the dynamics of specific risks in specific areas. For this reason, one of the first steps that we suggest in Chapter 3 is for deepfake risk assessments to be conducted in as many sectors or domains of activity as possible, so as to better map the landscape of potential vulnerabilities. In the meantime, however, we can begin to “triage” the various categories of deepfake, with a view to highlighting those areas in which a governance response is likely to be most beneficial. One way of doing this is with a simple framework that considers the following three dimensions of the risks posed by different categories of deepfake:

- Severity: the level of harm caused by the deepfake
- Scale: how widespread the harm is
- Resilience: the ability of the “target” to withstand the impact

If we use these three dimensions to consider the three levels of impact discussed in the previous section—individual, organizational and societal—it suggests a prima facie case for vigilance in relation to the individual and societal impacts of deepfakes. Where an individual is targeted by a deepfake, the impact is potentially severe and long-lasting (for example, where an individual’s face is swapped into a pornographic video which is then uploaded to the internet). Moreover, many individuals may not have the resilience or resources (financial, psychological, etc) to allow them to “bounce back” from a deepfake attack.

The opposite pattern applies with the society-wide category of deepfake impacts. Instead of individual targets being affected in an intense way, here the impact is diffuse to the point that it might cause a deterioration in the functioning of the entire system without necessarily having a sufficiently severe direct effect on enough people or organizations to galvanize

a system-wide response. Resilience to these societal impacts will vary among countries and communities, but if deepfakes cause social institutions such as trust and cohesion to begin to erode, there are no simple policy levers that can be pulled to restore them. (Time may be one of the differentiators here, between the individual and societal impacts of deepfakes. The impact on an individual is likely to be felt swiftly, whereas societies are vulnerable to the cumulative effect—on trust, truth, etc.—if deepfakes proliferate.)

Between these poles of individual and society-wide impacts are the many types of organization that might be affected by deepfakes. As noted before, the boundaries between these various categories are not rigid. Certain institutions have a particularly important societal role. So, for example, the widespread use of deepfakes to doctor criminal evidence would have systemic as well as narrower institutional implications. However, one tentative conclusion is that many organizations, particularly in the private sector, are likely to have existing resources and processes in place that could absorb deepfake impacts.

> *IF DEEPFAKES CAUSE SOCIAL INSTITUTIONS SUCH AS TRUST AND COHESION TO ERODE, THERE ARE NO SIMPLE POLICY LEVERS TO RESTORE THEM*

As an illustration, one hypothetical considered at our workshop was the use of deepfake images or videos to make fraudulent insurance claims. The central point here is that insurance companies already have sophisticated processes in place to deal with fraud. The introduction of deepfakes into the claims pipeline might cause initial problems for those anti-fraud processes, but the challenge would likely be one of recalibration rather than reinvention. This might lead to lower margins—for example, because of the introduction of deepfake-detection software, or the lowering of the threshold that triggers a fraud investigation—which would offset some of the efficiency gains from digitalization that most organizations have already achieved. However, unless one company was the victim of a disproportionate volume of deepfaked claims, the broad competitive landscape would be unchanged. The situation may be different in cases of reputational damage, where an isolated organization might be singled out for attack. But here too many organizations will have crisis-management and brand-management systems in place that could be adapted to respond to a deepfake attack.

Chapter 3

What can be done?

In this chapter, we suggest a range of steps that might be taken in order to understand the risks posed by deepfakes and how to respond to them. Different contexts or domains will likely require different mixes of responses. Our recommendations are offered with a view to stimulating further research and debate in an area that is new but rapidly evolving. We have grouped them into four categories: risk management, technology, legal and societal. Our hope is that this initial survey of governance responses to the deepfake phenomenon will spur more detailed work in the various fields it covers.

1.

Risk management

Two standard risk management approaches could make an important contribution to developing a better understanding of the nature and scale of the risks posed by deepfakes: risk assessments and incident reporting.

Granular assessments

The suggestion here is that a broad survey of deepfake risks, such as this report seeks to provide, needs to be followed by more detailed work to assess the impact that deepfakes might have across a range of different contexts: different countries, different economic sectors, different demographic groups, and so on. This phenomenon is so new and its rate of evolution now so rapid (particularly in the field of video), that more granular research is needed to assess the potential character and severity of deepfake risks in different contexts. In Chapter 2 we set out a simple framework for thinking about where deepfakes might have the greatest impact, but it is beyond the scope of this report to provide a detailed domain-by-domain assessment. We hope that other researchers and organizations with greater domain-specific knowledge will take up this challenge. This would enable the development of a better overall indication of the scope and depth of risks in this area. This process would not need to begin from scratch. There are numerous risk-assessment and horizon-scanning exercises undertaken each year across public sector, private sector and civil society organizations, which could be adapted to include deepfakes in their list of potential technological disruptors.

> *MORE DETAILED WORK IS NEEDED TO ASSESS THE IMPACT OF DEEPFAKES ACROSS A RANGE OF SPECIFIC DOMAINS*

Incident recording

Deepfake technologies are evolving rapidly, but it is not currently clear whether or how or where that might translate into a sharp escalation of the harm being caused. We risk missing such developments if we are not gathering as much data as possible about the deepfake incidents that are currently occurring. Some work in this area has begun, such as an annual study of the number of deepfake videos that can be found on the internet (Ajder et al., 2019). However, a more holistic approach is needed, with a view to capturing other kinds of deepfake (audio, text, etc) as well as incidents that may not involve the deepfake being publicly searchable. Our suggestion is for a two-stage process that leverages reporting processes that are already in place.

