weizenbaum journal OF THE DIGITAL SOCIETY

J\03 ISSUE 02

WEIZENBAUM JOURNAL OF THE DIGITAL SOCIETY
Volume 3 \ Issue 2 \ w3.2.3 \ 09-25-2023
ISSN 2748-5625 \ DOI 10.34669/WI.WJDS/3.2.3

Information on this journal and its funding can be found on its website:
https://wjds.weizenbaum-institut.de

KEYWORDS

deepfake detection
artificial intelligence
digital sovereignty
remote ID proofing

RESEARCH PAPER

Defending Informational Sovereignty by Detecting Deepfakes

Opportunities and Risks of AI-Based Deepfake Detection of Driven Disinformation and Illegal Activities

Milan Tahraoui [1*] \ Christian Krätzer [2] \ Jana Dittmann [2] \ Hartmut Aden [1]

[1] Berlin School of Economics and Law, Berlin Institute for Safety and Security Research (HWR/FÖPS Berlin)
[2] Magdeburg University
[*] Corresponding author, milan.tahraoui@hwr-berlin.de

ABSTRACT

This paper investigates possible contributions that AI-based detection mechanisms for deepfakes could make to the challenge of responding to novel cyber threats, including fraud and disinformation as anti-democracy tools. The paper investigates the implications of such a tool for the emerging European discourse on digital sovereignty in a global environment. While cybersecurity and disinformation are not new topics, recent technological developments around AI-generated deepfakes have increased the manipulative potential of online audio-visual content, making this a specific but important challenge in the global and interconnected information context.

1 Introduction

The term *deepfake* commonly refers to visual content that is artificially generated, manipulated, or distorted by using artificial intelligence tools to alter or replace a person or selected attributes of that person in the content. That content can be not only visual (i.e., pictures and videos) but also aural (i.e., sounds and noises). However, this paper's scope is limited to visual content, especially videos. Deepfakes can be used for numerous purposes, legitimate and illegitimate, including intentional manipulation of political decision-making processes (e.g., during electoral campaigns). In a report entitled "Tackling Deepfakes in European Policy," the European Parliamentary Research Service (2021) defines deepfakes as "manipulated or synthetic audio or visual media that seem authentic, and which feature people that appear to say or do something they have never said or done, produced by using artificial intelligence techniques, including machine learning and deep learning" (p. i; ***a***European Parliament, 2023, p. 141, (44d)).

This paper investigates the interplay between the generation of AI-based deepfakes and deepfake-detectors that respond to disinformation, fraud, and other potential threats that deepfakes may increasingly facilitate. This might include, for example, "fake" media content and the circumvention of identity authentication and verification systems. Threats associated with deepfakes (Hsu & Lee Mayers, 2023; Metz, 2023; Metz & Blumenthal, 2019) engage with pre-existing challenges that have the potential to impact issues around the emerging European conception of *informational digital sovereignty*. That concept describes the capacity to decide and act autonomously in the face of digital informational phenomena against the backdrop of a globally interconnected environment. In that context, the destabilizing potential of information-based threats (e.g., deepfakes) is becoming a growing source of concerns. The latest developments in digital information and communication technologies and the increased sensitivity to the destabilizing potential of disinformation campaigns have pushed various countries (including European nations) to follow this path by embracing their own version of informational digital sovereignty. Although disinformation is not a new topic, recent technological developments have increased the manipulative potential of video and audio-based content spreading online (***a***European Parliament, 2023, p 116, (40a) and p. 125, Annex III(8)(aa)).

Against this backdrop, this paper focuses on one specific dimension of the overall global context of digital transformation as accelerated by the ongoing artificial intelligence "revolution": the phenomenon of deepfakes, especially their potential to exert manipulative influence in a way that affects democratic and security issues. This paper relates to the research project FAKE-ID, an initiative driven by an interdisciplinary research team of IT, law, social, and cultural anthropology scholars working together in a consortium funded by the German Federal Ministry of Education and Research.[1] The project – and this paper specifically – consider applications that use AI-based tools to detect deepfakes while conducting remote identity controls or proofing methods (hereafter, "remote ID proofings"). Remote identity proofing methods "are a way to identify individuals without relying on physical presence" (***b***ENISA, 2021), capturing a diverse set of techniques and processes that "can be used in a variety of contexts where trust in the identity of a natural or legal person is essential – such as financial services, e-commerce, travel industry, human resources [and] public administrations" (***b***ENISA, 2021). Despite all the advantages of remote ID proofing solutions, which have experienced a boost in usage and user acceptance during the COVID-19 pandemic, such solutions are threatened by deepfakes and their ability to take over identities in live settings.

This paper first investigates possible contributions that AI-based detection of deepfakes can make to the challenges and threats associated with deepfakes. However, even if consensus is mostly lacking around fundamental questions relating to the concrete applicability of international law to cyber and digital phenomena, an undeniable trend among both states and major private corporations sees digital informational phenomena (such as emerging deepfakes) framed as potentially violating or interfering with digital sovereignty. Building on this background, this paper seeks to understand the implications of such a tool within the emerging European discourse on digital sovereignty in a global environment.

---

[1] The research project, FAKE-ID: Videoanalyse mit Hilfe künstlicher Intelligenz zur Detektion von falschen und manipulierten Identitäten [AI-Based Video Analysis to Detect False and Manipulated Identities] has been financed by the German Federal Ministry of Education and Research (BMBF) within the framework of the research programme Künstliche Intelligenz in der zivilen Sicherheitsforschung [AI in Civil Security Research] (FKZ: HWR/FÖPS 13N15737, OVGU 13N15736).

1.1 Approaches to the Regulation of Deepfakes

In April 2021, the European Commission published a proposal for a regulation meant to establish harmonized rules for the use of AI, including the generation and detection of deepfakes (***b***European Commission, 2021). Although this draft regulation remains in the law-making process, global technology companies have already started to establish their own guidelines and self-regulatory frameworks for deepfakes (***a***Twitter Help Center, 2023; ***b***Twitter Help Center, 2023; Vincent, 2023; LastBluejay, 2020; Bickert, 2020). Google has recently forbidden the use of its Colab service, one of the most popular online platforms for training machine-learning and AI systems with free computational resources to generate deepfakes (Fadilpašić, 2022; Google Research, 2022).[2] This exemplifies the risks increasingly perceived in association with deepfakes (Europol Innovation Lab, 2022; Kropotov et al., 2022), risks that have motivated public authorities, such as the European Commission (Chee, 2022) and the Cyber Administration of China (Baptista, 2022), and global leading private firms, such as Google and Meta (Bickert, 2020), to regulate deepfake generation and circulation online. Among the most common perceived risks with deepfakes – beyond so-called "revenge porn" (Delfino, 2019, pp. 895 – 898), fraud, and harmful application cases (Perset et al., 2023; Europol Innovation Lab, 2022) – is the anticipated facilitation and intensification of the dissemination of disinformation, among various forms of online manipulation (Ruth, 2023; Brooks et al., 2022, p. 23; US Cybersecurity and Infrastructure Security Agency, 2020, p. 4).

Despite the growing importance of deepfakes in the global cyber context (Hsu, 2023), there remains no prospect of a proper international framework regulating their generation, circulation, and detection. The difficulties in achieving a coherent regulatory framework at the international level can be explained by divergent political and economic interests. These interests range from defending liberal democracy against threats caused by disinformation to illiberal states attributing to intelligence services the mission of manipulating democratic decision-making in other countries. Although economic interests mostly speak against a restrictive regulatory framework to avoid limiting the economic opportunities that the use of AI presents, civil society groups tend to advocate for more restrictive rules to protect not only democracy but also the fundamental right of citizens to not be manipulated. The current global competition for leadership taking place in the field of artificial intelligence and machine-learning technologies, primarily involving the United States, China, the European Union (EU), and Russia, can be explained against the backdrop of these conflicting interests.[3] This competition for leadership spills over

---

2 "We prohibit actions associated with bulk compute, actions that negatively impact others, as well as actions bypassing our policies. The following are disallowed from Colab runtimes: […] creating deepfakes" (Google Research, 2022).

