UNIDIR UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH

the **2021**

innovations dialogue.

DEEPFAKES, TRUST & INTERNATIONAL SECURITY

CONFERENCE REPORT

ACKNOWLEDGEMENTS

Support from UNIDIR's core funders provides the foundation for all the Institute's activities. The 2021 Innovations Dialogue was the third edition of one of UNIDIR's flagship events organized by its Security and Technology Programme, which is funded by the Governments of Germany, the Netherlands, Norway, and Switzerland, and by Microsoft.

UNIDIR would like to thank all the speakers, moderators and participants for their presentations, interventions, and engagement in the discussions at the 2021 Innovations Dialogue, which have contributed to this report.

The Conference Report has been authored by Alisha Anand, Research Assistant and Belen Bianco, Graduate Professional in the Security and Technology Programme.

ABOUT UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR) is a voluntarily funded, autonomous institute within the United Nations. One of the few policy institutes worldwide focusing on disarmament, UNIDIR generates knowledge and promotes dialogue and action on disarmament and security. Based in Geneva, UNIDIR assists the international community to develop the practical, innovative ideas needed to find solutions to critical security problems.

CITATION

A. Anand and B. Bianco, *The 2021 Innovations Dialogue Conference Report: Deepfakes, Trust and International Security*, Geneva, Switzerland: UNIDIR, 2021.

NOTE

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views expressed in the publication are the sole responsibility of the individual authors. They do not necessarily reflect the views or opinions of the United Nations, UNIDIR, its staff members or sponsors.

ABOUT THE SECURITY AND TECHNOLOGY PROGRAMME

Contemporary developments in science and technology present new opportunities as well as challenges to international security and disarmament. UNIDIR's Security and Technology Programme (SecTec) seeks to build knowledge and awareness on the international security implications and risks of specific technological innovations and convenes stake-holders to explore ideas and develop new thinking on ways to address them.

ABOUT THE AUTHORS

Alisha Anand is the Research Assistant for Security and Technology at UNIDIR. Her areas of expertise include new and emerging technologies, arms control, and international security. Before joining UNIDIR, she worked on non-proliferation and export controls with the Disarmament and International Security Affairs Division of the Indian Ministry of External Affairs, the Manohar Parrikar Institute for Defence Studies and Analyses and the Federation of Indian Chambers of Commerce & Industry. She holds a Master's degree in Law and Diplomacy from the Fletcher School, Tufts University, where she specialized in international security and international law. Follow Alisha on Twitter @AlishaAnand912.

Belén Bianco is a Graduate Professional for the Security and Technology Programme at UNIDIR. She specializes in new technologies, non-proliferation, and international security. Previously, she worked at the Center for Security and Emerging Technologies, the Center for Security Studies at Georgetown University, and the Argentine Undersecretariat of Nuclear Energy, where she represented Argentina in nuclear technology negotiating platforms such as the Nuclear Suppliers Group. She also interned at the United Nations Office for Disarmament Affairs in New York and the Arms Control Association. Belén graduated from Georgetown University with a MA in Security Studies and a Graduate Certificate in Diplomatic Studies.

CONTENTS

ABOUT THE INNOVATIONS DIALOGUE

Launched in 2019, the Innovations Dialogue is one of UNIDIR's flagship events. The conference series was established pursuant to the 2018 General Assembly resolution[1] on the role of science and technology in the context of international security and disarmament. The Innovations Dialogue provides a unique multi-stakeholder forum—convening experts from the diplomatic and policy community, technical and scientific community, industry groups, and academia and civil society—to collectively examine developments in science and technology that have potentially radical and novel implications for international peace and security and disarmament. Through fact-based and balanced discussions, the dialogue aims to dispel myths about scientific and technological innovations and to build a shared understanding of the potential benefits, risks and policy challenges posed by such innovations.

The United Nations Secretary-General's May 2018 Agenda for Disarmament, *Securing Our Common Future*, and his 2018 and 2021 report on "Current Developments in Science and Technology and their Potential Impact on International Security and Disarmament Efforts" (A/73/177 and A/76/182) recognize UNIDIR's role as a source of knowledge and ideas, as well as a convener of multi-stakeholder dialogues, at the nexus of technology and security.[2]

The key objectives of the Innovations Dialogue are:

- **To collaboratively examine beneficial applications as well as new and converging challenges or risks** presented by advances in science and technology for international peace and security.
- **To promote multi-stakeholder engagement and to build new relationships** among a range of actors and tools that can contribute to mitigating potential harms, harnessing potential benefits, and promoting responsible innovation.
- **To explore how multi-stakeholder dialogue can facilitate policy responses** to developments in science and technology that have potentially radical and novel implications for international security and disarmament, with a view to identifying gaps or opportunities where early thinking on strategies for risk mitigation may be beneficial.

1 UNGA (2018b).
2 UNGA (2018a); UNGA (2021).

THE 2021 INNOVATIONS DIALOGUE

UNIDIR Director Dr. Robin Geiss delivering opening remarks at the 2021 Innovations Dialogue.

The world today is suffering from a 'trust deficit disorder', as noted by United Nations Secretary-General António Guterres, where trust among States, in institutions and the rules-based global order has weakened.[3] While some artificial intelligence (AI)-driven technologies could offer new tools to enhance trust, in a world where digital forms of communication are ubiquitous, the emergence of AI-enabled digital media fabrication and manipulation technology could further subvert trust at the individual, institutional and societal levels by providing new tools to generate malicious hyper-realistic synthetic media known as 'deepfakes'. Against the backdrop of growing concerns around the spread of false information and its disruptive consequences for societies and the stability and trustworthiness of the digital ecosystem, **the 2021 Innovations Dialogue unpacked the growing deepfake phenomenon and explored its implications for trust and international security and stability**.

Bringing together 20 expert speakers[4] **from government, international organizations, academia, and industry, and nearly 1,000 (virtual and in-person) participants** from around the world, the Dialogue illuminated how algorithmically generated synthetic media is created and disseminated, and how it could erode trust and present novel risks for international security and stability. The discussions also explored the key governance issues concerning deepfakes and the technical countermeasures and policy responses by which the technology's dangers could be addressed. Finally, the Dialogue reflected on how the international community can preserve and foster trust in the digital ecosystem going forward.

This report provides a summary of the key themes, issues, and takeaways that emerged from the 2021 Innovations Dialogue. Part 1 of the report introduces the growing deepfake phenomenon and provides an overview of its implications for trust, international security, and stability. It also considers how the international community can preserve and foster digital trust in the era of rapid digital transformation. Part 2 of the report explains the fundamentals of visual and textual synthetic media technology and examines their attendant technical countermeasures and governance responses.

---

3 United Nations Secretary-General (2018).
4 Geographic diversity: 15 nationalities; gender balance: female: 10, non-binary: 1 and male: 9.

Giacomo Persi Paoli, UNIDIR

Nina Schick, Tamang Ventures


Panel on Understanding the Implications for International Security and Stability

Izumi Nakamitsu, UN High Representative for Disarmament Affairs


Behind the scenes at the 2021 Innovations Dialogue


Participants at the 2021 Innovations Dialogue


Kaja Ciglic, Microsoft; Amandeep Gill, I-DAIR; Robin Geiss, UNIDIR

Arthur Holland Michel, UNIDIR; Laura Ellis, BBC; Giorgio Patrini, Sensity

HIGHLIGHTS

- Increasingly powerful deep learning algorithms accompanied by the rapid advances in computing power have enabled the generation of **hyper-realistic synthetic media; malicious synthetic media is commonly referred to as 'deepfakes'**.
- **Deepfakes include all forms of digital content—video, text, images, and audio**—that have been either manipulated or created from scratch using deep learning algorithms to primarily mislead, deceive or influence audiences.
- By portraying someone doing something they never did or saying something they never said, increasingly sophisticated **deepfakes could challenge and influence perceptions of reality**.
- The fabrication and manipulation of digital content is not a new phenomenon. **The growing deepfake phenomenon however represents a significant leap forward** from what has come before primarily because: a) the fidelity of synthetic media is unmatched; b) along with manipulation, synthetic content that did not exist before can be fabricated; c) synthetic media technology allows the manipulation or fabrication of all forms of digital media, not just images; and d) sophisticated synthetic media generation is becoming increasingly accessible through the emergence of user-friendly software tools and services.
- **Synthetic media technology is not inherently malign**, it has many beneficial applications across social and economic sectors ranging from advertising and education to fashion and entertainment.
- **Deepfakes are however emerging against a backdrop of growing trends towards the deliberate spread of false information and declining trust** in institutions and among actors in the international system.
- **As hyper-realistic false, misleading or malicious content, deepfakes have the capacity to intensify the erosion of norms related to truth and trust** at an individual, organizational and societal scale.
- **Deepfakes perpetuate the 'liar's dividend'**—in the era of truth skepticism, the mere fact that deepfakes exist could undermine even what is in fact true or authentic.
- From an international security and stability perspective, **ready access to increasingly sophisticated deepfake technology could lower the barriers to weaponizing information and delivering tailored harm or disruption** in society as well as in the political and military spheres.
- **Many technical countermeasures and policy approaches at industry, national and regional levels are emerging** to respond to the multifaceted risks posed by deepfakes, including media provenance solutions, deepfake detection tools, regulation across the deepfake life cycle, and media literacy.
- **The key question for the international peace and security community is how it could leverage and bolster technical countermeasures and governance approaches** to effectively address the risks presented to international security and stability.

1.1 Trust and International Security in the Era of Deepfakes[5]

Trust underpins every action, relationship, and transaction in society whether at the individual, communal, institutional, or inter-governmental level. In **international relations, trust is a central pillar of international security and stability**. It sets the foundation for cooperation, institution-building and multilateralism, and rests on the ability of actors to learn, perceive, and believe others' interests and actions through means of communication.

In the digital age, the rapid emergence and adoption of digital technologies in core societal functions, while having transformative benefits, is challenging traditional standards of whom and how we trust. Notably, the digitization of information and the ubiquitous digital means of rapid communication have facilitated the spread of false and misleading content, engendering disruption and distrust in society. In the current context, the widespread dissemination of false and even malicious information surrounding the COVID-19 pandemic has shown the extent to which it can hamper the ability of governments and frontline actors to manage and respond to crisis situations.[6]

The fabrication and manipulation of digital content is not a new phenomenon. The attention it has witnessed in recent years is owed to the ongoing revolution in AI that is not only changing the way that humans do commerce and communicate, but also the way that humans think about what is real or not. **Increasingly powerful deep learning algorithms**—a subset of machine learning—accompanied by **rapid advances in computing power have enabled the generation of hyper-realistic synthetic media; malicious synthetic media is commonly referred to as "deepfakes"** (a portmanteau of *deep* learning and *fakes*). Deepfakes include all forms of digital content—video, text, images, or audio—that have been either manipulated or created from scratch using deep learning algorithms to primarily mislead, deceive, or influence audiences.[7] In particular, deepfake technology is good at creating fake or synthetic media of people. Not only can it realistically impersonate people that exist (even if that person is already dead), but also create synthetic people that do not. In this way, **by portraying someone doing something they never did or saying something they never said, increasingly sophisticated deepfakes could challenge and influence perceptions of reality**.[8]

This ability of AI to generate synthetic media that is difficult to distinguish from reality is a new development that has only been within the realm of the possible for the last five years. The first deepfake video appeared in 2017,[9] when some Reddit users created fake pornographic videos by superimposing celebrities' faces onto real pornographic videos using open-source machine learning tools. **Since then, deepfakes have received significant attention due to the steady rise in their quality, quantity, and variety.**[10] According to Sensity, a platform that

5 This section is based on the keynote address delivered by Nina Schick, Director of Tamang Ventures; Schick (2021).
6 Butcher (2021).
7 Collins (2019).
8 Davis (2020).
9 Cole (2018).
10 Castro (2020).

tracks and detects deepfakes, the number of deepfake videos identified online has been doubling every six months since 2018.[11]

Media manipulation has a long and prolific history as a powerful tool to shape the collective human perception. This precedes digital media and has always played a role in politics and international relations, helping to shape geopolitical narratives. For example, in the twentieth century, photo doctoring was used to influence perceptions in a campaign of political repression. At the time, this was a laborious task that required careful labour by skilled artisans.[12] **Digital technology, however, has made media manipulation cheaper, easier, and more accessible.** In the 1990's an array of digital content editing tools surfaced including Adobe Photoshop and other photo and video editing smartphone applications.[13]

The growing deepfake phenomenon represents a significant leap forward from what has come before for primarily four reasons. First, the "fidelity of synthetic media is unmatched".[14] AI techniques underlying the generation of synthetic media have the ability to mimic human biometric indicators such as voice and facial features. As generation techniques become increasingly sophisticated, it will become nearly impossible for humans to distinguish between authentic and synthetic content. Second, synthetic media is not merely the manipulation of media. It also entails the fabrication of digital content that did not exist before, including the generation of entirely digital synthetic humans. Third, the fabrication and manipulation of digital content extends beyond images to video, text, and audio. AI techniques can already generate synthetic forms of all types of digital content. Finally, synthetic media generation is becoming universally accessible through the emergence of readily available and user-friendly deepfake software tools and services.[15] Although currently such easy-to-use deepfake applications are limited in their functionality, and the ability to generate highly sophisticated deepfakes is still largely in the hands of State actors and well-resourced groups or individuals. Nevertheless, this "limited functionality only exists due to technical constraints rather than ethical concerns".[16] As AI techniques underlying deepfakes become more sophisticated, these technical constraints are expected to be overcome.

Experts suggest that we are rapidly moving towards a synthetic future where 90–95 per cent of video content online will be synthetically generated by the end of the decade.[17] By 2030, individual content creators on platforms like YouTube will have the ability to generate high-quality synthetic content on par with special effects that we see in big-budget movies today. As a result, the popularity of deepfake applications will also grow. For example Reface, a face swapping deepfake application, has become one of the most downloaded applications around the world, and in its 18 months of existence its users have already generated over 3 billion videos.[18] The increasingly ubiquitous tools to generate hyper-realistic forgeries of all

11 De Saulles (2021).
12 Blakemore (2020); Schick (2021).
13 Schick (2021).
14 Schick (2021).
15 Collins (2019).
16 Schick (2021).
17 Schick (2021).
18 Schick (2021).

forms of digital media could bring a dramatic revolution in our information ecosystem, as digital media is becoming a critical medium for human communications. **This trend is further fuelled by the ever-growing unprecedented power of the Internet and social media platforms to effectively, rapidly, and virally disseminate digital content.**

Synthetic media generation technology is not inherently malign, it has many beneficial applications across social and economic sectors ranging from advertising, fashion, and entertainment to education and healthcare.[19] However, deepfakes are emerging in the context of the growing weaponization of information and the disruptive consequences of modern information warfare for the stability, integrity, and trustworthiness of institutions, the information ecosystem, and society more broadly. **Deepfakes have the capacity to intensify the erosion of norms related to truth and trust** at an individual, organizational, and societal scale. Furthermore, the fact that deepfakes first surfaced in the form of non-consensual pornography not only shows how malicious synthetic media can be a gendered phenomenon especially targeting women, but also that the unique power of AI techniques to mimic real humans is non-consensual.[20]

In context of trends such as the declining trust in institutions and among actors in the international system, **from an international security and stability perspective, deepfakes in combination with other disruptive digital technologies can be especially destabilizing**.[21] Ready access to increasingly sophisticated deepfake technology could lower the barriers to weaponize information and broaden the range of State and non-State actors that can engage in sophisticated disinformation campaigns and influence operations. **Deepfakes could also deliver tailored harm or disruption.** They provide novel tools for the conduct of malicious use of information and communication technologies (ICTs), particularly for social engineering and spear-phishing attacks.[22] In the political sphere, they could be used to influence public opinion, defame political leaders and spark violence and social unrest.[23] In an international military crisis, fabricated and manipulated media could be used for deception and degradation of enemy's situational awareness which may lead to escalation and miscalculations in high-pressure scenarios.[24] Even in peacetime, among other things, recent examples have shown how deepfakes could be used for deception during high-level political talks and negotiations.[25]

One of the most profound implications of deepfakes for trust however is a phenomenon referred to as the 'liar's dividend'[26]– **in the era of truth scepticism, the mere fact that deepfakes exist could undermine even what is in fact true or authentic**. Put simply, everything can be dismissed as fake. For example, bad actors wanting to dodge responsibility for their words or actions could denounce authentic content that may be incriminating as deepfakes. In this

[19] Boneh et al. (2020); Heaven (2020); Chitrakorn (2021).
[20] Schick (2021).
[21] Smith & Mansted (2020).
[22] BBC (2019b).
[23] BBC (2019a).
[24] Williams & Drew (2020).
[25] Cloutier (2021).
[26] Citron & Chesney (2019).

way, deepfakes can further perpetuate the subversion of trust that is the backbone of a stable society and international system. A recent alarming manifestation of this concerns claims that the video of George Floyd's death was a deepfake.[27] Although, in this particular case this claim did not garner widespread public support, it is unclear whether such certainties will hold true in the public's mind in the next decade, when our already noisy information ecosystem is even noisier with the emergence of synthetic media.[28] **Such developments will raise fundamental philosophical questions about how we interpret the nature of reality.** If there will be no way to distinguish between the authentic and the synthetic, especially in the digital realm that we increasingly inhabit, then how will we know what is real? "How can we ensure trust in human interactions that define commerce, politics, security, and society, if we cannot establish some layer of universally held truths or realities?"[29]

The advent of AI-enabled synthetic media technology highlights the necessity of fostering trust and protecting shared standards of truth in an increasingly digitized world. To this end, many efforts are being made to develop policy responses and technical countermeasures at both the industry and government levels to respond to the multifaceted risks posed by malicious synthetic media,[30] including media provenance solutions, deepfake detection tools, and regulation across the deepfake lifecycle.[31] **The key question for the international peace and security community is how it can leverage this emerging web of industry-led and national and regional government initiatives** and, if necessary complement it with multilateral governance responses, to effectively address the risks presented to international security and stability.

