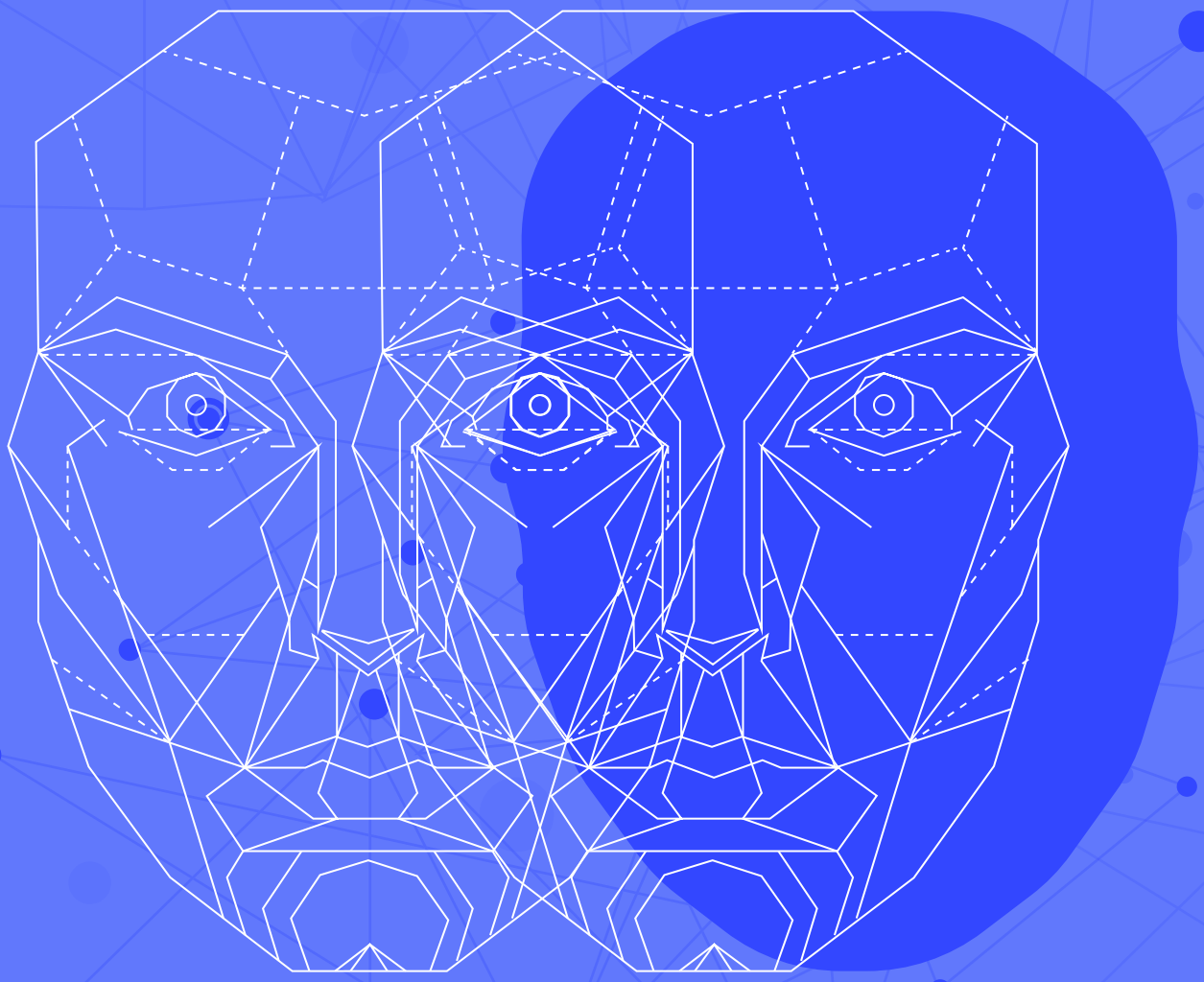
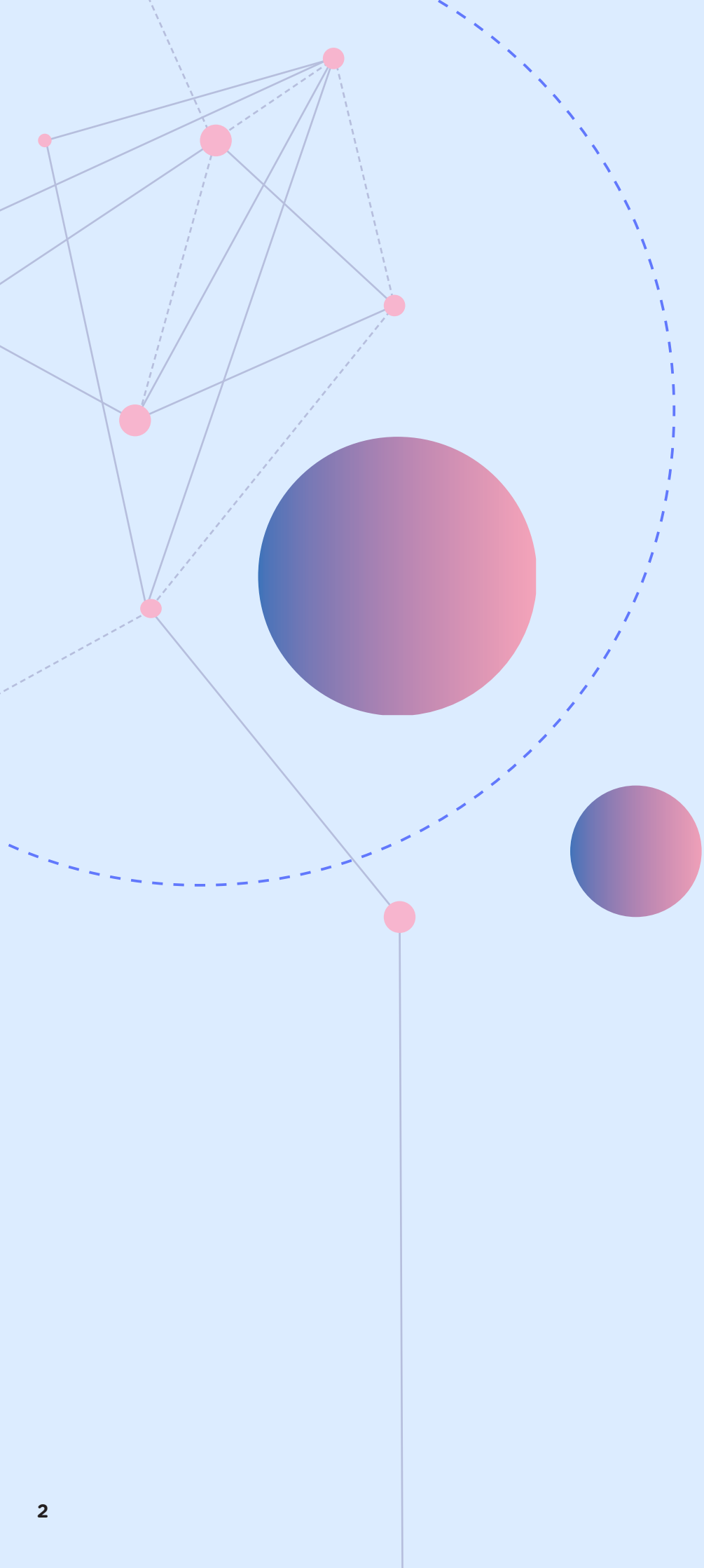
DEMOCRACY
REPORTING
INTERNATIONAL

DEEPFAKES

How prepared are we?

Multi-stakeholder perspectives and a recommendations roadmap





About Democracy Reporting International

Democracy Reporting International (DRI) strengthens democracy by shaping the institutions that make it sustainable. We support local ways of promoting democracy with impartial analysis and good practices, bringing international standards to life.

The belief that people are active participants in public life, not subjects of their governments, guides what we do. We work with local actors to protect and expand our shared democratic space in a polarized world, regardless of political opinions or personal beliefs.

Find out more at: http://www.democracy-reporting.org

Acknowledgements

This paper is based on interviews with experts at Adobe, Bellingcat, the Brookings Institution, the Carnegie Endowment for International Peace, the European Digital Media Observatory (EDMO), Facebook, the Global Public Policy Institute (GPPi), Human Rights Watch, Microsoft, NATO StratCom, the Oxford Internet Institute, Partnership on AI, TikTok, Twitter, UC Berkeley – Hany Farid, WITNESS and Mr. Bruno Sartori, a Brazilian satirical deepfake creator. Adeline Marquis supported in the preparation of the report structure and with the expert interviews.

This paper was designed by Forset.

In a three-part series, DRI is exploring deepfakes as an emerging disinformation threat. In the first paper, we provided an overview of the deepfake threat. In a final paper, DRI will make recommendations on deepfake disinformation threats in elections.

Date: November 2020

This paper is part of a project funded by the German Federal Foreign Office. Its contents in no way represent the position of the Foreign Office.