3 The international legal scholarship provides examples of how approaches between different regions of the world diverge regarding legal regulation of fundamental cyber phenomena in international law. For discussions, see Chander & Sun (2022) and Arner et al. (2021).

into the AI regulatory field, where the EU and China are at the forefront with non-sectorial AI regulatory frameworks (Heath, 2023; Keane, 2022). Due to divergent and conflicting interests that democratic and non-democratic political powers have with respect to the use of AI – for example, regarding AI in the context of remote biometric identification or freedom of speech – a global consensus on the purposes and aims of AI regulation is unlikely to occur in the near future.[4] Even if there is no global consensus on AI regulation, there have been some important international developments. One example is the adoption of the 2021 UNESCO recommendation on the ethics of artificial intelligence (UNESCO, 2021). Still, those UNESCO standards do not specifically address regulatory challenges associated with deepfakes.

The European Commission is currently trying to occupy this space – via its proposal for AI regulation – so that it can establish itself as a global standard-setter with a particular emphasis on a human-centered, ethical, and trustworthy model for regulating AI (European Commission, 2020; ***b***European Parliament, 2023). Deepfakes are among many phenomena covered by this future AI regulation. Notably, China has taken the lead at the global level in regulating AI-related fields. In September 2021, China's National New Generation Artificial Intelligence Governance Specialist Committee adopted a set of non-binding guidelines entitled "Ethical Norms for New Generation Artificial Intelligence." These guidelines offer general standards, but their enforcement enables some flexibility. Remarkably, these guidelines do not incorporate sanctions in the case of violations (Georgetown Center for Security and Emerging Technology, 2021). On January 10, 2023, China's special law on deepfakes, "Provisions on the Administration of Deep Synthesis Internet Information Services," entered into force. This Chinese law constitutes the first legislation worldwide to specifically focus on the regulation of AI-generated deepfakes and other AI-manipulated online content.[5] Most recently, China has circulated for comments the first draft of legislation that will apply to generative AI technologies, legislation including reference (in Art. 16) to Chinese legislation on deep synthesis media (DigiChina, 2023)[6]. Still, unsurprisingly (from a global perspective) challenges remain for deepfake regulation due to the differences between states and public authorities worldwide in terms of sensitivities and sensibilities around online content regulation. In contrast to most legal approaches applicable to online content regulation in the US and Europe, Art. 4(1) of that Chinese draft legislation requires

---

4 Regarding the diverging regulatory approaches of the US, China, and the EU, consider, for instance, Fung and Etienne (2022), Chan Chin et al. (2022), and De Gregorio (2022).

5 Deepfakes are addressed in this legislation as deep synthesis media or services, according to an unofficial translation of this law. See chinalawtranslate.com, 2022, Art. 2. and Hine and Floridi (2022).

6 This is the name that this Chinese legislation used for what is elsewhere called deepfakes: "Deep synthesis technology refers to the use of technologies such as deep learning and virtual reality, that use generative sequencing algorithms to create text, images, audio, video, virtual scenes, or other information; including but not limited to: […] (4) Technologies for generating or editing biometric features in images and video content, such as face generation, face swapping, personal attribute editing, face manipulation, or gesture manipulation;" (China Law Translate, 2022).

> that content generated through the use of generative AI shall reflect the Socialist Core Values, and may not contain: subversion of state power; overturning of the socialist system; incitement of separatism; harm to national unity; […] content that may upset economic or social order. (DigiChina, 2023, Art. 4(1))

2 Deepfake Detection in the Context of Digital Sovereignty

The concept of digital sovereignty remains controversial both within and outside the EU, especially regarding what it concretely entails.[7] If the applicability of the fundamental principle of sovereignty in cyberspace originally attracted many controversies on the international legal plane (Chapdelaine, 2021, p. 74; Lambach, 2020, pp. 496–498; Mueller, 2020, pp. 786–788; Goldsmith, 2019, p. 822), current international law heatedly debates its concrete scope of application – that is, how it applies, for which activities and with what legal consequences (Monyihan, 2019, p. 9, para. 22). A broadly accepted Western perspective once criticized the concept of cyber or digital sovereignty for not only undermining the protection of fundamental rights and freedoms but also for impeding the so-called international free flow of data. The fact that China was originally its main proponent in international law likely plays a role in some of the criticisms (Goldsmith, 2019, pp. 820–821; Chander & Sun, 2022, pp. 296–298). However, so too does the use of this principle of international law by Russia, India, and Brazil to impose data localization obligations on public and private actors deploying data activities within their territorial jurisdiction (Mainwaring, 2020). The concept of *informational* digital sovereignty that seeks to apply the fundamental international legal concept of state sovereignty to the governance and regulation of informational phenomena has experienced a similar fate: originally championed by countries such as China, Russia, and Iran and much criticized by Western countries, especially the United States, it now sees increasing political and legal popularity in various countries. One of the many drivers behind the growing adoption of the concept of digital sovereignty is its potential to be used "defensively" – from both an economic and security perspective – by asserting sovereignty in the face of foreign and external public and private phenomena that are deemed dangerous. Another factor that has arguably driven this global evolution is the growing awareness that the globally interconnected digital environment has strong potential to destabilize not only the internal affairs of states but also international relations. Although those concerns were at first more prominent among states willing to control information flowing through their countries in the name of the principle of

[7] To understand how conceptions of sovereignty diverge internationally: Chander and Sun (2022); Akande et al. (2022) and Jon Heller (2021).

sovereignty, certain Western countries have more recently also embraced that principle, especially because electoral processes have arguably been targeted by orchestrated disinformation campaigns.

Despite the lack of a unified European perspective on digital sovereignty, an emerging consensus within the EU about digital sovereignty concerns the notion of *strategic autonomy*,[8] which aligns with the EU's agenda of establishing itself as a global leader on the basis of its regulatory powers on digital matters and its worldwide influence via the appeal of its standards in related matters. In this context, the EU has increasingly established access rules for its market to influence the regulatory strategies of third countries. One of the most prominent examples is the General Data Protection Regulation (GDPR) that "applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union" (Art. 3(1) GDPR). The GDPR has motivated several other countries to pass legal rules for data processing that comply with the GDPR; this has been called the Brussels Effect (Bradford, 2020; Bradford, 2012; Savin, 2022, pp. 4–6; Bendiek & Stürzer, 2022, pp. 5–6; Pohle & Voelsen, 2022, pp. 20–21). The EU Commission's proposal for an AI regulation mentions similar ambitions (recitals (10)–(11)) (***b***European Commission, 2021, pp. 19–20, (10)–(11)). Furthermore, there is a trend in the EU towards ensuring informational privacy and self-determination for people and individuals, especially given the increased technology-driven possibilities of exerting manipulative influence over societies in the global digitalization process (Iliopoulou-Penot, 2022). This is among the core motivations behind the European Commission's proposal for an AI regulation as well as other recent EU regulations for the digital environment, such as the Digital Services Act (***b***European Commission, 2021, p. 21, (15); ***a***European Parliament & Council of the EU, 2022, Arts. 34(1)(c) and 35(1)(k); ***a***European Parliament, 2023, p. 116, (40a)–(40b) and p. 125, AnnexIII(8)(aa)).

Against this backdrop, the EU is increasingly moving away from its former reluctance to rely on a concept of cyber or digital sovereignty. This reluctance was partly due to its limited competencies – for instance, in the field of national security (European Union, 2012, Art. 4(2)(j)) – but also partly due to divergent views on digital sovereignty among member states. Still, there seems to be an evolution in which the EU is increasingly claiming (if merely implicitly for now) that some online informational phenomena threaten its strategic autonomy, and hence, on the legal plane, its sovereignty. AI-based informational phenomena boosted by the generative AI boom, especially deepfakes, promise to accentuate this trend for EU legal discourses and practices.

---

[8] This article uses the definition of the concept of digital sovereignty presented by Iakovelva (2022, p. 339): "the EU's *power to regulate* (ability 'to make its own choices, based on its values, respecting its own rules') is arguably the common denominator of the multiple definitions of a regulatory instrument advancing 'digital sovereignty' by EU political institutions in the area of data governance." For further discussion, see Chander and Sun (2022, pp. 298–299).