Although there are currently few public examples of large-scale harm, as with many emerging technologies, **the rate at which synthetic media technology is advancing is outpacing our ability to understand its implications and the required measures to respond to them**. At this stage, the policy and technology communities need to study the growing deepfake phenomnon and consider how the emerging web of technical and governance measures can be bolstered to harness the opportunities and contain the risks it presents. To this end, the 2021 Innovations Dialogue provided a multi-stakeholder forum to collectively demystify deepfakes, consider their attendant countermeasures, and examine their implications for trust, international security, and stability.

[27] Derysh (2020).
[28] Schick (2021).
[29] Schick (2021).
[30] Collins (2019).
[31] Van Huijstee et al. (2021).

1.2 Implications for International Security and Stability[32]

This panel examined how malicious synthetic media could potentially erode trust in international relations, and present novel risks for international peace, security, and stability.

Trust is the foundation on which most of our relations and communications in the information age are built, whether interpersonal, communal, societal, or international. Our ability to trust the verbal or nonverbal information that we receive, including its source, is a condition on which we are able to interact, trade, or make decisions in the digital age. Often information is in the form of data points which are easy to trust—the cost of an item is the cost of an item. The challenge however comes in where "you can have a layer of disassociation from a pure, simple scientific fact"[33]—that is where trust is vulnerable to breaking down even without the existence of deepfakes. And the anonymity of the online world only further enables this disassociation. Trust can nevertheless come from different sources—legitimacy through democratic processes or legitimacy through social cues that help one to determine the trustworthiness of a person or an information source.[34]

Where the rise of deepfakes comes into play and challenges our shared concepts of truth and trust is it perpetuates the 'liar's dividend'—not only can it make people believe things that are completely false, but it can also undermine trust in things that are otherwise objectively true. In this way, synthetic media can falsely cater to our very human belief that 'seeing is believing' or 'I heard it with my own ears'. In fact, in the age of synthetic media, the reverse of 'seeing is believing' is also true: "believing is seeing"[35]—disinformation campaigns generally target people's prejudices and if the content is in line with what people want to think, then they are less likely to question its authenticity. **Essentially any action, relationship, decision-making process or transaction that relies on the trustworthiness of information**[36] and trust between States, institutions, communities, or individuals is potentially vulnerable. Research suggests that those who engage with synthetic media or deepfakes on social media platforms become "highly sceptical of all kinds of information".[37] **If individual citizens start doubting even legitimate information, then societies will lose their basic shared standards of truth.** In recent years many real-world manifestations of the liar's dividend have already emerged in the political sphere where some political figures and their followers have claimed legitimate videos to be deepfakes either to avoid responsibility for wrongdoings, uphold their desired narrative or influence public perceptions.[38]

---

32 This section is based on the discussions that took place during the segment 'Understanding the Implications for International Security and Stability'; UNIDIR (2021b).
33 Drew (2021).
34 Drew (2021).
35 Vignard (2021).
36 Collins (2019).
37 Ahmed (2021).
38 Gregory (2021).

The phenomenon of synthetic media is emerging at a time where trust in governments and public institutions is at record lows. According to a 2019 study of States of the Organization for Economic Co-operation and Development (OECD), only 45 per cent of citizens trusted their governments.[39] Exacerbating this trust deficit in legitimate institutions around the world, there is a growth in extremist movements, science deniers, and conspiracy theorists seeking to spread their worldviews in the digital age where anyone can generate and rapidly disseminate information. On the other hand, populations are generally ill-equipped to consume the abundance of media they are subjected to with a critical eye. A combination of these trends means that deepfakes have a fertile terrain to flourish. And **while the advent of deepfakes did not create trust deficits in society and the international system, they can certainly compound them**.

Most current cases of malicious use of synthetic media technology are at the individual level, particularly targeting women through fake non-consensual pornography videos or targeting political figures.[40] However, **"as the capability of this technology develops, it's more likely to become attractive to those who are going to use it for more nefarious purposes on a grander scale".**[41] Parallels can perhaps be drawn with malicious use of information and communications technologies where initially cyberattacks were conducted against individuals primarily for monetary gain. Since then, cyber incidents have scaled up to the national and international levels. The growing deepfake phenomenon is likely moving in the same direction.[42]

While personal harm will remain a dominant threat, **nefarious use of synthetic media technology and its consequences for trust erosion present multifaceted risks for national and international peace, security, and stability**. The section below provides an overview of some of these potential implications.

Influence Operations

The intentional spread of false, misleading, or deceiving information commonly referred to as disinformation is a significant feature of influence operations whether domestic or international. As synthetic media technology and its increasing accessibility gives any motivated individual **the ability to create highly convincing forgeries of someone doing or saying something they never did, it not only lowers the barriers for a variety of actors to engage in disinformation campaigns but also provides them new sophisticated tools to this end**. In the Internet age, a majority of such campaigns are conducted online on social media platforms due to their ability to 'virally' disseminate content. A study by scholars from Massachusetts Institute of Technology has found that 'false news' spreads 10 to 20 times faster than 'real news' on Twitter.[43] As a powerful tool for disinformation, the deepfakes phenomenon is especially worrying.

[39] OECD (2021).
[40] Dunn (2021).
[41] Drew (2021).
[42] Vignard (2021); Drew (2021).
[43] Dizikes (2018).

Three factors make a disinformation campaign potent and deepfakes embody all three: a) accessibility, b) similarity and c) deniability.[44] Deepfakes are becoming increasingly accessible and easy to generate, even for free and by non-technical users due to the emergence of user-friendly applications and services. And through social media platforms, they are also easy to share widely. Furthermore, deepfakes exemplify similarity. They can be nearly indistinguishable from reality and in this way they can deceive viewers, in that people are more likely to trust audio-visuals because they have a higher resemblance to the real world in comparison to textual descriptions.[45] Lastly, deepfakes facilitate deniability. Not only is detecting synthetic media difficult, but its mere existence creates information uncertainty and scepticism in a time where trust in the information ecosystem is already eroding.

Earlier this year, the Federal Bureau of Investigation of the United States published a notification in which it warned that "Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations in the next 12–18 months".[46] An entire 'influence for hire'[47] industry is now emerging including companies which are essentially marketing firms that in some cases may also develop disinformation capabilities to carry out the political goals of their clients. These are the kind of actors that could see a significant opportunity in synthetic media technology.[48]

Cyber Operations

Not only can deepfakes aid influence operations, **they can also be used to deliver tailored harm and disruption by facilitating the malicious use of information and communications technologies, particularly through social engineering and 'spear-phishing'**. Recently, researchers from University College London released a report that ranked deepfakes as what experts believe is the most serious AI crime threat.[49] Similarly, the Federal Bureau of Investigation warned that "malicious cyber actors will use these techniques broadly across their cyber operations—likely as an extension of existing spear phishing and social engineering campaigns",[50] citing a 2020 joint Europol, United Nations Interregional Crime and Justice Research Institute and Trend Micro report.[51] Social engineering attacks more generally are a growing threat[52] and increasingly, deepfakes are becoming a part of wider social engineering strategies. For example, in 2019 AI was allegedly used to impersonate the voice of a CEO. This impersonation was sufficiently realistic that an employee approved the transfer of over €200,000, thinking he was speaking to his boss.[53] Researchers in Singapore have also found that deepfake text technology can craft highly tailored spear-phishing messages, which are

44 Ahmed (2021).
45 Sundar (2008).
46 FBI Cyber Division (2021).
47 Wallis et al. (2021).
48 Drew (2021).
49 Science Daily (2020).
50 FBI Cyber Division (2021).
51 Ciancagliani et al. (2020).
52 79% of surveyed Swiss business and technology experts believe that social engineering attacks are a likely threat over the coming months; see PwC (2021).
53 Stupp (2019).

otherwise more labour intensive to compose than mass phishing messages.[54] Moreover, researchers in South Korea have shown that deepfakes can fool facial recognition which is commonly used as a biometric authentication tool by companies to secure services and prevent fraud. **A combination of audio, video, and text deepfakes could therefore be used for cybercrime or to carry out cyberattacks with severe and widespread impact**.[55]

Synthetic media technology could also challenge the implementation of the international framework for responsible State behaviour in cyberspace by bolstering the capabilities and competencies of States to conduct more complex cyber operations. For instance, as it is becoming increasingly difficult to identify and attribute the source of deepfakes, they could challenge the cyber norm concerning attribution. Moreover, not all States will have equal capability to detect and attribute malicious synthetic media and respond to their negative impact. In this way the advent of deepfakes could augment the already existing differences in capacities of States to implement cyber norms, which creates vulnerabilities for the system at large.[56]

Military Decision-Making and Arms Control

Militaries and their political leadership rely heavily on the information environment to make decisions, especially during time-critical situations. **Weaponized deepfakes could be employed to 'poison' the information environment which could instigate a crisis or lead to inadvertent escalation and miscalculations during ongoing political and military tensions.** For instance, a deepfake claiming that a nuclear-armed State has put its nuclear arsenal on high alert could be generated, which could lead other nuclear-armed States to increase the alert status of their nuclear forces.[57] While deepfakes alone may not result in nuclear weapons use, they could undermine the political and security conditions surrounding strategic stability among nuclear-armed States and their allies that share adversarial relations.[58] Recently research has shown that it is possible to create synthetic satellite imagery.[59] As militaries rely heavily on satellite imagery for mission planning and decision-making, **geographic deepfakes could be employed to deceive adversaries and degrade their situational awareness**. Moreover, commercial satellite imagery is increasingly becoming an open-source verification technology for arms control.[60] **Deepfake satellite imagery could therefore also be used to falsely accuse adversaries of non-compliance with arms control agreements, or conversely, authentic satellite imagery-based evidence of violations could be discredited as a deepfake by guilty actors**.

[54] Newman (2021).
[55] Wiggers (2021a).
[56] Makumane (2021).
[57] Topychkanov (2021).
[58] Boulanin et al. (2020).
[59] Vincent (2021).
[60] Pabian et al. (2020).

Law Enforcement

Deepfakes **challenge law enforcement at both the domestic and international levels by not only giving rise to new forms of crime,**[61] **but also by risking the credibility of documentary evidence on which the policing and legal system is based**. They could be used to falsely incriminate individuals, institutions or States of wrongdoing, or discredit legitimate evidence of unlawful conduct. This was illustrated last year in a case where deepfake audio evidence was submitted in court to discredit the father in a child custody battle,[62] raising serious concerns regarding the reliability and admissibility of audio-visual content as electronic evidence.[63] In a broader sense, synthetic media could tarnish the image of law enforcement which requires legitimacy to carry out its function—for example, a manipulated video falsely showing a law enforcement official engage in wrongdoing could undermine an ongoing criminal investigation or lead to social unrest with international implications.[64]

Like with many other new and emerging technologies, advances in synthetic media technology are taking place at an unprecedented pace and scale. **The international community will have to undertake a combination of different measures to effectively contain the impact of the multifaceted risks it presents, particularly because synthetic media technology is not currently explicitly addressed under existing multilateral frameworks in the context of international security.** Responding to possible malicious use of synthetic media will therefore first and foremost require better understanding and awareness of the technology and its implications. As synthetic media have a highly accessible, open-source innovation landscape and their weaponization can target and impact different communities in different ways, the international peace and security community will have to engage with a range of synthetic media practitioners and stakeholders and jointly undertake education, awareness-raising and threat assessment activities. Such activities would be essential to build a common foundation on which possible multilateral governance tools could be built. In this spirit, INTERPOL together with the United Nations Interregional Crime and Justice Research Institute have been organizing their Global Meeting on AI for Law Enforcement to examine the challenges and opportunities of AI, including synthetic media. As an outcome of these meetings, they are also jointly developing a 'Responsible AI Innovation Toolkit for Law Enforcement' to raise awareness regarding potential threats as well as to "support and guide law enforcement in the design, development and deployment of AI".[65] Such an approach could be adopted in other areas of international peace and security. Furthermore, as many national[66] and regional[67] regulatory approaches to tackle deepfakes are emerging, the international community could leverage these to possibly develop internationally accepted normative and technical guidelines or standards[68] for the responsible use of synthetic media technology as well as norms against its malicious use.

---

61 INTERPOL & UNICRI (2020).
62 Swerling (2020).
63 Ciancagliani et al. (2020, 58)
64 Hazenberg (2021).
65 INTERPOL (2020).
66 Ferraro (2020); Jing (2019).
67 Van Huijstee et al. (2021).
68 In this regard, the JPEG standardization committee of the International Standardization Organization is actively exploring standards to guide the positive use and tackle malicious use of synthetic media; see JPEG Fake Media (2021).

1.3 Preserving and Fostering Digital Trust: The Way Forward[69]

In the era of digital transformation, digital technologies are revolutionizing all aspects of society. Our ability to leverage their transformative benefits for society, economy and the environment is dependent on preservation of trust in and the stability of the digital ecosystem. **While many digital technologies are providing novel tools to enhance this trust, some innovations such as AI-generated synthetic media can also erode trust in a world that is already suffering from a 'trust deficit disorder'**, as noted by the United Nations Secretary-General.[70] And as digital technologies now underpin core societal functions, international security and stability are reliant on the preservation of a trustworthy and stable digital ecosystem.

The COVID-19 pandemic has brought to light our collective vulnerability to the misuse of digital technologies and the disruption such misuse can cause. The widespread dissemination of misleading and even malicious information surrounding the COVID-19 pandemic in the past few months has fuelled public distrust in COVID response and mitigation measures,[71] including digital tools such as contact tracing applications.[72]

In the wake of broader trends such as the declining trust in institutions and among actors in the international system, misuse of innovations like synthetic media could sow discord and distrust in society and thus have an amplified destabilizing effect. **The international community therefore urgently needs to prioritize issues of digital trust and security and take concrete steps to protect and promote shared standards of truth to unlock the true potential of the digital domain**. To this end, this panel explored what digital trust entails and how it can be preserved and fostered.

What Does Digital Trust Entail?

Trust is a foundation of all human interactions and relationships whether within or outside the digital domain. It is an inherently elusive and relative notion that has social, emotional, and cognitive components. **In essence, general and digital trust are not distinct. However, the hyperconnectivity and anonymity that the digital domain embodies adds additional layers of complexity to the concept of trust.** Digital trust needs to be fostered on many different levels and across many different relationships to achieve an open, accessible, stable, secure, and peaceful information and communications technology ecosystem for all. For instance, at the user level there is a trust dynamic with the industry and technology providers—users need to have trust in the systems, networks, and software they are using, and trust that that their security and privacy are being ensured. The second dimension to this is the trust between

[69] This section is based on the discussions that took place during the segment 'Preserving and Fostering Digital Trust'; UNIDIR (2021e).
[70] United Nations Secretary-General (2018).
[71] Velásquez et al. (2021).
[72] NPR (2020).

individual citizens and the private sector and their governments—it is important for non-governmental actors in a society to trust that their government will protect their interests in cyberspace and will strike a right balance between national security, cybersecurity, and privacy. The third dimension is the trust among States in cyberspace, that they will maintain peace and security in this domain. This trust is essential as erosion of trust at the international level within and outside the digital sphere is increasingly causing overall fragmentation which could potentially result in escalations of tensions leading up to conflict.[73] The value of digital trustworthiness and trust-building at different levels was brought out by the COVID-19 pandemic. If governments and policymakers as respondents to the crisis cannot trust the data they are given, they cannot effectively allocate essential resources such as hospital beds and vaccines. Similarly, at the international and inter-State level, if there is no trust in the data a State or region provides regarding an outbreak, then the response measures are likely to be less than optimal.[74]

Apart from the levels of trust, another layer of complexity is added by the hierarchies of trust—actors in society place trust in some actors more than others.[75] While this is a human tendency, it is intensified in the information age. In the ongoing pandemic we have seen how individual citizens sometimes trusted their local physicians or leaders more than their governments. This perhaps calls for more traditional localized measures for trust-building.

Additionally, an important difference the digital world brings is the role that it gives to individual citizens to preserve and foster a trustworthy digital and information ecosystem,[76] beyond traditional custodians of trust—governments, international organizations, and the private sector. For example, as much of the digital content is also user-generated, individuals could also help to ensure that accurate information is shared. Therefore, the culture of preserving and fostering digital trust needs to be instilled at the individual level as well.

How can the International Community Foster Digital Trust and Mitigate the Potential Destabilizing Effects of Advances in Digital Technologies?

Fostering trust in the digital ecosystem will require more trust brokers and trusted brokers[77]—institutions or platforms where trust can be fostered. Many digital companies are automating building trust through consent management for data-sharing for example. However, building trust at the user level will require giving individuals more agency, in this regard, over what they are signing up for in terms of use of their personal data, so that they trust the platforms and technologies they are using. Furthermore, **there needs to be more diversity and interdisciplinarity in the network of trust brokers**. Traditional brokers such as institutions at both the national and multilateral levels need to become more inclusive and allow more actors (including the private sector, academia, and civil society) to take part in policy discussions.

[73] Nakamitsu (2021).
[74] Gill (2021).
[75] Gill (2021).
[76] Ciglic (2021).
[77] Gill (2021).

This is essential to ensure that top-down decision-making and initiatives are fitting to realities on the ground.

