BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Rolling in the deep(fakes)

by Sabina Marchetti

February 2022

Number 668

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Rolling in the deep(fakes)

by Sabina Marchetti

Number 668 – February 2022

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6627 (print)
ISSN 1972-6643 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

ROLLING IN THE DEEP(FAKES)

by Sabina Marchetti*

Abstract

Deepfakes are digital forgeries. They are highly credible multimedia representations of altered or fabricated events, created using sophisticated artificial intelligence (AI) techniques. Despite the remarkable contribution of the underlying technology to innovation in several fields, deepfakes per se are a powerful weapon for disinformation and fraudulent operations.
In the financial sector, the increasing importance of online platforms for payments and banking exposes consumers and retail investors to AI-enabled attacks. Moreover, at the macro level, malicious dissemination of deepfakes through information channels such as social media can sow distrust toward financial institutions, and ultimately have systemic effects.
In this paper, we describe the rapidly evolving deepfake technology, with a focus on the threats it poses to the financial sector. We then propose an analytical approach and a set of policy instruments for the effective countering of malicious deepfakes.

JEL Classification: O31, O32.
Keywords: deepfakes, artificial intelligence, disinformation, financial system.
DOI: 10.32057/0.QEF.2022.0668

Contents

* Bank of Italy, Directorate General Economics, Statistics and Research.

1. Introduction[1]

The financial sector has long been a prime target for malicious tools. Over the past years, the fast-paced shift of traditional payments and banking towards online platforms has increasingly exposed consumers and retail investors to fraudulent activities on cyberspace. With the advent of deepfakes, malicious operations run on digital channels have greatly improved in their effectiveness.

Deepfakes are digital forgeries created by artificial intelligence (AI) methods, providing highly credible altered or fabricated representations of multimedia contents.[2] Although not flawless, their realistic appearance, combined with individuals' natural aptitude to trust what they see and perceive with their senses to be factual, provides a remarkable contribution to innovation in several fields (Schetinger et al., 2017). When employed for deceptive purposes, deepfakes have the potential to be unsettling in their consequences, both for individuals and at the systemic level, especially when spread by social platforms (Chesney and Citron, 2019b). Nevertheless, synthetic contents are not illegal per se and, as of today, their malicious use might be prosecuted only if the conduct at stake falls under specific forms of offence (Bateman et al., 2021).

The present paper describes the emerging threat from deepfakes, and provides input on how authorities should respond to it. Section 2 outlines the different roles such synthetic media play to improve the disruptiveness of malicious initiatives targeting individuals as well as institutions. Our focus is on the extent to which AI-enabled operations relate to the financial sector, and affect it. We outline the main policy measures that are already in place against deceptive use of AI-enabled media in Section 3, and contribute to the debate over the regulation of the topic discussing how interventions should be improved or complemented in Section 4. Overall, we conclude that, in order to be effective, malicious operations leveraging on deepfakes ought to be countered envisioning a case-by-case approach, which encompasses the technical specificities of various applications.

2. Deep Impact

Technological advances of AI enhance highly flexible and realistic content generation, making it hard to discern synthetic media from real ones. Unlike *cheapfakes* - media altered with editing software tools - deepfakes are the primary outcome of an AI-based suite of methods[3] enabling malicious actors to spread digital disinformation.

To assess the threat severity of AI-enabled disinformation activities, it is useful to start from two dimensions: the primary aim and the possible collateral harms.[4] The first measure encompasses financial loss, reputational damage, undermining of trust in public institutions as well as terror, whereas the second captures the breadth of targeted audience – from single individuals to institutions to communities - that are affected. Along this second dimension, disinformation can be classified as either:

- *Narrowcast* (Individual Harm): target individual agents, consumers, public figures or companies;
- *Broadcast* (Collective Harm): exploit and magnify distrust in institutions, political and financial ones.

2.1 Narrowcast Initiatives

The spectrum of narrowcast disinformation ranges from commercial exploitation to identity theft for payments frauds, via scams, cyber-threats and blackmailing.

[1] The author is grateful to Claudia Biancotti, Oscar Borgogno, Michele Savini Zangrandi and Giovanni Veronese for their comments.

[2] One of the most famous early examples of AI-enabled synthetic content is a 2018 clip where a digital rendition of US President Barack Obama aptly states: "We're entering an era in which our enemies can make anyone say anything at any point in time". Available: https://www.youtube.com/watch?v=cQ54GDm1eL0 .

[3] See the Appendix for insights on the main technical aspects of deepfakes.

[4] Disinformation campaigns pursuing monetary profit can cause collateral reputational damage, or even chase terror while casting uncertainty and distrust among societies (Caldwell et al., 2020).