Auswärtiges Amt

Authors:

Rafael Goldzweig

Madeline Brady

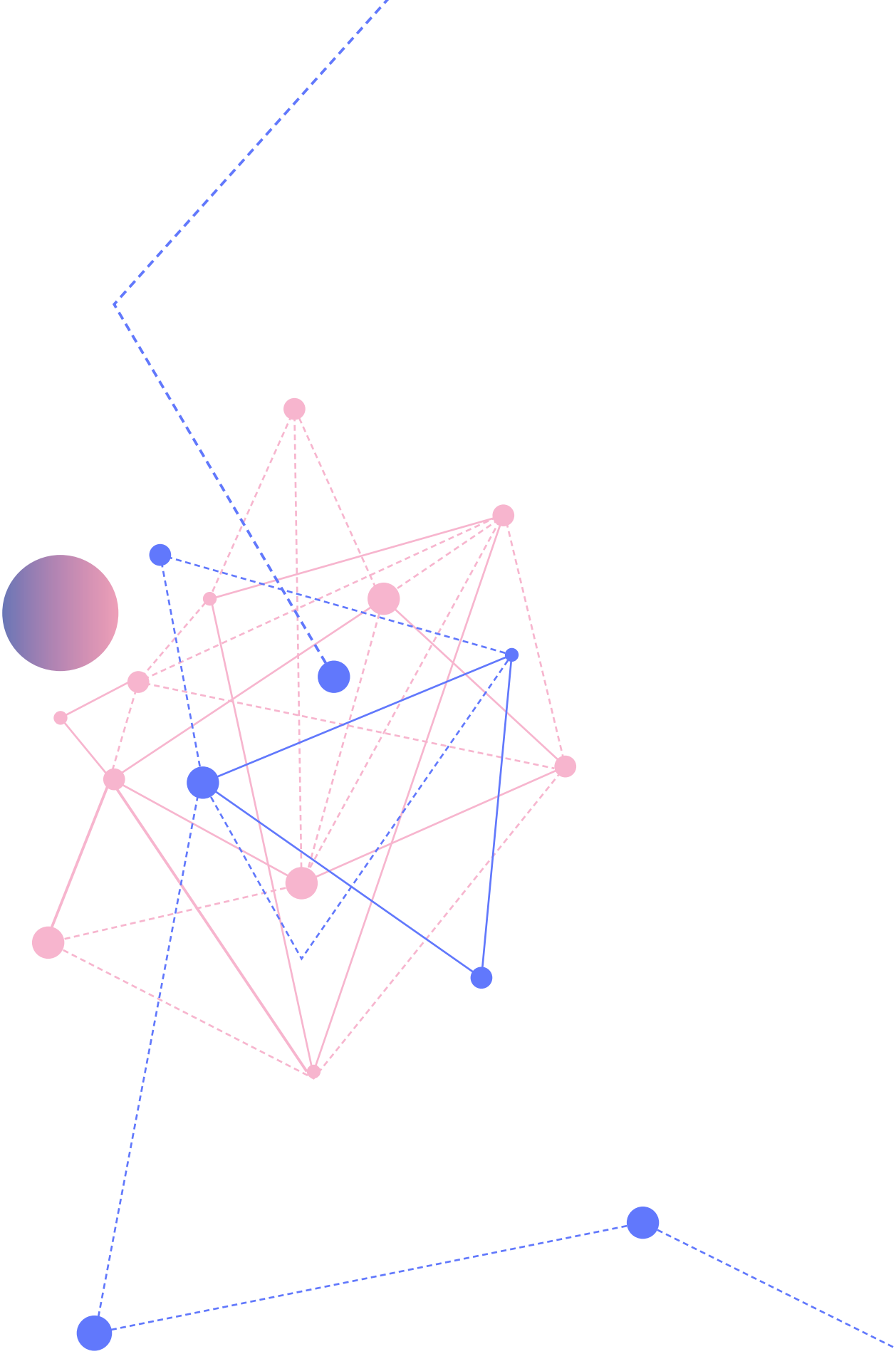


Table of Contents

Executive Summary

This research paper builds on the thoughts and suggestions from 22 interviewees directly involved with the topic of deepfakes, or indirectly responsible for tackling the challenges the phenomenon poses to democracies. We use "manipulated media" in this paper as an umbrella term to refer to video altered with the use of artificial intelligence (AI) – "deepfakes" – and video altered with less sophisticated or no tools – "cheapfakes". Here are the main takeaways:

What?

Cheapfakes, rather than deepfakes, dominate what experts are currently seeing when it comes to the use of visual media to disinform users. Focusing too much on the potential threat of deepfakes may distract from attention to this problem.

Who?

A successful perpetrator of a malicious deepfake will be a known disinformation actor, with both the technical capacity and local knowledge to make the deepfake appear realistic.

Are we prepared?

Technical, societal and platform preparedness will all be necessary to reduce the likelihood of the threat; one solution alone will not be sufficient. With regard to technical preparedness, most experts noted that more time will be needed, and technical solutions may lag behind for a period of time. There are two main technical approaches to being prepared: a collaborative, multi-stakeholder provenance solution (e.g., the use of a blockchain to time-print image sources), or algorithmic detection. Experts tend to find the former approach more promising, although there are considerable positive and negative considerations related to both. At the societal level, most experts agree that voters, media and governments are not prepared. Regarding platform preparedness, it is unclear whether they have the technical capacity to effectively identify and prevent deepfakes from going viral, especially given that experts and tech companies alike are seeing no deepfakes in practice. Despite the introduction of new policies toward manipulated media, the exact level of technical investment is unclear, and some might be better prepared than others.

When?

There is no agreement among experts as to when the threat of the perfect AI-generated deepfake will materialise, with most estimates ranging from several months to two years, and the most conservative at five years. The state of technology depends on the technique used.

Threat Scenarios:

The likelihood and impact of the threat will depend on the timing, target, speed of distribution and level of preparedness to counter the threat. Various potential scenarios could unfold, harming democratic discourse in the short, medium and long terms. In the short term, one undetected deepfake targeting a politician or vulnerable group, especially at a time of heightened social, political or geopolitical tensions, or during elections, may go viral on social media platforms if it is not detected in time. In the medium term, as the prevalence of this technology grows, news media and social media platforms may be overwhelmed with manipulated video or audio data, making factful reporting and online discourse challenging. This leads to the long-term threat, where citizens may no longer believe credible information in the course of their political decision-making, and thus lose trust in facts and democratic institutions.

Risk Assessment:

Overall, experts agree that, to a greater or lesser extent, we are not prepared in the areas of technical, societal and platform preparedness. Therefore, the risk of a successful deepfake influencing political behavior in the medium term should not be ruled out.

Recommendations

Work within the anti-disinformation framework: Use the existing disinformation frameworks to enact transparency requirements for platforms, get the conversation going through an EU policy paper or joint communication, promote collaboration between relevant actors, involve citizens in the conversation and deter violators.

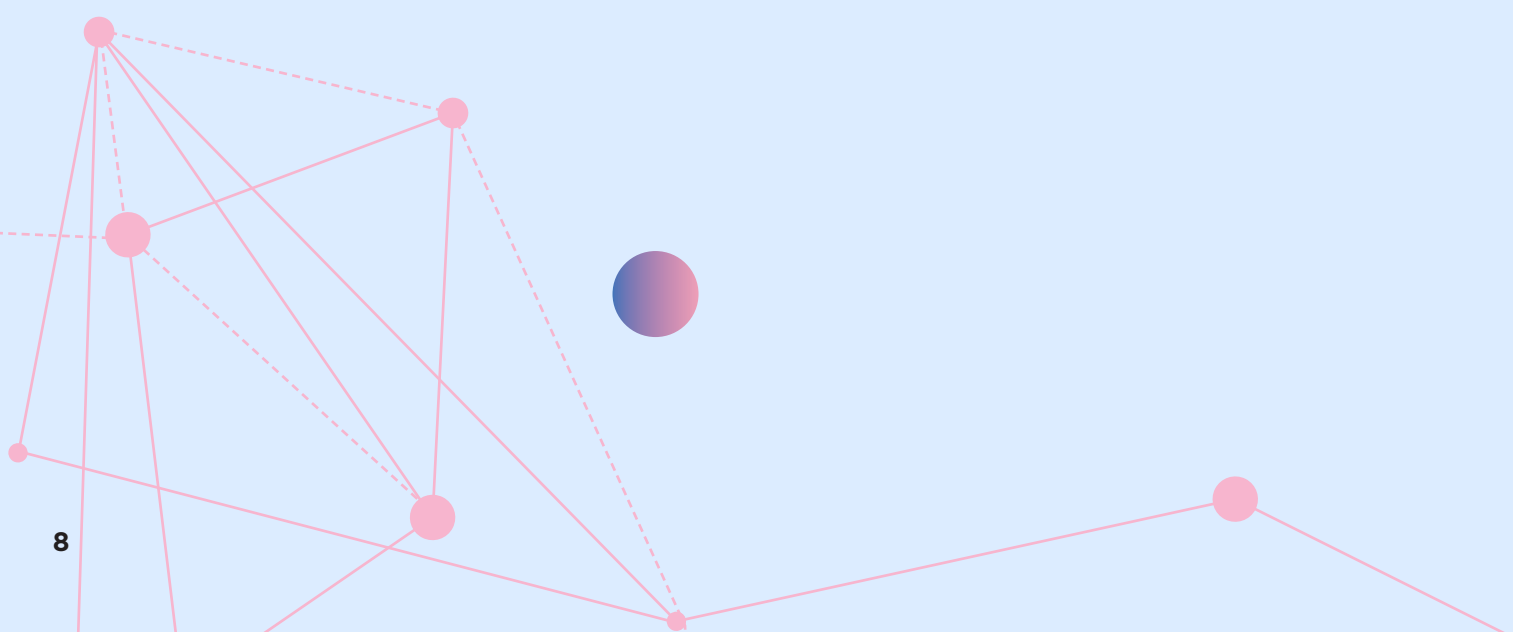
Foster a resilient ecosystem: Increase shared detection responsibilities, go beyond technical solutions, sustain support to civil society and research in the EU, and support media literacy programmes focusing on this topic.