At the intergovernmental level, in 2015 States adopted eleven norms of responsible State behaviour in cyberspace under the framework of the United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security.[78] In order **to foster trust among States, in the private sector, and between the private sector and governments in cyberspace, States must continue to clarify and build on the application and implementation of these norms** to the extent possible, especially in relation to attribution and capacity-building for responding to cyber incidents. Removing ambiguities and opaqueness in the international cyber norms framework that can be exploited by bad actors to sow mistrust in governments, private sector entities, and international institutions is essential for preserving security and stability in the cyber domain. In this regard, the Secretary-General's High-level Panel for Digital Cooperation has recommended the development of a Global Commitment on Digital Trust and Security "to shape a shared vision of digital stability and strengthen implementation of norms for responsible uses of technology".[79] Essentially this recommendation proposes capturing the eleven norms in a high-level political commitment including clarifications on essential contentious issues such as attribution. It is thought that such a commitment could set the building blocks for a multilateral policing capacity in the cyber context, just as the International Atomic Energy Agency holds on the nuclear front.[80]

As an important provider and operator of the digital domain, the private sector also has a key interest and responsibility in promoting trust, security, and stability in the digital domain. Many large technology companies and relevant civil society actors are now actively engaging in multi-stakeholder initiatives and discussions within and outside the United Nations framework to find pathways to a secure and trustworthy digital environment. The Paris Call for Trust and Security is one such significant external initiative that brings together a wide range of stakeholders to collaborate on the inclusive governance of cyberspace. The Paris Call has set out 9 principles and attendant working groups within which governments, industry and civil society actors come together to develop best practices on a range of issues from election security to protecting individuals and critical infrastructure from malicious cyber activities.[81] Such forums are necessary tools that can not only help to develop fresh thinking on the implementation of cyber norms, but also bolster United Nations cyber processes by feeding their outcomes into intergovernmental negotiations.[82]

78 UNGA (2015).
79 United Nations Office of the Secretary-General's Envoy on Technology (2021).
80 Gill (2021).
81 Paris Call (2021).
82 Ciglic (2021).

What is the Role of the United Nations in Promoting Trust, Security, and Stability in the Digital Ecosystem?

The key role of the United Nations has been to foster cooperation, multilateralism, substantive dialogue, and consensus on the most pressing and challenging issues. The issues of trust, security, and stability in cyberspace are no exception. A clear demonstration of this is the six meetings of the United Nations Group of Governmental Experts and the Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security that have been held to date. Through these processes States have collectively studied existing challenges and threats in the cyber domain and developed norms and action-oriented recommendations to address those challenges. As these intergovernmental negotiations continue, **the United Nations as a facilitator of dialogue and cosensus decision-making in the cyber context will continue to play a key role in fostering digital trust and security**.

The United Nations is now also trying to leverage its convening power to make the cyber processes more inclusive by bringing other stakeholders into multilateral deliberations. On the sidelines of the Open-Ended Working Group that recently concluded its work, the United Nations organized a stand-alone multi-stakeholder consultation that brought together 114 stakeholder representatives to discuss information and communications technology threats and ways to address them.[83] While States are the primary actors responsible for maintaining international security and stability, **in the cyber context non-State actors, particularly the private sector, play a defining role as primary providers, operators, users, and managers of cyberspace**. Preserving and fostering trust and security in the digital domain therefore requires a collaborative and coordinated multi-stakeholder approach. In this regard, **the United Nations has an important role to play in changing mindsets in a traditionally State-based international security system and encouraging the willingness to meaningfully include non-State stakeholders** in multilateral efforts to secure the digital domain. The United Nations Secretary-General's Roadmap for Digital Cooperation is one important step in this direction.[84] The United Nations will also need to continue to systematically engage with States and other stakeholders to foster a culture of accountability and adherence to the emerging web of norms, rules, and principles on responsible behaviour in cyberspace.[85]

How to Build a Resilient, Trustworthy Digital Ecosystem Moving Forward?

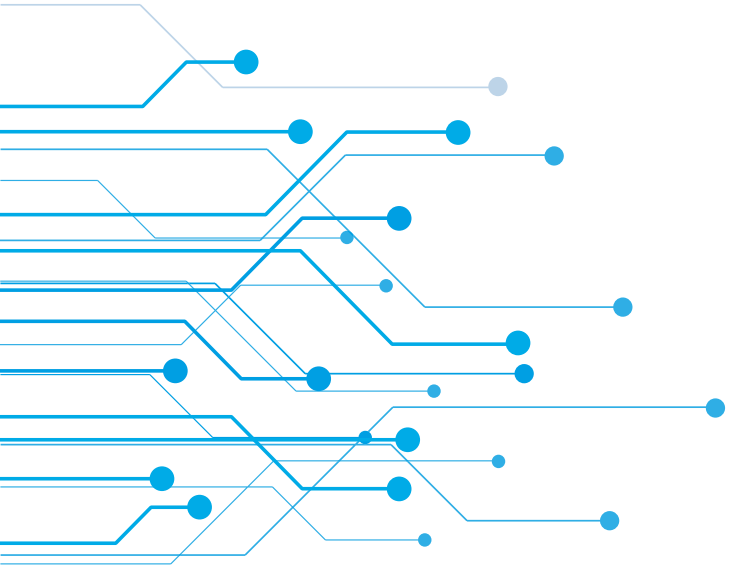
The technologies that empower the digital ecosystem are evolving rapidly, however the citizens of this ecosystem who utilize the technologies are not necessarily able to keep up with the pace of innovations and their consequences. For example, **it is becoming increasingly challenging for individuals to discern the authenticity of vast amounts of**

[83] Nakamitsu (2021).
[84] UNGA (2020).
[85] Nakamitsu (2021).

information that is generated and disseminated online, which is sowing distrust. To make digital citizens more resilient in a hyperconnected digital age, governments, industry, and civil society must **invest in digital media literacy and training individuals to engage with and critically evaluate the plethora of digital content they are exposed to online**.[86] Furthermore, the often-opaque inner workings of the digital domain need to become more transparent as openness is an essential pillar of trust-building.

In parallel, **trust is built when actors in the international system work together on big challenges**.[87] In the digital domain there are opportunities for stakeholders to collaborate despite any differences on cybersecurity in the international security context. For example, stakeholders can come together to leverage digital technologies to advance the Sustainable Development Goals, from improving healthcare and infrastructure to strengthening institutions. This could include collaboratively finding solutions to improve pandemic surveillance systems, build next generation chatbots to address global health worker shortages and pool resources to enhance computing capacities and algorithmic expertise. **Collaboration in these areas would likely foster trust between key stakeholders that could transcend into more contentious areas of cyberspace.**

Finally, a resilient and sustainable digital ecosystem cannot be achieved without inclusivity. Inclusivity not just in terms of including non-State stakeholders in multilateral deliberations on securing cyberspace, but also ensuring equal participation of women and substantively engaging young people who are natives of the digital age.[88]



[86] Ciglic (2021).
[87] Gill (2021).
[88] Nakamitsu (2021).

2.1 Creation and Dissemination[89]

This panel provided a technical overview of how visual and textual synthetic media are created and disseminated. Through presentations and an interactive discussion, this panel dscribed the underlying technologies of synthetic media, explained how they could be used, which actors currently have the means to create and disseminate them, and shared foresight on a range of future advances concerning synthetic media.

2.1.1 Visual Synthetic Media[90]

Synthetic videos and images are the most prominent forms of deepfakes that have gained significant attention in recent years. Take for example the viral deepfake video of Tom Cruise playing golf that emerged on TikTok and generated headlines across the world due to its quality.[91] **The technology underlying visual deepfakes can manipulate or fabricate a video or an image showing someone doing or saying something they never did.** While such capabilities are not new in principle, what is novel is the use of sophisticated machine-learning techniques which have made it possible to generate hyper-realistic synthetic videos and images much more quickly and easily through open-source applications and services.

Visual synthetic media can be created in many ways, including face swapping, facial re-enactment, and face generation. There are three main technological factors that underpin all these processes—deep learning algorithms (a subset of machine learning based of artificial neural networks),[92] high-performance computing resources powerful enough to run deep neural networks,[93] and training data on which the deep neural network can be trained. Below is a technical overview of different ways in which visual synthetic media can be created.

Face Swapping

Face swapping is the original and most common technique of generating visual synthetic media. **It entails the task of swapping a face in a video or an image by transferring one individual's identity from a source picture onto a target image of another person**, while still maintaining the rest of the bodily and environmental features.[94] It is carried out using a deep learning system or model called deep generative model, and an encoder-decoder architecture.

---

89 This section is based on the discussions that took place during the segment 'Unpacking Deepfakes—Creation and Dissemination of Deepfakes'; UNIDIR (2021c).

90 This section is primarily based on the presentation delivered by Hao Li, Co-Founder and CEO of Pinscreen and Distinguished Fellow at University of California, Berkeley; Li (2021). For enhanced readability and understanding, the authors advise watching Hao Li's presentation in tandem with reading this section.

91 Brown (2021).

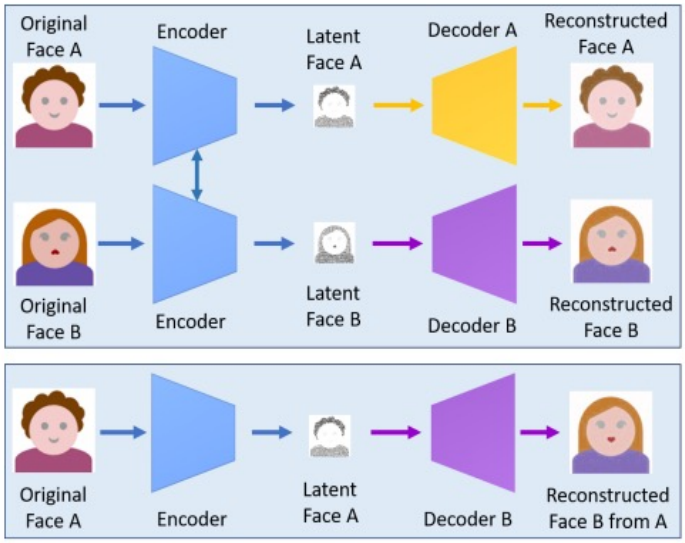
92 Scharre & Horowitz (2018).

93 "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"; see Collins (2019, 7).

94 Ding et al. (2020).

The encoder and decoder are components of the deep generative model. The encoder is the part in which the model learns how to "compress the input data into an encoded representation",[95] and the decoder is the part in which "the model learns how to reconstruct the data from the encoded representation to be as close to the original input as possible".[96] To create a face swap, the image of an individual's (A) face on which one wants to swap someone else's face is extracted from an original video frame. This original image is given as input to a common encoder which creates a latent face of A including only dormant features such as expressions, pose, and the lighting information, leaving out the identity of A. This latent face with only dormant features is given as input to A's specific decoder which creates a reconstructed image of A. The same is done for the target person (B), who's face is to be swapped on to A. Then the common encoder takes A's original face image and creates a latent face with only dormant features of A. This latent face is given as input to B's specific decoder which then does the face swap by reconstructing B's face from A. In this way by using a generic encoder and a person-specific decoder, a face swap of a person can be generated.[97] Face swapping techniques to generate synthetic videos and images have become increasingly accessible, even for relatively less-skilled individuals. This is owed to the emergence of various open-source face swapping software tools such as the Zao[98] and Reface[99] applications. On one hand, this technique is being used in the entertainment industry, for example to insert an actors' face in a movie scene in which they never appeared,[100] but on the other, it is being used to create non-consensual pornography.[101]

Figure 1: *Process of creating a face swap*[102]

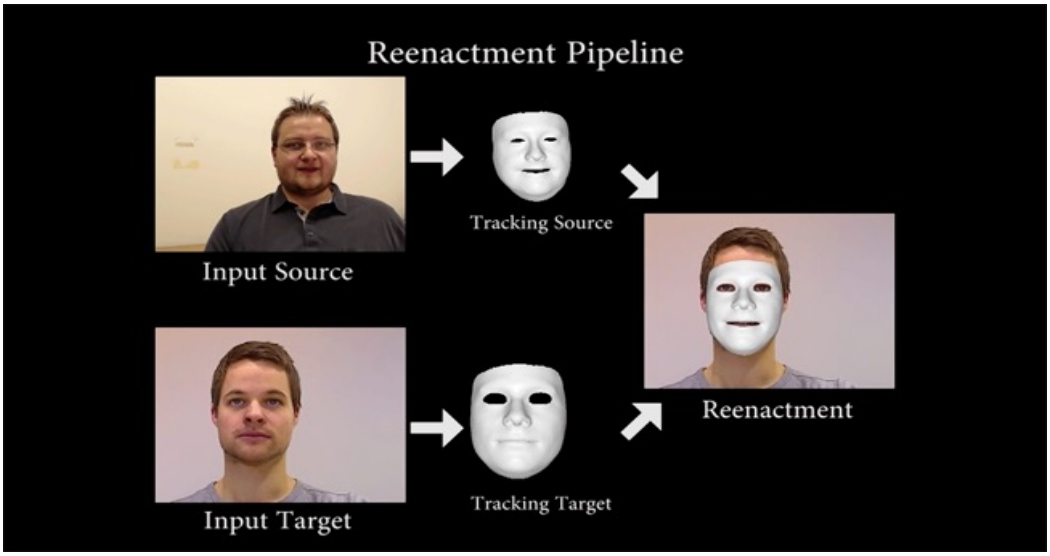


95 Badr (2019).
96 Badr (2019).
97 Li (2021); Nguyen et al. (2021, 3).
98 Porter (2019).
99 Reface (2021).
100 Boylan (2018).
101 Harwell (2018).
102 Nguyen et al. (2021, 3).

Facial Re-enactment

Another prominent method is facial reenactment also referred to as 'puppet-master technique', **which takes an existing video of the target person, but another performer can take control of the facial performance**. Essentially a target person is "animated (head movements, eye movements, facial expressions) by a performer" acting out what they want the target person or 'puppet' to say or do.[103] The famous deepfake video of Mark Zuckerberg that surfaced in 2019 employed this technique.[104]

Figure 2: *Facial Reenactment Pipeline*[105]



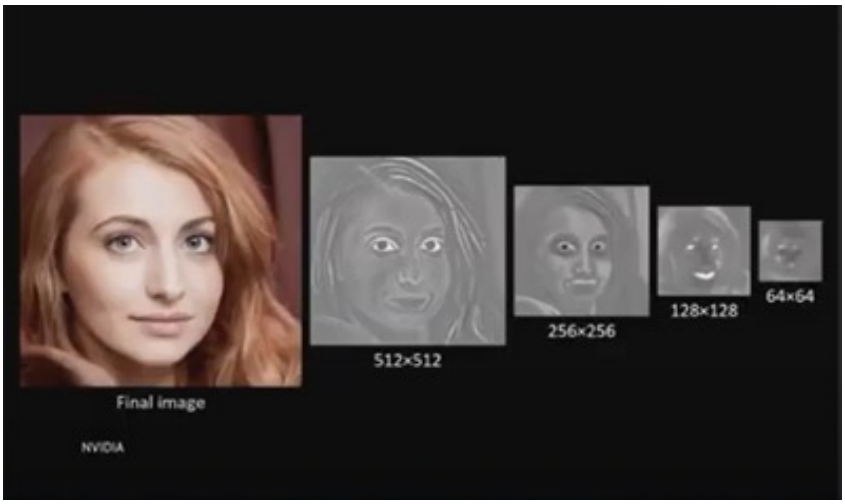
Face Generation

This technique employs generative adversarial networks (GANs), a special type of deep neural network that can create hyper-realistic fabricated images of non-real people. It entails a network that improves itself to generate realistic synthetic people. In the GAN process, two neural networks are trained by competing with each other. The first neural network, the 'generator', is tasked with creating fake content using a sample dataset to learn the characteristics of the content being faked. The second neural network, the 'discriminator', receives the fake content from the generator and assesses their quality by comparing them to the original dataset. Then the decision of the discriminator regarding the 'authenticity' of the sample is fed back to the generator, and the generator in turn informs the discriminator whether the submitted samples were realistic or inauthentic. This reciprocal feedback loop where the generator crates new samples and the discriminator assesses whether they are fake or authentic helps the GAN system to improve its performance to generate very convincing synthetic media. The generator tries to develop more realistic samples, and the discriminator tries to distinguish the realistic from inauthentic with greater accuracy.[106]

103 Agarwal et al. (2019, 1).
104 Rea (2019).
105 Niessner (2016).
106 Rea (2019).

Over the years there have been huge advancements in terms of the resolution and quality of images that are being generated using this technique. For example, Nvidia Corporation's StyleGAN system is able to generate high resolution portraits of people that are nearly flawless, so one would have to look very closely to identify certain flaws or artifacts—details that deepfakes fail to reproduce realistically.[107]

Figure 3: *Face generation using StyleGAN*[108]



This technique has also been used by researchers at Pinscreen to generate facial expressions, rather than a new subject—based on a single input photo of the target person, compelling facial expressions can be generated based on the likeness of the person in real time.[109] Another method for image animation, called the 'first order motion model'[110] further allows the generation of head movements of a person using only a single photo of them. More recent advances in this area have made it possible to not only mimic head movements but also to generate full-body movements of a person using only a single input image.[111] This technology for example can make anyone appear to dance like a professional dancer.

Audio-Driven Synthetic Media

Synthetic Media can also be audio-driven where audio signals can be used to generate a performance of a target person using just a video of them and an audio clip you want them to follow. **In this technique, referred to as 'lip syncing', an audio file of the subject's speech is converted into realistic mouth shapes, "which are then grafted onto and blended with the head of that person from another existing video".**[112] Researchers from University of Washington developed one of the first variants of this technique and demonstrated it by generating a hyper-realistic video of US President Barack Obama talking about terrorism and other topics using his real video addresses that were originally on a different topic.[113]

107 Karras et al. (2020).
108 Li (2021).
109 Li (2021).
110 Siarohin et al. (2019).
111 Jingles (2020).
112 Langston (2017).
113 Suwajanakorn et al. (2017).