Within e-commerce and online activities, several operations leverage on social media for microtargeting[5]. In their mildest form, they pursue personalised design of synthetic content, to attract clicks and encourage traffic toward a web page (Vaccari and Chadwick, 2020). Such practices usually entail *click-baiting*, defined as luring end users in exchange for login information or private data, whose effectiveness may benefit from AI-enhanced sensationalised narratives, as well as fake video footages for misleading advertising (Kietzmann et al., 2020). As for consumer exploitation, deepfakes support manipulative marketing, orienting consumers' perceived needs toward specific products or services.[6]

More frequently, deepfakes are also deployed in support of narrowcast cyber-attacks, mostly:

i) Spear-phishing: electronic communications that are falsely presented as coming from a certain sender (spoofing), targeting specific individuals, organizations or business with the purpose of stealing sensitive information or spreading malware.[7] Deepfakes were reported to increase the success rate of spear-phishing attacks from 60-70 percent to 100 percent.[8]

ii) Identity theft: Cyber-enabled theft of personal information and/or credentials to impersonate the victim. Such practice pursues facilitation or funding other criminal activities, like terrorism. With deepfakes, it may lead to impersonation for payment frauds;

iii) Blackmailing / Cyber-Extortion: Payment request upon threatening the victim of diffusion of slanderous material. In the case of deepfakes, this is typically a compromising video feed.

The COVID-19 pandemic, by accelerating the shift from traditional to electronic payments and banking, has also magnified the threat from deepfake technology.[9] The reduction of face-to-face interactions forced by the quarantine regimes revealed financial system vulnerabilities to impersonation frauds, ranging from voice cloning to fake video footage. Ghost, new-account and synthetic identity frauds are among the most frequently reported. Ghost frauds rely on deepfakes to impersonate deceased persons and access to services and benefits on their behalf, from creditworthiness to pension and other economic benefits. New-account and synthetic identity frauds exploit stolen or fabricated data, respectively, to apply for credit cards or loans, as well as to bolster and improve additional fake customers' creditworthiness.[10] The effectiveness improvement enhanced by AI methods builds on their ability to deceive biometric-based authentication protocols. This prompted several financial institutions to increase the sophistication in their authentication protocols to ward off fraudsters pursuing customer impersonation.[11]

2.2 Broadcast Initiatives

Broadcast weaponisation of deepfake technology usually wedges into pre-existing narratives or cultural and social faultlines within societies. Their aim is some form of societal subversion[12] and financial destabilisation.

When employed for subversion, disinformation campaigns either leverage floods of deceptive contents (propaganda) or require long-term strategies and investments (political interference). Deepfakes improve

[5] Microtargeting initiatives track consumers' browsing habits, e.g. with web "cookies", to build up profiles and deliver highly personalised advertisements.

[6] See: https://www.voguebusiness.com/companies/how-deepfakes-could-change-fashion-advertising-influencer-marketing .

[7] The term usually entails email communications, whereas attacks carried out via phone calls, SMS or web platforms are referred to as, respectively, vishing, smishing and baiting.

[8] See: https://www.ft.com/content/8a5fa5b2-6aac-41cf-aa52-5d0b90c41840 .

[9] See: https://nilsonreport.com/publication_chart_and_graphs_archive.php .

[10] Authentication protocols and identity check routines against impersonation frauds with deepfakes are already available for use to banking and payments operators. As of January 2021, Anna Money, ABN Amro, Aegon, Caixa Bank, Chase, HSBC, ING, Mastercard and Rabobak adopted digital identity verification technologies based on AI. See for details: https://deepware.ai/deepfake-backed-financial-crimes-spread-more-easily-during-pandemic/ .

[11] See: https://www.consumeraffairs.com/finance/identity-theft-statistics.html .

[12] Within the current framework, we refer to subversion as either pursuing i) political interference via propaganda or smear campaigns, or ii) dissemination and promotion of fake content, to promote conspiracy theories and cause reputational damage to public actors. Subversion may include other types of initiatives that shall be classified according to their level of severity (Kastner and Wohlforth, 2021).

effectiveness of operations for both: they enable fabrication of defamatory content to inflict reputational damage upon public figures as well as impersonation, to increase uncertainty and distrust among societies (Chesney and Citron, 2019a).

Political interference typically relies on deepfakes to imbue vulnerable communities with the desired messages on social media platforms, and successively control and hijack them. This kind of socio-political subversive initiatives aim at distracting and/or weakening communities and at gaining influence. Dedicated actors, like the well-known *Cambridge Analytica*, usually carry out the enacting of broadcast political disinformation strategies.[13] In the geopolitical chessboard, Russia established itself as the leader in the use of information as a "weapon of psychological warfare" for interference[14]. Russia first promoted anti-US conspiracy theories on AIDS with *Operation Infektion* in the 1980s, and pursued influence of foreign elections throughout the 2010s, with the *Internet Research Agency* (Galeotti, 2019, White, 2016).[15]

Beside propaganda and political interference, deepfakes for broadcast disinformation could also serve as a weapon in the hand of terrorists or cyber-criminals, pursuing systemic subversion by sowing panic[16], undermining diplomatic initiatives (Aftergood, 2017) and jeopardizing public safety (Allen and Chan, 2017).