Be ready to respond rapidly: Pay attention to virality, share threat signs among news media and social media industry participants, involve stakeholders to implement solutions and develop response action plans and scenarios.

I. Introduction

The aim of this paper is to provide a realistic assessment of how the manipulation of media may threaten democratic discourse. We evaluate the current prevalence of related technologies in online spaces and assess possible developments, threat scenarios and levels of preparedness for countering the use of this dynamic, harmful technology. We propose a roadmap of action to assist stakeholders in building their resilience to the misuse of deepfake technologies.

To inform this analysis, we conducted interviews with experts from around the world who work directly with deepfakes and the manipulation of media. The interviews were conducted between 13 August and 24 October 2020, with 22 experts from 8 countries (12 men and 10 women), coming from diverse fields of work: academia, tech companies, civil society and think tanks, and technical deep fake creators.

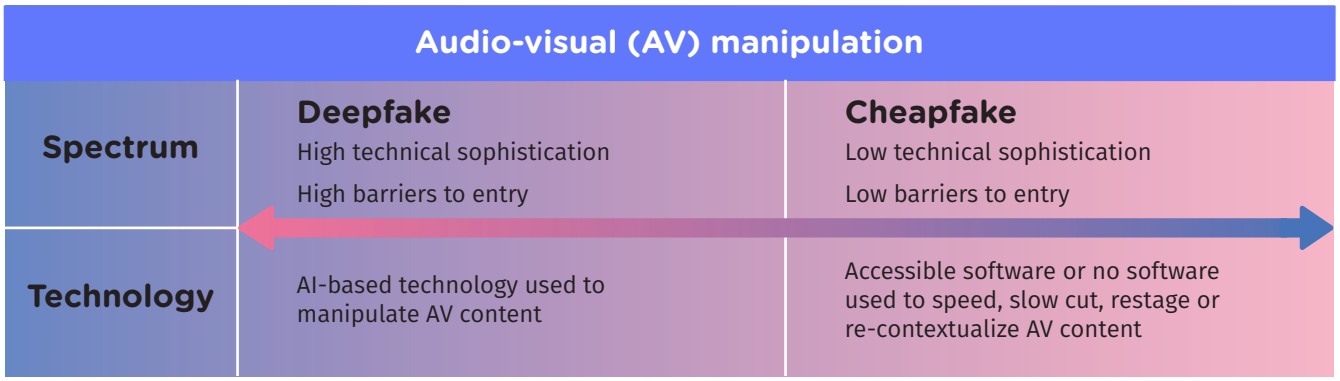


II. Context

What are deepfakes?

It is now possible to use artificial intelligence (AI) and less sophisticated means to manipulate video and audio. The technology for this manipulation varies broadly both with regard to ease of use and stage of development. Specific categories and definitions may be helpful in understanding both the problem and its solutions at a more precise level. Within the expert community, the nomenclature in this field is still being developed, so the definitions and terms found in this paper may not be those that ultimately pass into common usage. This is because the line between automated video manipulation and simple software is blurry. With this in mind, it is important to consider video manipulation across a broad spectrum. The distinctions involved here may not actually matter to the average internet user, but are important in assessing the options available to perpetrators and the risks they pose when it comes to manipulating public opinion. A low-level manipulated video might be just as effective in manipulating a viewer's perception as a much more technically sophisticated manipulation.

For the purposes of this paper, our focus is on the use of manipulated media (audio-visual manipulation) as an umbrella term to include both deepfakes and cheapfakes[1]. Cheapfakes are media products manipulated with a low level of technical sophistication, in which the speeding up, slowing, cutting, re-staging or re-contextualisation of media content can be performed with little to no use of a sophisticated software. Such manipulation might require no real use of technology at all, such as simply sharing a video with a misleading or false caption. Deepfakes require a higher level of technological sophistication, using AI-based technology, which is growing and evolving by the day.



Why might deepfakes harm society?

Experts and media have expressed concern over a "perfect" deepfake being released, meaning that forensic scientists and AI would be unable to reliably verify its authenticity. Experts have pointed out that while there are many people working on the further development of this technology, relatively few are working on means of detection[2]. Additionally, video manipulation technology is becoming relatively inexpensive and much easier to use[3]. As mentioned, little to no software may be needed to manipulate a video, and the resulting cheapfakes are being used to manipulate public perceptions. With this in mind, we can expect to see more sophisticated ways of using manipulated media, with the same purposes[4].

Other global disinformation trends may accelerate the threat factor. First, state actors known to be active in carrying out foreign influence campaigns have used increasingly sophisticated AI technology. Second, the popularity of video-sharing platforms such as TikTok is on the rise, making deepfakes an attractive medium for disinformation. Third, private messaging platforms are some of the most popular social media worldwide, meaning disinformation can spread quickly, unnoticed and at scale, reducing the ability to take such content down quickly – or at all.

At the same time, deepfake technology has not yet been used for high-profile, nefarious political purposes. The technology has been used in the film industry, and also in political satire, but without the aim to deceive users. The largest, most real threat at present is non-consensual pornography, targeting women with the potential for use in cyber-crime, particularly in the financial sector. These threats were discussed in greater detail in DRI's first paper on deepfakes.[5]

---

[1] Britt Paris and Joan Donovan, "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence", Data & Society, 18 September, 2019, https://datasociety.net/library/deepfakes-and-cheap-fakes/
[2] Will Knight, "Deepfakes Aren't Very Good. Nor Are the Tools to Detect Them", Wired, 12 June 2020, https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/
[3] Paris and Donovan, "Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence", op. cit., note 1.
[4] Giorgio Patrini, "Mapping the Deepfake Landscape", Sensity, 10 July 2019, https://sensity.ai/mapping-the-deepfake-landscape/
[5] Democracy Reporting International, "Deepfakes: A New Disinformation Threat?", August 2020, https://democracy-reporting.org/wp-content/uploads/2020/08/2020-09-01-DRI-deepfake-publication-no-1.pdf.

III. Threat Assessment

A) Cheapfakes: The Threat of Today

Cheapfakes, rather than deepfakes, currently predominate what experts are seeing online. Experts are seeing little to no use of AI-generated deepfakes for harmful political purposes, with their absence during the campaign the 2020 elections in the United States as a prime example.

It remains unclear, however, whether we are seeing a measurable quantitative increase in the use of cheapfakes. One researcher interviewed for this paper commented on the difficulty of monitoring video at scale. Unlike with data on text-based posts, analyzing video requires more technical tools, qualitative skill sets and time. This makes it more difficult to understand the scope of the cheapfake phenomenon.

Trained digital forensic experts who work with video data tell us that it is quite easy to spot low-level manipulation. In most cases, they have either seen the original video material or are able to gather videos from different angles to assess the authenticity of the item qualitatively. This may not be as easy for untrained eyes, especially when content goes viral on social media platforms. Another expert noted that monitoring cheapfakes at a large scale may be even more difficult for platforms than monitoring deepfakes. With deepfakes, platforms can use AI to detect AI manipulation, thus providing a clear target.
Cheapfakes, however, require those operating and monitoring platforms to know what they are looking for, and have to be identified more subjectively (e.g., real videos with misleading captions).

While there is a clear need to focus on the future threat of deepfakes, this should not distract from the real threat today – cheapfakes. For everyday users, cheapfakes may already be real enough to manipulate perceptions. If these videos go unchecked on social media platforms, the effect on users will surely contribute to lower trust in facts and institutions over time.

Based on this, "disbelief-by-default" culture[6] poses the biggest current threat to democratic discourse. Disbelief-by-default culture describes a world where users no longer believe what they see and are unable to critically evaluate information and facts. Here, even the possibility that a video might be manipulated creates doubt in viewers' minds. It also allows politicians and other public figures to claim any video is fake, thus reducing the public's ability to hold governments accountable. As in the case with disinformation in general, this level of cynicism and loss of trust could impact political behavior and attitudes towards democratic institutions.

B) Deepfakes: The Threat of Tomorrow

When it comes to deepfakes, there is much uncertainty as to when, from whom and how the threat to democratic discourse will materialise:

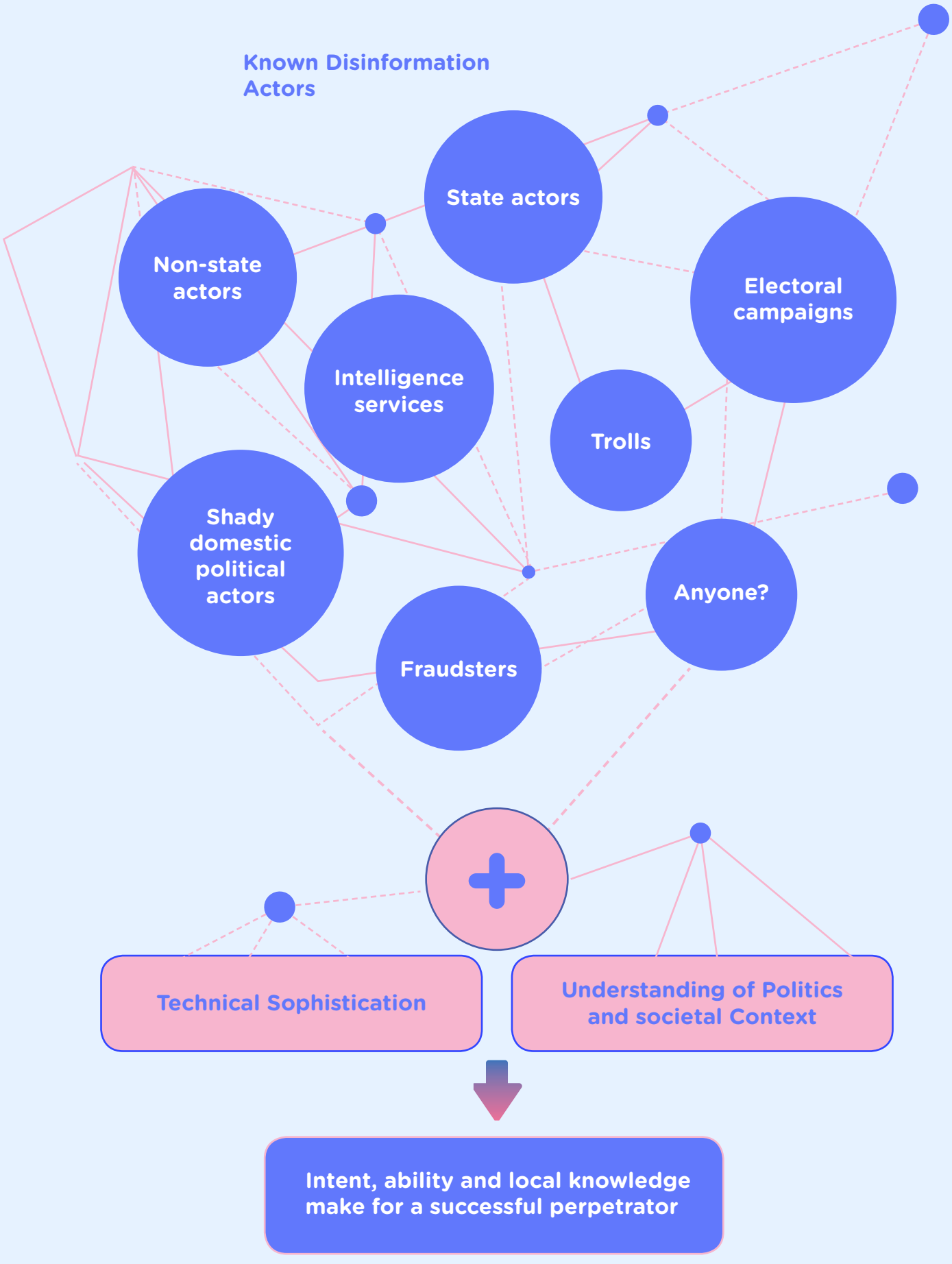
When will the deepfake threat materialise? Most experts provided estimates of from several months to two years for deepfakes to be perfected and used by nefarious actors. The longest estimate was up to five years. There is thus no agreement on the threat horizon within the expert community.

According to a (satirical) deepfake creator, the current level of development varies depending on the type of audio-visual manipulation (lip-sync dubbing is at a different stage than audio manipulation, etc.). He was surprised at the time of the interview that deepfakes had not been used to try to influence the United States elections, despite the current state of this technology.

Who would be the likely perpetrators? As deepfake technology offers a new tool for disinformation campaigns, those currently involved in disinformation activities should be considered. Such actors may include intelligence services, electoral campaigns, shady domestic political actors, fraudsters, trolls or any other nefarious actors.

Creating a realistic, convincing deepfake will, however, require a strong technical background, in addition to political and social knowledge of the specific context. As a result, understanding the local context and both the intent and ability of potential perpetrators is critical.

A phrase used by Sam Gregory from WITNESS; Sam Gregory, "Ticks or it Didn't Happen", WITNESS, December 2019, https://lab.witness.org/ticks-or-it-didnt-happen/


Known Disinformation Actors
State actors
Non-state actors
Electoral campaigns
Intelligence services
Trolls
Shady domestic political actors
Anyone?
Fraudsters
Technical Sophistication
Understanding of Politics and societal Context
Intent, ability and local knowledge make for a successful perpetrator

How might this threat to democratic discourse materialise? During the expert interviews, it became clear that the threat level varies, based on several factors:

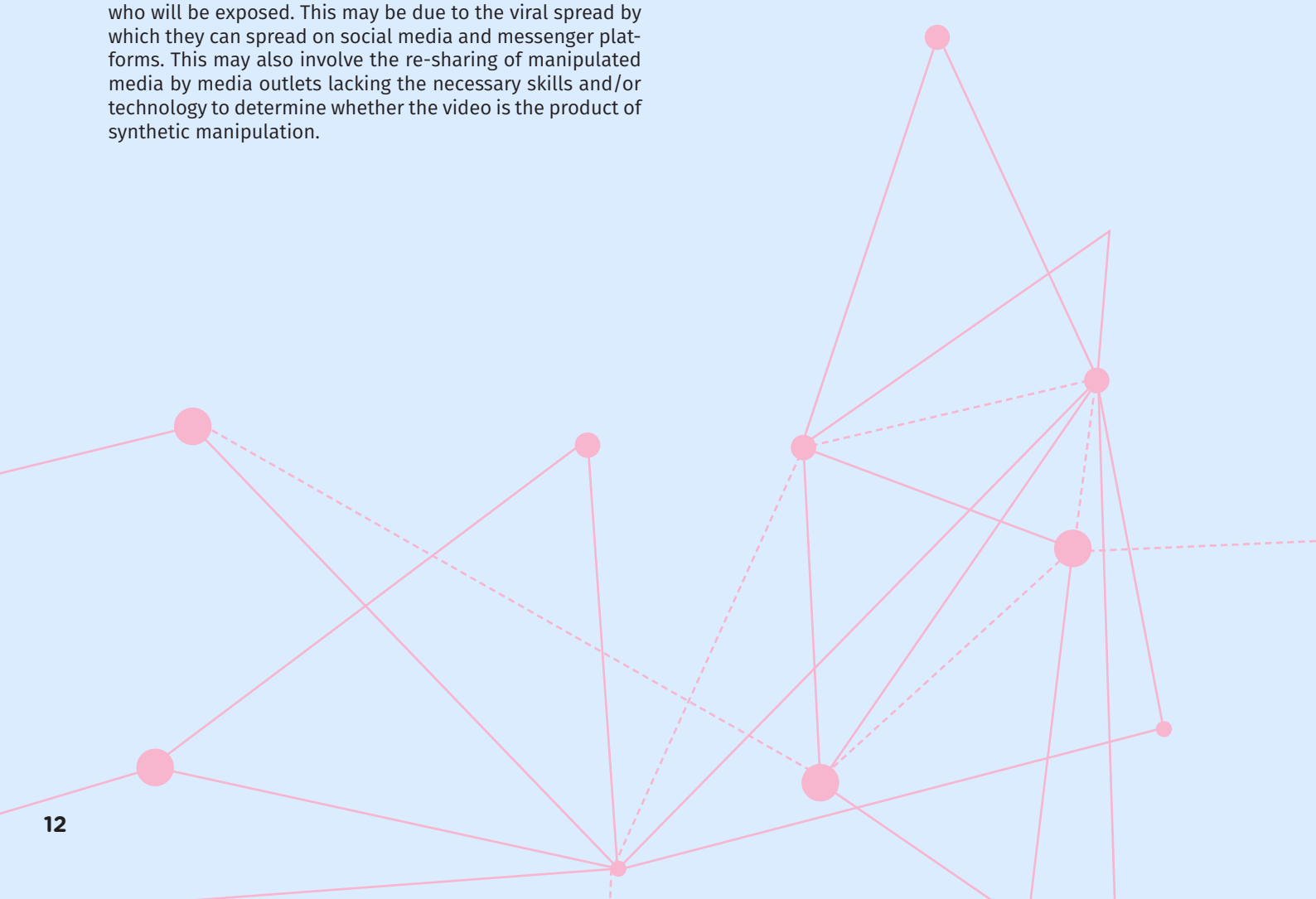
Timing: Certain moments might be highly sensitive, creating a situation where an unverifiable video could have immediate and immense impact. Examples include elections and moments of high social, economic, political or geopolitical tensions. Such moments create the potential for impacting the outcome of elections, sparking protest, causing violence, raising diplomatic complications, or even leading to armed conflict.

Targets: Along with "high-level" targets (e.g., national politicians or other high-profile and influential people), lower-profile individuals (e.g., local politicians, journalists), as well as more vulnerable groups (e.g., women, ethnic/religious minorities, LGBTQI+, the less-educated) may also be targeted by nefarious actors.

Speed of distribution: The faster a harmful video can circulate online, the greater the number of people who will be exposed. This may be due to the viral spread by which they can spread on social media and messenger platforms. This may also involve the re-sharing of manipulated media by media outlets lacking the necessary skills and/or technology to determine whether the video is the product of synthetic manipulation.

Volume: Many of the experts interviewed expressed concerns over the appearance of one viral, undetected and harmful deepfake. Although, if the problem were to manifest itself via volume attacks, involving the dissemination of multiple deepfakes, the potential impact would increase dramatically. If nefarious actors were interested in volume attacks and detection technology were not developed in time, such a scenario could be possible.

Preparedness: Both technical and societal preparedness impact the threat level. Technical preparedness is needed to identify and discredit manipulated media. Social preparedness is needed to maintain a critical eye for such content and create long-term resilience.