The synthetic media technology landscape is evolving rapidly as described above. **In the future, significant improvement in the quality of output generated through these techniques is expected, particularly with respect to their resolution.** Existing visual synthetic media are often imperfect with some blurring or artifacts appearing on faces for example. The quality is likely to keep improving to the point where it will become impossible to distinguish synthetic from real.[114] The use of synthetic media technology is also expected to become increasingly ubiquitous as it is becoming highly accessible and relatively easy to use—many techniques require rather small amounts of data and computing power,[115] and there are a growing number of open-source applications and tools that are making the creation and dissemination of synthetic media increasingly accessible.[116]

In an era where video conferencing and other digital forms of communication have become dominant, this accessibility poses a significant threat to security and stability at the individual, organizational, societal, national, and international levels. In the political sphere, one could create a forgery of a world leader admitting to committing an illegal activity or saying racially insensitive things which could lead to civil unrest or even a constitutional crisis. Deepfakes could also be weaponized to spread false, manipulative, or deceiving information during election season or in a military crisis. Further, they could be employed in the conduct of sophisticated cyberattacks—for instance deepfakes could be potentially used to fool biometric authentication or for social engineering in phishing attacks against individuals.[117] In this respect, businesses can also be likely targets. Deepfakes are particularly dangerous in contexts where content is not regulated in real time. Although there are exceptions, generally on social media platforms it is easy for anyone to post content at any time, and at times before any malicious content can be detected it is already disseminated widely. Another example is video conferencing where one can pretend to be someone else in real time. Nonetheless, malicious actors do not always need sophisticated deepfake technology to sow discord and distrust. They can employ routine techniques to manipulate content, referred to as 'cheapfakes'. A famous example of this is the viral video of Nancy Pelosi, speaker of the United States House of Representatives, in which a malicious actor made her appear drunk by merely slowing down the pace of the video. [118]

Not all applications of visual synthetic media are malicious, this technology also has profound beneficial applications across sectors. For instance, AI-generated synthetic media can be used in assistive navigation tools to aid individuals with vision impediments. In the education sector, synthetic media can help to have better learning outcomes by revolutionizing storytelling—for instance, important historical figures can be brought to life to teach students history.[119] In the entertainment industry synthetic media technology is being used for visual effects in movies and music videos.[120] Similarly, this technology is also being used to develop the next generation of human-machine interfaces.[121]

114 Li (2021).
115 Agarwal et al. (2019).
116 EG (2021).
117 Wiggers (2021a).
118 Reuters (2020).
119 Jaiman (2021).
120 Murray (2021).
121 Pinscreen (2021).

2.1.2 Synthetic Text [122]

Visual deepfakes such as videos and images have inspired much worry in the policymaking community in recent years, however deepfake text is no less concerning. **Advances in AI-techniques cannot only challenge our trust in what we see, but also our trust in what we read or who wrote it**.

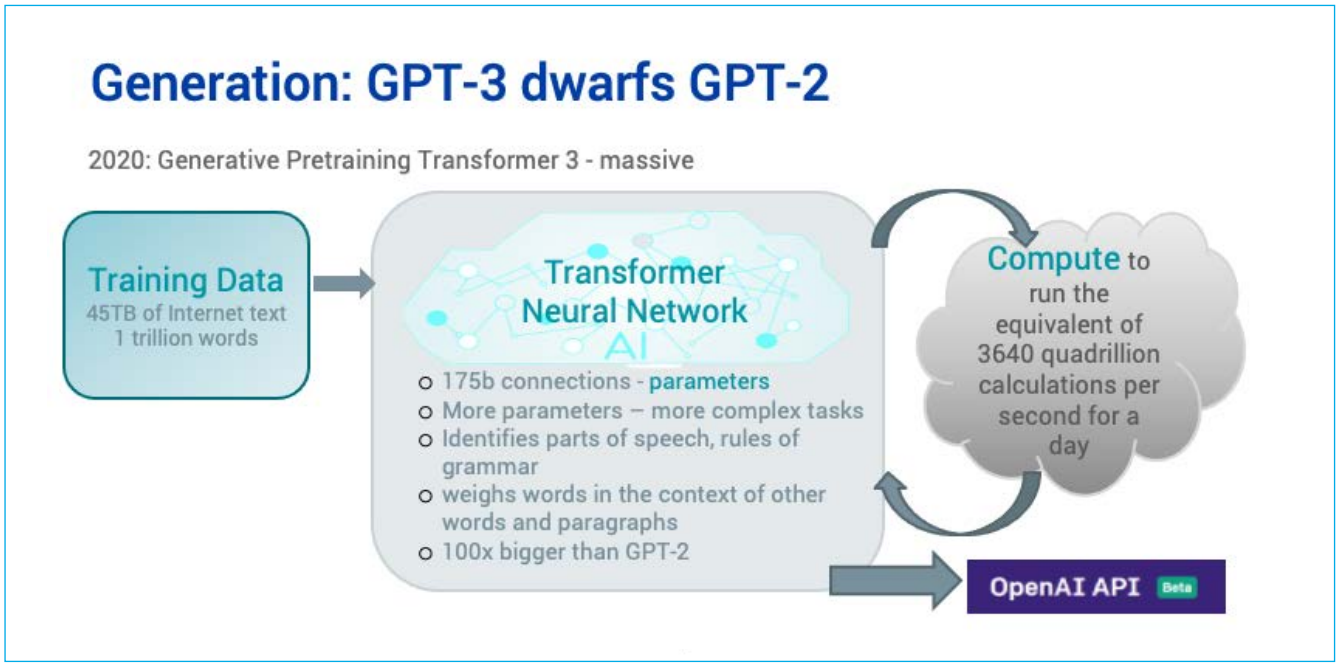
The advent of synthetic text is a product of two research areas in machine learning—unsupervised deep learning, which refers to a machine's ability to learn from large amounts of unlabelled training data,[123] and natural language processing, which refers to a machine's ability to analyse, manipulate, and generate human language.[124] Natural language generation systems, or models, are the underlying technology of synthetic text driven by the so called 'AI triad'—abundance of data, innovative algorithms and massive amounts of computing power.[125] **This ability to generate synthetic text using natural language generation models came to light in February 2019 when OpenAI unveiled a powerful language model called Generative Pre-Trained Transformer or GPT-2.** In doing so it ushered in an era of large language models. These systems use innovative algorithm architecture, a transformer which is a deep learning model that selectively focuses on segments of input text that it predicts to be the most important.[126] In this way, a transformer is particularly adept at analysing sequences in text and generating language. GPT-2 is pretrained on a large quantity of text—8 million web pages of training data—[127]with the objective "to predict the next word, given all of the previous words within some text".[128] When given a prompt, it generates synthetic text with close resemblance to human writing by assigning probabilities to words occurring in a particular sequence in sentences and then predicting the most likely next word. **Put simply, it is a "cutting-edge writing system that takes in a prompt, a few words or a sentence, and completes it".**[129] It essentially works like autocomplete on mobile phone messaging applications, except on a much bigger scale where it can assist in generating paragraphs one phrase at a time. What makes GPT-2 powerful is that it is adaptive to the style and content of the conditioning text which allows the user to create realistic and coherent continuations to a text on any topic. It thus can assist in performing a wide range of language tasks including text summarization and translation.[130]

This ground-breaking innovation of 2019 was however dwarfed in just 18 months as OpenAI introduced an updated version, GPT-3. Although GPT-3 is built on the same transformer architecture as GPT-2, **its novelty comes from the enormity of its size.** It is trained on 45 terabytes (nearly a trillion words of human writing) of Internet text, which spans nearly the entirety of the text available on the open web. The transformer learns through a large neural network with 175

123 Scharre & Horowitz (2018).
124 Shetty (2018).
125 Sedova (2021).
126 Maxime (2019).
127 OpenAI (2019).
128 OpenAI (2019).
129 Sedova (2021).
130 OpenAI (2019).

billion connections called 'parameters' which help the network to store and process data. This size of the neural network helps GPT-3 to process data much more efficiently, and thus the model can perform more complex tasks. GPT-3 can identify parts of speech, rules of grammar, whether words are synonymous, and resolve word meaning by the context of words around them.[131] Like autocomplete, given a prompt of a few short sentences, it emulates its style of writing and generates multiple paragraphs, dialogues, and whole stories, nearly indistinguishable from human-generated text. It can do so because in its training it analysed most of the content available on the open web, and "assigned the probabilities and weighed the relative importance of words and word combinations".[132] Therefore the process of generating synthetic text using GPT-3 is as follows: A human operator defines instructions, gives the model a sample and tells it how much to write. The model starts writing until the output fulfils the specified parameters. And then the human operator may iterate and refine the prompts. **With its sheer enormity GPT-3 has proven its ability to produce op-eds,[133] write Shakespearean poetry[134] and generate computer code,[135] all in a human-like manner.[136]**

Figure 4: *Generating synthetic text using GPT-3[137]*



131 Sedova (2021).
131 Scharre & Horowitz (2018).
132 Sedova (2021).
133 The Guardian (2020).
134 Merzmensch (2020).
135 Sagar (2020).
136 Heaven (2020).
137 Sedova (2021, slide 4).

This powerful ability of large language models like GPT-2 and GPT-3 to generate hyper-realistic synthetic text makes them an all-purpose technology with dual-use applications. These advances are increasing the productivity of writers and revolutionizing translation, **but on the other hand their unprecedented ability to mimic human writing could potentially be misused to increase fakery in the information ecosystem, generate disinformation, skew public opinion, and undermine trust both domestically and internationally.**[138]

To understand the possible misuse of large language models like GPT-3, recently academic institutions have conducted studies with the support of OpenAI. For example, the findings of a report published by the Middlebury Institute for International Studies suggests that GPT-3 demonstrates the ability to generate synthetic text that "accurately emulates interactive, informational, and influential content that could be utilized for radicalizing individuals into violent far-right extremist ideologies and behaviours".[139] The study also suggests that in the "absence of safeguards, successful and efficient weaponization that requires little experimentation is likely".[140] Similarly, the Center for Security and Emerging Technology at Georgetown University conducted a study to examine GPT-3's potential to write disinformation, such as misleading tweets, compelling fake news stories and polarized opinion pieces. They tested GPT-3 on six key tasks associated with information manipulation and found that with some curation it was quite capable at performing what they had tasked it to do—generating tweets to advance a given theme, re-writing articles to suit a political slant, devising new narratives in a style of a conspiracy theory, writing tailored messages to target specific racial and religious groups and writing messages tailored to political affiliation on foreign policy and international security issues, among others.[141] One of the most worrying insights that came out of this experiment was that **GPT-3 was particularly adept at generating content with a slant rather than neutral content thereby making it an attractive tool for manipulators** who seek to exacerbate divisions at the individual, organizational, or societal scale. While large language models like GPT-3 generate output best in conjunction with a human operator in the loop who can craft and iterate on prompts to improve the output, these systems have the potential to increase the scale of false and deceptive content, particularly by freeing the cognitive resources of the human operator to experiment and see what content gets more traction.

This explains how deepfake text could further the erosion of trust and threaten security and stability at the individual, organizational, national, and international levels. For example, it could be used to flood social media with misleading and polarizing content during election season or in an international crisis. It could also be used in social engineering and phishing attacks, for instance by impersonating a known figure to the target and getting them to click on a malicious link. Recognizing these risks of possible misuse, Open AI themselves attempted to limit the public access to the large GPT-2 model.[142] However in short order the model was replicated by others which has led to the emergence of easily accessible GPT-2-based applications. Access to GPT-3 is still restricted and available only to vetted researchers and developers

138 Sedova (2021).
139 McGuffie & Newhouse (2020).
140 McGuffie & Newhouse (2020).
141 Sedova (2021); Buchanan et al. (2021).
142 OpenAI (2019).

for a fee.[143] However, other efforts are underway to replicate and release GPT-3-size language models openly.[144] This raises concerns as to the range of actors that could have the means to create or access synthetic text technology.

Large language models require massive amounts of computing power. As the models learn from vast amounts of data and encode insights into their neural networks in an iterative process, they must perform quadrillions of mathematical calculations requiring an enormous amount of computing power. Training a model of the size of GPT-3 requires access to very high-performance computing resources—the cost of training GPT-3 is estimated to have been 12 million dollars.[145]

Despite the costs, the diffusion of these large language models is underway—any motivated or well-resourced State or non-State actor has access to varying degrees of this capability.[146] Once GPT-2 was trained, it has become a service run in the cloud and available to researchers and developers on a subscription basis.[147] The high costs come in when one is trying to build a novel generative language model. Once they are built and released openly, the remaining costs are associated with its adoption, use, and the computing power required to run it. Any actor who has enough computer processing power and technical skill to work with open source GPT-2 applications can do so to generate millions of messages. Although GPT-3 has not been released openly and is only accessible to vetted researchers, open source replicas of GPT-3 are already emerging which could be accessed by well-resourced State or non-State actors.[148] EleutherAI for example has recently released a smaller, 6 billion parameter, version of GPT-3.[150] **While computationally expensive, motivated and well-resourced actors could use open-source replicas to generate human-like synthetic text.** The high costs have also not prevented actors from developing their own large language models. Since OpenAI first released GPT-3, more States and technology companies have announced large language models, and in a variety of languages such as Chinese, Russian, Korean, and French.[151] Some of these language models are even larger and more complex than GPT-3, meaning that the quality of their output may be as or more impressive than that of GPT-3. In the coming years, the diffusion of these powerful language models is likely to continue as efforts are being made to reduce the costs of computing[152] and to provide infrastructure to train even larger models.[153]

143 The Economist (2021).
144 Leahy (2021).
145 Wiggers (2020).
146 Solaiman et al. (2019).
147 For example, a GPT-2 text generation product is available at AWS Marketplace; see: https://aws.amazon.com/marketplace/pp/prodview-cdujckyfypprg
148 Williams (2021).
149 Romero (2021).
150 Alford (2021).
151 Wiggers (2021b); Kyoung-Son (2021); PAGnol (2021).
152 GlobaNewswire (2009).
153 Rella & Sharma (2021).

The main concern this proliferation raises is that some of these emerging models may be openly accessible and **at the disposal of malicious actors with resources and some technical skill, motivated to clutter the information environment with false and manipulated content**. While it may be cheaper to employ human content farms than to use large language models that require massive computing power to generate content, a determined State or non-State actor may still find employing this powerful technology attractive. Moreover, now computing power is also increasingly being made more accessible as it is being offered as a service through the cloud[154] and through laptops powered with high-performance computing. Additionally, large language models are general purpose in the sense that they are trained on massive amounts of data. However, if a malicious actor wants to exploit it for a specific purpose, they can retrain it on a small, curated dataset of very specific terminology using an AI technique called fine tuning.[155] Lastly, the proliferation of large language models in different languages could remove barriers for non-native speakers to infiltrate the information ecosystems of other regions or States for both legitimate or malicious purposes.[156]

154 Microsoft (2021).
155 Solaiman & Dennison (2021); Solaiman et al. (2019).
156 Sedova (2021).

2.2 Technical Coutermeasures[157]

This panel provided an overview of the technical countermeasures for deepfakes and reflected on their effectiveness to combat the risks posed by them, particularly in light of the pace at which synthetic media technology is advancing. It focused mainly on two categories of measures aimed at making malicious uses more difficult and costly: deepfake detection and media provenance.

2.2.1 Deepfake Detection Tools[158]

The increase in the number of synthetic media published online has given **rise to a field of research and organizations that study the evolving capabilities and threats of deepfakes and develop technical countermeasures to detect them**.[159] Deepfake detection techniques typically work by identifying 'artifacts'—errors or details that the deepfake fails to reproduce in a realistic manner—and inconsistencies in synthetically created pieces of media. However alternative approaches also exist.[160]

Detecting Visual Deepfakes

Early attempts to spot visual deepfakes were "based on handcrafted features obtained from artifacts and inconsistencies"[161] of the synthesis process, while most recent methods depend on deep learning to automatically analyse the visual piece of media. Detection technologies of visual deepfakes are primarily categorized as fake image detection methods and fake video detection methods.[162] The first mainly use deep learning techniques to analyse the statistical features of images to determine whether they are synthetic or manipulated.[163] This and most image detection methods generally cannot be used for videos because of the degradation effect video compression has on the video frame data.[164] Synthetic video detection techniques focus instead either on temporal features across different frames of a video such as eye blinking, or visual artifacts that can be found in one frame such as the contour of faces, eyes, or teeth.[165] Microsoft's Video Authenticator falls in this last category. It analyses the blending boundary and fading and greyscale elements of each video frame, which are usually not detectable by the human eye.[166]

---

157 This section is based on the discussions that took place during the segment 'Managing the Deepfake Phenomenon—Counter-Deepfake Technologies'; UNIDIR (2021a).
158 This section draws from the contributions of Giorgio Patrini, Ashish Jaiman, Hao Li and Katerina Sedova; Patrini (2021); Jaiman (2021); Li (2021); Sedova (2021).
159 Patrini (2019).
160 Collins (2019, 7).
161 Nguyen et al. (2021, 4).
162 Nguyen et al. (2021, 4).
163 Almars (2021, 26).
164 Nguyen et al. (2021, 5).
165 Nguyen et al. (2021, 4 – 5 & 7).
166 Burt & Horovitz (2020).

The aforementioned techniques are used for example by the Netherlands-based company Sensity, which creates deep learning and cybersecurity-based techniques to create products that organizations can use to protect themselves from and identify and respond to generative technologies.[167] The services and products that deepfake detection companies develop are used by a wide variety of organizations and professionals, including expert witnesses, reporters, media and law enforcement organizations, and private and public agencies that employ biometric tests to confirm the identity of their users.[168] They have also served to identify AI-generated pictures used to create fake social media accounts which then facilitate scam operations or disinformation campaigns.[169] **In the field of international security, detection-based technologies and forensic techniques have been applied to high-profile cases involving speeches and public appearances of political figures.**[170] Considering the impact deepfakes can have on international peace and security, technology-based detection techniques are necessary to analyse visual material to prevent potential severe consequences of responses stemming from fake or doctored media.