The first stage would focus on incident recording rather than reporting and would seek to ensure that any involvement of a deepfake is logged when incidents of various relevant kinds are being recorded and reported. For example, if an insurance claim is rejected because of the use of a deepfaked image, are there systems in place to register the role of the deepfake? Similarly, if cases involving the use of sexual images to harass individuals are being recorded (for example, by the police), is a distinction being logged when the incident involves a deepfake image rather than an authentic one? The second stage would involve collating incident data from as many sources as possible. The idea here would be to create a central hub or clearing house to aggregate existing reports that flag the involvement of a deepfake, rather than invent a new reporting system through which all deepfakes would have to go at the outset. (If this exercise makes it clear that the scale and severity of deepfake risks are sufficient, then such a standalone deepfake reporting system could be advocated.) Each country could designate a suitable body to fulfil this “deepfake hub” role — perhaps an existing official entity, perhaps a trusted and neutral NGO — and authorize it to receive suitably anonymized data about deepfake incidents that have been recorded by existing reporting schemes (for example, for cyber-attacks or fraud) or by the police.

2.

Technology

Technological responses to deepfakes may not be able to succeed in a definitive sense, as adversaries with sufficient skills, resources and determination will be able to match whatever defences are put in place. But to say that a technological arms race over deepfakes cannot be won outright is not to say that it is not worth pursuing up to a point. If there are measures that can be taken that make it more difficult and costly to use deepfakes harmfully, then implementing them may well push all but the most sophisticated adversaries out of the field, thereby greatly reducing the level of harm being done (Farid, 2018).

Detection

Deepfake-detection techniques remain at the heart of current technological efforts to combat deepfakes. They are also the clearest example of the arms-race difficulties mentioned above. For any set of detection-based defences, there will always be ways to train deepfake algorithms to outwit them. This dynamic is even built into the generator-discriminator loop within the GANs that create deepfakes. As new digital-forensics techniques for detecting deepfakes are developed and disseminated, they can be incorporated into the GAN, pushing the algorithm to make even better deepfakes (Carlini & Wagner, 2017). For example, at one point the lack of realistic eye-blinking was a way of distinguishing fake from authentic videos. This weakness was quickly incorporated into deepfake-producing GANs, leading to new videos with realistic eye-blinking, and rendering the technique redundant for forensic purposes. However, ongoing investment in research in this area should keep deepfake-detection capabilities ahead of all but the most state-of-the-art adversaries.

An additional measure that has been suggested as a way of keeping ahead of sophisticated adversaries is to place greater limits on the dissemination of cutting-edge deepfake-detection research, so that advances made by digital forensics researchers are not immediately outpaced by their adversaries (Ovadya & Whittlestone, 2019). Some researchers are proactively doing this (Farid, 2018), but there may be scope for developing community-wide norms or codes of conduct. Aviv Ovadya (2019) has also mooted the creation of an International Media Authenticity Council, which would have privileged access to non-public detection technologies.

There is another governance question that arises in the context of these detection techniques, relating to how and where they should be used. There is a potentially important distinction here between (i) ex post processes designed to establish what has happened and attribute responsibility for any harm caused, and (ii) ex ante processes designed to prevent the dissemination of deepfakes in the first place. Such ex ante “filtering out” of deepfakes might be uncontroversial in some contexts, such as ruling fabricated evidence inadmissible in court. But it might be highly contentious in others. For example, prohibiting the upload of deepfakes to social media and similar platforms raises particular issues in countries with strong protections for freedom of expression, notably including the United States. (Blitz, 2018).

One promising development in relation to deepfake-detection relates to collaborative efforts between different stakeholders to create pooled resources that are more than the sum of their parts. Training machine-learning models to detect manipulated content requires access to large sets of deepfakes. Recently, researchers at the Technical University of Munich (TUM) worked with Google and Jigsaw to release FaceForensics++, a forensics dataset containing over 1.8 million visual deepfakes (Rossler, et al., 2019) (Google, 2019b).

While detection in the context of deepfakes typically refers to efforts to identify artefacts that have already been created, there may be other possibilities. One

> *FOR ANY SET OF DETECTION-BASED DEFENCES, THERE WILL ALWAYS BE WAYS TO TRAIN DEEPFAKE ALGORITHMS TO OUTWIT THEM*

suggestion for further exploration is the possibility of companies that provide cloud-computing services being able to detect—for example, from a distinctive metadata signature—that their processors are being used for the large-scale creation of deepfakes.

Provenance

This refers to techniques designed to verify the origin and integrity of a digital artefact, in order to establish that it has not been fabricated or manipulated. It covers a wide range of potential approaches. One would be a “trusted hardware” scheme, whereby cameras and other devices would produce a digital signature for each image or video created, securely storing it so that there is a trusted record of the original, authentic image against which subsequent versions of the image can be verified. A second avenue of provenance research is in the area of metadata preservation[3]. One of the obstacles to image or video verification is that when content is uploaded to internet platforms, it is transformed in ways that strip it of its original metadata. Work is under way—for example, in the JPEG committee—to allow transformations that would preserve metadata. A third area of provenance research focuses on analyzing artefacts to see whether they reveal traces of training datasets used to create deepfakes (Zhou, Li, & Tian, 2017) (Moreira, et al., 2018). Other research searches open-access and proprietary media collections for close matches of the artefact in question, on the basis that deepfakes are often altered versions of publicly available images or video (Apostolidis, Apostolidis, Patras, & Mezaris, 2019). As with detection, no provenance solution is going to be perfect, but it could raise the bar significantly, and might allow the development of new tools or conventions for online content, such as labels to distinguish videos that have gone through a verification process from those about which nothing is known.