Deepfakes could increase the effectiveness of broadcast disinformation initiatives, leveraging the growing influence of social media in various areas of social and economic life, including finance.[17] Although deployment of systemic AI-enabled operations has not been documented thus far, as on January 2021 the Gamestop-WallStreetBets saga put on a spotlight on the extent to which coordination on social media platforms can influence the stock market (Biancotti and Ciocca, 2021). Smear campaigns could pursue sabotage of a company, to put it out of business while wreaking havoc in the markets. Such operations may consist in the fabrication and diffusion of synthetic media displaying market-moving events. They may also combine deepfakes with bots, i.e. pieces of computer code that post content on social platforms, to spread rumours, altering the sentiment of consumers and investors. The effectiveness of such initiatives in causing malicious flash crashes would require their framing within a social media operation, to convey fake media across platforms. This is not an easy goal to attain - nevertheless, the achievability of the altered media generation process *per se* might be worth the attempt and cause several fake news poisoning the perception (and sentiment) of investors. AI-enhanced disinformation could also target banking institutions, pursuing destabilisation of the system, to the point of triggering runs on cash. Analogously, doctored media could threaten stability of the electronic payment system, in force of the substantial weight it has gained throughout the last decades.

Finally, deepfakes could be weaponised against regulatory institution, by simulating imminent enforcement of government measures or policy shifts. Analogously to the case of systemic market manipulation, no disinformation campaign based on deepfakes has been publicly documented yet in this respect. Nevertheless, vulnerability of public institutions to rumours spread on the web has been repeatedly exposed over the last years[18], and it was evident even on traditional media. Deepfakes can also be leveraged for *astroturfing* - the practice of devising virtual communities showing overt grassroots support over a particular regulatory

[13] See: https://www.bbc.com/news/av/world-43472347 .

[14] The term "disinformation" is traced back to the 19th century. Remarkably, its origin is often attributed to the Soviet "special disinformation office" for propaganda, constituted in 1923. The word would just be the literal translation of the Russian word for "misinformation": дезинформация ("dezinformatsiya").

[15] Some contend that during the Covid-19 pandemic both the US and China engaged in online disinformation activities. Other emerging countries active in the political use of Deepfakes for broadcast disinformation initiatives include Iran, Saudi Arabia, United Arab Emirates, North Korea, as well as some communities internal to the US (Schick, 2020).

[16] E.g. In 2014 the Muslim minority of Myanmar was attacked after fake news on a child's raping episode circulated on social media Facebook. In 2021, fake news spreading on instant messaging platform WhatsApp caused harmful behaviours yielding provisional suspension of the service and accounts banning operations in India.

[17] See: https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/?sh=68e697e72fac . See also: https://www.cnbc.com/2021/01/29/elon-musks-tweets-are-moving-markets.html .

[18] E.g. https://www.dnaindia.com/business/report-buzz-of-430-billion-loss-top-banker-defection-freaks-out-china-1431780 .

initiative on social media platforms, to influence the rulemaking processes (Forest, 2021). Table 1 provides a summary of possible uses of deepfakes for cyber operations in the financial system, as illustrated by the Carnegie Endowment for International Peace (CEIP).

Table 1

Target	Scenario	Role of Synthetic Media	Key Malicious Technique
Individuals	1. Identity theft	Voice cloning or face-swap video is used to impersonate a wealthy individual and initiate fraudulent transactions. Alternatively, it is used to impersonate a corporate officer and gain access to databases of personal information, which can enable larger-scale identity theft.	
	2. Imposter scam	Voice cloning or face-swap video is used to impersonate a trusted government official or family member of the victim and coerce a fraudulent payment.	
	3. Cyber extortion	Synthetic pornography of the victim is used for blackmail.	
Companies	4. Payment fraud	Voice cloning or face-swap video is used to impersonate a corporate officer and initiate fraudulent transactions.	
	5. Stock manipulation via fabricated events	Voice cloning or face-swap video is used to defame a corporate leader or falsify a product endorsement, which can alter investor sentiment.	
	6. Stock manipulation via bots	Synthetic photos and text are used to construct human-like social media bots that attack or promote a brand, which can alter investor perception of consumer sentiment.	
	7. Malicious bank run	Synthetic photos and text are used to construct human-like social media bots that spread false rumors of bank weakness, which can fuel runs on cash.	
Markets	8. Malicious flash crash	Voice cloning or face-swap video is used to fabricate a market-moving event.	
Regulatory Structures	9. Fabricated government action	Voice cloning or face-swap video is used to fabricate an imminent interest rate change, policy shift, or enforcement action.	
	10. Regulatory astroturfing	Synthetic text is used to fabricate comments from the public on proposed financial regulations, which can manipulate the rulemaking process.	