2.1 Challenges for Digital Sovereignty

Various scandals have contributed to making the concept of digital sovereignty more relevant within the EU, including the Snowden revelations on US global intelligence practices, the Cambridge Analytica scandal, the allegations that the 2016 US presidential elections took place under the influence of manipulative data-driven campaigns, and disinformation campaigns during the global COVID-19 pandemic. For instance, in its October 2020 Conclusions, the Presidency of the Council of the EU stated:

> The COVID-19 pandemic has shown more clearly than ever that Europe must achieve digital sovereignty in order to be able to act with self-determination in the digital sphere and to foster the resilience of the European Union. (Council of the EU, 2020, p. 3)[9]

This statement exemplifies the EU's growing openness to the necessity of either establishing, ensuring, or defending its digital sovereignty, including informational dimensions of control and power that the EU and its member states can exercise over digital forms of information (***a***Council of the EU, 2022, p. 34; Iakovelva, 2022). There are especially growing concerns about the need to safeguard the integrity of electoral processes against rising digital means of influence over political processes, as this EU strategic compass for security and defense also indicates (***a***Council of the EU, 2022, p. 13, para. 26), as does the EU's recently adopted *Digital Services Act* (***a***European Parliament & Council of the EU, 2022, (69), (84), (95), (104), Art. 35(1)(k)).

The growing importance and attention attributed to informational sovereignty have become even clearer since the Russian invasion of Ukraine in early 2022, as also evidenced by the EU strategic compass for a stronger security and defense in the next decade, formally approved by the Council of the EU in March 2022:

> As its institutions are subject to an increasing number of cyberattacks or attempts to intrude [on] their systems, the EU needs to enhance the protection of its most critical processes, assets, and information and ensure that it can rely on robust and trustworthy information and adequate European communication systems. To this end, we will streamline security rules and regulations as well as bolster the common approach by the Member States, EU Institutions, bodies, and agencies, as well as CSDP missions and operations, to the protection of information, infrastructure, and communication systems. Building on the EU Cybersecurity Strategy, we call upon the EU institutions, agencies, and bodies to adopt additional standards and rules on information and cyber security, as well as on the protection of EU classi-

[9] See also, ibid. (p. 5, p. 7).

> fied information and sensitive non-classified information, thus facilitating secured exchanges with Member States. (***a***Council of the EU, 2022, p. 21)

Remarkably, the regulation of deepfakes not only integrates the emerging EU perspective on digital sovereignty on the international law and geopolitical plane but also relates at a more concrete level to new regulatory frameworks, for which one of the main objectives is to promote the use of online identification means by ensuring their secure use (among other goals). These new developments also play a role in the further development of the EU's approach to digital sovereignty, which – beyond security issues – seeks to preserve the EU's political autonomy in the global interconnected digitalized environment.

2.2 Deepfake Detection and Remote ID Proofing as Part of the Emerging EU Approach to Digital Sovereignty

Remote ID proofing or verification is not only performed by public administrations. Private operators including banks, financial institutions, and digital service providers are using these services with increasing frequency (***a***ENISA, 2021, p. 21). Remote ID proofing procedures are based on several categories of data that are collected from various sources and third-party databases, privately or publicly owned, that serve as a template or reference for verifying a person's identity (***a***ENISA; 2021, p. 25). This has important implications for data protection and privacy (***a***ENISA, 2021, pp. 39–40; Pohle & Voelsen, 2022, p. 22) that can have consequences for the compliance of an AI-based deepfakes detection tool (Masood et al., 2023) for law enforcement purposes, which entail requirements relating to the protection of fundamental rights and the rule of law (***b***European Commission 2021, Art. 52(3), pp. 26–28, (32), (38), (40); ***f***Council of the EU, 2022, p. 136, Art. 52(3) and pp. 199–201, Annex III). Recent survey papers highlight the wide variance currently seen in deepfake detection technologies, identifying differences in the analyzed artefacts and the approaches used to implement detection and the performance of these extant mechanisms in terms of technically different deepfake categories (Masood et al., 2023). In summary, none of the methods currently used have already obtained 100% reliable detection performance and most of the methods discussed in the recent literature demonstrate a significant drop in performance if the material used for training and the material under evaluation differ in terms of either or both their content and (encoding) characteristics. These technical limitations must be considered when addressing how deepfakes detection is used in the context of remote ID proofing.

Several methods of remote ID proofing exist (***b***ENISA, 2021, pp. 25–26). A combination of various methods ("breed methods") (***b***ENISA, 2021, p. 25), including the use of AI and human intervention to operate verifications or final controls, is currently the most reliable existing approach, and it also complies with the ethical and emerging legal requirement to use only "weak" AI with a human as final de-

cision-maker. Advanced technology is usually far from enabling confirmation of identity with attribution of an absolute score (i.e., YES/NO). This means that the typical outcome of remote ID proofing is the issuance of a proof of authenticity for a person's identity (based on a confidence level as a percentage or a likelihood ratio) or the assignment of identification credentials (***b***ENISA, 2021, p. 15).

The trend to develop and implement remote ID proofing is rapidly taking off, accelerated by the COVID-19 crisis, which contributed to the spread of remote identity verification procedures worldwide (***d***European Commission, 2021, para. 1.53; ***b***ENISA, 2021, p. 4). However, even before the pandemic, the EU had already moved towards establishing a European digital identity model, with verification control mandated via several major legislative initiatives.

One such initiative was the 2014 adoption of Regulation (EU) 910/2014 on electronic identification and trust services for electronic transactions in the internal market, known as the 4th eIDAS Regulation. One of the overall goals of the 4th eIDAS Regulation was to create a "European internal market for electronic trust services – namely electronic signatures, electronic seals, time stamp, electronic delivery service and website authentication – by ensuring that they will work across borders and have the same legal status as traditional based processes" (European Commission / eIDAS Observatory, 2016). This regulation is of importance not only for the financial and banking sector but also for e-governance in the public sector (***a***European Commission, 2021, p. 16; Zetzsche et al., 2020, p. 350). The overall aim of the 4th eIDAS Regulation was to facilitate "mutually recognized digital identity for cross-border electronic interactions between European citizens, companies and government institutions" (Zetzsche et al., 2020, pp. 349 – 350). That said, the highly technical nature of this regulatory regime might explain why this ambition failed before the advent of the COVID pandemic: The use of digital identity documents under the framework of the eIDAS Regulations was not widespread amongst the public, with specialized public or private institutions more likely to adopt such procedures (***d***European Commission, 2021, Explanatory Memorandum, p. 1).

Despite failing to meet its original ambition, the overall objective of fostering the development and wide use of secure digital forms of identity remains unchanged (European Commission, 2023). This is manifest in the plan to establish a European digital identity "wallet" common to all EU citizens (***c***European Commission, 2021) via the ongoing negotiation of a 5th eIDAS Regulation based on the European Commission's Proposal from June 3, 2021 (***d***European Commission, 2021). That proposal aims to harmonize remote ID proofing of EU identities, both online and offline (***d***European Commission, 2021, p. 2). In this context, detecting deepfakes will become of increased importance for the EU's objective to provide for cross-border activities "access to highly secure and trustworthy electronic identity solutions […] that public and private services can rely on trusted and secure digital identity solutions" (***d***European Commission, 2021, p. 1). Another EU objective relevant in this context involves

empowering and facilitating the use of digital identity solutions by natural and legal persons (***d***European Commission, 2021, p. 1; Zetzsche, 2020, p. 351), and another core regulatory target involves facilitating secure online business transactions as well as secure access to public services (Council of the EU, 2021, pp. 19–20, Art. 6A; ***d***European Commission, 2021, pp. 19, (34) and pp. 23–25). In "2030 Digital Compass: the European Way for the Digital Decade" (2021), the European Commission declares that one objective of the European digital identity system is the need of the people "to have easy access to digital public services on [the] basis of a universal digital identity (***a***European Commission, 2021). Additionally, among the diverse purposes of a pan-European digital identity system in relation to public services (***d***European Commission, 2021, (34) and Art. 54b), one important consideration is ensuring the cybersecurity of election infrastructures when verifying the identity of people voting online.