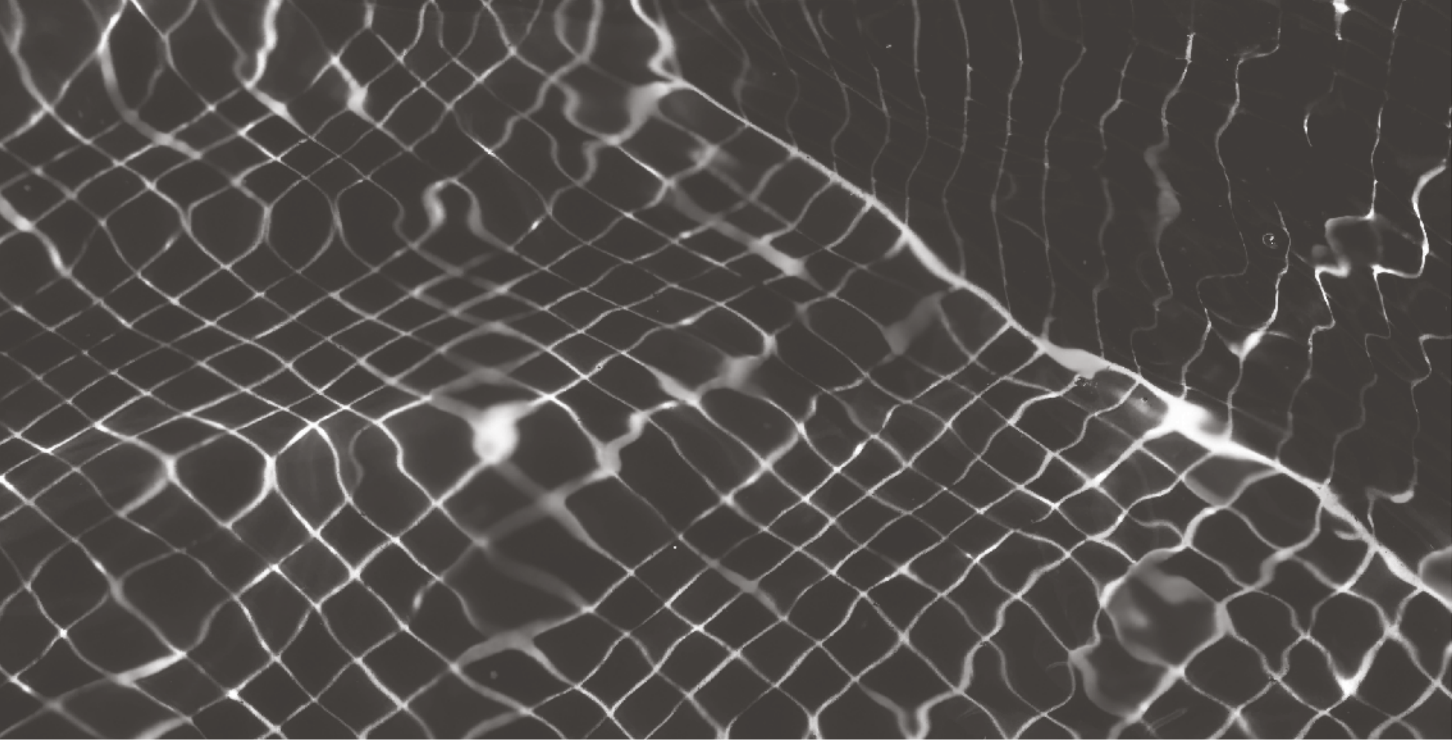
Image rights and control

While much of the focus of technological responses to deepfakes focuses on the processes of creation and dissemination, there may be steps that can be taken to give individuals greater control over online content that relates to them. There are analogies here with copyright protection and the systems that internet platforms have put in place for flagging and taking down material that is found to be in breach. Could similar rights and processes be extended so that, for example, an individual could demand that images of them are taken down? Or could that image control be implemented as a technological default, so that images uploaded to internet platforms are processed so that the faces in them are blurred until the person depicted has consented to de-blurring? This would not be straightforward. It would entail complicated trade-offs involving privacy and freedom of expression (for example, relating to truthful representation, public figures and celebrity parodies). However, advances in cryptographic technology mean that these trade-offs do not need to be zero-sum.

Digital corroboration

The idea of corroboration is familiar in the context of eye-witness testimony. In contexts such as court proceedings or journalism, there is a strong pressure to provide the testimony of multiple witnesses or sources who will attest to the same version of events. In the context of digital content, more could be done to find analogies for this process of corroboration. This would entail using evidence from multiple digital sources—for example, CCTV cameras, mobile phone cameras, digital assistants, car sensors—to help establish a baseline of truth against which a potential deepfake could be assessed. For example, if a doctored image or video is used to substantiate an insurance claim for damage to a car or house, are there additional sources of digital evidence that could provide a comparable record of the same place at the same time? This digital corroboration approach is not without problems. For example, the devices would need to be independent to provide reliable corroboration, which might not be the case if, say, several CCTV cameras were used but all were under the control of a single person or organization. Nevertheless, this is an underdeveloped area, where further work would be helpful.

[3] For a discussion of the use EXIF metadata, see (Huh, Liu, Owens, & Efros, 2018). Iuliani et al (2019) demonstrate how variations in the camera manufacturer’s implementation of the MP4 video standard can be used to identify manipulations in image content. Similarly, Bunk et al (2017) demonstrate how other software manipulations can be detected using the meta data generated during the image encoding process.



Secure digital processes

This category refers to making improved use of existing technologies to provide people with better tools for establishing when trust is and is not warranted. One of the deepfake examples highlighted earlier involved the use of a manipulated voice instruction within a company designed to trigger a fraudulent money transfer. It highlights a general tendency for people to be overly dependent on, or overly trusting of, voice and video authentication. Part of the solution to a case like that might be training people to be careful of the instinct to trust what they hear over the phone. (Even recent innovations such as comparing the voice of someone on the phone to a recorded sample of their voice will be increasingly vulnerable when the quality of audio deepfakes improves.) However, there are other technological possibilities. One example would be end-to-end digital authentication for communication between key parties (such as the CEO and CFO in a company): the device would tell the CFO that he or she can be certain that it is the CEO's device on the other end of the line and not an unauthenticated number masquerading as the CEO. This would allow for a baseline of trust to be established before a word has been spoken.