Deepfake voice phishing | Fabricated private remarks | Synthetic social botnet | Narrowcast | Broadcast

Source: Bateman (2020)

3. Countering deepfakes

Albeit synthetic media are not illegal per se, their producers and distributors can be prosecuted for infringing copyright, breaching data protection law, and defamation, depending on the nature and purpose of the content. These tools, however, are often ineffective in countering the malicious use of deepfakes (Bateman et al., 2021). This is imputable to i) the limited portion of the spectrum of applications tackled by existing legislation, ii) the distance of regulators and governing bodies from the technological debate, iii) the two-fold role of tech platforms as content moderators and stakeholders, and, finally, iv) enforcement delays.

The general-purpose nature of deepfake technology prevents a comprehensive approach from being effective. A first step in unravelling the inherent complexity of the task requires establishing an essential value chain of fake and real media content (Pavis, 2021).[19] Stakeholders are, in this account:

[19] A remarkable initiative in this direction is the 2021 study by the European Parliament on "Tackling Deepfakes in European Policy".

- the *deepfaked* persons whose biometric data are used as input to generate deepfakes,
- deepfakes producers, generating synthetic contents,
- deepfakes amplifiers, that actively contribute to validate, distribute and disseminate deceptive content,
- deepfakes receivers, i.e. the audience exposed to deepfakes.

Furthermore, the nature and target of measures deployed require separately envisioning of the strands of interventions. These entail the *governance of truth*, including content moderation, *national security, transparency* and *privacy against identity theft and fraudulent impersonation.*

Governance of truth engages policy- and lawmakers on technological and ethical issues. Overall, a coordinated approach to effective enforcement actions, either incentivising good actors or sanctioning bad ones, has been lacking. In Europe, policy makers have undertaken light touch regulatory initiatives to nudge social media platforms self-policing.[20] While debunking and fact-checking practices are acknowledged as effective practices against disinformation, the main responses currently address continuous adjustment of the Terms of Service agreements for content moderation, with the consequential adoption of countermeasures against fake or seemingly deceiving content.[21] Such content moderation practices, however, present several limitations. The first is the substantial one of *delayed action of content detection*: any measure addressing the generation and distribution phase fails in protecting the portion of people eventually exposed to deepfakes between release and intervention (Davis et al., 2020). Moreover, accurate detection of synthetic media is a challenging task both in technological and ethical terms. As of today, tech platforms mostly rely on dedicated teams for manual detection of misleading information. Beside the stakeholders' interest in contrasting deceptive practices beyond deepfakes, the unreliable performance of technological detection tools withholds adoption of automated tools.[22] On ethics, inaccurate detection of deepfakes – and privately enforced content moderation in general – might raise concerns on freedom of expression whenever disinformation is not deliberate nor framed within a malicious initiative. When regulating the financial systems, we reckon most ethical issues would likely not apply, and stakeholders could effectively enforce provisions.

Continuous exposure to misleading information enhanced by social media platforms is increasingly perceived as a matter of concern for democracy and national security by most countries.[23] Existing countermeasures against threats to national security mainly tackle fake content producers (Chesney and Citron, 2019b). Single jurisdictions already deployed coercive responses against ascertained use of AI to run disinformation campaigns, ranging from sanctions to covert action.[24] However, a first critical point in addressing national security lies in the stakeholders operating at an international scale. This makes it hard to legislate on a unilateral basis, since perpetrators residing under a different jurisdiction may not be persecuted in most cases.

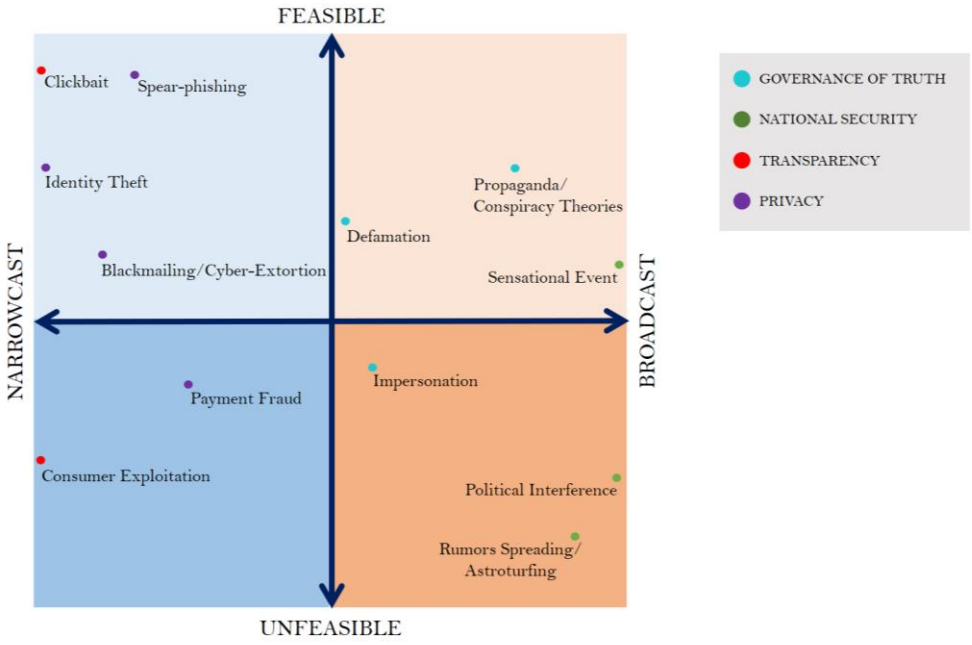
The difficulty in distinguishing deepfakes from real media, poses several challenges to obviate or mitigate risks of exploitation. On this basis, a second strand of interventions entails transparency of AI-enhanced content. Law enforcement of transparency obligations engages producers and amplifiers, depending on the context. Since 2019, the Cyberspace Administration of China has announced and issued regulation requirements on AI-enhanced synthetic content generation and use.[25] In a similar direction, the recent European proposal on AI Governance envisions disclosure of "image, audio or video content that appreciably resembles