The Defense Advanced Research Projects Agency of the United States is currently working on an alternative detection technique which relies on 'semantic technologies' for analysing media. Their starting point is that existing synthetic media generation and manipulation technology tends to make contextual errors, such as creating individuals with mismatched earrings, which statistical detection methods tend to overlook. Hence, the Agency's forensics approach aims to include semantic detection, attribution, and characterization algorithms that could allow their technology to spot semantic inconsistencies to detect deepfakes.[171] The Agency is also working towards empowering their detection technology to identify if the manipulated content is being used for malicious purposes.[172] Such forensic technologies **could be useful in the international security context, since they can be customized for specific individuals, such as government officials or representatives.** By identifying the distinct patterns in speech, facial and head movements of a person, this detection method could distinguish one individual from another and spot deepfake impersonators.[173]

Detecting Text Deepfakes

Early text deepfake detection technologies were artifact-based and focused on errors or inconsistencies in writings. However, the increased sophistication of output of synthetic text generation technologies such as GPT-3 has encouraged new methods which analyse the use of highest probability words by reverse engineering the system of synthetic text generators. Humans writing patterns are more random than that of generative machines, which 'write' by putting together the most likely sequence of words. Exploiting this difference between the writing approach of humans and machines, non-artifact-based text detection technologies

167 Patrini (2021).
168 Patrini (2021).
169 Alba (2019).
170 Patrini (2021).
171 Turek (2021).
172 Turek (2021).
173 Agarwal et al. (2019, 2).

analyse patterns of words in messages to classify pieces of written content that reflect a tight predictability pattern as machine authored.[174]

The detection of text deepfakes is however quite challenging due to various factors. First, the high quality of the output of existing synthetic text-generating technologies exhibits relatively few clues of origin or authorship that could signal they are fake.[175] Second, the few artifacts that can be found in text deepfakes are usually related to contextual errors that human operators can easily spot and edit before the text is posted. Therefore, the discriminative features would only surface if the generation system is not supervised and is able to publish online directly. **Finally, influence operations have become increasingly cross-platform where malicious content is also published off of mainstream social media platforms, remaining below the radar of more robust threat-hunting and content-moderation efforts.**[176]

The fast pace at which synthetic media technology is improving and the inherent adversarial nature of some of the creation techniques has incentivized a 'cat and mouse' type of game between the offence and the defence.[177] This could prevent detection tools from always succeeding in a definite way. A sufficiently skilled adversary with enough resources will most certainly find ways to match the defences put in place, and for every detection exploit found, there will most likely be ways to train algorithms to overcome them.[178] In sum, there are different ways to detect deepfakes from a technological standpoint, but the aforementioned arms race-type difficulties might mean that **there will not be a guaranteed solution for detecting deepfakes. This underscores the need for multimodal approaches to address the risks deepfakes pose to international peace, security, and stability**.[179]

2.2.2 Media Provenance[180]

Media provenance refers to the origin and history of a digital object, be it an image, video, audio recording, or a document. This information comes in the form of 'metadata'—information about how, who, when, and where a piece of media was created and edited—and could allow users to verify the authenticity of a piece of content and its alleged source.[181] Deepfake countermeasures reliant on media provenance are typically focused on technical standards and technologies that could certify the integrity of metadata, providing digital consumers with tools to identify synthetic or doctored pieces of content.

174 Sedova (2021).
175 Sedova (2021).
176 Sedova (2021).
177 Florin (2021).
178 Collins (2019, 17).
179 Jaiman (2021).
180 Ellis (2021).
181 Project Origin (2021).

The concept of media provenance was born in 2019, when organizations concerned about the threat of the malicious use of synthetically manipulated media and about mis- and disinformation started working on developing tools to counter them. Universities, research institutes, private firms, and other institutions have created programmes to develop technologies to authenticate metadata and have undertaken some joined efforts to tackle the issue in a collective manner.[182] For instance, the Coalition for Content Provenance and Authenticity (C2PA) is a product of the collaboration of initiatives and organizations in the private sector which aims to set up industry standards and technologies for provenance of media content.[183]

There are different technological approaches to establishing a system to authenticate metadata. For instance, some organizations have focused on distributed ledger—or blockchain—technologies to record modifications of content.[184] The Coalition instead is currently developing a 'tamper-proof media' technology which involves the use of digital signature technology to provide evidence of tampering.[185] As per their proposed architecture, the metadata of a digital object would first be wrapped in a digitally signed manifest, which would be embedded in the file as a package or stored remotely to be reunited with the digital object when it is consumed. The provenance of the file would then be authenticated by a cryptographic hash, a unique 'digital fingerprint' for each existing file.[186] Editing the file would change that fingerprint, proving that the digital object has been altered. The whole package would then be signed with the identification key of the source of that media object. Signature reading machines or 'validators', in the form of a browser extension for instance, would check the certificates or hashes, verify the source, and confirm that the content has or has not been altered.

There are however challenges that provenance-based countermeasures have to overcome. First, the process of editing digital content can remove its metadata. Hence, the proposed technologies will need to ensure that the integrity of digital objects is maintained throughout their life cycle. Second, the process of creating authenticated metadata would have to be simple, accessible, and easily embedded in workflows if it is to be used for every piece of media and by everyone. Third, consumers must be able to understand how to use provenance to verify the source and authenticity of media. Therefore, emphasis would have to be placed on phrasing and framing the provenance information to ensure that users understand how to interpret metadata. And lastly, there can be many different valid reasons why a trustworthy piece of content might lack its metadata. For example, a content author or provider might prefer to

182 For instance, the BBC, CBC/Radio Canada, Microsoft, and the New York Times created 'Project Origin', with a focus on creating a tracking process for news publishing. Adobe, the New York Times, and Twitter founded the 'Content Authenticity Initiative' which aims at developing systems to provide history and context for any piece of digital media. Dublin City University, Trinity College Dublin, the Graz University of Technology, Fundación Cibervoluntarios, Newswhip Media, the Institute of Information Theory and Automation at the Academy of Sciences of the Czech Republic, and NTT Data Spain are members of the consortium 'Provenance', which conducts research on digital content verification; Project Origin (2021); C2PA (2021c); Provenance (2021a).

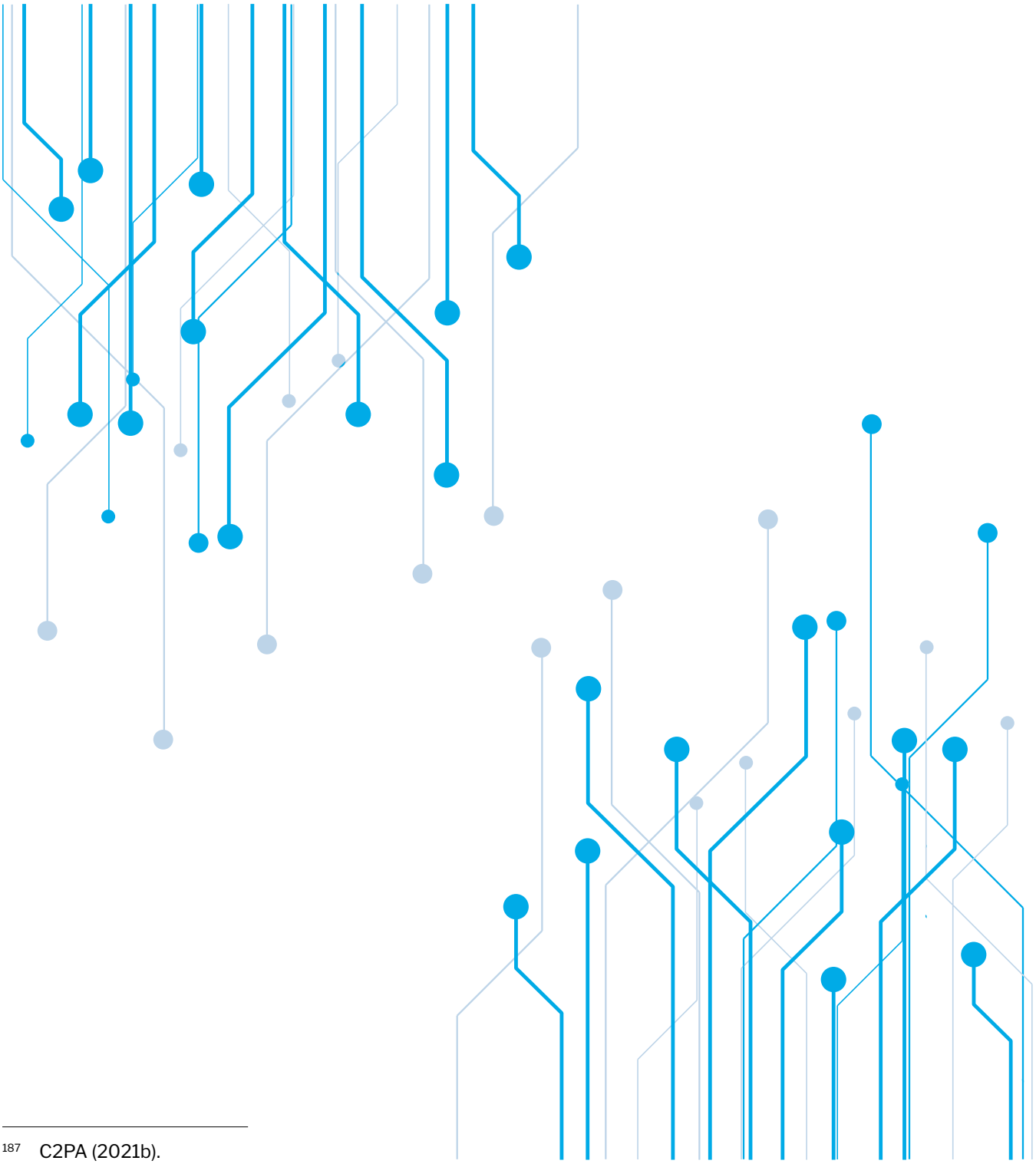
183 C2PA was born as a collaboration between Project Origin and the Content Authenticity; C2PA (2021c).

184 Provenance (2021b).

185 Ellis (2021); C2PA (2021a, 2).

186 Ellis (2021).

remain anonymous, or it might have not been possible to embed the signature in the digital content. Therefore, it is crucial to clearly communicate to consumers that media provenance is optional and that content authors, providers, and publishers might have valid reasons to refrain from signing a media object or decide not to provide certain information.[187] As with other technological deepfake countermeasures, **successful media provenance approaches might require additional investment in media literacy, education, and awareness-raising initiatives**.



187 C2PA (2021b).

2.3 GOVERNANCE RESPONSES[188]

This panel explored the key governance issues concerning deepfakes and the existing and required industry-led, national, regional, and multilateral governance responses to address this phenomenon. Panelists underscored the **centrality of governance to complement the deepfake technological countermeasures and the need for multimodal initiatives**.[189] Considering the potential global impact of the malign use of synthetic media, an inclusive **and comprehensive approach to governance responses, including a combination of measures and involving all stakeholders is crucial**.[190]

2.3.1 Societal-Level Resilience Measures

Awareness-raising and Knowledge Development

There is a general lack of awareness of the potentially destabilizing consequences of the malicious use of synthetic media at the individual, institutional, and societal levels.[191] **Governance responses should therefore first focus on raising awareness of deepfakes, the risks associated with them, and the existing tools to detect them.**

At the international level, the United Nations plays a critical role in underscoring how technological advances such as synthetic media technology could amplify growing political instability and mistrust. In this regard, the United Nations Secretary-General has publicly acknowledged the issue of the erosion of public trust and expressed concern about the potential use of deepfakes by violent extremist groups and terrorists to spread conspiracy theories, and called on Member States to take action to minimize the risks.[192] United Nations entities such as the United Nations Interregional Crime and Justice Research Institute, the United Nations University, the Office of Counter-Terrorism, and the International Telecommunication Union have also publicly recognized the international security dangers of the potential malicious use of synthetic media technology, and have started building public knowledge and understanding on the topic. [193]

---

188 This section is based on the discussions that took place during the segment 'Managing the Deepfake Phenomenon—Governance Issues and Responses'; UNIDIR (2021d).
189 Roth (2021).
190 McCarthy (2021); Florin (2021).
191 Florin (2021).
189 Roth (2021).
190 McCarthy (2021); Florin (2021).
191 Florin (2021).
192 On 22 January 2020, the United Nations Secretary-General delivered a speech in which he categorized the "growing global mistrust" and the "downside of technology"—and the abuse of new technologies to fake information in particular—as two of the four horsemen that can "can jeopardize every aspect of our shared future". On 28 June 2021, the United Nations Secretary-General commented at the High-Level Conference of Heads of Counter-Terrorism Agencies of Member States on the risk of terrorists using deepfakes to stoke conspiracy theories and to defeat facial and voice identification systems; United Nations Secretary-General (2020); United Nations Secretary-General (2021).
192 For instance, UNOCT & UNICRI (2021a); UNOCT & UNICRI (2021b); ITU (2021); Pauwels (2021); McCarthy (2021).
193 Roth (2021).
193 Ahmed (2021).

The private sector also has an important role to play in raising awareness of malicious synthetic media and their impact. Not only does deepfake dissemination take place online, but many users committed to information integrity as hobbyists identify manipulated media online and share their findings in real time. **Telecommunication companies can help to amplify the voices and the findings of those engaged in source verification, making users aware that not every piece of media found online can be trusted**.[194]

Media Literacy

Education must be invested in by governments, industry, civil society, and the international community to ensure individuals and organizations can critically evaluate the media content they find online. As deepfakes can deceive even the most skilled individuals,[195] education strategies should be targeted at all levels, not just at the most vulnerable groups. On the technology creation side, preventing the nefarious use of synthetic media technology will benefit from instilling a culture of responsible and ethical innovation among the machine learning community.[196]

Acknowledging the paramount importance of media literacy, private organizations have become increasingly involved, collaborating with universities and institutes to develop tools to educate online users and create curriculums to train professionals on the topic.[197] Furthermore, by publishing education material online, telecommunication and social media organizations can help to train users in the use of technological tools for source verification and media authenticity and encourage them to engage with these as an everyday practice.[198] Nevertheless, it is important to note that **developing critical thinking of users might not be a panacea**. The speed and amount of information users encounter online and the fact that viral content tends to appeal to emotional rather than rational drivers might challenge the sufficiency of enhancing critical engagement.[199]

From an international security perspective, awareness-raising and media literacy initiatives put forward by governments, regional, and international organizations aimed at security and law enforcement professionals could prove critical in mitigating the impact of weaponization of synthetic media.

194 Florin (2021).
195 Ahmed (2021).
196 Florin (2021).
197 For instance, Microsoft has partnered with the Center for Informed Public at the University of Washington, Sensity and USA Today to create a 'Spot the Deepfake' quiz, a media literacy tool in the form of an interactive experience. Microsoft has also partnered with the Poynter Institute to develop trainings for journalists on hybrid threats, including deepfakes, and tools to tackle them; see https://www.spotdeepfakes.org/en-US/quiz; Snapp (2021).
198 Roth (2021).
199 Collins & Ebrahimi (2021, 3).

2.3.2 Regulatory Frameworks

While awareness-raising and building media literacy are important undertakings, **effectively governing synthetic media technology will also require regulatory approaches to establish parameters of acceptable behaviour in the creation, use, and dissemination of the technology**.[200] Regulatory responses for mitigating the risks of deepfakes could cover different phases of the their 'lifecycle': a) the technology dimension; b) the creation dimension; c) the circulation dimension; c) the target dimension; and d) the audience dimension.[201] The panel discussions focused in particular on the dissemination of deepfakes.

Laws and Regulation

As with other governance responses, there may not be a simple legal solution for the risks deepfakes pose.[202] **There are various complications linked to regulating deepfakes**, including the fact that many diverse legal instruments may be applicable to the different dimensions and uses of synthetic media technology.[203] On the target dimension, deepfakes might be used to commit crimes such as non-consensual and child pornography, identity theft, breach of privacy, and extortion which are not all regulated through the same existing legal tools.[204] On the technology dimension, as an application of AI technology, synthetic and manipulated media falls under the umbrella of broader AI regulation. **Clarifying how existing laws and regulations apply to deepfakes could be a very important first step towards achieving a robust regulatory framework to tackle the risks of synthetic or doctored media.**[205]

The increasing sophistication of synthetic media technology and growing cases of misuse have motivated the development of national and domestic legislation to prevent and mitigate the use of deepfakes for non-consensual pornography, election interference, and disinformation.[206] For instance, in June 2021 came into effect in South Korea a revision of an act which makes it a crime to produce or distribute fabricated or manipulated non-consensual pornographic videos, images, or voice clips.[207] Furthermore, a bill to facilitate the deletion of deepfake pornography was recently proposed at the South Korean National Assembly.[208] In October 2019 the state of California in the United States banned the dissemination of deepfakes of political candidates within 60 days of an election.[209] Similarly, the US National Defence Authorization

200 Collins (2021, 21).
201 Van Huijstee et al. (2021, 58).
202 Collins (2019, 20).
203 McCarthy (2021); Florin (2021); Collins (2019, 20).
204 Collins (2019, 20).
205 0McCarthy (2021); Florin (2021). Collins (2019, 20)
206 Collins & Ebrahimi (2021, 3).
207 The Act on Special Cases Concerning the Punishment, etc. of Sexual Crimes (Act. No. 14412); Soo-Yeon (2020).
208 Kim (2021).
209 California Assembly Bill (AB) 730 Elections: Deceptive Audio or Visual Media; see https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730

Act for Fiscal Year 2020 included deepfake legislation in the areas of weaponization of deepfakes, deepfake disinformation activities, and deepfake detection technologies.[210] And in January 2020 it became illegal in China to publish and distribute fake news created with AI technologies.[211]

Nevertheless establishing hard law to regulate deepfakes has challenges. First, laws that tackle the circulation dimensions of deepfakes could be confronted with the need to balance restrictions with civil rights such as freedom of expression.[212] Second, as synthetic media technology continues to develop, some governmental actors might struggle to find ways to regulate nefarious uses without adopting 'excessive regulation' that could hinder technological development.[213] Third, the high volume and diversity of deepfakes and the fact that many of them are shared off main telecommunications platforms might make enforcement particularly cumbersome.[214] There are also jurisdiction difficulties: even if a State adopts legislation in any phase of the deepfake lifecycle, deepfakes might be created and disseminated beyond its borders but still cause harm within its national boundaries, making enforcement even more difficult.[215] Despite these challenges, **regulation sets societal boundaries of responsible behaviour and could therefore help prevent the spread of synthetic media that could cause demonstratable harm**.[216]

Regional approaches to deepfake governance could aid in overcoming some of the challenges mentioned above. For instance, regional regulation towards deepfake creation and dissemination might have global results. Compelling organizations and individuals to adopt policies and behaviours to comply with regulations in one region might incentivize them to adopt the same policies and behaviours in other regions, resulting in a harmonization effect.[217] The proposed European Union's AI regulatory framework could be a good example of this. Without completely banning the creation of synthetic media, it blacklists deepfakes created to insfluence a person's behaviour that could cause psychological or physical damage. It also institutes "transparency obligations for systems designed to create deepfakes".[218] Since systems that are placed and used in the EU market will have to comply with this regulation regardless of where they are designed and manufactured, companies might decide to adopt the EU's regulations for the products and systems they commercialize worldwide, which could yield global benefits.[219]

At the international level, currently, there is no coordinated effort with respect to the governance of deepfakes,[220] which will be essential given their potential implications for international peace, security, and stability. To this end, as a first step, State and non-State stakeholders

---

210 US National Defense Authorization Act for Fiscal Year 2020; see https://www.congress.gov/116/plaws/publ92/PLAW-116publ92.pdf.
211 Reuters (2019).
212 McCarthy (2021).
213 Kim (2021).
214 Ahmed (2021); Sedova (2021).
215 Jaiman (2021).
216 Collins & Ebrahimi (2021, 3).
217 Heikkilä (2021).
218 Collins & Ebrahimi (2021, 3); Heikkilä (2021).
219 Heikkilä (2021).
220 McCarthy (2021).

need to collaboratively develop a global shared understanding of the AI security landscape. **Since synthetic media is an application of AI and machine learning technologies, developing any international governance framework will first require establishing international AI governance structures.** Considering the distributed nature of deepfake creation and dissemination and the widespread consequences they could cause, this endeavour will demand a multi-stakeholder and multimodal approach that includes a web of top-down governance measures as well as bottom-up initiatives coming from the private sector and AI research communities.