These examples demonstrate the importance of ensuring secure digital means of remote (online) identification. In this context, the reliable and trustworthy detection of deepfakes within the overall EU regulatory framework for AI systems will constitute an important aspect of (for example) the European digital identity wallet and a European conception of digital sovereignty in general.

3 Deepfakes as a Potential Threat to Digital Sovereignty and the European Union's Regulatory Reactions

Having established the legal, political, and sociological foundations for this study's focus, it is possible to consider whether deepfakes constitute a cybersecurity threat for an emerging EU conception of digital sovereignty. That question requires a nuanced answer. First, various cases for misuse or abuse of deepfake technologies exist, but it is important to note that deepfakes can also be used for artistic (Snow, 2021), educational (Ciftci, 2023), entertainment (Bradshaw, 2019), commercial (Simonite, 2020), or medical purposes (European Parliamentary Research Service, 2021). Consequently, they cannot always be considered a security threat. Nonetheless, there are two constellations wherein deepfakes may contribute to deepening "cyber security threats," namely, disinformation and identity manipulation or thefts (***b***ENISA, 2023, p. 25). Both constellations can interrelate in the particular context of elections, where digital identity now plays a crucial role, either formally to verify the authenticity of national identity – and hence, whether an individual can vote in an election – or informally to prevent that content from being spread online via accounts using fake identities to manipulate electoral processes in the increasingly digitalized dimensions of public debates.[10] The second constella-

[10] See the eight scenarios developed to illustrate ethical harms that could be generated using deepfakes in the context of elections (Diakopoulos & Johnson, 2021; Dobber et al., 2021). Regarding allegations concerning the use of false identity online in the context of the 2016 US presidential elections, see Schmitt (2018, p. 36).

tion invokes delicate issues given the anonymity that users on online platforms often utilize due to privacy concerns. From a public or private law perspective, deepfakes can fall under the emerging EU cybersecurity regulatory framework to impose disclosure and notification obligations to both public and private actors in the face of cyber threats. In any case, in a report about cybersecurity threats and challenges for 2030, the EU Agency for Cyber Security has recently identified deepfakes among the top-priority cybersecurity threats for their potential to tamper with verification software supply chains:

> By 2030, deepfake technology will be widely used. It may be used as a form of harassment, evidence tampering, and provoking social unrest. Although there will likely be a rapid influx of verification software that analyses videos and voice to verify the identity of individuals, the urgent market demand leads to programmers cutting corners. This software will be highly targeted by anyone wishing to use deepfakes for illegal or unethical purposes. (***a***ENISA, 2023, p. 22)

3.1 Deepfakes as Cybersecurity Threat to EU Digital Sovereignty

Identity manipulation is increasingly perceived as a potential threat to digital sovereignty due to the possibility that it could materialize at a general level that could endanger EU laws, interests, and values (ENISA, 2022). Although such a claim might appear exaggerated in the contemporary context, as digital forms of identification processes and related verification mechanisms gradually grow, so too will such threats, especially those based on deepfakes (Brooks et al., 2022; US Department of Homeland Security, 2021; US Cybersecurity and Infrastructure Security Agency, 2020). Even if such major deepfake-based threats have yet to make their presence felt, regulatory efforts to control deepfakes are anticipating the possibility that deepfake-based forms of security threats can rapidly metastasize because they are increasingly easier to create and deploy (Satarino & Mozur, 2023; US Congress, 2022). The logic increasingly at play here—which links digital sovereignty and identity control in the process of digitalization (Leese, 2022) – also echoes the phenomenon of smart or virtual borders (Shachar, 2020; Püschmann, 2022). In this context, AI-based identification technologies correspond to a restrictive policy attempting to prevent "undesired" migrants from entering EU territory (Aden, 2020; Vavoula & Özkul, 2023, fn 43), and deepfakes may then be used to circumvent these technologically strengthened external borders by creating new attack vectors for identity-proofing protocols. Several EU legal instruments have already been adopted for that purpose (European Commission, 2023). These developments should be assessed critically because they arguably participate to frame the figure of the "foreigner" as an indirect threat to EU sovereignty, especially given widespread securitization discourses in EU policies and legislation (Liboreiro, 2022).

Remote ID verification increasingly relates to the emerging EU approach to digital sovereignty, as understood in its minimalistic conception in terms of strategic autonomy and cybersecurity, which can include so-called informational threats. Art. 24(1) of the eIDAS Regulation is a good example of how remote ID proofing *already relates* to the exercise of state sovereignty through digital means by relativizing the traditional importance of physical powers exercised territorially by sovereign states in international society. This provision mandates the following:

> When issuing a qualified certificate for trust service, a qualified certificate for a trust service provider shall verify, by appropriate means and in accordance with national law, the identity and, if applicable, any specific attributes of the natural or legal person to whom the qualified certificate is issued. (European Parliament & Council of the EU, 2014, Art. 24(1))

Nonetheless, the digitalization and datafication of interactions in that context have the legal effect of inciting states to justify the exercise of power over persons in the digital realm, based on both territoriality and personality principles under international law. Identity control and verification procedures are now increasingly exercisable without the individuals subjected to them having to be physically present on the territory of the state in question. Protection against identity manipulation and theft refers to the general concept of sovereignty, insofar as it deeply concerns the nationality principle under international law and the emerging legal concept of digital citizenship in EU law. Under general international law, one of the traditional core prerogatives of states is to attribute nationality and associated status, such as the status of a legal resident. This also confers on states the power to verify their validity by various means. Most importantly for our purpose, this also entails the power to verify the official identity of physical persons under their jurisdiction. We submit that in the process of digitalizing major processes, the nationality principle and its derivatives constitute – together with the territoriality principle – the main legal basis upon which states exercise powers in the cyber/digital environment to verify the identity of persons under their jurisdiction. However, international law has "no coherent, accepted definition of nationality in international law and only conflicting legal approaches exist under the different municipal laws of states" (Shaw, 2014, p. 479). If this is not the only constellation within which deepfakes can be framed as a security threat, it constitutes a particularly important case study that intervenes at the intersection of states' general powers to determine and verify identity online and challenges relating to the preservation of the integrity of electoral processes in an increasingly digitalized and datafied global environment (van Dijck, 2014; ***a***ENISA, 2023). For this reason, the European Parliament has proposed an amendment to Annex III of the European Commission's proposal for an AI regulation that would qualify as high-risk AI systems that are "intended to be used for influencing the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referenda" (***a***European Parliament, 2023, p. 116,

(40a) and p. 125, Annex III(8)(aa)). If this amendment is finalized at the end of the trilogue between the three core EU political institutions after the European Parliament's adoption of a report on the AI regulation in its Plenary session of June 14, 2023 (***d***European Parliament, 2023) this qualification could apply to certain deepfakes.

More generally, the legal status of deepfake detection remains in flux in the ongoing legislative process at the EU level because negotiations between the European Commission, the Council of the EU, and the European Parliament continue. For this reason, there remains no definitive version of the AI regulation at the time of writing. The original proposal was published in April 2021 by the European Commission (***b***European Commission, 2021). In November 2022, the Council published an amended version of the text entitled "General Approach" (***f***Council of the EU, 2022). Then, in May 2023, the European Parliament issued a partially amended version of the text for the future EU regulation entitled "Draft Compromise" (***a***European Parliament, 2023). In the original version of the so-called AI Act as drafted by the European Commission, deepfake detection was clearly classified per se in Art. 52(3) as a medium-risk AI system mainly subject to transparency obligations:

> Users of an AI system that generates or manipulates image, audio, video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated. (***b***European Commission, 2021, p. 69, Art. 52(3); ***f***Council of the EU, 2022, p. 136, Art. 52(3))

However, this legal qualification changes when deepfake detection is used in applications representing high levels of risk, leading to a listing in a technical annex to the (proposed) EU Regulation (***b***European Commission, 2021, pp. 4–5, Annex III). For instance, when deepfake detection is used for law enforcement purposes, as provided, inter alia, in Annex III(6) of the future AI regulation (ibid.). Subsequently, the Council of the EU – which represents the interests of the governments of the EU member states – has amended the text of the Regulation to downgrade the risk-based classification of deepfake detection when used for law enforcement purposes, essentially removing it from the list in Annex III(6) (***f***Council of the EU, 2022, p. 5, para. 1.4 and p. 200, Annex III(7)(c)). Finally, the European Parliament adopted its own amendments with its Draft Compromise (published on May 11, 2023), which was formally adopted in its plenary session on June 14, 2023 (***d***European Parliament, 2023).