[20] In 2018, the European Commission released the Code of Practice on Disinformation. While the code was signed by stakeholders like Twitter, Facebook and Google, its contribution was evaluated far from being effective (Plasilova et al., 2020).
[21] E.g. In 2019 Twitter introduced tagging of fake content.
[22] In 2020, Facebook has launched the DeepFake Detection Challenge to develop AI-based tools for Deepfakes detection (Dolhansky et al., 2020). See the Appendix on Deepfake detection tasks and limitations.
[23] See: https://www.allianceofdemocracies.org/initiatives/the-copenhagen-democracy-summit/dpi-2021/ .
[24] See: https://www.justice.gov/usao-sdny/pr/russian-hacker-sentenced-12-years-prison-involvement-massive-network-intrusions-us.
[25] E.g. https://www.reuters.com/article/us-china-technology-idUSKBN1Y30VU .

existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful" to consumers.[26]

Deepfakes threaten privacy of individuals and companies, whose personal and biometric data and identifiers could be employed for impersonation and reputational damage.[27] Legal measures are already enforced on data collectors – including tech platforms -, when it comes to privacy, acting on data transfers and purposes of usage, leveraging on existing regulations. In Europe, we acknowledge that the General Data Protection Regulation already enables prosecution of illicit deepfake content generation and distributions, while guaranteeing victims with the right to intervene on their personal data.

Figure 1



Source: Our elaboration

According to a recent survey by CEIP, since 2018 over one out of 5 research studies on countering broadcast influence operations explicitly referred to AI or deepfakes in their abstract.[28] Half of those entailed governance of truth via content moderation (among those, three out of four as primary intervention), whereas 37.5 percent of interventions addressed disclosure and transparency (over 58 percent of those as primary intervention).

Overall, no single strand of intervention is expected to provide a comprehensive solution, and strands of interventions ought to be coordinated into an effective action. In Figure 1 we display the likely most relevant strand of intervention pertaining to each disinformation initiative, ranked in terms of feasibility (mainly technological effort) and breadth of harnessed audience.

[26] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206, Art. 52(3). Previous interventions by the European Parliament involve the Resolution on Online platforms and the Digital Single Market, invoking transparency requirements against fake content, not necessarily AI-enabled, within the framework of online commerce. For the US, see https://www.congress.gov/bill/116th-congress/house-bill/3230 .

[27] Criminal offences targeting more tightly deepfakes, as it is the case with revenge porn, are already being defined by legislation.

[28] Carnegie's Partnership for Countering Influence Operations Baseline Datasets. Available at: https://ceip.knack.com/pcio-baseline-datasets .

4. Recommendations

We envision a case-by-case tailored approach to tackle AI-enabled disinformation initiatives, as opposed to a comprehensive approach. Moreover, we deem the contribution of technological measures critical in complementing and supporting effective countering of deepfakes.

Within the governance of truth, we envision the outsourcing of content curation services from dominant tech platforms to an intermediate layer of so-called *middleware* companies (Fukuyama et al., 2021). Following Fukuyama and co-authors, this would require enforcing the adoption of externally provided recommendation systems and content moderation algorithms by digital platforms. *Ad hoc* creation of a layer of companies, operating between content creators and amplifiers to offer middleware products (both software and services), would foster innovation and, as a by-product, introduce competition into a market that is currently dominated by few actors, namely the tech platforms. Concurrently, appointing such middleware companies responsible for content moderation would enable disentangling of the ambiguity characterising the initiatives of stakeholders in this account.