Soft Law

Soft law approaches such as industry standards, codes of conduct, and policies are vital components of the governance framework, especially in the early stages of the development of synthetic media technology and while regulatory frameworks are being established and refined.[221] As the technology is advancing rapidly, soft laws have the benefit of being flexible and adapting to new realities.[222] Moreover, developing principles and codes of conduct to guide research and industry work can further a culture of responsibility and accountability within these communities.

Since deepfakes are mostly created and disseminated using private sector tools and software, the soft law that research communities and technology companies adopt are the first level of protection against the nefarious uses of synthetic media. On the creation dimension, this could translate into codes of conduct or standards for the industry and research communities.[223] For the dissemination phase, it has taken the form of policies of technology companies.[224] Twitter, Facebook, YouTube, Microsoft, and others are taking technological and policy development steps to address deepfakes. In general terms, these do not entail the complete removal of all synthetic media that is shared through their products.[225] Instead, the approach of most organizations has been to label content identified as doctored or synthetic and only remove it if shared in a deceptive manner and if it is likely to cause harm.[226]

While this approach can to an extent help tackle the spread of malicious synthetic media, it is not foolproof. The removal and labelling criteria might overlook potentially dangerous deepfakes, such as doctored satellite images. **'Deepfake geography' might not be a threat for the average digital user, but it could have dire consequences for international and national security and stability**.[227] Furthermore, **labelling content might not work as a blanket strategy because** some users could overlook or disregard labels or even discredit them in favour of their personal biases, limiting the impact of the policy.[228] As with other deepfake countermeasures, **soft law might not be enough on its own to mitigate the risk posed by deepfakes, reinforcing the need for multimodal and multi-stakeholder approaches to tackle their threat**.

221 Kim (2021); Collins (2019, 21).
222 Collins (2019, 21).
223 Florin (2021); Collins (2019, 21).
224 Roth (2021).
225 Roth (2021). Bickert (2020).
226 Twitter (2021).
227 Ahmed (2021).
228 Ahmed (2021).

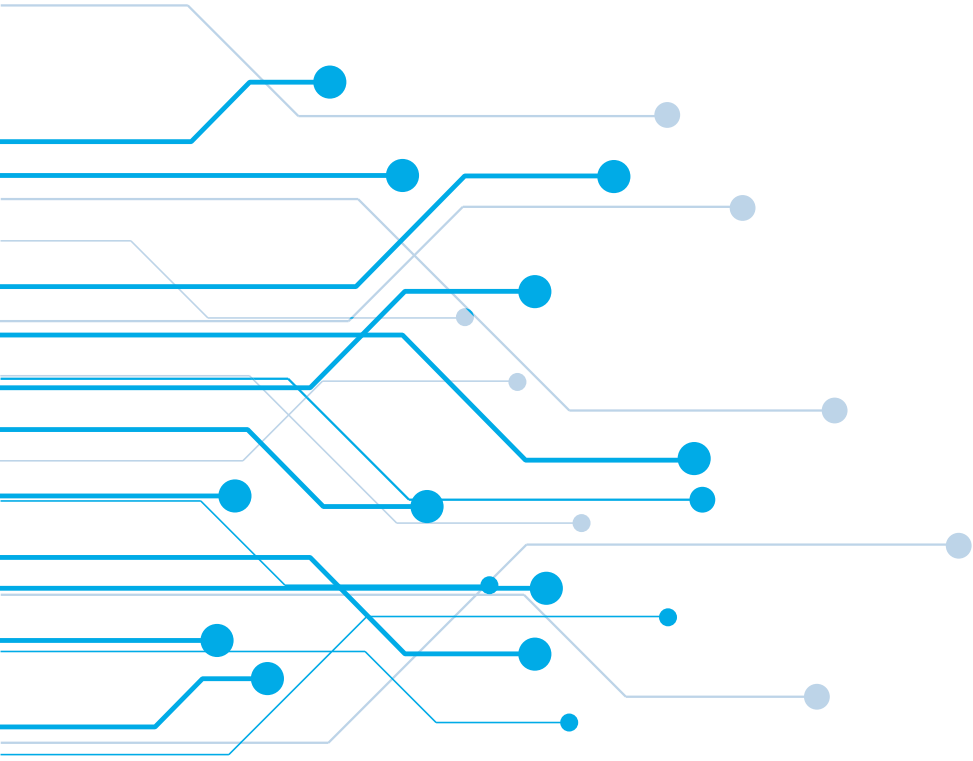
CONCLUSION

Trust is the foundation on which most of our relations and communications in the information age are built, whether interpersonal, communal, societal, or international. Particularly, in international relations, trust is a central pillar of international security and stability underpinning cooperation, multilateteralism, and institution-building. As digital means of communication are becoming ubiquitous, the emergence of **AI-driven hyper-realistic synthetic media generation technology could further subvert trust at the individual, institutional, and societal levels in a time when the international system is already experiencing declining trust among actors and in institutions**. While synthetic media-generation technology is not inherently malign, it allows for the creation of malicious synthetic media or deepfakes that can portray someone doing something they never did or saying something they never said to primarlity mislead, deceive or influence audiences. What makes the advent of deepfakes especially concerning is the steady rise in their quality, quantity, and variety as well as the increasing democratized ability to generate them through the emergence of user-friendly software tools and services.

Most current cases of use of deepfakes are mainly at the individual level, particularly targeting women through fake non-consensual pornography videos. While individual harm will remain a dominant threat, **malicious use of synthetic media technology and its consequences for trust erosion present multifaceted risks for national and international peace, security, and stability**. Ready access to deepfake technology could lower the barriers for a range of State and non-State actors to engage in sophisticated disinformation campaigns and influence operations, and also provide them novel tools to this end. They could also deliver tailored harm or disruption by faciliting cyberattacks through social engineering and spear-phisihing. Furthermore, deepfakes could be employed to polute the information environment which could instigate a crisis or lead to inadvertent escalation during ongoing political and military tensions. One of the most profound implications of deepfakes for trust however is that it perpetuates the 'liar's dividend'–undermining trust in things that are otherwise objectively true–making any action, relationship, decision-making process or transaction that depends on the trustworthiness of information and trust between States, institutions, communities, or individuals potentially vulnerable.

Emerging against the backdrop of growing concerns around the spread of false or malicious information and declining trust in goverments and public institution, deepfakes have a fertile terrain to flourish. They therefore highlight the necessity of fostering trust and protecting shared standards of truth to unlock the true potential of the digital domain. Since synthetic media technology is not currently explicitly addressed under any existing multilateral frameworks in the context of international security, **the international peace and security community will need to undertake a combination of measures to address the multifaceted risks deepfakes present**. This would require a comprehemsive multi-stakeholder and multimodal approach, engaging a range of synthetic media practitioners and stakeholders and collaboratively undertaking education, awareness-raising and threat assessment activities. Such an approach would

be essential to not only inculcate societal resiliency and a culture of responsible innovation, but also build a common foundation on which possible multilateral governance tools could be built– normative and technical guidelines or standards for the responsible use of synthetic media technology as well as norms against its malicious use. Moreover, since synthetic media is an application of AI and machine learning technologies, developing any international governance framework will first require establishing international AI governance structures.



REFERENCE LIST

Agarwal, Shruti, Hany Farid, Yuming Gu, Yuming Gu, Mingming He, Koki Nagano & Hao Li. 2019. 'Protect World Leaders against Deep Fakes.' In I*EEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019: https://farid.berkeley.edu/downloads/publications/cvpr19/cvpr19a.pdf.

Ahmed, Saifuddin. 2021. 'Understanding the Implications for International Security and Stability', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=eIKXPrxeTZk&list=PLEQ2SvONI8gxxwmw-vmLF2eB-WnQtHJrd3&index=5.

Alba, Davey. 2019. 'Facebook Discovers Fakes that Show Evolution of Disinformation.' *The New York Times*, December 20. As of 29 October 2021: https://www.nytimes.com/2019/12/20/business/facebook-ai-generated-profiles.html.

Alford, Anthony. 2021. 'EleutherAI Open-Sources Six Billion Parameter GPT-3 Clone GPT-J.' InfoQ, July 13. As of 30 October: https://www.infoq.com/news/2021/07/eleutherai-gpt-j.

Almars, Abdulqader M. 2021. 'Deepfakes Detection Techniques using Deep Learning: A Survey.' *Journal of Computer and Communications*, 2021, vol. 9: https://www.scirp.org/pdf/jcc_2021051813373227.pdf.

Badr, Will. 2019. 'Auto-Encoder: What Is It? And What Is It Used For? (Part 1).' *Towards Data Science*, 22 April. As of 20 October 2021: https://towardsdatascience.com/auto-encoder-what-is-it-and-what-is-it-used-for-part-1-3e5c6f017726.

BBC. 2019a. 'Deepfake Videos Could "Spark" Violent Social Unrest.' *BBC News*, 13 June. As of 2 May 2021: https://www.bbc.com/news/technology-48621452.

———. 2019b. 'Fake Voices Help Cyber-Crooks Steal Cash.' *BBC News*, 8 July. As of 6 May 2021: www.bbc.com/news/technology-48908736.

Bickert, Monika. 2020. 'Enforcing against Manipulated Media.' *Facebook Newsroom*, January 6. As of 29 October: https://about.fb.com/news/2020/01/enforcing-against-manipulated-media.

Blakemore, Erin. 2020. 'How Photos Became a Weapon in Stalin's Great Purge.' *History.com*, April 8. As of 6 October 2021: https://www.history.com/news/josef-stalin-great-purge-photo-retouching.

Boneh, Dan, Andrew Grotto, Patrick McDaniel & Nicolas Papernot. 2020. *Preparing for the Age of Deepfakes and Disinformation*. Human-Centered Artificial Intelligence (HAI) Policy Brief, Stanford University: https://fsi.stanford.edu/publication/preparing-age-deepfakes-and-disinformation.

Boulanin, Vincent, Lora Saalman, Petr Topychkanov, Fei Su & Moa Peldán Carlsson. 2020. *Artificial Intelligence*, Strategic Stability and Nuclear Risk. SIPRI: https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

Boylan, Jennifer Finney. 2018. 'Will Deep-Fake Technology Destroy Democracy?' *The New York Times*, October 17. As of 15 October 2021: https://www.nytimes.com/2018/10/17/opinion/deep-fake-technology-democracy.html.

Brown, Lee. 2021. '"Deepfake" Tom Cruise Goes Viral on TikTok with over 11 Million Views.' *New York Post*, 2 March. As of 13 October 2021: https://nypost.com/2021/03/02/deepfake-tom-cruise-goes-viral-on-tiktok-with-over-11m-views.

Buchanan, Ben, Andrew Lohn, Micah Musser & Katerina Sedova. 2021. *Truth, Lies, and Automation*. Center for Security and Emerging Technology: https://cset.georgetown.edu/publication/truth-lies-and-automation.

Burton, Tom & Eric Horovitz. 2020. 'New Steps to Combat Disinformation.' Microsoft, 1 September. As of 29 October 2021: https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator.

Butcher, Paul. 2021. 'COVID-19 as a Turning Point in the Fight against Disinformation.' *Nature News*, 25 January. As of 5 October 2021: https://www.nature.com/articles/s41928-020-00532-2.

Castro, Daniel. 2020. 'Deepfakes Are on the Rise—How Should Government Respond?' *GovTech*, January/February. As of 6 October 2021: https://www.govtech.com/policy/deep-fakes-are-on-the-rise-how-should-government-respond.html.

Chitrakorn, Kati. 2021.'How Deepfakes Could Change Fashion Advertising.' *Vogue Business*, 11 January. As of 7 October 2021: https://www.voguebusiness.com/companies/how-deep-fakes-could-change-fashion-advertising-influencer-marketing.

Ciancagliani, Vincenzo, Craig Gibson, David Sancho, Odhran McCarthy, Philip Aman & Aglika Klayn. 2020. *Malicious Uses and Abuses of Artificial Intelligence*. Trend Micro Research, United Nations Interregional Crime and Justice Research Institute & Europol's European Cybercrime Centre, 19 November: https://www.europol.europa.eu/publications-documents/malicious-uses-and-abuses-of-artificial-intelligence.

Ciglic, Kaja. 2021. 'Preserving and Fostering Digital Trust: The Way Forward', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=hYkN3BGJzbQ&list=PLEQ2SvONl8gxxwmw-vmLF2eBWnQtHJrd3&index=8.

Citron, Danielle K. & Robert Chesney. 2019. Deep Fakes: *A Looming Challenge for Privacy, Democracy, and National Security*. 107 California Law Review 1753: https://scholarship.law.bu.edu/faculty_scholarship/640.

Cloutier, Christopher. 2021. 'Much Unclear about "Deepfake Conversation" of MPs with Employee Navalny.' *Netherlands News Live*, 24 April. As of 7 October 2021: https://netherlandsnewslive.com/much-unclear-about-deepfake-conversation-of-mps-with-employee-navalny-2/141736.

Content Authenticity Initiative. 2019. 'Introducing the Content Authenticity Initiative.' Content Authenticity Initiative, 4 November. As of 31 October: https://contentauthenticity.org/blog/test.

Coalition for Content Provenance and Authenticity (C2PA). 2021a. *C2PA Technical Specification*. Public Draft: https://c2pa.org/public-draft/C2PA_Specification.pdf.

———. 2021b. 'FAQ.' As of 31 October: https://c2pa.org/faq.

———. 2021c. 'Overview.' As of 31 October: https://c2pa.org.

Cole, Samantha. 2018. 'Deepfakes Were Created as a Way to Own Women's Bodies—We Can't Forget that.' *Vice*, June 18. As of 31 October: https://www.vice.com/en/article/nekqmd/deepfake-porn-origins-sexism-reddit-v25n2.

Collins, Aengus. 2019. *Forged Authenticity: Governing Deepfake Risks*. EPFL International Risk Governance Center: https://www.epfl.ch/research/domains/irgc/specific-risk-domains/projects-cybersecurity/forging-authenticity-governing-deepfake-risks.

Collings, Aengus & Touradj Ebrahimi. 2021. 'Risk Governance and the Rise of Deepfakes.' EPFL, 12 May: https://www.epfl.ch/research/domains/irgc/spotlight-on-risk-series/risk-governance-and-the-rise-of-deepfakes.

Davis, Raina. 2020. *Technology Factsheet: Deepfakes*. Belfer Center for Science and International Affairs, Harvard Kennedy School: https://www.belfercenter.org/sites/default/files/2020-10/tappfactsheets/Deepfakes.pdf.

Derysh, Igor. 2020. 'GOP House Candidate Publishes Lengthy Report Claiming George Floyd's Killing Was a 'Deepfake' Hoax.' *Salon*, June 24. As of 7 October 2021: https://www.salon.com/2020/06/24/gop-house-candidate-publishes-lengthy-report-claiming-george-floyds-killing-was-a-deepfake-hoax.

De Saulles, Martin. 2021. 'How Deepfakes are a Problem for Us All and Why the Law Needs to Change.' *Information Matters*, 26 March: https://informationmatters.net/deepfakes-problem-why-law-needs-to-change.

Ding, Xinyi, Zohreh Raziei, Eric C. Larson, Eli V. Olinick, Paul Krueger & Michael Hahsler. 2020. *Swapped Face Detection Using Deep Learning and Subjective Assessment. EURASIP Journal on Information Security*, vol. 6: https://jis-eurasipjournals.springeropen.com/articles/10.1186/s13635-020-00109-8.