Threat Scenarios

Based on these factors, we can see the different ways in which deepfake technology may be used by nefarious actors to harm democratic discourse. These scenarios serve as an exercise to better understand the threat, rather than a definite timeline for what is to come:[7]

Impact	Scenario	Target	Description
Short-term ↓ Long-term	**One viral, undetected and harmful manipulated media**	 1. Politician, government or campeign	One false video released during an election or moment of high political or economic tensions
		 2. Vulnerable groups or segments of polarized society	One false video released during a moment of high societal tensions, when violence may be likely
		 3. Politician or government on the geopolitical stage	One false video released during a moment of high geopolitical tensions
	False report by credible source based on harmful, manipulated media	 4. News outlets	A news organisation makes a false report based on false video footage, and loses credibility
		 5. Civil society organisations (CSOs) or human rights groups	Human rights organisation makes a false report based on false video data and loses credibility
	"Disbelief by default"	 6. Society	Potential to manipulate leads society to lose trust and assume all information is false. This may result from an actual flood of manipulated video online, or even before this actually happens

[7] Note that threat scenarios are not mutually exclusive. Multiple cases of 1-6 could lead to an increased threat of "disbelief by default".

1. Political Attack

A political attack scenario would entail one viral, undetected and harmful deepfake targeting a government (e.g., at a moment of high political tension) or a political party or candidate (e.g., during an election campaign).

This scenario would have an immediate impact on a society, with the potential to influence the outcome of an election, to spark violence, to inspire protests or even to encourage a coup d'état. The damage could already be done before the content could be verified, and the impact would be even more severe in cases where trust in institutions and media is already low, as people may not trust the government or the platforms responsible for verifying the content in question. This is of particular concern given the fact that there are currently many different experimental detection algorithms, and there is no single proven effective tool or authoritative verification body.

In terms of preparedness for this scenario, academics at the University of California, Berkley, for example, are focusing on detection algorithms to determine the authenticity of videos of high-profile world leaders. For example, a specific detection algorithm may be produced for a world leader's face, knowing that they might be a likely target. Other experts believe such an approach has limitations, as it focuses on typical situations (the person standing at a podium) but may be less prepared for a less typical set-up (the leader sitting on a sofa, where their posture could be entirely different from that in the more official situation). Such an approach might, to some degree, increase preparedness in detecting manipulated video of high risk, but it is only a small part of a comprehensive solution.

2. Attack on Vulnerable Groups

In this case a viral, undetected and harmful deepfake targeting vulnerable groups or certain segments of a polarized society would be released at a time when societal tensions are already high. As seen with discriminatory disinformation more broadly (e.g., anti-migrant disinformation in the EU over the past five years), specific groups can be targeted and harmed.

This scenario could have an immediate impact on a society, with the potential to spark protests or conflict between different groups. A targeted deepfake video could generate strong emotions based on local social dynamics. Such a video may be powerful even after it has been identified as a fake, especially if trust in government and media is low.

3. Geopolitical Attack

A state or non-state actor may publish and distribute a false video targeting a politician or government in an attempt to influence the geopolitical situation. For example, a manipulated audio track or video might show a president ordering troops to prepare to invade a disputed territory, with the resulting immediate implications for international politics. This could lead the country holding the territory in question to react, even if the content either hasn't or can't be verified as a fake, due to a lack of the tools necessary to make such a determination. The window for verification in such a scenario would be very small, making the existence of effective technical detection mechanisms essential. In the aftermath of the explosion in Lebanon on 4 August 2020, caused by a large storage of ammonium nitrate,[8] cheapfakes were shared showing a missile hitting the port of Beirut. Michel Aoun, president of Lebanon, suggested at one point that the explosion could have been caused by just such an attack.[9] This shows that the capacity for effective manipulation in this area already exists, even if only through the use of cheepfakes. More developed video manipulation techniques are likely to make this problem even more complex.

[8] Nik Waters, "The Beirut Explosion – Is It a Bird? Is It a Plane? Is It a Video of a Faked Missile?", Bellingcat, 7 August 2020, https://www.bellingcat.com/news/mena/2020/08/07/the-beirut-explosion-is-it-a-bird-is-it-a-plane-is-it-a-faked-video-of-a-missile/

[9] Campbell MacDiarmid, "Beirut Blast: Missile May Have Been to Blame, Says Lebanon's President Michel Aon, The Telegraph, 7 August 2020, https://www.telegraph.co.uk/news/2020/08/07/beirut-blast-missile-may-have-blame-says-lebanons-president/

4. False Report by News Media

In this scenario, a mainstream news source provides reporting based on a deepfake. This may result from a one-off mistake, where journalists rely on manipulated content as their main source. In the digital era, news outlets are under pressure to publish information faster than ever. Not only could outlets be fooled, but their brand could be hijacked through the distribution of a deepfake video bearing the media outlet's logo. Additionally, news outlets may be expected to respond to a possible case of manipulated media with authoritative information regarding its validity. If different media outlets produce different reports – based on different determinations of the authenticity of a video, for example – this will only contribute to further confusion. Outlets with fewer resources may be more vulnerable to such attempts, due to a lack of human and technical capacity. False or inconsistent reporting by media outlets will clearly have a negative impact on their credibility. How serious this impact would be will depend the level of public trust in the media.

5. False Report by CSOs or human-rights groups

A CSO or human-rights group that uses video materials to draw attention to abuses could be fooled into making a report based on manipulated media. This is a particular danger for organisations who work with non-traditional data sources, including crowdsourced video and audio, and threatens to have a negative impact on their credibility with the public and their overall reputation. This could also reduce their ability to continue using image and video data until a verification solution can be developed.

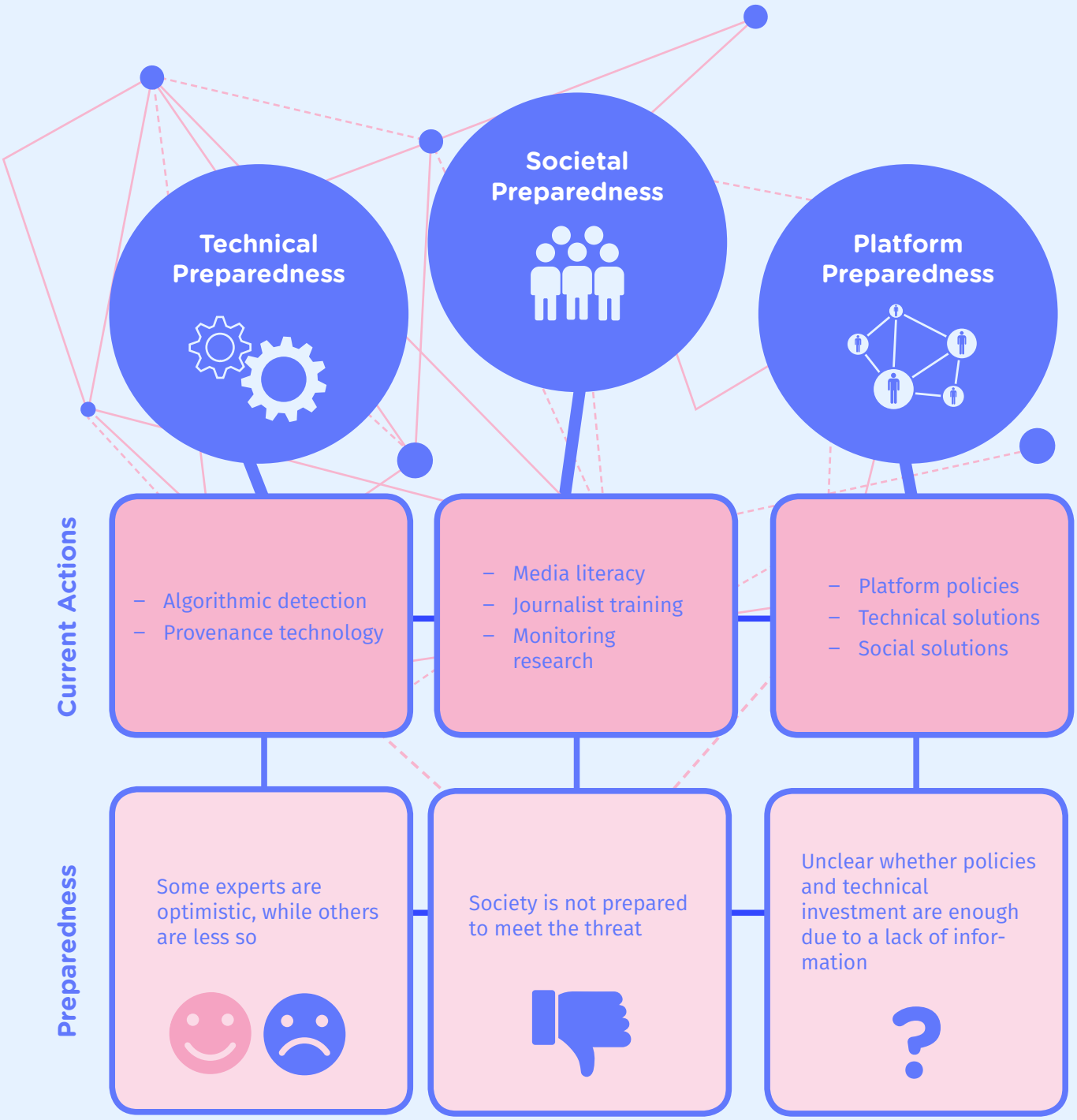
In terms of preparedness, representatives of some of the organisations interviewed highlighted the need for more training and tools, such as verification tools used by governments and police. In the even that deepfakes become more prevalent, these organisations will need to adapt in their work. Cheapfakes currently pose a low-level threat to such organisations, as they are often familiar with the source video, so spotting a fake is not too difficult. On the positive side, unlike news media, many of these organisations have more time to examine sources, which allows for a more thorough review.