Technological measures stand to play a critical role also in the validation of truthfulness of information, before this spreads onto platforms. Recent proposals, originating within the framework of data sharing and reconciliation among parties, highlight the relevance of blockchain technology in this regard (Welfare, 2019, Fraga-Lamas and Fernàndez-Caramés, 2020). Enforcement of approval protocols, to validate trustworthiness of content or sources, would enable automatized suppression of fabricated sensationalised contents, and curb negative externalities caused by delayed action. In the aftermath of blockchain technology adoption, the same tools might also support backtracking deepfakes production along the transmission chain.[29] On deployment, we argue single jurisdictions ought to individually contribute to design and enforcement of requirements, to meet their internal demand. A coordinated effort might prove more effective in guaranteeing the technological supervision and monitoring of the validation infrastructure.[30] Recommendations on the technical aspects of deployment, however, are beyond the scope of this paper.

A global approach pursuing achievement of a code of conduct is unlikely to succeed in countering deepfakes, when it comes to national security. Not only we urge the international community to acknowledge the issue of deepfakes weaponisation: we once again recommend they recognize it as a mainly technological one. While moral suasion could be pursued, leveraging on middleware parties tackling the governance of truth, regulators ought to be prepared, to lead the technological debate upfront. Indeed, not all actors are expected to prove equally compliant, due the geopolitical implications in the generation and release of deepfakes. Operationally, existing measures could already benefit from technological measures analogous to those outlined for content moderation and truthfulness validation. Within the financial sector, we envision appointing international regulating authorities responsible for surveillance and, were applicable, deployment of middleware solutions.

On transparency obligations, education and training for content consumers could complement countering of deceptive deepfakes. While companies have been increasingly investing in cybersecurity awareness training for their employees, widespread educational campaigns for informed citizens and consumers might take a leap into dissemination initiatives. This would enhance reaching of consumers against exploitative online practices, particularly within advertisement. In respect of the latter, policymakers could also leverage on the existing privacy legislation, to limit microtargeting.

On privacy, regulators ought to improve the effectiveness of interventions countering identity theft for impersonation by complementing legal measures with technological instruments, to prevent exposure of

[29] See: https://www.ibm.com/blogs/industries/blockchain-protection-fake-news-deep-fakes-safe-press/ .
[30] Existing blockchain-based certification protocols could also contribute to re-define the role of truthful content within formal justice and the legal system (Donald and Hedges, 2020).

personal data. *Privacy enhancing tools* like encryption of sensitive data (including biometrics), non-fungible tokens or zero-knowledge proof technology could support replacement of current identification processes.[31]

5. Conclusions

General-purpose deepfake technology has the potential to magnify the disruptiveness of disinformation initiatives. More relevantly, it pioneers novel types of frauds and offences, thanks to the opportunity it provides to create new identities and communities from scratch on digital platforms. Beyond societal and individual damage, the impact of narrowcast and broadcast operations can also profoundly affect the economy. For these reasons, we argue regulators and governing bodies ought to step into and lead what has so far been chiefly a debate among technologists.

As of today, AI-enabled applications are characterised by several degrees of complexity, ranging from the contextual underlying implications to the technological ones. On the one hand, this prevents a comprehensive approach from being effective. On the other, acknowledgment of deepfake technology as a sophisticated instrument envisions a case-by-case approach, leveraging on technological measures to tackle the specificities of each application, alongside education initiatives, to improve responsiveness against disinformation.

Appendix

One of the first appearances of the term "deepfake" dates back to November 2017, when a discussion forum was started on online platform Reddit under the name "r/deepfakes". In full accordance with Rule #34 of the Internet[32], synthetic content generation initially addressed pornographic material.[33] Nonetheless, it took a short time to understand the great potential underlying their use for broader purposes.

On April 2018 the Buzzfeed website already called out for vigilance when it released the aforementioned video hoax displaying US President Barack Obama dropping several out-of-character statements while making a public announcement from the Oval Office.[34] While the video hoax merely aimed at underscoring the potential disruptive influence of synthetic media, the first use for fraudulent purposes was eventually documented in 2019. At the end of August, the Wall Street Journal reported of a fraudulent phone call to a British energy company. Fraudsters had relied on an AI-based voice cloning tool to impersonate the chief executive officer of a parent company, and obtain "urgent" wiring of over £200,000 to a supplier.[35]

A.1 Generation Tasks

Due to the effectiveness of deepfakes in challenging human perception, most generation tasks address audio-visual contents. Those discussed by the Reddit community, for instance, addressed the task of *identity swap*, which consists in the superimposition of a target face onto a source image or video content. Further applications include *face synthesis, attribute manipulation and re-enactment*. See Table A1.

[31] Blockchain technology could also effectively used to mitigate the risk of copyright violation – e.g., use of audio-visual media for deepfakes generation - via Non-Fungible Tokens.
[32] See: https://rulesoftheinternet.com/ .
[33] According to a report from Deeptrace, in 2019 deepfakes addressing pornographic materials amounted to the 96 percent of the circulating content. See https://www.wired.com/story/most-deepfakes-porn-multiplying-fast/ .
[34] Cf 2.
[35] See: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402. See also https://www.clearskysec.com/operation-dream-job/ .