The European Parliament also provides that although deepfake detection falls generally into the medium-risk category, it partially reintroduces the high-risk qualification when deepfake detection is used in, for example, the forensic work of law-enforcement agencies as established in an amended version of the clause in Annex III(6) (***a***European Parliament, 2023, pp. 122–125, Annex III). The European Parliament has also introduced new obligations for foundational

AI models (i.e., generative AI models) – which are classified as high-risk AI systems that are subject to the labeling obligations provided for in Art. 52(1) of the Regulation (ibid., 2023, p. 40, Art. 28(b)) – while expending transparency obligations foreseen in Art. 52 to a wider range of AI systems (ibid., p. 41, Art. 29(5), p. 143, Art. 4a(2) and p. 42, Art. 29(6a)). In addition, the European Parliament wants to introduce a new exemption, according to which biometric AI-based systems for certain identity verification purposes ("1 to 1") do not fall within the category of high-risk AI systems that should otherwise apply to them (ibid., p. 112, (33) and p. 122, Annex III(1)(aa)). However, Annex III(7)(c)–(d), as amended by the European Parliament, still classifies AI systems as high-risk when used for identity verification in the context of migration management, asylum, and border control (ibid., pp. 124–125). In brief, not only do legal qualification and corresponding obligations under the future AI regulation remain in flux but current approaches followed by the three main EU institutions involved in the legislative process do not fully cohere at present.

However, despite current uncertainties, we can advance that deepfake detection in the context of remote ID proofing might fall under the high-risk AI category that subjects the development, sale, deployment, or use of such AI systems to stricter legal obligations under the future AI regulation. The various existing techniques rely on biometric data and can even entail forms of "emotion" recognition based on the analysis of biometric features. This could imply the application of Art. 6 and Title III of the future AI regulation, which describe "high-risk AI systems" (***b***European Commission, 2021, pp. 45–58).

Despite their apparent technicality, these legal questions are central, with digital identity verification and control increasingly participating in important aspects of digital sovereignty. The link between remote ID proofing and sovereignty is even stronger for certain use cases for which EU law mandates the verification of the identity of persons in online transactions for anti-money laundering or countering terrorism financing purposes, as foreseen by the Anti-Money Laundering/Countering the Financing of Terrorism (AMT/CFT) directives. The AML/CFT 5th Directive was adopted to strengthen the possibilities of the EU controlling financial transactions, including the identity of persons involved in those transactions, especially with respect to third countries considered to present risks, due to an insufficient level of control over money laundering and terrorism financing (European Commission, 2018). However, this logic is not limited to the relevant supervisory authorities in the banking and financial sectors because the AML/CFT 5th Directive "grants the general public access to beneficial ownership data of EU-based companies" (Zetzsche et al., 2020, p. 352).

Art. 9 of the AML/CFT 5th Directive seeks to protect the integrity of the European financial system (European Parliament & Council of the EU, 2015; Savin, 2022, p. 2). The exploitation of cybersecurity vulnerabilities via identity manipulation can engender the materialization of threats. These threats are not only potentially damaging for parties involved in online transactions or communications but more broadly for countries beyond a certain threshold (*a*ENISA, pp. 45–46). This is particularly possible in the case of governmental or economic and financial activities that are in the process of digitalization because of the high stakes involved (European Parliament, 2020, Art. 9(2)–(3) and Arts. 10(1)–(2)), should cybersecurity threats and losses within the context of digital cross-border activities materialize.[11]

Deepfakes potentially impact EU digital sovereignty because they represent tools that can be used to manipulate digitalized identity or digital means of identity verification, such as by employing artificially generated faces to circumvent digital identity verification or the demand for other ID documents (Europol European Cybercrime Centre, 2022, pp. 54–65). This might include cases where deepfakes use so-called morphing attacks to create new face images by morphing or combining the face images of two (or more) persons. These morphed images have sufficiently high biometric similarity to all persons in this morph set, which would make a travel document generated for one of these persons also usable for every other person in that group. This includes official identification mechanisms, such as passports. Deepfakes may also correspond to external "informational threats,"[12] a pressing issue due to rapid technological developments that enable the generation of increasingly elaborate forms of deepfakes (Kropotov et al., 2022; Europol Innovation Lab, 2022).

3.2 The EU's Regulatory Reactions in the Cybersecurity Field

As discussed, the precise contours of the EU perspective on digital and informational sovereignty remain disputed. Nonetheless, there is, at the least, agreement that sovereignty in the digital context can be equated with the objective of ensuring strategic autonomy, mostly against external threats, including informational threats (European Parliament, 2023). Against this backdrop, one important constellation for achieving strategic autonomy involves ensuring a satisfying level of cybersecurity for the EU, thereby connecting sovereignty and cybersecurity (Savin, 2022, p. 4). Several EU digital policy milestones have emerged in connection to cybersecurity issues, as demonstrated by the strength-

[11] There is indeed, according to some authors, the ambition to render identification verification procedures within data-driven processes at the EU and international level as more secure than their pre-digital counterparts. For a discussion, see Zetzsche et al. (2020, p. 352).

[12] Deepfakes can qualify as a cyber threat according to the definition provided by the 2019 EU Cyber Security Act (European Parliament & Council of the EU, 2019, Art. 2(8)). Also, deepfakes can fall both under the qualification of cybersecurity incident and cyber threat of the US SEC Proposed Rule on Cybersecurity Risk Management, 2022, p. 41.

ened role attributed to the EU Agency for Cybersecurity (ENISA) following the adoption of the EU Cybersecurity Regulation in 2019 (Bendiek & Stürzer, 2022, pp. 3–4). Since then, ENISA has indeed been entrusted with the task of "contribut[ing] to the development and implementation of Union policy and law, by "supporting […] the development and implementation of Union policy in the field of electronic identity and trust services" (European Parliament & Council of the EU, 2019, Art. 5). Deepfakes can constitute a threat to all the policies still in the making in relation to electronic identification and trust services at the EU level.

Under the currently applicable version of the eIDAS Regulation (IV), the supervisory bodies of EU member states must already inform "other supervisory bodies and the public about breaches of security or loss of integrity per Art. 19(2) of this Regulation" (European Parliament & Council of the EU, 2014, Art. 17(3)(c)). Art. 19(2) requires trusted service providers to notify competent supervisory bodies or other relevant bodies "without undue delay but in any event within 24 hours after having become aware of it," in case "of any breach of security or loss of integrity that has a significant impact on the trust service provided or on the personal data maintained therein" (ibid., Art. 19(2)).

This indicates that European institutions are willing to strengthen the role of ENISA in this constellation, with the new version of the eIDAS Regulation currently under negotiation, especially regarding the notification obligation for cybersecurity breaches within the EU (Veale & Brown, 2020; US Security and Exchange Commission, 2022). This version of notification of security breaches and other incidents constitutes a major new cybersecurity tool,[13] adding to existing obligations under EU law, including Art. 33 of the GDPR, which already demands notification when cybersecurity breaches compromise personal data (European Parliament & Council of the EU, 2016, Art. 33(1) and Art. 33(3)). Furthermore, the European Commission has made explicit its ambition to holistically enhance cybersecurity at the EU level while strengthening the EU's digital sovereignty with remote ID proofing procedures, to which deepfakes constitute a growing threat. Importantly, the amended version of Art. 17(4)(c) stresses the importance of security threats for the public interest. This demonstrates a willingness to treat security threats targeting electronic identification mechanisms as a matter of public and EU law and no longer as a mere question of cooperation between EU member states (***e***Council of the EU, 2022, p. 46; ibid. pp. 46–47, Art. 17(4)(f)).