> *"TECHNICALLY, THE TWEAK WAS SMALL, BUT POLITICALLY ITS EFFECT WAS ENORMOUS"*

Platform nudges

Later in this chapter we suggest the need for far-reaching discussions on the way in which the overall online ecosystem is governed. The role of the internet platforms would be a major element of such discussions. However, at a much more granular level there may be modifications that the platforms could make in order to alter the way harmful deepfakes can circulate on the internet. One suggestion at our workshop was for an across-the-board reduction in the speed at which content can be shared across internet platforms, the idea being that what is lost in terms of immediacy would be more than offset by the opportunity for greater reflection about what is being shared. Another idea is the use of "nudges" that would prompt users to reconsider before forwarding content that may be manipulated. This is an approach that is already used in other contexts. Since 2017 Instagram has been using a feature that can detect when a user is about to send a bullying message, prompting a light warning to them that they may want to reconsider. The feature does not prevent the user from proceeding to send the message, but is intended to encourage them to make that decision for themselves (Ravenscraft, 2019). A more thorough attempt to alter the dynamics of online sharing has been attempted in Taiwan, where a platform designed to tackle polarizing political issues was engineered so as to promote the visibility of messages that received broad support across opposing groups, while messages that divided the groups were demoted by the algorithm. "Technically, the tweak was small, but politically its effect was enormous" (Miller, 2019).

3.

Law and regulation

As in the technological field, there is no prospect of a simple legal fix for deepfakes. Complications attend every stage of the legal process, starting with the fact that multiple legal principles and instruments may apply to aspects or applications of deepfake technologies. A non-comprehensive list might include: defamation, breach of privacy, fraud, identity theft, electoral crime, image-based sexual abuse ("revenge porn"), harassment, extortion, appropriation of personality, misinformation and child pornography. Nevertheless, there is a range of legal and regulatory steps within and between countries which may have some traction with respect to deepfakes and which ought to be further explored.

Awareness-raising

The first task for improving the way deepfakes are dealt with by the law is to foster greater knowledge and understanding throughout the legal system. There are two elements to this. The first is simple awareness-raising. If lawyers, judges and court officials are not informed about what deepfakes are, then the legal treatment of deepfakes is necessarily going to be deficient. This applies directly to the integrity of the criminal justice system. One of the potential examples of malicious use of deepfakes we cited in Chapter 2 was the fabrication or manipulation of evidence in court. Anecdotal evidence suggests that, in the US at least, there is little awareness within court systems that deepfakes pose this kind of threat. The digital manipulation of evidence long predates deepfakes (Paul, 2008). But deepfakes bring the potential for both higher quality and higher volumes, and so require vigilance. This is particularly true because of the impact that video or other documentary evidence can have during court proceedings. People tend to trust what they see. They are also particularly likely to trust that evidence presented in a court has been vetted and verified. That is not necessarily the case, however. In the US for example, there has been a move in the past decade towards self-certification of evidence because of the costs involved in having it authenticated by experts (Grimm, Capra, & Joseph, 2017). We are not aware of any cases in which deepfake evidence has been shown to have been introduced in court. But if it had been, it is not clear that current systems are robust enough to have spotted it.

> *PEOPLE ARE PARTICULARLY LIKELY TO TRUST THAT EVIDENCE PRESENTED IN A COURT HAS BEEN VETTED AND VERIFIED*

Legal guidance

If the first legal task is to raise awareness of the fact that deepfakes have legal implications, the second task is to develop greater clarity as to how deepfakes fit within current legal frameworks. As noted above, there are multiple laws that may cover harms caused by deepfakes. Greater clarity is needed on whether and how this is the case, as well as how the treatment of deepfakes is likely to differ from jurisdiction to jurisdiction. A possible point of comparison here is the Tallinn Manual, which assess how traditional laws of war apply in cyberspace (Schmitt, 2017). Might it be possible to assemble an international body of experts to produce a similar mapping of how current legal frameworks apply to deepfakes? In Europe a possible starting point might the EU's General Data Protection Regulation (GDPR), which on the face of it contains a number of provisions that would appear relevant to deepfakes, such as constraints on the processing of personal information or the "right to be forgotten."

[4] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, HR3230, 116th Cong. (2019).

Hard law

There is evidence that a growing number of legal restrictions on deepfakes are being considered or enacted. As noted above, the state of Virginia recently amended its laws to ensure that deepfakes would be covered by its prohibition of harassment via the sharing of sexual images and videos. In October 2019 the state of California amended its laws to prohibit the distribution of deepfakes portraying political candidates within 60 days of an election. And in China new rules take effect from January 2020 prohibiting the publication of fake news that has been created using deepfake technology.

Where harm can be carefully delineated, there is a strong case for legal restrictions. This remains the case even if identifying and prosecuting culprits may be difficult (which could be the case because of numerous factors, such as inter-jurisdictional issues or the greater ease of masking one's identity online than in real life). The legal code plays an important role in signaling where societal boundaries of acceptable behavior lie. However, as proposals to restrict deepfakes become broader, they are increasingly likely to come into tension with the value of free speech. This would be particularly true of a blanket ban, which Chesney & Citron (2018) assert would "chill experimentation in a diverse array of fields, from history and science to art and education."

> *WHERE HARM CAN BE CAREFULLY DELINEATED, THERE IS A STRONG CASE FOR LEGAL RESTRICTIONS*

A slightly different approach is taken by draft legislation published in mid-2019 in the US – the so-called DEEPFAKES Accountability Act.[4] Rather than prohibit deepfakes, if enacted this would make it a crime not to mark deepfakes with "irremovable digital watermarks, as well as textual descriptions". This approach of favoring increased transparency in the information ecosystem is worthy of consideration, even if in practice it would be difficult to enforce against a technologically sophisticated individual who is creating malicious deepfakes.

Penalties

Where hard law provisions are in place that apply to deepfakes work is needed to ensure that penalties for online crimes reflect their severity. This reflects a wider point: online and offline crimes that may appear very similar in terms of intent may be very different in terms of their impact on the target. This is particularly true when the target is an individual. Cases of intimidation or harassment are an example. In the digital space, a one-off episode cannot be assumed to have a transitory impact: if an abusive image or video has been uploaded to the internet, then it may last indefinitely, increasing the intensity and duration of the harm caused. Penalties have not evolved in response to this shift. We are not proposing an answer here, but suggesting that questions of proportionality and deterrence deserve attention given the extent to which digitalization can disrupt traditional patterns of crime and punishment.