Table A1

Task	***Identity Swap***	***Face Synthesis***	***Attribute Manipulation***	***Re-Enactment***
Description	Replaces the target face in the source image	Generates non-existent biometrics	Edits existing characteristics, E.g. gender, age, hair color	Replaces the target facial expression in the source image
Alters Biometrics	Hard ones (Highest risk)	No (Lowest risk)	Yes, Soft	Yes, Soft
Generation Process Requires Target Face	Yes (Single face image may be used)	No	No	Yes (Videos and/or several images required)

Face synthesis consists in generating new features from scratch[36], whereas attribute manipulation and re-enactment are used, respectively, to alter existing biometrics or facial expressions. This characterises them as lower risks practices, compared to identity swap, that could be used to hijack a target face into an existing image or video for malicious purposes, such as blackmailing or extortion.[37] All deepfake generation tasks rely on artificial neural network models[38], whose training process might be more or less demanding, depending on the task. In this account, generation of re-enacted media requires several observations of the target face as reference, to obtain a realistic outcome, whereas face synthesis and attribute manipulation can make up features to be added by themselves.

Albeit less commonly recognized, further remarkable strands of applications tackle audio and textual content generation for disinformation. The former is usually combined with synthetic video footage. For malicious purposes, deepfakes typically result from the audio counterpart of re-enactment, dubbed *voice cloning*. Such generation task greatly relies on input data to enhance an accurate performance. As for textual content, deepfake generation typically supports broader disinformation campaigns run on social media platforms, to reinforce and subvert narratives. In combination with bots, they may pursue generation short textual variations (e.g., tweets) of a sourced content to hijack a viral topic or hashtag, boost its spread across social media platforms, or expose as many users as possible to a narrative (*reiteration*). Moreover, they can craft fake news from scratch on a given topic (*elaboration*)[39], spin a content to undercut a perspective (*manipulation*), re-elaborate it to devise new narratives or cast doubt (*seeding*) or to match ideology of targeted readers (*persuasion*), and create divisive content to wedge divides open and polarize opinions among vulnerable communities (*wedging*).

[36] See https://thispersondoesnotexist.com/ for some examples.
[37] Generation of damaging content to blackmail an individual person and/or launch a smear campaing against public figures is also known as *kompromat* (Choy, 2020).
[38] See the Appendix for details.
[39] Based on sourced tones, narrative elaboration heavily depends on the relevance of the data used to train the model, and requires human operators to provide adequate background so as to make the generated content realistic.

A.2 Modelling overview

The term deepfakes combines the notions of "deep learning" and "fake". The former term refers to a branch of AI, aimed to learn sequences of complex representation rules from *training* data. It is mainly based on Neural Network models, that may come with different shapes and characteristics, called *architectures*. Thanks to the open-source collaborative nature of research in the AI field, Neural Networks achieved great accuracy in performing increasingly complex tasks on unstructured data - like images, audios, videos and texts - in a relatively short period of time.

As of today, the deepfakes generation constitutes an actively growing area of AI. A recent survey higlights how, up to the end of 2020, all papers published on top conferences and journals on the subject only accounted for a third of the total, witnessing the open-source nature of the topic (Juefei-Xu et al., 2021).

From a technical perspective, the first deepfake media were images, obtained using a class of Neural Network architectures called *autoencoders*. Broadly, these models comprise two stacked symmetric components: an *encoder*, to project a multi-dimensional input toward a smaller latent dimension, and a *decoder*, mapping the compressed output of the former toward original size. By the time deepfakes gained popularity, however, a further class of architectures took over their generation: Generative Adversarial Neural Networks (GAN).
These models were introduced in 2014 by seminal paper of Goodfellow et al. (Goodfellow et al., 2014) and brought a key contribution to the research in the whole deep learning area. In a nutshell, GAN comprise two components, dynamically interacting with each other during the training phase: a *generator* (G) and its *adversarial* (A, also referred to as *discriminator*). Both components pursue optimization of an objective function *f*, measuring the amount of correctly classified outcomes from a generation process: G's goal is to learn how to deliver realistic outcomes so as to minimize the number of those correctly spotted as fake (minimize *f*). Conversely, A must learn to correctly detect G's fake outcomes (maximize *f*). The adversarial nature of learning enhances faster convergence and more accurate performance of the generating process.[40]
Features of the GAN for deepfakes generation differ from task to task. For instance, image generation usually relies on a G *autoencoder architecture*, with convolutional layers. Processing of the image requires sliding along its dimensions, to enhance local elaboration of groups of pixels. Quality of the outcome has increased over time, enhancing generation of highly realistic images. With textual content, both G and A are usually language models, with recurrent layers to cope with sequences of elements. Language models represent textual content in form of statistical probabilistic distributions over a multi-dimensional space. They can be used to assess similarity among words or phrases, and to generate text sequences.