Cybersecurity concerns and related "risk management, reporting, and information sharing" are also relevant in relation to the recently adopted NIS2 Directive for a high common level of cybersecurity in the EU (European Parliament & Council of the EU, 2023; European Parliament, 2023). The NIS2 Directive focuses on the operators of essential service providers, a category encom-

[13] Notably, there is a trend toward developing strengthened notifications for cyber threats, including deepfake-based cyber threats in the regulation of the private sector (e.g., under securities law). For a discussion, see Trautman and Newman (2022).

passing public administration entrusted with the task of verifying the identity of persons for public law purposes, but also some major public and private companies, provided they qualify as critical entities (***b***European Parliament & Council of the EU, 2022, Arts. 2 and 3). The European Commission has issued another proposal for a regulation on horizontal cybersecurity requirements for products with digital elements, with a potentially broad scope of application because all products with digital elements will have to comply with these future rules – not only critical entities (European Data Protection Supervisor, 2022, pp. 5–6, para. 8).[14] This future regulation will also entail obligations to give notification concerning risks related to products with digital components, alongside other information and cooperation obligations aimed at ensuring a satisfying level of security.

For all these reasons, several recent EU initiatives demonstrate the ambition of more deeply considering cybersecurity and informational threats for the whole EU and not just for individual member states.[15] Deepfakes could intervene as one such cybersecurity and informational threats, given that (by definition) they technologically enable identity distortion and manipulation both online and offline.

The fact that consensus regarding digital sovereignty within the EU mostly appears in relation to notions of strategic autonomy and cybersecurity threats is illustrated by several recent Council of the Union conclusions that all emphasize the importance of cybersecurity and informational self-determination for the EU approach to digital sovereignty (***c***Council of the EU, 2021; ***d***Council of the EU, 2022). Developments at the EU level clearly show a willingness to move forward with the establishment of a common cybersecurity strategy that serves the whole EU's digital and informational sovereignty. Indeed, they make manifest the fact that securing digital identity mechanisms and establishing cybersecurity processes aimed specifically at protecting the integrity of decision-making processes are increasingly influential in the emergence of a minimal consensus European understanding of digital sovereignty.

4 Conclusion

In conclusion, the high number of regulatory initiatives around cybersecurity at the EU level demonstrates that EU policymakers have identified cyber-related threats as highly relevant to the region's emerging digital sovereignty. To the

14 However, these two EU legal instruments (the NIS2 Directive and the future EU regulation on horizontal cybersecurity requirements for products with digital elements) should not be perceived as mutually exclusive, given the EU's willingness to foster a complementary enforcement of the two (ibid., p. 9, paras. 25–27).

15 European Parliament & Council of the EU (2019, (7), (10)) with ***e***Council of the EU (2022, p. 46, p. 52, Arts. 24(2)(e) and (fb), p. 55, Art. (31)a). Regarding systemic risks for the EU associated with the activities of very large online platforms, especially those related to "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security," see ***a***European Parliament & Council of the EU (2022, Arts. 34(1)(c) and 35(1)(k)); ***a***European Parliament (2023, p. 125, Annex III(8)(ab) and p. 116, (40b)).

extent that deepfakes constitute cyber threats and disinformation tools in various constellations, they will be covered by several existing or future EU legal instruments that are indirectly participating in the development of an EU conception of digital sovereignty. However, this does not automatically mean that deepfake detection can effectively prevent all potential related threats. Undoubtedly, deepfake detection constitutes one tool among many for mitigating risks related to deepfakes that can only potentially have some effect if certain conditions are fulfilled, such as AI literacy increasing among the general public (***a***European Parliament, 2023, p. 89, Art. 56b(s), p. 136, (9b), pp. 143 – 144, Art. 4d). Furthermore, more clarification is needed regarding the legal status of deepfake detection in the various settings in which it can operate. That is not currently the case with the ongoing negotiations for the EU's future AI regulation.

In this context, reliable and trustworthy deepfake detection would exert a narrow but nonetheless important influence. Deepfakes are increasingly perceived as being able to threaten decision-making processes in the global context of digitalization while posing threats to the security of persons and societies within the EU. For this reason, deepfake detection and its use for ID remote verification integrate the emerging EU approach to digital informational sovereignty, which is currently mostly focused on ensuring security and strategic autonomy while protecting fundamental rights, democracy, and the rule of law.

References

Aden, H. (2020). Interoperability between EU policing and migration databases: Risks for privacy. *European Public Law*, 26(1), 93 – 108.

Akande, D., Coco, A., & de Souza Dias, T. (2022). Drawing the cyber baseline: The applicability of existing international law to the governance of information and communication technologies. *International Law Studies*, 99, 4 – 36.

Arner, D. W., Castello, G. G., & Selga, E. K. (2021). The transnational data governance problem. *University of Honk Kong Faculty of Law Research Paper No. 2021/039*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3912487

Baptista, E. (2022, January 28). *China issues draft rules for fake in cyberspace*. Reuters. https://www.reuters.com/world/china/china-regulator-issues-draft-rules-cyberspace-content-providers-2022-01-28/

Bendiek, A., & Stürzer, I. (2022, April 30). Die digitale Souveränität der EU ist umstritten [Advancing European Internal and External Digital Sovereignty]. *SWP-Aktuell 2022/A*. https://www.swp-berlin.org/publikation/die-digitale-souveraenitaet-der-eu-ist-umstritten

Bickert, M. (2020, 6 January). *Enforcing against manipulated media*. Meta. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

Bradford, A. (2020). *The Brussels Effect: how the European Union rules the world*. Oxford University Press

Bradford, A. (2012). The Brussels effect. *Northwestern University Law Review*, 107(1), 1 – 67.

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

Brooks, T., Daniel, C. P., H., J., Kim, S., R., M., Sahin, B., S., J., T., O., & V., R. (2022). *Phase 2: Increasing threats of deepfakes identities – Mitigation measures*. US Department of Homeland Security. https://www.dhs.gov/sites/default/files/2022-10/AEP%20DeepFake%20PHASE2%20FINAL%20corrected20221006.pdf

Ciftci, U. A., Yuksek, G., & Demir, I (2023, November 2). My face my choice: Privacy enhancing deepfakes for social media anonymisation. https://arxiv.org/pdf/2211.01361v1.pdf

Chan Chin, Y., Park, A., & Li, K. (2022). A comparative study on false information governance in Chinese and American social media platforms. *Policy & Internet*, 14(2), 262 – 283.

Chander, A., & Sun, H. (2022). Sovereignty 2.0. *Vanderbilt Journal of Transnational Law*, 55(2), 283–324.

Chapdelaine, P., McLeod Rogers, J. (2021). Contested sovereignties: States, media, platforms, peoples, and the regulation of media content and big data in the networked society. *Laws*, 10(3), 66–97.