Dizikes, Peter. 2018. 'Study: On Twitter, False News Travels Faster than True Stories.' *MIT News*, March 8. As of 9 October: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308.

Drew, Alexi. 2021. 'Understanding the Implications for International Security and Stability', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=eIKXPrxeTZk&list=PLEQ2SvONl8gxxwmw-vmLF2eB-WnQtHJrd3&index=5.

Dunn, Suzie. 2021. 'Women, Not Politicians, are Targeted Most Often by Deepfake Videos.' *Centre for International Governance Innovations*, March 3. As of 8 October: https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos.

EG, Meenu. 2021. 'Try These 10 Amazingly Real Deepfake Apps and Websites.' *Analytics Insight*, 19 May. As of 17 October 2021: https://www.analyticsinsight.net/try-these-10-amazingly-real-deepfake-apps-and-websites.

Ellis, Laura 2021. 'Managing the Deepfake Phenomenon—Counter-Deepfake Technologies', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=jcaoexTzr2A&list=PLEQ2SvONl8gxxwmw-vmLF2eB-WnQtHJrd3&index=5.

Federal Bureau of Investigation (FBI) Cyber Division. 2021. *Malicious Actors Almost Certainly Will Leverage Synthetic Content for Cyber and Foreign Influence Operations*. Private Industry Notification: 210310-001, 10 March: https://www.ic3.gov/Media/News/2021/210310-2.pdf.

Ferraro, Matthew F. 2020. 'Congress's Deepening Interest in Deepfakes.' *The Hill*, 29 December. As of 10 October 2021: https://thehill.com/opinion/cybersecurity/531911-congresss-deepening-interest-in-deepfakes.

Florin, Marie-Valentin. 2021. 'Managing the Deepfake Phenomenon—Governance Issues and Responses', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=_1T8IJ5KWDQ&t.

Gill, Amandeep Sigh. 2021. 'Preserving and Fostering Digital Trust: The Way Forward', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=hYkN3BGJzbQ&list=PLEQ2SvONl8gxxwmw-vmLF2e-BWnQtHJrd3&index=8.

GlobaNewswire. 2009. 'New NVIDIA Tesla GPUs Reduce Cost of Supercomputing by a Factor of 10.' *GlobeNewswire*, November 16. As of 30 October 2021: https://www.globenewswire.com/news-release/2009/11/16/1149797/0/en/New-NVIDIA-Tesla-GPUs-Reduce-Cost-of-Supercomputing-by-a-Factor-of-10.html.

Gregory, Sam. 2021. 'Authoritarian Regimes Could Exploit Cries of "Deepfake".' *Wired*, 14 February. As of 8 October: https://www.wired.com/story/opinion-authoritarian-regimes-could-exploit-cries-of-deepfake.

Harwell, Drew. 2018. 'Scarlett Johansson on Fake AI-Generated Sex Videos: "Nothing can Stop Someone from Cutting and Pasting My Image".' *The Washington Post*, December 31. As of 15 October 2021: https://www.washingtonpost.com/technology/2018/12/31/scarlett-johansson-fake-ai-generated-sex-videos-nothing-can-stop-someone-cutting-pasting-my-image.

Hazenberg, Anita. 2021. 'Understanding the Implications for International Security and Stability', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=eIKXPrxeTZk&list=PLEQ2SvONl8gxxwmw-vmLF2eB-WnQtHJrd3&index=5.

Heaven, Will Douglas. 2020. 'OpenAI's New Language Generator GPT-3 is Shockingly Good—and Completely Mindless.' *MIT Technology Review*, July 2020. As of 29 October 2021: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp.

Heaven, Will Douglas. 2021. 'People are Hiring out Their Faces to Become Deepfake-Style Marketing Clones.' *MIT Technology Review*, 27 August. As of 7 October 2021: https://www.technologyreview.com/2021/08/27/1033879/people-hiring-faces-work-deepfake-ai-marketing-clones.

Heikkilä, Juha. 2021. 'Managing the Deepfake Phenomenon—Governance Issues and Responses', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=_1T8IJ5KWDQ&t.

International Criminal Police Organization (INTERPOL). 2020. 'Artificial Intelligence and Law Enforcement: Challenges and Opportunities.' 1 December. As of 10 October 2021: https://www.interpol.int/en/News-and-Events/News/2020/Artificial-Intelligence-and-law-enforcement-challenges-and-opportunities.

International Criminal Police Organization (INTERPOL) & United Nations Interregional Crime and Justice Research Institute (UNICRI). 2020. *Towards Responsible AI Innovation: Second INTERPOL–UNICRI Report on Artificial Intelligence for Law Enforcement*. http://www.unicri.it/towards-responsible-artificial-intelligence-innovation.

International Telecommunications Union (ITU). 2021. 'AI for Good: 2041 Envisioned: AI-Driven Futures According to Kai-Fu Lee.' As of 31 October 2021: https://aiforgood.itu.int/2041-vision-ai-for-good-kai-fu-lee.

Jaiman, Ashish. 2021. 'Unpacking Deepfakes—Creation And Dissemination of Deepfakes', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=nAZ5d6sEVnE&list=PLEQ2SvONl8gxxwmw-vmLF2e-BWnQtHJrd3&index=3.

Jing, Meng. 2019. 'China Issues New Rules to Clamp Down on Deepfake Technologies used to Create and Broadcast Fake News.' *South China Morning Post*., 29 November. As of 29 October 2021: https://www.scmp.com/tech/apps-social/article/3039978/china-issues-new-rules-clamp-down-deepfake-technologies-used.

Jingles, Hong Jing. 2020. 'Realistic Deepfakes in 5 Minutes on Colab.' *Towards Data Science*, 31 March. As of 15 October 2021: https://towardsdatascience.com/realistic-deepfakes-colab-e13ef7b2bba7.

JPEG Fake Media. 2021. 'Exploration on JPEG Fake Media.' *JPEG*. As of 10 October 2021: https://jpeg.org/jpegfakemedia/index.html.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaako Lehtinen & Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition: https://arxiv.org/pdf/1912.04958v2.pdf.

Kim, Yoo-Hyang. 2021. 'Managing the Deepfake Phenomenon—Governance Issues and Responses', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=_1T8IJ5KWDQ&t.

Kyoung-Son, Song. 2021. 'Naver Claims Big Korean-Language Machine-Learning Win.' *Korea JoongAng Daily*, May 25. As of 30 October 2021: https://koreajoongangdaily.joins.com/2021/05/25/business/tech/naver/20210525184900307.html.

Langston, Jennifer. 2017. 'Lip-Syncing Obama: New Tools Turn Audio Clips into Realistic Video.' *UW News*, 11 July. As of 16 October 2021: https://www.washington.edu/news/2017/07/11/lip-syncing-obama-new-tools-turn-audio-clips-into-realistic-video.

Leahy, Connor. 2021. 'Why Release a Large Language Model?' EleutherAI, June 2. As of 29 October 2021: https://blog.eleuther.ai/why-release-a-large-language-model.

Li, Hao. 2021. 'Unpacking Deepfakes: Creation & Dissemination of Video Deepfakes', presentation, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=nAZ5d6sEVnE&list=PLEQ2SvONl8gxxwmw-vmLF2eBWnQtHJrd3&index=4.

Makumane, Moliehi. 2021. 'Understanding the Implications for International Security and Stability', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=eIKXPrxeTZk&list=PLEQ2SvONl8gxxwmw-vmLF2eBWnQtHJrd3&index=5.

Maxime. 2019. 'What is a Transformer?' *Inside Machine Learning*, 4 January. As of 18 October 2021: https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04.

McCarthy, Odrhan. 2021. 'Managing the Deepfake Phenomenon—Governance Issues and Responses', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=_1T8IJ5KWDQ&t.

McGuffie & Newhouse. 2020. 'The Radicalization Risks of GPT-3 and Neural Language Models.' Middelbury Institute of International Studies at Monterrey Center on Terrorism, Extremism, and Counterterrorism, September 9: https://www.middlebury.edu/institute/academics/centers-initiatives/ctec/ctec-publications/radicalization-risks-gpt-3-and-neural-language.

Merzmensch, Vlad Alex. 2020. '2020 Review (GPT-3) | AI as a Poet, Novelist, and Dramaturg.' Medium, February 1, As of 29 October: https://medium.com/merzazine/2020-review-gpt-3-ai-as-a-poet-novelist-and-dramaturg-6cf9fff1c21.

Microsoft. 2021. 'Microsoft and Cloud Computing.' 30 October: https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/#benefits.

Murray, Robin. 2021. 'Travis Use Deep Fake Tech for "Waving at the Window".' *Clash*, 10 March. As of 18 October 2021: https://www.clashmusic.com/videos/travis-use-deep-fake-tech-for-waving-at-the-window.

Nakamitsu, Izumi. 2021. 'Preserving and Fostering Digital Trust: The Way Forward', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=hYkN3BGJzbQ&list=PLEQ2SvONl8gxxwmw-vmLF2e-BWnQtHJrd3&index=8.

Newman, Lily Hay. 2021. 'AI Wrote Better Phishing Emails than Humans in a Recent Test' *Wired*, 7 August. As of 10 October 2021: https://www.wired.com/story/ai-phishing-emails.

Nguyen, Thanh Thi, Quoc Viet Hung Nguyen, Cuong M. Nguyen, Dung Nguyen, Duc Thanh Nguyen & Saeid Nahavandi. s2021. 'Deep Learning for Deepfakes Creation and Detection: A Survey'. *arXiv*: https://arxiv.org/pdf/1909.11573.pdf.

Niessner, Matthias. 2016. 'Face2Face: Real-Time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral).' *YouTube*. As of 26 October 2021: https://www.youtube.com/watch?v=ohmajJTcpNk.

NPR. 2020. 'Conspiracy Theories Aside, Here's What Contact Tracers Really Do.' *NPR*, 14 July: https://www.npr.org/sections/health-shots/2020/07/14/890628203/conspiracy-theories-aside-heres-what-contact-tracers-really-do

PwC. 2021. '2021 Global Digital Trust Insights.' As of 10 October 2021: https://www.pwc.ch/en/insights/cybersecurity/global-digital-trust-insights.html.

OpenAI. 2019. 'Better Language Models and Their Implications.' 14 February. As of 18 December 2021: https://openai.com/blog/better-language-models.

Organization for Economic Co-operation and Development (OECD). 2021. 'Trust in Government.': https://www.oecd.org/gov/trust-in-government.htm.

PAGnol. 2021. 'PAGnol: French Generative Models.' As of 30 October 2021: https://lair.lighton.ai/pagnol.

Paris Call. 2021. 'The 9 principles.' As of 29 October 2021: https://pariscall.international/en/principles.

Patrini, Giorgio. 2019. 'Mapping the Deepfake Landscape.' Sensity, 7 October. As 29 October: https://sensity.ai/blog/deepfake-detection/mapping-the-deepfake-landscape.

———. 2021. 'Managing the Deepfake Phenomenon—Counter-Deepfake Technologies', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=jcaoexTzr2A&list=PLEQ2SvONl8gxxwmw-vmLF2eB-WnQtHJrd3&index=5.

Pauwels, Eleonor. 2021. 'How Can Multilateralism Survive the Era of Artificial Intelligence?' *UN Chronicle*. As of 31 October: https://www.un.org/en/chronicle/article/how-can-multilateralism-survive-era-artificial-intelligence.

Pabian, Frank V., Guido Renda, Rainer Jungwirth, Lance K. Kim, Erik Wolfart, Giacomo G. M. Cojazzi, Willem A. Janssens. 2020. 'Commercial Satellite Imagery: An Evolving Tool in the Non-proliferation Verification and Monitoring Toolkit', in Niemeyer I., Dreicer M., Stein G. (eds). *Nuclear Non-proliferation and Arms Control Verification*. Springer, Cham: https://link.springer.com/chapter/10.1007%2F978-3-030-29537-0_24.

Porter, Jon. 2019. 'Another Convincing Deepfake App Goes Viral Prompting Immediate Privacy Backlash.' *The Verge*, 2 September. As of 15 October 2021: https://www.theverge.com/2019/9/2/20844338/zao-deepfake-app-movie-tv-show-face-replace-privacy-policy-concerns.

Pinscreen. 2021. 'The Most Advanced AI-Driven Virtual Assistants.' As of 17 October 2021: https://www.pinscreen.com.

Project Origin. 2021. 'About.' As of 31 October: https://www.originproject.info/about.

Provenance. 2021a. 'Consortium: Consortium Members.' As of 31 October 2021: https://www.provenanceh2020.eu/consortium/consortium-members.

———. 2021b. 'Providing Verification Assistance for New Content. Fact Sheet.' *European Commission CORDIS*. As of 31 October: https://cordis.europa.eu/project/id/825227.

Rea, Naomi. 2019. 'Artists Create a Sinister "Deepfake" of Mark Zuckerberg to Teach Facebook (and the Rest of Us) a Lesson About Digital Propaganda.' *Artnet New*, 12 June. As of 15 October 2021: https://news.artnet.com/art-world/mark-zuckerberg-deepfake-artist-1571788.

Reface. 2021. 'About.' As of 15 October 2021: https://hey.reface.ai/about.

Reuters. 2019. 'China Seeks to Root Out Fake News and Deepfakes with New Online Content Rules.' *Reuters*, 29 November. As of 31 October 2021: https://www.reuters.com/article/us-china-technology/china-seeks-to-root-out-fake-news-and-deepfakes-with-new-online-content-rules-idUSKBN1Y30VU.

———. 2020. 'Fact check: "Drunk" Nancy Pelosi Video is Manipulated.' *Reuters*, 3 August. As of 17 October 2021: https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated-idUSKCN24Z2BI.

Rella, Sirisha & Siddhartha Sharma. 2021. 'Announcing Megatron for Training Trillion Parameter Models & NVIDIA Riva Availability.' *Nvidia Developer Blog*, April 12. As of 30 October 2021: https://developer.nvidia.com/blog/announcing-megatron-for-training-trillion-parameter-models-riva-availability.

Romero, Alberto. 2021. 'Can't Access GPT-3? Here's GPT-J—Its Open-Source Cousin.' *Towards Data Science*, June 24. As of 30 October 2021: https://towardsdatascience.com/cant-access-gpt-3-here-s-gpt-j-its-open-source-cousin-8af86a638b11.

Roth, Yoel. 2021. 'Managing the Deepfake Phenomenon—Governance Issues and Responses', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=_1T8IJ5KWDQ&t.

Sagar, Ram. 2020. 'OpenAI's GPT-3 Can Now Generate the Code for You.' *Analytics India Magazine*, 20 December. As of 29 October 2021: https://analyticsindiamag.com/open-ai-gpt-3-code-generator-app-building.

Scharre, Paul & Michael Horowitz. 2018. 'Artificial Intelligence: What Every Policymaker Needs to Know.' *Center for a New American Security*, June 19. As of 18 October 2020: https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know.

Schick, Nina. 2021. 'Deepfakes and the Age of Synthetic Media', keynote address, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=R4hVYhh4EIM&list=PLEQ2SvONI8gxxwmw-vmLF2eBWnQtHJrd3&index=3.

Science Daily. 2020. '"Deepfakes" Ranked as Most Serious AI Crime Threat.' *ScienceDaily*, 4 August: https://www.sciencedaily.com/releases/2020/08/200804085908.htm.

Sedova, Katerina. 2021. 'The 2021 Innovations Dialogue: Unpacking Text Deepfakes', presentation, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=nAZ5d6sEVnE&list=PLEQ2SvONl8gxxwmw-vmLF2e-BWnQtHJrd3&index=3.

Shetty, Badreesh. 2018. 'National Language Processing (NLP) for Machine Learning.' *Towards Data Science*, 24 November. As of 18 October 2021: https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b.

Siarohin, Aliaksandr, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci & Nicu Sebe. 2019. *First Order Motion Model for Image Animation*. Advances in Neural Information Processing Systems 32: https://proceedings.neurips.cc/paper/2019/hash/31c0b36aef265d9221af-80872ceb62f9-Abstract.html.

Smith, Hannah & Katherine Mansted. 2020. *Weaponised Deep Fakes—National Security and Democracy*. Australian Strategic Policy Institute, 20 April: https://www.aspi.org.au/report/weaponised-deep-fakes.

Snapp, Mary. 2021. 'An Update on our Effort to Help Preserve and Protect Journalism.' Microsoft, 16 June: https://blogs.microsoft.com/on-the-issues/2021/06/16/microsoft-journalism-initiative-pilots-update.

Soo-Yeon, Yoon. 2020. 'Drag and Drop: Deepfakes Create a New Kind of Crime.' *Korea JoongAng Daily*, 17 May. As of 30 November 2021: https://koreajoongangdaily.joins.com/2020/05/17/features/deepfake-artificial-intelligence-pornography/20200517190700189.html.

Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford et al. 2019. 'OpenAI Report: Release Strategies and the Social Impacts of Language Models.' *OpenAI*, November 2019. As of 30 November 2021: https://d4mucfpksywv.cloudfront.net/papers/GPT_2_Report.pdf.

Solaiman, Irene & Christy Dennison. 2021. *Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets*. *OpenAI*: https://cdn.openai.com/palms.pdf.

Stupp, Catherine. 2019. 'Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case.' *The Wall Street Journal*, 30 August. As of 10 October 2021: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

Sundar, S. Shyam. 2008. 'The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility', in Miriam J. Metzger and Andrew J. Flanagin (eds). *Digital Media, Youth, and Credibility*: https://www.issuelab.org/resources/875/875.pdf.

Suwajanakorn, Supasorn, Steven M. Seitz & Ira Kemelmacher-Shlizerman. 2017. 'Synthesizing Obama: Learning Lip Sync from Audio'. ACM Transactions on Graphics, vol. 36, no. 4: https://doi.org/10.1145/3072959.3073640.