Soft law

The process of adapting traditional legal and regulatory instruments and institutions to the rapidly evolving digital landscape is likely to be protracted and contested. In the interim, this creates incentives for the use of "soft law" methods to establish norms and practices related to deepfakes. These might include industry codes of practice or sets of standards issued by bodies such as the ISO or IEEE. Advantages of soft law approaches include the fact that they are typically international in scope and easier to modify than hard law instruments. However, there are significant disadvantages, including limited transparency and accountability, and the absence of direct governmental enforcement mechanisms. Nevertheless, these approaches have worked before in areas that may be of relevance to deepfakes. In the US in the 1990s, the threat of government-imposed television rating standards led to the introduction of an industry code of conduct in the same area. Might a similar approach lead to the introduction of stronger deepfake provisions in the terms of service agreements used by the internet platforms, which play a pivotal role in determining what content is disseminated. Interestingly, there is already some movement in this direction, such as the October 2019 announcement by Twitter that it is developing "a new policy to address synthetic and manipulated media" (Harvey, 2019). In the US in particular, where first amendment considerations shape to a large degree what is feasible in the realm of hard law, some observers see terms of service (TOS) agreements as the area to focus on: "Today's most important speech fora, for better or worse, are the platforms. And TOS agreements determine if speech on the platforms is visible, prominent, or viewed, or if instead it is hidden, muted, or never available at all. TOS agreements thus will be primary battlegrounds in the fight to minimize the harms that deep fakes may cause" (Chesney & Citron, 2018). A similar way of imposing standards might be via platform app stores, which could refuse to distribute apps that create deepfakes unless they meet requirements related to watermarks or consent, for example.

4. Society

The potential risks, challenges and trade-offs raised by deepfakes are symptomatic of, and inextricable from, wider questions relating to the societal costs and benefits of the world's digital transformation. The purpose of this report is not to grapple directly with those wider issues, but they impact directly on deepfake risks and so we conclude this chapter by highlighting briefly two areas of possible action.

Education

More effective steps in the area of digital education are needed. Discussion of digital education or digital literacy typically focuses on the "demand side" of the ecosystem, by encouraging individuals to become more critical and engaged consumers of digital content. There is interesting work being done in this area. One line of research suggests that a powerful way of making people more vigilant about fake and fabricated content is to give them experience of creating it.[5]

However, the limitations should be acknowledged of relying on education to solve problems in the information ecosystem. First, while efforts to foster greater knowledge and critical engagement are to be welcomed, part of the problem with the viral circulation of harmful content is that it appears to appeal to emotional rather than rational drivers (Vosoughi, Roy, & Aral, 2018). Second, it is possible that in some digital contexts the flow of information is too rapid to allow for effective assessment – humans may not be capable of evaluating accuracy or authenticity in the available time, and that may be something that education cannot simply "fix". Third, in some contexts increased knowledge can compound rather than solve problems related to false content. For example, one problem faced when trying to correct online falsehoods is that the correction can serve to amplify the falsehood without actually succeeding in rebutting it (Wang & Aamodt, 2008).

> *THE VIRAL CIRCULATION OF HARMFUL CONTENT APPEARS TO APPEAL TO EMOTIONAL RATHER THAN RATIONAL DRIVERS*

A fourth consideration is that encouraging more critical consumption of digital content may be counterproductive if it is not done carefully. If one of the purposes of increased digital education is to encourage people to be skeptical about the digital content they encounter, it is possible that this will exacerbate wider problems of declining trust. Unless critical thinking is taught in a rounded sense, digital education may make things worse and not better. Skepticism is necessary but not sufficient for digital responsibility – it risks eroding trust unless accompanied with a focus on proactive steps such as corroboration using trustworthy sources.

The "supply side" of the ecosystem should not be neglected. Part of the educational task involves developing and sustaining an ethos of ethical and social engagement among developers and content producers. In this, we echo a recent call for a "culture of responsibility" in the wider field of artificial intelligence (Brundage, et al., 2018).

[5] For an overview of work in this area, see (Arguedas Ortiz, 2018). The "Bad News" game, which involves setting up a fake news operation can be found at getbadnews.com. For the research underlying the game, see (Roozenbeek & van der Linden, 2019).

> *RISK GOVERNANCE FOR DEEPFAKES CANNOT EASILY BE ISOLATED FROM WIDER QUESTIONS ABOUT THE GOVERNANCE OF THE INTERNET MORE GENERALLY*

Digital governance

Risk governance for deepfakes cannot easily be isolated from wider questions about the governance of the internet more generally, including prevailing incentive structures (light regulation coupled with current business models) that facilitate the viral circulation of false and harmful content. We lack a fit-for-purpose system of governance for the online ecosystem. This is not to argue for or against any specific system of governance (top-down, bottom-up, national, global, etc). It is to make the more basic point that a more coherent system of some sort is required given the dramatic impact—both positive and negative—that the internet and related technologies are having on all aspects of contemporary life.

As noted earlier, governance refers to the totality of actors, rules, traditions and institutions by which authoritative decisions are taken. It seems increasingly clear that a greater degree of careful, deliberate decision-making is required in relation to the evolution of the online ecosystem. Managing the risks associated with deepfakes presents a similar challenge in microcosm. There are decisions to be taken about how to balance various factors: privacy against freedom of expression, for example, or technological innovation versus societal precaution.