Chee, F. Y. (2022, June 14). *Exclusive: Google, Facebook, Twitter to tackle deepfakes or risk EU fines*. Reuters. https://www.reuters.com/technology/google-facebook-twitter-will-have-tackle-deepfakes-or-risk-eu-fines-sources-2022-06-13/

China Law Translate (2022, January 28). *Provisions on the administration of deep synthesis internet information services (Draft for solicitation of comments)*. https://www.chinalawtranslate.com/en/deep-synthesis-draft/

Council of the EU (2020, 21 October). *Presidency conclusions – the charter of fundamental rights in the context of artificial intelligence and digital change* [Report 11481/20]. https://www.consilium.europa.eu/media/46496/st11481-en20.pdf

Council of the EU (2021). *Proposition de règlement du Parlement européen et du Conseil modifiant le règlement (UE) n°910/2014 en ce qui concerne l'établissement d'un cadre européen relatif à une identité numérique – Deuxième proposition de compromis [Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity. – Second Compromise Proposal]* [Report 9200/22].

aCouncil of the EU (2022, March 21). *A strategic compass for security and defense* [Press release]. https://www.consilium.europa.eu/en/press/press-releases/2022/03/21/a-strategic-compass-for-a-stronger-eu-security-and-defence-in-the-next-decade/

bCouncil of the EU (2022). *Council conclusions on a Framework for a coordinated EU response to hybrid campaigns* [Report 10016/22]. https://data.consilium.europa.eu/doc/document/ST-10016-2022-INIT/en/pdf

cCouncil of the EU (2022). *Council conclusions on the Special Report of the European Court of Auditors No 05/2022 entitled 'Cybersecurity of the EU Institutions, bodies and agencies: Level of preparedness overall not commensurate with the threats'* [Report 10504/22]. https://data.consilium.europa.eu/doc/document/ST-10016-2022-INIT/en/pdf

dCouncil of the EU (2022). *Council conclusions on Foreign Informational Manipulation and Interference (FIMI)* [Report 11429/22]. https://data.consilium.europa.eu/doc/document/ST-11429-2022-INIT/en/pdf

eCouncil of the EU (2022). *Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity – Fifth compromise proposal* [Report 13700/2].

fCouncil of the EU (2022). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – General approach* [Report 14336/22].

De Gregorio, G. (2022). Digital constitutionalism across the Atlantic. *Global Constitutionalism*, 11(2), 297–324.

Delfino, R. A. (2019). Pornographic deepfakes: The case for federal criminalization of revenge porn's next tragic act. *Fordham Law Review*, 88(3), 887–938.

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, 23(7), 2072–2098.

DigiChina (2023). Translation. Measures for the management of generative artificial intelligence services (Draft for Comment)–April 2023. Stanford University. https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, 26(1), 69–91.

aENISA (2021, March 11). *Remote ID proofing analysis of methods to carry out identity proofing remotely.* https://www.enisa.europa.eu/publications/enisa-report-remote-id-proofing

bENISA (2021, July 16). *Remote identity proofing: How to spot the Fake from the Real?* https://www.enisa.europa.eu/news/enisa-news/remote-identity-proofing-how-to-spot-the-fake-from-the-real

ENISA (2022, December 8). *Cybersecurity & Foreign Interference in the EU Information Ecosystem* [Press release]. https://www.enisa.europa.eu/news/cybersecurity-foreign-interference-in-the-eu-information-ecosystem

aENISA (2023, March). *Identifying emerging cyber security threats and challenges for 2030*. enisa.europa.eu. https://www.enisa.europa.eu/publications/enisa-foresight-cybersecurity-threats-for-2030/

bENISA (2023, June 7). *Artificial intelligence and cybersecurity research*. https://www.enisa.europa.eu/publications/artificial-intelligence-and-cybersecurity-research

European Commission, eIDAS Observatory (2016, June 28). *eIDAS Regulation (Regulation (EU) N° 910/2014)*. https://ec.europa.eu/futurium/en/content/eidas-regulation-regulation-eu-ndeg9102014.html

European Commission (2018, July 9). *Strengthened EU rules to prevent money laundering and terrorism financing*. https://ec.europa.eu/info/files/fact-sheet-main-changes-5th-anti-money-laundering-directive_en

European Commission (2020). *On artificial intelligence – A European approach to excellence and trust* [White paper]. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0065

aEuropean Commission (2021). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions 2030 Digital Compass: the European way for the Digital Decade* [Report COM(2021) 550 final]. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX-:52021DC0550

bEuropean Commission (2021, April 21). *Proposal for a Regulation of the European Parliament and of the Council Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts,* COM(2021) 206 final, 2021/0106(COD) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

cEuropean Commission (2021, May 28). *European Digital Identity*. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-digital-identity_en#digital-identity-for-all-europeans

dEuropean Commission (2021, June 3). *Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX-:52021PC0727

European Commission (2023). Travel – Digitalsing travel documents to make travelling easier. Public Consultation: Consultation period: 5 April 2023 – 28 June 2023. ec.europa.eu. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13514-Travel-digitalising-travel-documents-to-make-travelling-easier_en

European Data Protection Supervisor (2022, November 9). *Opinion 23/2022 on the Proposal for a Regulation of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020*. https://edps.europa.eu/data-protection/our-work/publications/opinions/2022-11-10-horizontal-cybersecurity-requirements-products-digital-elements_en

European Parliament & Council of the EU (2014, August 28). *Regulation (EU) 910/2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC, OJ L 257/73*.

European Parliament & Council of the EU (2015, May 20). *Directive (EU) 2015/849 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing, amending Regulation (EU) No 648/2012 of the European Parliament and of the Council, repealing Directive 2005/60/EC*.

European Parliament & Council of the EU (2016, April 27). *Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and of the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*.

European Parliament & Council of the EU (2019). *Regulation (EU) 2019/881 on ENISA (the European Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act)*, PE/86/2018/REV/1.

aEuropean Parliament & Council of the EU (2022, October 27). *Regulation (EU) 2022/2065 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ L 277/1*.

bEuropean Parliament & European Council (2022, December 27). *Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 210/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive), OJ L 333/80*.

European Parliament (2020, November 10). *Digital finance: Digital Operational Resilience Act (DORA), European legislative resolution on the proposal for a regulation of the European Parliament and of the Council on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009*, (EU) No 648/2012, (EU) No 909/2014 (COM(2020)0595 – C9-0304/2020 – 2020/0266 (COD)), P9_TA(2022)0381, Arts. 9(2)-(3) and 10(1)-(2).

aEuropean Parliament (2023, 11 May). *Draft compromise amendments on the draft report, Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD))', KMB/DA/AS*. https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence

bEuropean Parliament (2023, May 11). *AI Act: a step closer to the first rules on Artificial Intelligence*. https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence

cEuropean Parliament (2023, June 1). *Foreign interference in all democratic processes in the European Union, including disinformation. European Parliament resolution of 1 June 2023 on foreign interference in all democratic processes in the European Union, including disinformation (2022/2075(INI)). P9_TA(2023)0219*.

dEuropean Parliament (2023, June 14). *MEPs ready to negotiate first-ever rules for safe and transparent AI*. https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai

European Parliament Think Tank (2023, February 8). *The NIS2 Directive: A high common level of cybersecurity in the EU*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)689333

European Parliamentary Research Service (2021, July). *Tackling deepfakes in European policy, PE 690.039*. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf

European Union (2012). *Consolidated version of the Treaty on the Functioning of the European Union* [Report C 115/47]. Official Journal of the European Union. https://eur-lex.europa.eu/resource.html?uri=cellar:41f89a28-1fc6-4c92-b1c8-03327d1b1ecc.0007.02/DOC_1&format=PDF

Europol Innovation Lab (2022, April 28). *Facing reality? Law enforcement and the challenge of deepfakes*. https://www.europol.europa.eu/media-press/newsroom/news/europol-report-finds-deepfake-technology-could-become-staple-tool-for-organised-crime

Europol European Cybercrime Centre, United Nations Interregional Crime and Justice Research Institute (UNICRI), & Trend Micro (2020, November 19). *Report on Malicious Uses and Abuses of Artificial Intelligence (AI)*. https://eucrim.eu/news/report-on-malicious-uses-and-abuses-of-artificial-intelligence/

Fadilpašić, S. (2022, May 31). *Google is cracking down hard on deepfakes. Tech Radar*. https://www.techradar.com/news/google-is-cracking-down-hard-on-deepfakes

Fung, P. & Etienne, H. (2022). Confucius, cyberpunk and Mr. Science: comparing AI ethic principles between China and the EU. *AI and Ethics 3*, 505–511. https://doi.org/10.1007/s43681-022-00180-6

Georgetown Center for Security and Emerging Technology (2021, October 21). Translation. *Ethical norms for new generation artificial intelligence released*. https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/

Goldsmith, J. (2019). Sovereign difference and sovereign deference on the internet. *Yale Law Journal Forum*, 128, 818–826.