6. Disbelief by default

In the long term, an accumulation of the issues listed above could result in a culture of disbelief by default. Unlike the other scenarios, this would have a long-term effect on society. When disbelief by default becomes a common attitude toward information, democratic discourse loses its grounding in commonly agreed facts. This includes an increased ability on the part of politicians to simply deny any proof of wrong-doing as manipulation.

IV. How prepared are we?

Whenever the deepfake threat ultimately materialises, experts agree that the level of preparation will be key to reducing the potential impacts. Technical, societal, governmental and platform preparedness will be vital in dealing with all of the above scenarios. The important question, therefore, is how are we preparing and how do experts assess the success of these initiatives so far.



How prepared are we technically?

Technical preparation will be needed to authenticate potentially manipulated content and/or provide users with information about the source of media. At present, there are two approaches being taken:

- **Algorithmic detection:** This is the use of machine learning techniques to detect whether or not content has been manipulated. Algorithms are being trained to learn from a dataset of videos. For example, they may learn to understand the specific manipulated qualities within the video (e.g., greyscale elements) or the behavior of a specific subject's face (e.g., how Angela Merkel blinks).

- **Provenance technology:** At the time any media product is created, additional information about its origin (e.g., time, date) would be included in the media product itself or added as metadata. Through the use of blockchain technology, an immutable cryptographic signature is created for any original piece of media.[10] This includes dozens of automatic checks to ensure the media product does not already exist with different data.[11] For more information on the specifics, see WITNESS' overview of provenance tools. This would make it easy to identify the source of any photo or video, and to detect any changes being made. This solution can be implemented either through software (e.g., an app) or hardware (e.g., hardcoded into a camera).

	Algorithmic Detection	Provenance Technology
Cons	– Not reliable to date – Will raise the expectation of proof – Will not solve the human factor (sharing without reading) – Niche solution that cannot be adopted by everyday people – Only useful when there is a continual line of ownership – Will not change the scale and speed that content circulates on platforms	– People probably will not use random apps, so the best solutions will be – hardcoded into cameras which requires time and industry collaboration – Requires collaboration with tech – companies as they currently strip – metadata from videos and image data – Cannot be used alone and social – solutions are just as important – Privacy concerns must be – accounted for
Pros	– Creating successful algorithms when the time comes might not be so difficult – The more deepfakes, the better detection algorithms will be	– Puts information into the hands of users

[10] Gregory, "Ticks or It Didn't Happen", op. cit., note 6.
[11] Ibid

Regarding algorithmic detection, many of the experts interviewed for this paper were skeptical of this approach, although others were more optimistic. The skeptics noted that such detection is not reliable to date, with a number of different algorithms in use. For example, the winning model in the Deepfake Detection Challenge competition, launched by Facebook in December 2019, managed only 65 per cent accuracy.[12] Without an extremely high rate of algorithmic accuracy, the risk of different reports on a material's authenticity is high, undermining trust in detection. Some of the experts also noted that algorithmic detection is a niche solution, and an expensive one that cannot be used by the average person. It will also raise the expectation on the part of the public that they have to verify any content, meaning the burden could fall on users to always have to determine what is authentic or not. It also will not solve the human factor, in the practice of sharing without reading. Further, making the code from detection models widely available poses the risk that nefarious actors can use it to game detection systems and improve their own deepfake models. To avoid this risk, further technical developments are typically published without the code in academic articles. This all makes algorithmic detection a rather high-level solution, and there is typically a continuous line of ownership of detection models (e.g., by one academic team or startup). The more optimistic assessments from the experts included the belief that developing an effective algorithm will not ultimately prove to be difficult. A few noted that the existence of more deepfakes in circulation could improve detection algorithms, as they would provide more video data as learning material. Nonetheless, they said it would be better to see such a situation in a lab setting first. These experts also stress that algorithmic detection would not have to be used by normal users to be effective. If platforms or news media used this approach effectively, this would allow them to remove deepfakes, or at least to slow their spread.

Provenance technology seems to be the preferred solution, although such technology will not be ready for years and there are a number of considerable downsides. A provenance solution would apply upstream, right at the source of production of a video or photo, rather than trying to detect a deepfake in a sea of information later (like the proverbial "needle in a haystack"). Implementing such a solution requires not only the development of the associated technologies, but also collaboration among many stakeholders to ensure the technology is implemented systematically. Such an approach will take a significant amount of time. Most experts agree that this technology would not be effective as an add-on software feature (such as an app that people would download for their cameras), but that it would have to be hardcoded into cameras to make it ubiquitous. Forensic experts pointed out that social media platforms currently remove metadata from image and video content, so platforms may need to be involved in development of this solution, to ensure it is successful. Such an approach also requires close work with civil society, in order to prevent unintended consequences and to address privacy concerns. WITNESS has summarised potential downsides of provenance solutions in their Ticks or It Didn't Happen report, and has provided a critical review of Adobe's Content Authenticity initiative. Included among the issues raised were that these tools may ultimately be used to surveil people, that too much trust might lie in blockchain rather than humans, and technical restraints may cause roadblocks in countries where these tools are most needed.[13]

The experts interviewed were in agreement that more time is needed to improve the effectiveness of both approaches. Several suggested that, at least for a while, these technical solutions would be outrun by deepfake innovations. It is not clear that these solutions could stop the scale and speed of distribution on social media platforms, as has been seen in attempts to deter other forms of disinformation.

A broad range of actors are leading the development of technical solutions. Academia and tech startups are leading the way on algorithmic detection. The Deepfake Detection Challenge is a notable multi-stakeholder initiative. With regard to provenance technology, Adobe and Microsoft are the leaders on this front, with new international startups also working towards a solution. Facebook also appears to be making some investment in provenance technology, via co-authorship with academia.

[12] Ather Fawaz, "Facebook Announces the Results of Its Deepfake Detection Challenge, Winner Hits 65% Accuracy", Neowin, 12 June 2020, https://www.neowin.net/news/facebook-announces-the-results-of-its-deepfake-detection-challenge-winner-hits-65-accuracy/
[13] Gregory, "Ticks or It Didn't Happen", op. cit., note 6.

Technical Solution**	Organisation	Organisation Type	Initiative	Description
Algorithmic Detection	UC Berkeley – Hany Farid	Academia	Research	Developing techniques to detect manipulated media, with a focus on content depicting world leaders. Manipulating videos of world leaders arguably poses the largest threat to democracy and society.
	Facebook, Amazon Web Services, Microsoft, Partnership on AI, Kaggle	Multi-stakeholder	Deepfake Detection Challenge	Companies created a dataset of over 1,000,000 videos and asked experts from around the world to benchmark their deepfake detection models, try new approaches and learn. Prizes were awarded to the top five winners, with the most accurate model reaching 65-percent accuracy.[14]
	Microsoft	Tech Company (United States)	Microsoft Video Authenticator	Can analyze a still photo or video to provide a percentage of confidence that the media has been artificially manipulated. For video, this includes a per-frame percentage as the video plays. The technology is based on subtle fading or greyscale elements that might not be detectable by the human eye.
	FakeNetAI*	Tech Company (United States)	Detection tool	Web app and application programming interface (API) that can scan content through a machine learning model trained to detect a variety of alteration techniques. The company is currently partnering with Facebook, the International Fact-Checking Network and Newsmobile.
	Sentinel*	Tech Company (Estonia)	Detection tool	Online upload and API that can scan content through a machine learning model trained to detect AI manipulation.
	Sensity*	Tech Company (The Netherlands)	Detection tool and database of detected deepfakes	Collects data from the open and dark web and provides a database of high-target industries and countries. The company also provides an API for face manipulation (images/videos) and GAN-generated faces (images).
	Quantum + Integrity*	Tech Company (Switzerland)	Detection tool	Online upload that can scan content through a patented machine learning model trained to detect AI manipulation.
Provenance Technology	Adobe (with the New York Times Company, Twitter, Truepic, WITNESS, CBC/Radio Canada and the BBC)	Tech Company leading a multi-stakeholder group	Content Authenticity Initiative	Goal is to create an industry standard for digital content attribution. This includes "a system to provide provenance and history for digital media, giving creators a tool to claim authorship and empowering consumers to assess whether what they are seeing is trustworthy".
	The BBC (with Microsoft, CBC/Radio Canada, the New York Times)	Multi-stakeholder approach	Project Origin	Multi-stakeholder development of a provenance solution to attach a digital watermark to media indicating when content has been manipulated. The goal is to include a message on the content or in the browser and, ultimately, contribute to an automated signal warning of false media.
	Facebook	Tech Company (United States)	Academic research	Facebook has made some investment in the development of provenance solutions through academic co-authorship and research. The amount of investment in and progress of such initiatives is unclear.
	Truepic	Tech Company (United States)	Control capture technology	Creates a cryptographic signature at the moment an image is created, which is then verified, stored and recorded as an immutable ledger. They have successfully integrated their technology into the hardware of a prototype mobile device.
	OARO*	Tech Company (Spain)	OARO Media provenance product	Uses blockchain technology to issue a certificate with user ID, content timestamp and GPS coordinates. Product targets home insurance claims.
	Axon Enterprise Inc.*	Tech Company (United States)	Police bodycams with provenance technology	Produces police bodycams for United States law enforcement agencies which includes blockchain technology authenticating media from the source.[15]
	Factom Protocol*	Tech Company (United States)	Provenance tool	Blockchain provenance solution to provide source data to media.
Algorithmic Detection and Provenance Technology	Amber Authenticate*	Tech Company (United States)	Detection tool and provenance software	Offers a software-based provenance solution with the Amber Authenticate product. Also offers a "signal processing" and AI tool to detect deepfakes when the source is unknown.