Swerling, Gabriella. 2020. 'Doctored Audio Evidence Used to Damn Father in Custody Battle.' *The Telegraph*, 31 January. As of 10 October 2021: https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence.

The *Economist. 2021. 'AI is Transforming the Coding of Computer Programs.'* The Economist, 10 July. As of 29 October 2021: https://www.economist.com/science-and-technology/2021/07/07/ai-is-transforming-the-coding-of-computer-programs.

The Guardian. 2020. 'A Robot Wrote this Entire Article. Are You Scared Yet, Human?' *The Guardian*, 8 September. As of 29 October: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

Topychkanov, Petr. 2021. 'Understanding the Implications for International Security and Stability', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=eIKXPrxeTZk&list=PLEQ2SvONl8gxxwmw-vmLF2eBWnQtHJrd3&index=5.

Turek, Matt. 2021. 'Semantic Forensics (SemaFor).' Defense Advanced Research Projects Agency. As of 29 October 2021: https://www.darpa.mil/program/semantic-forensics.

Twitter. 2021. 'Synthetic and Manipulated Media Policy.': https://help.twitter.com/en/rules-and-policies/manipulated-media.

United Nations General Assembly (UNGA). 2015. *Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. UN document A/70/174, 22 July.

———. 2018a. *Current Developments in Science and Technology and Their Potential Impact on International Security and Disarmament Efforts*. UN document A/73/177, 17 July.

———. 2018b. *Role of Science and Technology in the Context of International Security and Disarmament*. UN document A/RES/73/32, 11 December.

———. 2020. *Roadmap for Digital Cooperation: Implementation of the Recommendations of the High-Level Panel on Digital Cooperation*. UN document A/74/821, 29 May.

———. 2021. *Current Developments in Science and Technology and Their Potential Impact on International Security and Disarmament Efforts*. UN document A/76/182, 19 July.

United Nations Institute for Disarmament Research (UNIDIR). 2021a. "Managing the Deepfake Phenomenon—Counter-Deepfake Technologies", panel discussion, *The 2021 Innovations Dialogue: Deepfakes, Trust and International Security*, Geneva, 25 August 2021: https://www.youtube.com/watch?v=jcaoexTzr2A&t.

———. 2021b. “Understanding the Implications for International Security and Stability”, panel discussion, *The 2021 Innovations Dialogue: Deepfakes, Trust and International Security*, Geneva, 25 August 2021: https://www.youtube.com/watch?v=eIKXPrxeTZk&t.

———. 2021c. “Unpacking Deepfakes—Creation and Dissemination of Deepfakes”, panel discussion, *The 2021 Innovations Dialogue: Deepfakes, Trust and International Security*, Geneva, 25 August 2021: https://www.youtube.com/watch?v=nAZ5d6sEVnE&t.

———. 2021d. “Managing the Deepfake Phenomenon—Governance Issues and Responses”, panel discussion, *The 2021 Innovations Dialogue: Deepfakes, Trust and International Security*, Geneva, 25 August 2021: https://www.youtube.com/watch?v=_1T8IJ5KWDQ&t.

———. 2021e. “Preserving and Fostering Digital Trust”, panel discussion, *The 2021 Innovations Dialogue: Deepfakes, Trust and International Security*, Geneva, 25 August 2021: https://www.youtube.com/watch?v=hYkN3BGJzbQ&t.

United Nations Office of Counter-Terrorism (UNOCT) & United Nations Interregional Crime and Justice Research Institute (UNICRI). 2021a. *Countering Terrorism Online with Artificial Intelligence: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia*: http://213.254.5.198/sites/default/files/2021-06/Countering%20Terrorism%20Online%20with%20AI%20-%20UNCCT-UNICRI%20Report.pdf.

———. 2021b. *Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes*: http://unicri.it/sites/default/files/2021-06/Malicious%20Use%20of%20AI%20-%20UNCCT-UNICRI%20Report_Web.pdf.

United Nations Office of the Secretary-General’s Envoy on Technology. 2021. ‘High-Level-Panel for Digital Cooperation Launches Report & Recommendations for Building an Inclusive Digital Future.’ As of 11 October 2021: https://www.un.org/techenvoy/es/news/HLP%20report%20launch.

United Nations Secretary-General. 2018. ‘Secretary-General’s Address to the General Assembly.’ 25 September: https://www.un.org/sg/en/content/sg/statement/2018-09-25/secretary-generals-address-general-assembly-delivered-trilingual.

———. 2020. ‘Secretary-General’s remarks to the General Assembly on his priorities for 2020.’ 22 January: https://www.un.org/sg/en/content/sg/statement/2020-01-22/secretary-generals-remarks-the-general-assembly-his-priorities-for-2020-bilingual-delivered-scroll-down-for-all-english-version.

———. 2021. ‘António Guterres (UN Secretary-General) on the High-Level Conference of Heads of Counter-Terrorism Agencies of Member States.’ 28 June: https://media.un.org/en/asset/k14/k14mhg667h.

Van Huijstee, Mariëtte, Pieter van Boheemen, Djurre Das, Linda Nierling, Jutta Jahnel, Murat Karaboga & Martin Fatun. 2021. *Tackling Deepfakes in European Policy*. Scientific Foresight Unit, European Parliamentary Research Service: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.

Velásquez, N. Leahy R., Restrepo, N. J. et al. 2021. 'Online Hate Network Spreads Malicious Covid-19 Content outside the Control of Individual Social Media Platforms.' *Sci Rep* 11, 11549: https://doi.org/10.1038/s41598-021-89467-y.

Vignard, Kerstin. 2021. 'Understanding the Implications for International Security and Stability', panel discussion, The 2021 Innovations Dialogue, Geneva, 25 August 2021: https://www.youtube.com/watch?v=eIKXPrxeTZk&list=PLEQ2SvONI8gxxwmw-vmLF2eB-WnQtHJrd3&index=5.

Vincent, James. 2021. 'Deepfake Satellite Imagery Poses a Not-so-Distant Threat, Warn Geographers.' *The Verge*, 27 April. As of 10 October 2021: https://www.theverge.com/2021/4/27/22403741/deepfake-geography-satellite-imagery-ai-generated-fakes-threat.

Wallis, Jacob, Ariel Bogle, Albert Zhang & Hillary Mansour. 2021. *Influence for Hire: The Asia-Pacific's Online Shadow Economy*, Australian Strategic Policy Institute, 10 August: https://www.aspi.org.au/report/influence-hire.

Wiggers, Kyle. 2020. 'OpenAI's Massive GPT-3 Model is Impressive, but Size isn't Everything'. *VentureBeat*. As of 29 October 2021: https://venturebeat.com/2020/06/01/ai-machine-learning-openai-gpt-3-size-isnt-everything.

———. 2021a. 'Study Warns Deepfakes can Fool Facial Recognition.' *VentureBeat*, 5 March. As of 10 October 2021: https://venturebeat.com/2021/03/05/study-warns-deepfakes-can-fool-facial-recognition/.

———. 2021b. 'Huawei Trained the Chinese-Language Equivalent of GPT-3.' *VentureBeat*, 29 April. As of 30 October 2021: https://venturebeat.com/2021/04/29/huawei-trained-the-chinese-language-equivalent-of-gpt-3.

Williams, Brad D. 2021. 'Researchers Warn Of "Dangerous" Artificial Intelligence-Generated Disinformation at Scale.' *Breaking Defense*, September 30. As of 30 October: https://breakingdefense.com/2021/09/researchers-warn-of-dangerous-artificial-intelligence-generated-disinformation-at-scale.

Williams, Heather & Alexi Drew. 2020. *Escalation by Tweet: Managing the New Nuclear Diplomacy*. Canter for Science & Security Studies, King's College London: https://www.kcl.ac.uk/csss/assets/escalation-by-tweet-managing-the-new-nuclear-diplomacy-2020.pdf.

the 2021
innovations dialogue.

DEEPFAKES, TRUST & INTERNATIONAL SECURITY

25 AUGUST 2021 | 09:00-17:30 CEST

WELCOME AND OPENING REMARKS — 09:00

Robin Geiss — *United Nations Institute for Disarmament Research*

KEYNOTE ADDRESS: TRUST AND INTERNATIONAL SECURITY IN THE ERA OF DEEPFAKES — 09:10

Nina Schick — *Tamang Ventures*

The scene-setting keynote address will explore the importance of trust for international security and stability and shed light on the extent to which the growing deepfake phenomenon could undermine this trust.

UNPACKING DEEPFAKES – CREATION AND DISSEMINATION OF DEEPFAKES — 09:40

MODERATED BY:

Giacomo Persi Paoli — *United Nations Institute for Disarmament Research*

FEATURING:

Ashish Jaiman — *Microsoft*
Hao Li — *Pinscreen and University of California, Berkeley*
Katerina Sedova — *Center for Security and Emerging Technology, Georgetown University*

This panel will provide a technical overview of how visual, audio and textual deepfakes are generated and disseminated. The key questions it will seek to address include: what are the underlying technologies of deepfakes? How are deepfakes disseminated, particularly on social media and communication platforms? What types of deepfakes currently exist and what is on the horizon? Which actors currently have the means to create and disseminate them?

COFFEE BREAK — 11:15

UNDERSTANDING THE IMPLICATIONS FOR INTERNATIONAL SECURITY AND STABILITY — 11:30

MODERATED BY:

Kerstin Vignard — *United Nations Institute for Disarmament Research*

FEATURING:

Saiffudin Ahmed — *Nanyang Technological University*
Alexi Drew — *RAND Europe*
Anita Hazenberg — *INTERPOL Innovation Centre*
Moliehi Makumane — *United Nations Institute for Disarmament Research*
Carmen Valeria Solis Rivera — *Ministry of Foreign Affairs of Mexico*
Petr Topychkanov — *Stockholm International Peace Research Institute*

This panel will examine the uses of deepfakes and the extent to which they could erode trust and present novel risks for international security and stability. The key questions it will seek to address include: how could deepfakes be used? What is the potential geopolitical and societal impact of deepfakes on the stability, integrity and trustworthiness of institutions, the information ecosystem and society more broadly? What are the existing risks to international security and stability, and what future risks may arise?

LUNCH BREAK — 13:00

MANAGING THE DEEPFAKE PHENOMENON – COUNTER-DEEPFAKE TECHNOLOGIES — 14:00

MODERATED BY:

Arthur Holland Michel — *United Nations Institute for Disarmament Research*

FEATURING:

Laura Ellis — *BBC*

Giorgio Patrini — *Sensity*

This panel will give an overview of the technological countermeasures being developed and reflect on their effectiveness to combat the risks posed by the malicious uses of deepfakes, particularly in light of the pace at which deepfake technology is advancing.

MANAGING THE DEEPFAKE PHENOMENON – GOVERNANCE ISSUES AND RESPONSES — 15:00

MODERATED BY:

Giacomo Persi Paoli — *United Nations Institute for Disarmament Research*

FEATURING:

Marie-Valentine Florin — *International Risk Governance Center, EPFL*

Juha Heikkilä — *European Commission*

Yoo Hyang Kim — *National Assembly Research Service of the Republic of Korea*

Odhran McCarthy — *United Nations Interregional Crime and Justice Research Institute*

Yoel Roth — *Twitter*

This panel will explore the key governance issues concerning deepfakes and discuss what governance measures are needed to respond to them. The key questions this panel will seek to address include: what governance challenges do deepfakes present? What industry-led, national and regional governance responses against deepfakes are emerging? How can we create synergies between the various bottom up and top down governance measures? What new, if any, multilateral and multi-stakeholder tools are needed to fill governance gaps?

COFFEE BREAK — 16:10

PRESERVING AND FOSTERING DIGITAL TRUST: THE WAY FORWARD — 16:20

MODERATED BY:

Robin Geiss — *United Nations Institute for Disarmament Research*

FEATURING:

Kaja Ciglic — *Microsoft*

Amandeep Singh Gill — *International Digital Health & AI Research Collaborative*

Izumi Nakamitsu — *United Nations Office for Disarmament Affairs*

In the era of digital transformation, digital technologies now underpin core societal functions. Our ability to unlock the true potential of digital technologies and leverage their transformative benefits for society, economy and the environment is dependent on preservation of trust and the stability of the digital ecosystem. And as international security and stability are increasingly dependent on digital security and stability, the international community urgently needs to prioritize issues of digital trust and security and take concrete steps to protect and promote shared standards of truth to harness the transformative benefits of the digital domain. Through an open discussion, this panel will reflect on what digital trust entails. What are the challenges to preserving and fostering digital trust? What initiatives can be taken by the international community to build digital trust and mitigate the potential destabilizing effects of advances in digital technologies?

CLOSING REMARKS — 17:20

Robin Geiss — *United Nations Institute for Disarmament Research*

CONFERENCE MATERIALS

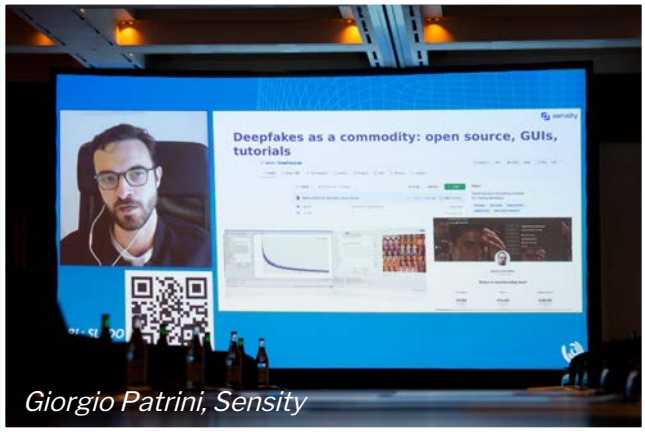
The video recording of the conference is available on UNIDIR's website here.

Panel on Managing the Deepfake Phenomenon – Governance Issues and Responses

Yoo Hyang Kim, National Assembly Research Service of the Republic of Korea

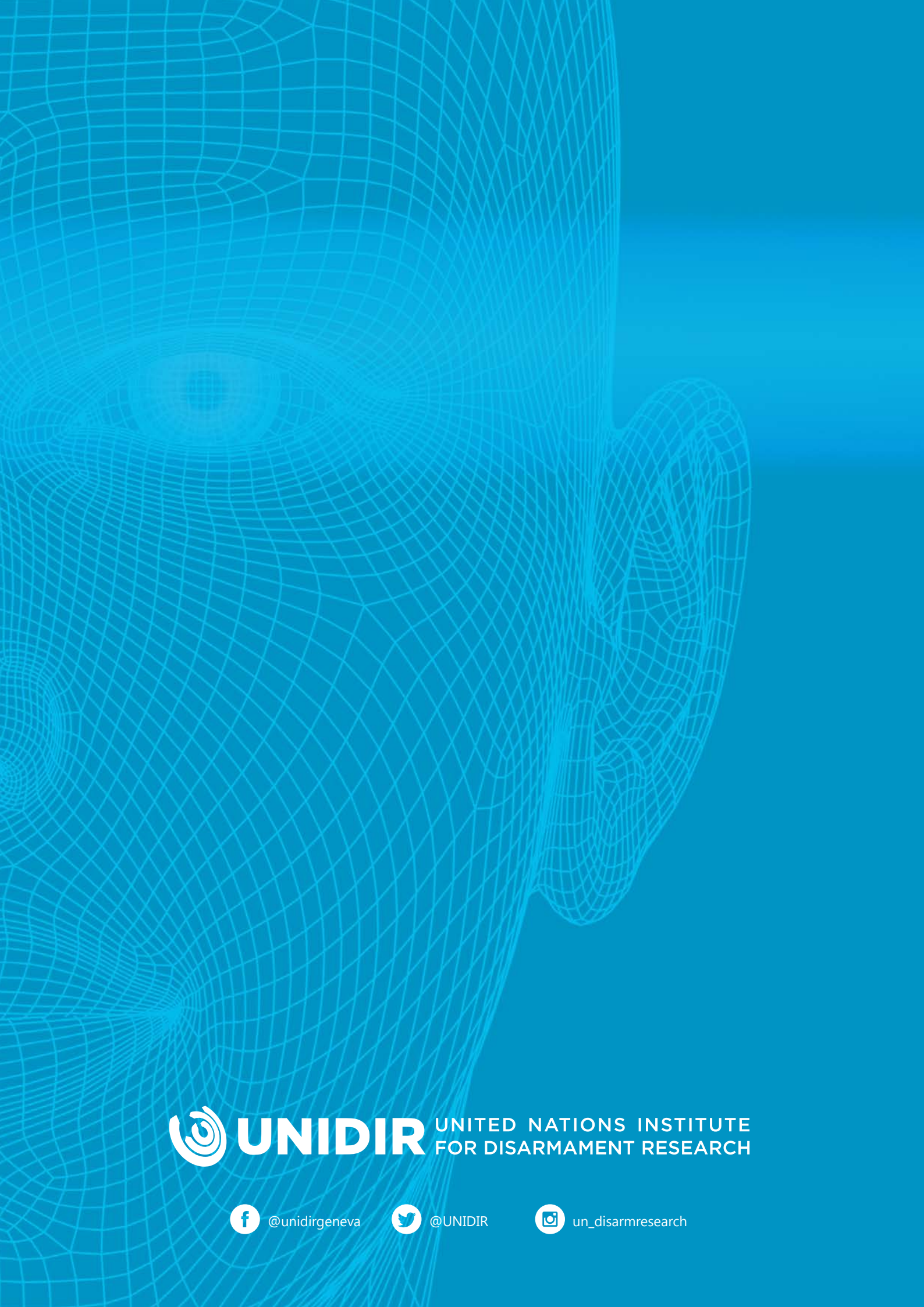
Alexi Drew, RAND Europe and Kerstin Vignard, UNIDIR

Katerina Sedova, CSET



Giorgio Patrini, Sensity

Marie-Valentine Florin, International Risk Governance Center at EPFL



UNIDIR UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH

@unidirgeneva @UNIDIR un_disarmresearch