Google Research (2022, August 28). *Colaboratory: Frequently asked questions*. https://research.google.com/colaboratory/faq.html

Heath, R. (2023, May 8). *China races ahead of US on AI regulation*. Axios. https://www.axios.com/2023/05/08/china-ai-regulation-race

Hine, E., & Floridi, L. (2022). New deepfake regulations in China are a tool for social stability, but at what cost? *Nature Machine Intelligence*, 4, 608–610.

Hsu, T., & Lee Myers, S. (2023, April 8). *Can we no longer believe anything we see?* New York Times. https://www.nytimes.com/2023/04/08/business/media/ai-generated-images.html

Iakovelva, S. (2022). On digital sovereignty, new European data rules, and the future of free data. *Legal Issues of European Integration*, 49(4), 339–348.

Iliopoulou-Penot, A. (2022). The construction of a European digital citizenship in the case law of the Court of Justice of the EU. *Common Market Law Review*, 59(4), 969–1006.

Jon Heller, K. (2021). In defense of pure sovereignty in cyberspace. *International Law Studies*, 97, 1432–1499.

Keane, J. (2022, May 6). *China and Europe are leading the push to regulate A.I. – one of them could set the global playbook*. CNBC. https://www.cnbc.com/2022/05/26/china-and-europe-are-leading-the-push-to-regulate-ai.html

Kropotov, V., Yarochkin, F., Gibson, C., & Hilt, S. (2022, September 27). *How underground groups use stolen identities and deepfakes*. Trend Micro. https://www.trendmicro.com/en_us/research/22/i/how-underground-groups-use-stolen-identities-and-deepfakes.html

Lambach, D. (2020). The territorialization of cyberspace. *International Studies Review*, 22(3), 482–506.

LastBluejay (2020, January 9). *Updates to Our Policy Around Impersonation* [Online Forum Post]. Reddit. https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation/

Lawrence, J., Trautman, L. J., & Newman, N. (2022, 29 April). A proposed SEC cyber data disclosure advisory commission [Research paper No. 22]. *Texas A&M University School of Law Legal Studies*. https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=4097138

Leese, M. (2022). Fixing state vision: Interoperability, biometrics, and identity management in the EU. *Geopolitics*, 2022, 27(1), 113–133.

Liboreiro, J. (2022, November 20). *Joseph Borell apologises for controversial 'garden vs jungle' metaphor but defends speech*. Euro News. https://www.euronews.com/my-europe/2022/10/19/josep-borrell-apologises-for-controversial-garden-vs-jungle-metaphor-but-stands-his-ground

Mainwaring, S. (2020). Always in control? Sovereign states in cyberspace. *European Journal of International Security*, 5(2), 215–232.

Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, 53, 3974–4026. https://doi.org/10.1007/s10489-022-03766-z.

Metz, C. (2023, April 4). *Instant videos could represent the next leap in A.I. technology*. New York Times. https://www.nytimes.com/2023/04/04/technology/runway-ai-videos.html

Metz, C., & Blumenthal, S. (2019, June 7). *How A.I. could be weaponized to spread disinformation*. New York Times. https://www.nytimes.com/interactive/2019/06/07/technology/ai-text-disinformation.html

Monyihan, H. (2019, December). *The application of international law to state cyberattacks: Sovereignty and non-intervention* [Research paper]. Chatham House. https://www.chathamhouse.org/sites/default/files/publications/research/2019-11-29-Intl-Law-Cyberattacks.pdf

Mueller, M. (2020). Against sovereignty in cyberspace. *International Studies Review*, 22(4), 779–801

Perset, K., Plonk, A., & Russel, S. (2023, April). *As language models and generative AI take the world by storm, the OECD is tracking the policy implications*. OECD AI Policy Observatory. https://oecd.ai/en/wonk/language-models-policy-implications

Pohle, J., & Voelsen, D. (2022). Centrality and power. The struggle over the techno-political configuration of the Internet and the global digital order. *Policy & Internet* 14(1), 13–27.

Püschmann, J. (2022, March 18). *Book review: The shifting border: legal cartographies of migration and mobility*. University of Oxford Faculty of Law Blogs. https://blogs.law.ox.ac.uk/research-subject-groups/centre-criminology/centreborder-criminologies/blog/2022/03/book-review

Ruth, M. (2023, 11 May). *Turkish presidential candidate quits race after release of alleged sex tape*. theguardian.com. https://www.theguardian.com/world/2023/may/11/muharrem-ince-turkish-presidential-candidate-withdraws-alleged-sex-tape

Satariano, A., & Mozur, P. (2023, 7 February). The People Onscreen Are Fake. *The Disinformation Is Real*. nytimes.com. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

Savin, A. (2022, April 4). *Digital sovereignty and its impact on EU policymaking* [Research paper 22-02]. Copenhagen Business School Law. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4075106

Schmitt, M. N. (2018). "Virtual" disenfranchisement: Cyber election meddling in the grey zones of international law. *Chicago Journal of International Law*, 19(1), 30–67.

Shachar, A. (2020). *The shifting border: Legal cartographies of migration and mobility*. Manchester University Press.

Simonite, T. (2020, July 7). *Deepfakes are becoming the hot new corporate training tool.* Wired. https://www.wired.com/story/covid-drives-real-businesses-deepfake-technology/

Snow, J. (2021, October 24). *These historical artefacts are totally faked*. Wired. https://www.wired.co.uk/article/fake-artefacts-ai

aTwitter Help Center (2023, April). *Synthetic and manipulated media policy*. https://help.twitter.com/en/rules-and-policies/manipulated-media

bTwitter Help Center (2023, April). *Misleading and deceptive identities policy.* https://help.twitter.com/en/rules-and-policies/twitter-impersonation-and-deceptive-identities-policy

UNESCO (2021, November). *Recommendation on the ethics of artificial intelligence*. [Report No. 66582]. https://unesdoc.unesco.org/ark:/48223/pf0000380455

US Congress (2022, June 3). *Deep Fakes and National Security. US Congressional Research Service*. https://crsreports.congress.gov/product/pdf/IF/IF11333

US Cybersecurity and Infrastructure Security Agency (2020, July 28). *Critical Infrastructure Security Agency and Resilience Note*. CISA. https://www.cisa.gov/sites/default/files/publications/cisa-election-infrastructure-cyber-risk-assessment_508.pdf

US Department of Homeland Security (2021). *Increasing threat of deepfake identities*. US Department of Homeland Security. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

US Security and Exchange Commission (2022, March 9). *Proposed Rule on Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure* [Report RIN 3235-AM89]. https://www.sec.gov/news/press-release/2022-39

van Dijck J. (2014). Datafication, dataism, and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2), 197–208.

Vavoula, N., & Özkul, D. (2023, March 29). *Submission to European Commission consultation on "security-related information sharing."* StateWatch. https://www.statewatch.org/analyses/2023/submission-to-europe-an-commission-consultation-on-security-related-information-sharing/

Veale, M., & Brown, I. (2020). Cybersecurity. *Internet Policy Review*, 9(4). https://doi.org/10.14763/2020.4.1533

Vincent, J. (2023, March 21). *TikTok bans deepfakes of nonpublic figures and fake endorsements in rule refresh*. The Verge. https://www.theverge.com/2023/3/21/23648099/tiktok-content-moderation-rules-deepfakes-ai

Zetzsche, D., Arner, D., Buckley, R., & Weber, R. H. (2020). The evolution and future of data-driven finance in the EU. *Common Market Law Review*, 57(2), 331–360.

Acknowledgements

The work in this paper is partly funded by the German Federal Ministry of Education and Research (BMBF) under the FAKE-ID project (grant numbers OVGU FKZ: 13N15736 and HWR/FPÖS FKZ: 13N15737). We would like to thank all project partners for their contribution to fruitful discussions and exchanges in the context of the project. We thank the participants in our panel during the 2022 4th Weizenbaum Conference *Practicing Sovereignty. Interventions for open digital futures*. Finally, we would like to thank Mario Petoshati (HWR) for his thorough review and edits.

Date received: March 2023
Date accepted: July 2023