*DRI did not interview this group
** Note: List is not comprehensive and other actors may be missing

[14] Soumyarendra Barik, "Results from Facebook's Deepfake Detection Challenge Show How Difficult It Is to Detect Deepfake Videos", Medianama, 15 June 2020, https://www.medianama.com/2020/06/223-facebook-deepfake-detection-challenge/
[15] Lucas Cacioli, "Axon Explores Blockchain to Fight Body-Cam Deepfake Videos", Blockchain News, 4 October 2019, https://blockchain.news/news/axon-explores-blockchain-to-fight-body-cam-deepfake-videos

How prepared are society and democratic institutions?

Society

Reasons why society is not prepared

- Preparedness requires trust in media, which is difficult to guarantee
- Voters lack awareness on the issue
- Policymakers lack sufficient awareness of the issue to implement the necessary initiatives and investment

Societal preparedness will be just as important as technical solutions to ensure citizens' trust and resiliency. The previously described scenarios demonstrate that even if a video can be authenticated through technical solutions, people may not trust the verification body.

There was an overwhelming consensus amongst the experts interviewed that society is not currently prepared for the threat. Some demographics may be particularly vulnerable to the threat of deepfakes, especially where media literacy levels and trust in government are low. Journalists and media outlets are also important actors, and their skillsets must be strengthened. Governmental preparedness will also be key in cases of a political attack. This includes determining who will be the authoritative voice authenticating a video for the public, especially if the government itself is the target? It is unclear whether governments have action and response plans to prepare for such scenarios. Some of the experts interviewed believed that governments are still in the process of familiarising themselves with this issue.

A number of initiatives are bringing stakeholders together and sharing knowledge with relevant groups. Many such initiatives include the training of journalists or newsrooms in general.

Organisation	Initiative	Target Audience	Aim	Tools
Carnegie Endowment for International Peace	Carnegie Silicon Valley and FinCyber Project	AI industry, SM platforms, researchers, politicians	To identify high-probability threats and interventions, and to build channels engaging policymakers and technologists	Joint problem-solving and independent analysis
WITNESS	Prepare, don't panic	Communities at risk in the global south, the United States and, to a lesser extent, the EU and United Kingdom	To address the issue of deepfakes, to avoid panicked reactions, and to address communities at risk	Meetings, workshops, internet website targeting communities in the global south, mainstream and civic journalists; work also with researchers and tech platforms
Partnership on AI	Manipulated media work	Affecting practitioners (not only large technology platforms, but informing the academic community and policymakers).	Closing the gap of understanding between tech companies and society and shaping how people think about it. They also work on understanding user behavior in platforms, and how they react and consume disinformation	Workshops, publications, Steering Committee to work together
Reuters (and Facebook)	Reuters E-Learning Course (in collaboration with the Facebook Journalism Project)	Newsrooms worldwide	Helping newsrooms identify deepfakes and manipulated media	Free online training course
MIT Media Lab*	Detect Fakes Challenge	Internet users Internet users Internet users (focus on Europe)	Challenging users to spot the difference between real and AI manipulated video as a media literacy tool, as well as research.	Online challenge
Microsoft (and the University of Washington, USA Today and Sensity)	SpotDeepfakes.org	TikTok users	Providing a media literacy tool to teach users how to spot deepfakes.	Online quiz
Centre for Research and Technology Hellas, MODUL Technology, Universitat de Lleida, Exo Makina, WebLyzard Technology Gmbh, Condat AG, APA-IT, Agence France-Presse and Deutsche Welle (Funded by the European Union)	InVid Verification Plugin		Offering a plugin to help journalists verify content on social networks by quickly providing contextual information about a video or image. The tool has a number of features, including reverse image search, magnifying lens and accessing metadata about the content.	Web plugin
TikTok (and the National Association for Media Literacy Education)	"Be Informed" series		Media literacy campaign targeting users in their feeds, in the style of TikTok videos. One topic included, "question the graphics", challenges users to analyze images.	In-app media literacy campaign

*DRI did not interview this group
** Note: List is not comprehensive and other actors may be missing

How prepared are social media platforms?

Social Media Platforms

Unanswered questions

- Are technical capacities strong enough to enforce policy fast enough?
- Is removing of labelling content a better approach?
- Will platforms make enforcement data available to researchers?
- Are other issues more pressing and worth investing in?

More than any other actors, social platforms are critical to reducing the speed of distribution of harmful deepfakes. On the technical side, they will need to successfully and quickly detect manipulated media. On the social side, they need to enforce an effective policy that respects freedom of speech, while also preventing the spread of disinformation.

Many platforms have already started addressing the threat by implementing new policies over the course of 2020. Some platforms directly remove content, while Twitter, for example, labels some content. To enforce these policies, companies appear to use a combination of algorithmic detection and human involvement.

Company**	Policy	Criterion	Response Mechanism
Facebook	Manipulated Media policy	Media that have been edited or synthesised – beyond adjustments for clarity or quality – in ways that aren't apparent to an average person and would likely mislead someone into thinking that the subject of the video said things that they did not actually say. And: The item is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic.	Content is removed (Some content that does not meet the threshold will go to fact checkers). Manipulated videos that do not meet this standard are generally eligible for fact-checking and receive a specific rating for "altered" content. This policy complements other community standards (e.g., bullying and harassment, graphic violence), so if a video does not meet the Manipulated Media policy, it may be removed under a different policy.
	Altered media policy (for fact checkers)	Third-party fact-checkers have the option to rate content as "altered", meaning image, audio, or video content has been edited or synthesised beyond adjustments for clarity or quality, in ways that could mislead people	Some actions may include reduced distribution, sharing warning, sharing notifications, misinformation labels and removing incentives for repeat offenders
Twitter	Synthetic and manipulated media policy	The content is significantly and deceptively altered or fabricated	Content may be labeled
		Content is shared in a deceptive manner	Content may be labeled
		Content is likely to impact public safety or cause serious harm	Content is likely to be labeled or may be removed. (Note: Other Twitter rules might already apply)
		Content is significantly and deceptively altered or fabricated AND shared in a deceptive manner	Content is likely to be labeled.
		Content is significantly and deceptively altered or fabricated AND shared in a deceptive manner AND likely to impact public safety or cause serious harm	Content is likely to be removed.
TikTok	Synthetic or manipulated content policy	Prohibits synthetic or manipulated content that misleads users by distorting the truth of events in a way that could cause harm.	Uses a mix of technology and human moderation to enforce these policies, including by removing content, banning accounts, and making it more difficult to find harmful content in recommendations or search.
YouTube*	Manipulated Media policy	Content that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm.	Content is removed.

*DRI did not interview this group
** Note: List is not comprehensive and other actors may be missing

On the technical question of enforcing these policies, some platforms are investing in solutions. Facebook, for example, contributed to the Deepfake Detection Challenge to develop a better algorithmic detection solution. As mentioned previously, the winning model in the challenge achieved only a 65-percent accuracy level.[16] They are also partnering with a startup called FakeNetAI, which offers an algorithmic detection solution. Recently, they called for further cross-sector collaboration and investment in technical research on manipulated media and deepfakes.[17] Twitter has partnered with Adobe to enable an opt-in provenance system to include attribution data to content.[18] It is unclear how, and how much, companies are investing on the technical side beyond these public initiatives. Additionally, it is unclear whether these technologies are effective in addressing both deepfakes and cheapfakes.