> *AN AMBITIOUS APPROACH MIGHT BE TO EMBARK ON A FORMAL DELIBERATIVE PROCESS, WHETHER NATIONALLY OR INTERNATIONALLY*

One area of particular importance in framing an overarching system of online governance is the balance between online anonymity and verifiable identity. Current internet governance norms lean very strongly towards privileging anonymity over identity verification. Given the significance of the consequences that flow from that trade-off, it should at least be reviewed periodically to assess how the balance of costs and benefits is evolving. In addition, we recommend that greater consideration is given to emerging technologies which may be able to alter the terms of the trade-off, preserving anonymity but in ways that allow for some individual accountability and sanctions

An ambitious approach to these kinds of governance decisions might be to embark on a formal deliberative process, whether nationally and/or internationally. This would aim to (i) agree what the desired outcome is, given the balance of the rights and interests of all the affected stakeholders, and (ii) work back from that goal to identify and implement the various rules, norms, institutions and incentives might help to realise it. If successful, it might edge us towards something like a social contract updated for the digital world.

Conclusion

The complexity of the risk-governance landscape for deepfakes is instructive. Deepfakes represent one tiny field of application of machine-learning technologies, which in turn are embedded in wider and deeper processes of digital transformation. The way our simple framing of key deepfake risks and responses has unfolded into a pattern of overlapping domains, motivations, impacts and trade-offs highlights in microcosm how difficult the task of governing technological risks has become.

In this report, we have suggested a framework or heuristic for categorizing and prioritizing the potential risks posed by deepfakes to individuals, organizations and whole societies. We have also listed 15 potential governance responses, which we believe will help to better understand and mitigate these risks. And we have stressed the need for more detailed work that would bring domain-specific expertise to bear on the specific risks posed by deepfakes in specific contexts.

If it has been difficult to map even the basic contours of the risk landscape for deepfakes, then it would be much more complicated to develop a risk-governance perspective with relevance across emerging technologies more generally. However, given the central role now played by these technologies in contemporary societies, we believe that such an overarching perspective would be valuable and is worth attempting.

Some of the building blocks needed for such an over-arching risk-governance approach to emerging technology may be contained in previous IRGC work. Many emerging and converging technologies share characteristics that we have dealt with in our work on the governance of emerging and systemic risks (IRGC, 2015, 2018). Perhaps these earlier insights can be developed, honed and applied more specifically to new technologies.

It remains to be seen how the deepfake phenomenon will evolve, and whether the risks associated with it start to crystallize more and more disruptively or instead fade as a concern. Either way, one thing that is sure to recur is the pattern of rapid technological development and dissemination coupled with significant uncertainty about direct and indirect consequences. If general principles of emerging-technology risk governance can be formulated, then they would provide a valuable head start in tackling the next such cases.

References

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting World Leaders Against Deep Fakes. *The IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) Workshops, 38-45.

Ajder, H., Cavalli, F., Patrini, Giorgio, & Cullen, L. (2019). *The State of Deepfakes: Landscape, Threats, and Impact*.

Apostolidis, E., Apostolidis, K., Patras, I., & Mezaris, V. (2019). Video Fragmentation and Reverse Search on the Web. In V. Mezaris, L. Nixon, S. Papadopoulos, & D. Teyssou (eds.), *Video Verification in the Fake News Era*, Springer.

Arguedas Ortiz, D. (2018). *Could this be the cure for fake news?* Retrieved from BBC Future: www.bbc.com/future/article/20181114-could-this-game-be-a-vaccine-against-fake-news

BBC. (2019, July 8). *Fake voices 'help cyber-crooks steal cash'*. Retrieved from www.bbc.com/news/technology-48908736

Blitz, M. (2018, 71:1). Lies, line drawing, and deep fake news. *Oklahoma Law Review*, pp. 59-116.

Bonfanti, M. (2018). An Intelligence-based Approach to Countering Social Media Influence Operations. *Romanian Intelligence Studies Review*(19), 47-67.

Bradshaw, T. (2019, October 10). *Deepfakes: Hollywood's quest to create the perfect digital human*. Retrieved from Financial Times: www.ft.com/content/9df280dc-e9dd-11e9-a240-3b065ef5fc55

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., . . . Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Retrieved from maliciousaireport.com

Bunk, J., Bappy, J. H., Mohammed, T. M., Nataraj, L., Flenner, A., Manjunath, B., . . . Peterson, L. (2017). Detection and Localization of Image Forgeries Using Resampling Features and Deep Learning. *arXiv*: 1707.00433 [cs.CV]

Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: bypassing then detection methods. *arXiv*: 1705.07263 [cs.LG]

Chesney, R., & Citron, D. K. (2018). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *SSRN Electronic Journal*. Retrieved from papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954

Cole, S. (2017, December 11). *AI-Assisted Fake Porn Is Here and We're All Fucked*. Retrieved from Vice: www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn

Ctrl Shift Face. (2019, May 10). *Bill Hader impersonates Arnold Schwarzenegger [DeepFake]*. Retrieved from YouTube: www.youtube.com/watch?v=bPhUhypV27w

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. CVPR. Retrieved from image-net.org/papers/imagenet_cvpr09.pdf

Farid, H. (2018, June). Digital forensics in a post-truth age. *Forensic Science International*, pp. 268-269. Retrieved from www.sciencedirect.com/science/article/pii/S0379073818303013

Finn, C., Abbeel, P., & Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv*: 1703.03400 [cs.LG]

Giles, M. (2018, March/April). *The GANfather: The man who's given machines the gift of imagination*. Retrieved from MIT Technology Review: www.technologyreview.com/s/610253/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative Adversarial Networks. *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (pp. 2672-2680). Montreal, Canada: MIT Press.

Google. (2019a). *Cloud TPU*. Retrieved from cloud.google.com/tpu/

Google. (2019b). *Contributing Data to Deepfake Detection Research*. Retrieved from Google AI Blog: ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

Grimm, P., Capra, D., & Joseph, G. (2017). Authenticating Digital Evidence. *Baylor Law Review*, 69(1).