A.3 Deepfake detection

Deepfake detection methods constitute an evolving research area themselves and vary depending on the content they target and the signal they address to evaluate it. Focus may either be on extracted features of an image or audio content - E.g. robustness to perturbation attacks, investigation of local patches, frequency domain analysis - or biological signals - Visual/audio inconsistencies, e.g. lip-sync violations in videos, unnatural facial features in images - or lexical consistency (textual data). As for textual data, their quality massively relies on training data. If used to generate content on new topics, they usually manage to deliver compelling narratives in a fictional manner rather than repeating existing facts. As a consequence, they make up elements that require careful human monitoring and intervention to avoid detection. Furthermore, textual deepfakes often deliver quirk content, enhancing their detection. Finally, a further sub-topic entails deepfakes detection-evasion methods. These tackle specific features addressed by detection methods, to deceive them.

[40] Objective function *f* settles on its global maximum when G can no longer fool A, which in turn has reached maximum accuracy.

References

Aftergood, S. (2017). *Cybersecurity: The cold war online.* Nature, 547(7661), 30-31.

Allen, G., & Chan, T. (2017). *Artificial intelligence and national security*. Cambridge, MA: Belfer Center for Science and International Affairs.

Bateman, J. (2020). *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace.

Bateman, J., Hickok, E., Shapiro, J.N., Courchesne, L., & Ilhardt, J. (2021). *Efficacy of Influence Operations Countermeasures: Key Findings and Gaps From Empirical Research*. Carnegie Endowment for International Peace.

Biancotti, C., & Ciocca, P. (2021), *Financial Markets and social media: lessons from information security.* Carnegie Endowment for International Peace.

Caldwell, M., Andrews, J. T. A., Tanay, T., & Griffin, L. D. (2020). *AI-enabled future crime.* Crime Science, 9(1), 1-13.

Chesney, R., & Citron, D. (2019a). *Deepfakes and the new disinformation war: The coming age of post-truth geopolitics.* Foreign Aff., 98, 147.

Chesney, B., & Citron, D. (2019b). *Deep fakes: A looming challenge for privacy, democracy, and national security.* Calif. L. Rev., 107, 1753.

Choy, J. P. (2020). *Kompromat: A theory of blackmail as a system of governance*. Journal of Development Economics, 147, 102535.

Davis, R., Wiggins, C., & Donovan, J. (2020). *Tech Policy Factsheet: Deepfakes*. Belfer Center for Science and International Affairs.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). *The deepfake detection challenge (dfdc) dataset.* arXiv preprint arXiv:2006.07397.

Donald, B., & Hedges, R.J. (2020). *Deepfakes Bring New Privacy and Cybersecurity Concerns*. Corporate Counsel Business Journal.

Forest, J. J. (2021). *Digital Influence Warfare in the Age of Social Media*. Praeger.

Fraga-Lamas, P., & Fernández-Caramés, T. M. (2020). *Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality.* IT Professional, 22(2), 53-59.

Fukuyama, F., Richman, B., & Goel, A. (2021). *How to Save Democracy from Technology: Ending Big Tech's Information Monopoly.* Foreign Aff., 100, 98.

Galeotti, M. (2019). *The mythical 'Gerasimov Doctrine'and the language of threat*. Critical Studies on Security, 7(2), 157-161.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). *Generative adversarial nets.* Advances in neural information processing systems, 27.

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2021). *Countering malicious deepfakes: Survey, battleground, and horizon*. arXiv preprint arXiv:2103.00218.

Kastner, J., & Wohlforth, W. C. (2021). *A Measure Short of War: The Return of Great-Power Subversion*. Foreign Aff., 100, 118.

Kietzmann, J., Mills, A. J., & Plangger, K. (2020). *Deepfakes: perspectives on the future "reality" of advertising and branding*. International Journal of Advertising, 1-13.

Pavis, M. (2021). *Rebalancing our regulatory response to Deepfakes with performers' rights*. Convergence, 27(4), 974-998.

Plasilova, I. (2020). *Study for the Assessment of the Implementation of the Code of Practice on Disinformation.* European Commission.

Schetinger, V., Oliveira, M. M., da Silva, R., & Carvalho, T. J. (2017). *Humans are easily fooled by digital images*. Computers & Graphics, 68, 142-151.

Schick, N. (2020). *Deep Fakes and the Infocalypse: What You Urgently Need To Know*. Hachette UK.

Vaccari, C., & Chadwick, A. (2020). *Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news*. Social Media+ Society, 6(1), 2056305120903408.

Welfare, A. (2019). *Commercializing Blockchain: Strategic Applications in the Real World*. John Wiley & Sons.

White, J. (2016). *Dismiss, distort, distract, and dismay: Continuity and change in Russian disinformation.* Institute for European Studies: Vrije Universiteit Brussel, 13.