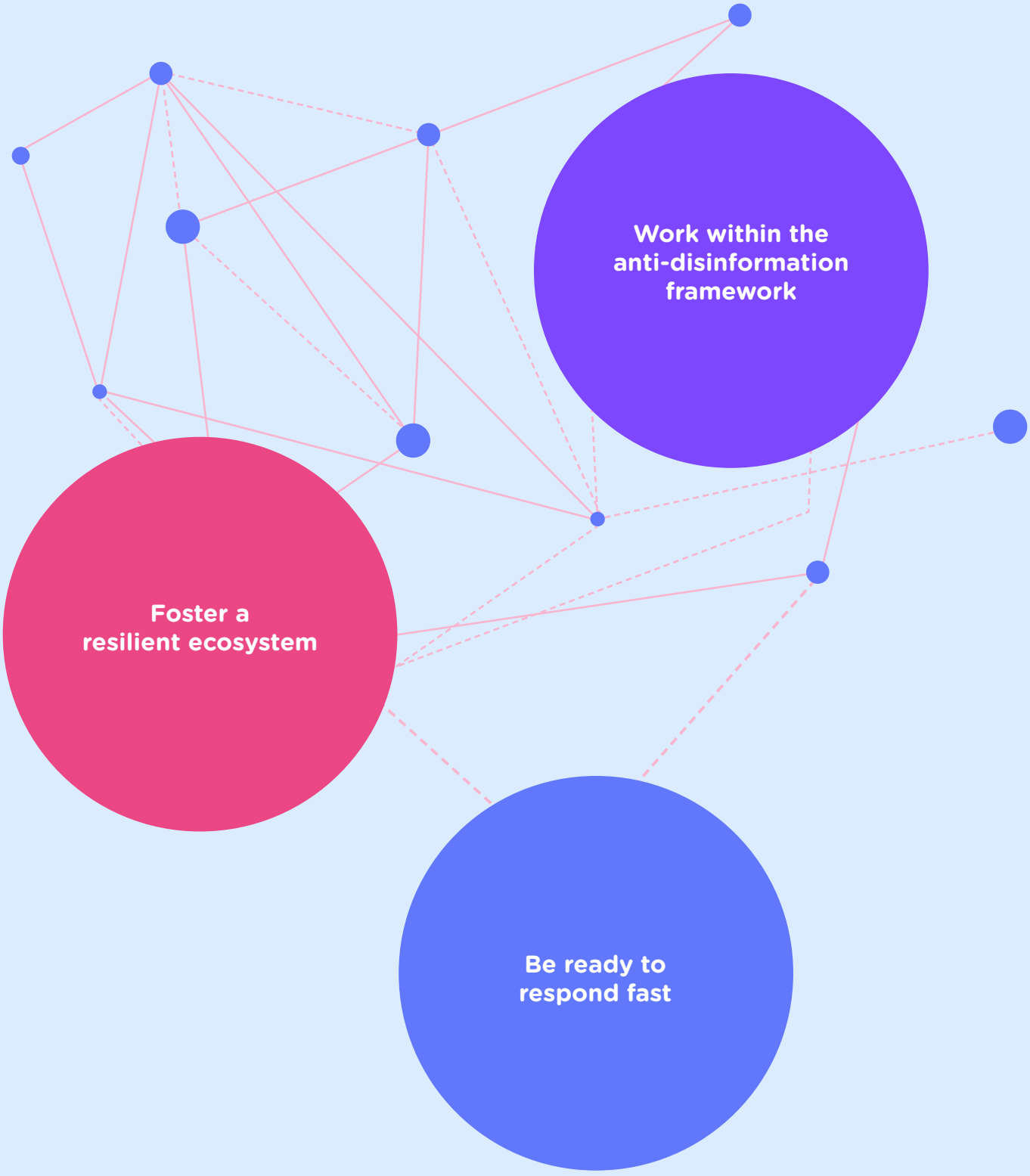
Is the current level of technical investment and development of new policies enough? It is unclear whether either of these have been tested, given that experts have identified no deepfakes so far. It is difficult to evaluate the success of these policies without access to policy enforcement data, via transparency reports. For policy enforcement, it is also unclear whether some policies introduced by companies to address the deepfake threat will overlap with or unnecessarily duplicate other existing policies in practice. For example, for platforms that already have policies in place related to misinformation, a harmful political deepfake might already be covered under these pre-existing policies. It also would be interesting to understand how policies that apply to manipulated media are being enforced and the quantity of content platforms are seeing.

Private messaging apps (e.g., WhatsApp, Telegram, Facebook Messenger) do not appear to have any relevant detection or enforcement policies in place. These platforms could be ripe environments in which unchecked disinformation could spread.

[16] Soumyarendra, "Results from Facebook's Deepfake Detection Challenge Show How Difficult It Is to Detect Deepfake Videos", op. cit., note 14.
[17] Nathaniel Gleicher, Recommended Principles for Regulation or Legislation to Combat Influence Operations, Facebook, 8 October 2020, https://about.fb.com/news/2020/10/recommended-principles-for-regulation-or-legislation-to-combat-influence-operations/
[18] Ina Fried, "Adobe, Twitter, NYT Launch Effort to Fight Deepfakes", Axios, 4 November 2019, https://www.axios.com/adobe-twitter-nyt-launch-effort-to-fight-deepfakes-4a865394-764f-418d-a444-bb0fe9bebe18.html

V. What should be done next?

Reflecting on the different perspectives and experiences in this field, we divided the recommendations into three blocks of different approaches, acknowledging that the fight against any form of disinformation should be a societal effort, focusing on short term answers but building resilience in the longer term.



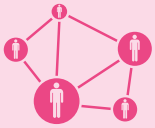
A. Work within the anti-disinformation framework:

The challenges posed by deepfakes relate to the work against disinformation in general, and other technological advances in audio, text, video and other forms of media manipulation can be abused in different ways. In that sense, constantly adapting and perfecting existing frameworks should be encouraged. This can be done by:

	**Increase knowledge on platform's response:** We need a better understanding of the prevalence of manipulated materials on social media platforms and how platforms are implementing their relevant policies. Therefore, transparency requirements should be legally mandated, so researchers and the broader public can better understand this issue.
	**Establish a framework of discussion:** The European Commission should publish a communication laying out its position on deepfakes in more detail, in order to establish definitions and language for a European discussion.
	**Mandate cross-platform cooperation:** The threat of deepfakes should be taken as seriously as other threats, such as child pornography or terrorism (deepfakes may play a role in relation to these issues as well). Legislation should mandate collaboration across social media platforms to identify and respond to potential deepfake threats.
	**Public consultation:** Deepfakes will affect everybody, and public political discourse as a result. Therefore, it should not be left to private sector actors alone to address the threat. There should be a public discussion of the issue and consultations on possible responses. These will indirectly serve to bolster public media literacy.

B. Foster a resilient ecosystem

As with existing forms of disinformation, the more prepared actors in civil society are to identify and counter manipulation attempts, the less the chance they will be surprised and affected by them. Protecting ourselves from future disinformation threats can be achieved by making our societies more resilient, which can be achieved through:

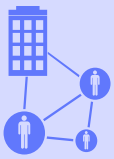
	**Shared detection responsibilities:** Collaborative efforts like the Deepfake Detection challenge should be fostered, as detection rates are low and need to be improved. The related technologies need to be shared with expert organizations (including forensics training for CSOs and journalists). Enhancing these skills to react quicker in identifying and debunking manipulated media can help increase resilience and trust in the long run.
	**Going beyond technical solutions:** Cross-sector collaboration is needed. Bringing together media experts, representatives of civil society and policymakers to discuss potential responses to threat scenarios is essential to preparing for potential threats. Platforms should coordinate policies and responses when it comes to the deepfake threat. Standardized rules and responses will help to identify such content and to react in a more coordinated manner.
	**Sustained funding to civil society & research in the EU:** Strengthening civil society to work on the impacts of AI on society is key. Providing funding, training and exchanges among different stakeholders can help create a critical mass of understanding, similar to what the Partnership on AI is doing. There must be a forum for discussions on how the development of new technologies can impact society in Europe, in parallel to a Silicon Valley perspective.
	**Media literacy:** Programmes need to focus on raising awareness on different forms of manipulated media, from cheapfakes to deepfakes. The content of such programmes should not only look back at traditional disinformation strategies, but also prepare for those of the future.

C. Be ready to respond fast

The most successful disinformation campaigns have occurred when everyone was caught by surprise. We are at a stage where we know that actors with malicious intentions and resources exist, so there is a need to prepare scenarios and processes to react to malicious uses of new technologies. For this, a rapid response is crucial, and can be achieved through:

Paying attention to virality at key moments: Slowing the distribution of potentially harmful deepfakes before it is fact-checked or authenticated would reduce the post reach. This could be done by linking content virality to a platform's content moderation practices. Such an approach would be especially important during elections or in crisis situations.

Sharing threat signs amongst platforms: The sharing of information between tech companies can help in tackling a potential deepfake threat scenario. The coordination of malicious activities happens in different forums and different platforms, and companies enforce their standards differently. A code of conduct on countering malicious uses of new technologies across different platforms, including what to do with reported deepfakes, can help shape a single, standard approach and create protocols on how to deal with this threat.



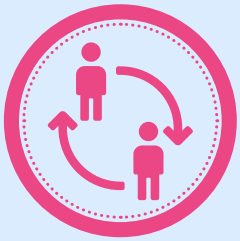
Involving stakeholders: Informing intelligence services, expert civil society and media about these threats can also support better coordination and understanding of the issue among the public, being mindful not to create panic.

Developing action plans and scenarios on how to deal with deepfakes. Define steps to avoid unintentionally amplifying their effects and define a body with the authority to check it. Cooperation between news organizations is also important to avoid approaching the matter in a sensationalized manner, thus creating a culture of disbelief by default.

VI. Conclusion

The threat posed by deepfakes is not limited to direct harm to an individual, group or organisation. As the threat scenarios show, there is a greater fear that new technologies can do even more harm to the relationship between citizens and information, blurring the lines between what is true and false and decreasing trust in journalism, facts and institutions even further. In that sense, some aspects of the solution are linked to the overall actions needed to tackle disinformation as a societal problem in general, while others are more specific to threats related to the use of AI in content manipulation.

Two structural recommendations are central to actions by all stakeholders:



Collaboration is key: Trust in institutions and journalism has declined over recent years, and societies have grown increasingly polarised – there is no social agreement on what acceptable and legitimate discourse is. In this type of environment, disinformation thrives, and new technologies can widen the divide between citizens and facts even further. Only a strong ecosystem that follows developments closely and informs both public discussions and regulatory approaches can create societal resilience. General suggestions to achieve this include the strengthening of multi-stakeholder and politically diverse dialogue and working together on contingency plans in case deepfakes are published: How should the media report on instances of manipulated media? Is there a protocol to follow? What about the victim? How should politicians react? We already know enough to not be caught by surprise.

Closing the knowledge gap: One of the worrying trends identified was the increased gap between technology development and skills to monitor and identify new forms of manipulated and AI-generated content. In recent years, different sectors have had to catch up in their understanding of risks posed by new technologies, in order to be able to respond to them. When it comes to the risks posed by deepfakes, closing the knowledge gap among different communities is essential to reducing the risk of this technology being used to harm public discourse. Media organisations need to improve their detection skills so that they don't report and amplify such content; civil society needs to be ready to identify and measure the impact on human rights and democracy; researchers need to develop new detection techniques, as well as to understand the impacts of such technologies on society; and policymakers need to be able to understand such uses to come up with meaningful legislation, where necessary.