Harvey, D. (2019). *Help us shape our approach to synthetic and manipulated media*. Retrieved from Twitter: blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv*:1512.03385 [cs.CV]

Huh, M., Liu, A., Owens, A., & Efros, A. A. (2018). Fighting Fake News: Image Splice Detection via Learned Self-Consistency. *arXiv*:1805.04096 [cs.CV]

IRGC. (2015). *IRGC Guidelines for Emerging Risk Governance*. Lausanne: EPFL International Risk Governance Center.

IRGC. (2018). *IRGC Guidelines for the Governance of Systemic Risks*. Lausanne: EPFL International Risk Governance Center.

Iuliani, M., Shullani, D., Fontani, M., Meucci, S., & Piva, A. (2019). A Video Forensic Framework for the Unsupervised Analysis of MP4-Like File Container. *IEEE Transactions on Information Forensics and Security archive, Volume 14, Issue 3.*

Knight, W. (2019, August 16). *The world's top deepfake artist is wrestling with the monster he created*. Retrieved from MIT Technology Review: www.technologyreview.com/s/614083/the-worlds-top-deepfake-artist-is-wrestling-with-the-monster-he-created/

Li, Y., Chang, M.-C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv*.

Lin, H. (2019, 75:4). The existential threat from cyber-enabled information warfare. *Bulletin of the Atomic Scientists*, pp. 187-196 . doi: doi.org/10.1080/00963402.2019.1629574

MacDougald, P. (2017). *The Unflattering Familiarity of the Alt-Right in Angela Nagle's Kill All Normies*. Retrieved from New York Magazine: nymag.com/intelligencer/2017/07/angela-nagles-kill-all-normies-the-alt-right-and-4chan.html

Maras, M.-H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof*. Retrieved from journals.sagepub.com/doi/abs/10.1177/1365712718807226

McMahon, L. (2018, November 9). *"Deepfakes": a looming nightmare for insurers?* Retrieved from Insurance Information Institute: www.iii.org/insuranceindustryblog/deepfakes-nightmare-insurers/

Miller, C. (2019). *Crossing Divides: How a social network could save democracy from deadlock*. Retrieved from BBC News: www.bbc.com/news/technology-50127713

Moreira, D., Bharati, A., Brogan, J., Pinto, A., Parowski, M., Bowyer, K. W., . . . Scheirer, W. J. (2018). Image Provenance Analysis at Scale. *arXiv*:1801.06510 [cs.CV]

Mullin, J. (2017, June 30). *Ars Technica*. Retrieved from arstechnica.com/tech-policy/2017/06/facebook-and-twitter-could-be-fined-up-to-57-million-under-new-german-law/

Nvidia. (2019, August 8). *Cloud & Data Center*. Retrieved from www.nvidia.com/en-us/data-center/dgx-1/

OpenAI. (2019, December 9). *Better Language Models and Their Implications*. Retrieved from openai.com/blog/gpt-2-1-5b-release/

Ovadya, A. (2019). *Proposal: International Media Authenticity Council (v 0.8)*. Retrieved from thoughtfultech.org/authenticity-council

Ovadya, A., & Whittlestone, J. (2019, July). Reducing Malicious Use Of Synthetic Media Research: Considerations And Potential Release Practices For Machine Learning. *Draft Paper.*

Paul, G. (2008). *Foundations of Digital Evidence*. American Bar Association.

Ravenscraft, E. (2019). *Instagram's New Anti-Bullying Nudges Could Actually Work*. Retrieved from OneZero: onezero.medium.com/instagrams-new-anti-bullying-nudges-could-actually-work-9811ef41b8cb

Robertson, A. (2019). *Virginia's 'revenge porn' laws now officially cover deepfakes*. Retrieved from The Verge: www.theverge.com/2019/7/1/20677800/virginia-revenge-porn-deepfakes-nonconsensual-photos-videos-ban-goes-into-effect

Roozenbeek, J., & van der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*. Retrieved from www.tandfonline.com/doi/full/10.1080/13669877.2018.1443491

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *arXiv*:1901.08971 [cs.CV]

Rothman, J. (2018, November 5). In the Age of A.I., Is Seeing Still Believing. *The New Yorker*. Retrieved from www.newyorker.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-still-believing

Schmitt, M. (Ed.). (2017). *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. Cambridge: Cambridge University Press.

US House of Representatives Permanent Select Committee on Intelligence. (2019). *The National Security Challenge of Artificial Intelligence, Manipulated Media, and "Deepfakes"*. Retrieved from intelligence.house.gov/news/documentsingle.aspx?DocumentID=657

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

Wang, S., & Aamodt, S. (2008, June 29). *Your brain lies to you*. Retrieved from New York Times: www.nytimes.com/2008/06/29/opinion/29iht-edwang.1.14069662.html

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. *arXiv*: 1905.08233 [cs.CV]

Zhou, W., Li, H., & Tian, Q. (2017). Recent Advance in Content-based Image Retrieval: A Literature Survey. *arXiv*: 1706.06064 [cs.MM]

Acknowledgements

IRGC thanks the following individuals who participated at our expert workshop on governing deepfake risks, 9–10 September 2019: **Lisa Bechtold**, Zurich Insurance Company; **Matteo E. Bonfanti**, ETH Zurich; **Eric Durand**, Swiss Re; **John Egan**, BNP Paribas; **Rainer Egloff**, Swiss Re; **Kelsey Farish**, DAC Beachcroft; **François Fleuret**, IDIAP Research Institute; **Bryan Ford**, EPFL; **Priscilla Guo**, Tsinghua University; **Jerry Gupta**, Swiss Re; **Nikolas Hunt**, Zurich Insurance Company; **Käsper Kivisoo**, Hybrid COE; **Mihalis Kritikos**, European Parliament Research Service; **James Larus**, EPFL; **Andre Loesekrug-Pietri**, Joint European Disruptive Initiative; **Marie-Helen Maras**, John Jay College of Criminal Justice; **Gary Marchant**, Arizona State University; **Vasileios Mezaris**, Centre for Research and Technology Hellas & Information Technologies Institute; **Ian Molloy**, IBM Watson Research Center; **Aaron Rosa**, Fraunhofer ISI; **Gunter Roth**, NVIDIA; **Alexandre Rouxel**, European Broadcasting Union; **Stephan Schreckenberg**, Swiss Re; **Marc Stampfli**, NVIDIA; **Denis Teyssou**, Agence France-Presse; **Pierre Vandergheynst**, EPFL; **Martin Weymann**, Swiss Re.

The views and recommendations contained in this report do not necessarily represent the views of individual workshop participants or their employers.

The author of this report is **Aengus Collins** (Deputy Director, IRGC), with contributions from **Marie-Valentine Florin** and **Anshuman Saxena**. Responsibility for the final content of this report rests entirely with IRGC.

We are extremely grateful to the following workshop participants and external experts for their insightful inputs while the report was being drafted: Shahar Avin, Centre for the Study of Existential Risk; Alison Bewick, Nestlé; Maren Bodenschatz, Swiss Re; Eric Durand, Swiss Re; Rainer Egloff, Swiss Re; Priscilla Guo, Tsinghua University; Clément Guitton, Swiss Re; Jerry Gupta, Swiss Re; Käsper Kivisoo, Hybrid COE; Mihalis Kritikos, European Parliament Research Service; Gary Marchant, Arizona State University; Vasileios Mezaris, Centre for Research and Technology Hellas & Information Technologies Institute; Ian Molloy, IBM Watson Research Center; Aviv Ovadya, Thoughtful Technology Project; Jean-Marc Rickli, Geneva Centre for Security Policy; Gunter Roth, NVIDIA; Bernd Stahl, De Montfort University; Christopher Tucci, EPFL; Denis Teyssou, Agence France-Presse; Jeremy West, OECD.

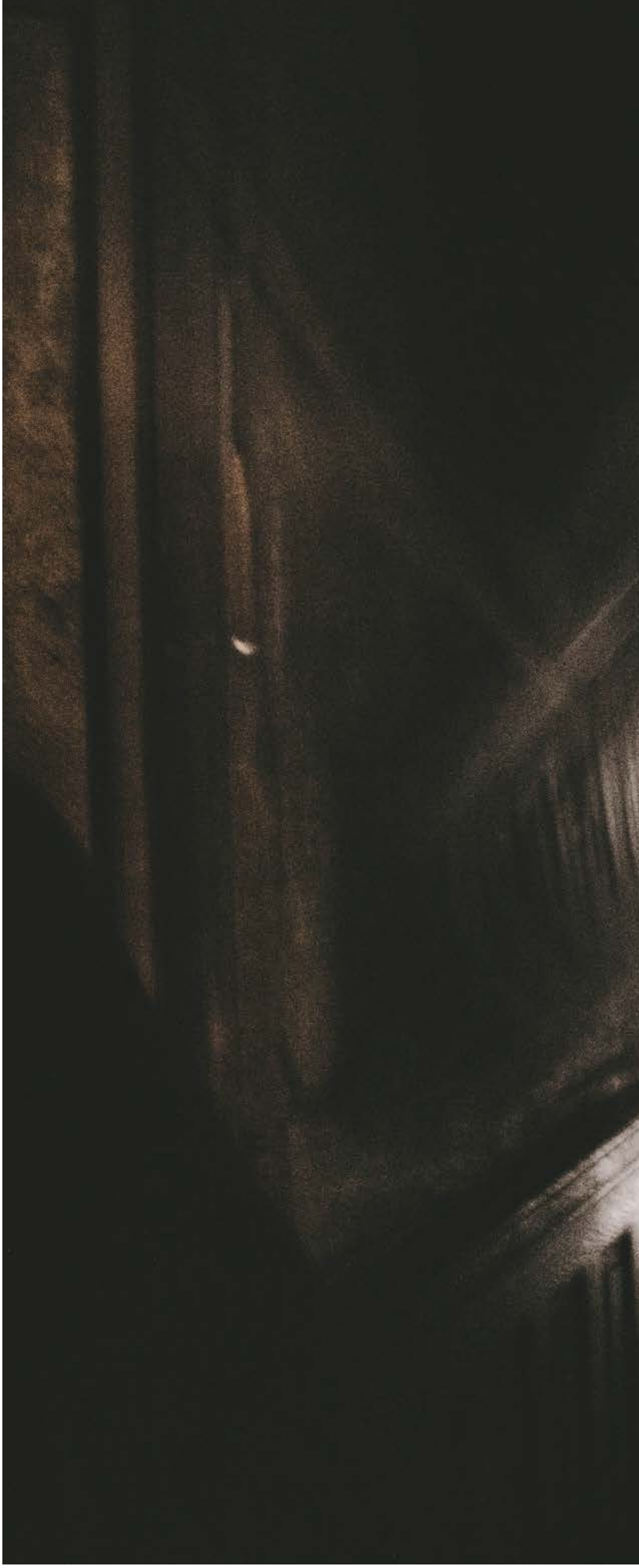
All graphic design work on this report was done by **Anne-Sylvie Borter**. Copy-editing was done by **Anca G. Rusu**.

IRGC wishes to acknowledge and thank the Swiss Re Institute for providing a financial contribution towards the cost of the workshop.

About IRGC

The International Risk Governance Center at EPFL (Ecole polytechnique fédérale de Lausanne) helps to improve the understanding and governance of systemic risks that have impacts on human health and safety, the environment, the economy and society at large. IRGC's mission includes developing risk governance concepts and providing risk governance policy advice to decision-makers in the private and public sectors on key emerging or neglected issues. It emphasises the role of risk governance for issues marked by complexity, uncertainty and ambiguity, and the need for appropriate policy and regulatory environments for new technologies where risk issues may be important.

irgc.epfl.ch



EPFL International Risk
Governance Center

EPFL IRGC
Station 5, BAC
1015 Lausanne
Switzerland

+41 21 693 82 90

irgc@epfl.ch
irgc.epfl.ch