https://doi.org/10.6092/issn.1825-1927/18872 ISSN 1825-1927

Data collection via web scraping: privacy and facial recognition after Clearview

Fabrizio Lala*

Abstract.

Web scraping, intended as the series of available techniques for automatically gathering data publicly available on the internet is ubiquitous in today's society, characterised by an ever increasing amount of data on the web. This contribution aims to shed light on the existing legal framework for fighting illegal scraping personal data, with a focus on biometric data. The recent judgment of the ECtHR in Glukhin v. Russia and the Clearview AI "saga", are crucial in the analysis, considering their impact on the proposed EU AI act as amended by the European Parliament, which included the untargeted scraping of facial images from the internet in the category of prohibited high-risk activities.

Keywords:

Web scraping, publicly available information, personal data, biometric data, facial recognition, AI, human rights, Clearview.

1 Introduction

The information economy has seen a steady, astonishing increase of the amount of data available on the internet. As part of the "everyday revolution" humans are experiencing in this era[1], all kinds of data are published online, most of which is publicly available, thus free for people to extract it. The amount of information within easy reach is so important that it would be almost impossible for an individual, although interested in a specific topic, to gather it all manually timely, without the help of specific technological tools.

Such tools are represented by web scrapers, which access and download large volumes of data available on websites and web crawlers, bots which browse and index web pages systematically[2]. Scrapers have become ubiquitous on the internet: recent statistics show that almost half of worldwide internet traffic comes from bots[3].

In 2023, web scraping has often made the headlines, with Twitter (now X) announcing "emergency measures" (rate limiting) to deal with scraping[4], and major tech companies such as Google, Microsoft and Meta reported to grab part of the vast amount of data collected through their services to train their artificial intelligence tools[5].

*University of Turin, Department of Law; ✉ fabrizio.lala@unito.it

1. Durante, Massimo. *Computational power: the impact of ICT on law, society and knowledge*. Routledge, 2021.

2. See definition of web scraping by Zhao, B., in Schintler, Laurie A., and Connie L. McNeely, eds. *Encyclopedia of big data.* Cham: Springer International Publishing, 2022.: "Web scraping, also known as web extraction or harvesting, is a technique to extract data from the World Wide Web (WWW) and save it to a file system or database for later retrieval or analysis. Commonly, web data is scraped utilizing Hypertext Transfer Protocol (HTTP) or through a web browser. This is accomplished either manually by a user or automatically by a bot or web crawler".

3. According to the annual Bad Bot Report, bots accounted for 47.4% of total internet traffic in 2022, indicating a 5.1% increase from the previous year. Imperva 2023 Bad Bots Report, available at: https://www.imperva.com/resources/resource-library/reports/bad-bot-report/.

4. Washington Post. "Twitter bars unregistered lurkers from peeking at tweets for now". 30 June 2023, available at: https://www.washingtonpost.com/business/2023/06/30/twitter-bars-unregistered-lurkers-peeking-tweets-now/.

5. Fowler, Geoffrey A., "Your Gmail and Instagram are training AI. There is little you can do about it", Washington Post, Available at: https://www.washingtonpost.com/technology/2023/09/08/gmail-instagram-facebook-trains-ai/.

This contribution aims to analyse the legal instruments applicable to web scraping, with a focus on privacy and data protection, adopting a European perspective.

Web scraping can be considered as a set of technical means for extracting vast amounts of information from web pages, which can serve various goals. For instance, it is used as a tool for research, statistics, for the analysis of firms' websites, of prices of products and services, of customers' habits and preferences, for target advertising, for criminal investigations.

The ubiquity of scraping determines that depending on the specific use of this powerful tool, different outcomes and different consequences can derive. While this study investigates harmful uses of web scraping and possible remedies, scraping can have indeed also positive implications and serve the public good.

From a legal perspective, as far as we know, there is no general law specifically tailored for web-scraping. This derives from the above mentioned multitude of applications of this instrument, and from its definition itself as a technical instrument[6].

Instead, website owners - and, more generally, companies or individuals who may be affected by scraping - can rely on a variety of legal measures, depending on the interests at stake and on the concrete and specific circumstances of the case: contractual liability, non contractual liability, unfair competition law, copyright law, sui generis database protection and privacy and data protection law are all areas of law that are relevant to the fight against illegal uses of web scraping.

The following section of this study is devoted to the analysis of the privacy and data protection implications of a specific kind of web scraping: the one concerning biometric data, and, in particular, of data gathered online for creating and operating facial recognition technologies (FRTs). This analysis is carried out firstly through an introduction about the general implications of scraping publicly available personal data, based on the application of ground principles of the GDPR, such as lawfulness, consent, data protection by design and by default and accountability, with examples of decisions issued by national data protection authorities (DPAs). Secondly, the peculiarities of biometric data and its application to scraping are evaluated.

Section 3 looks at some recent important decisions that shed light on the privacy and human rights issues associated with facial images scraping: first, the landmark ruling of the European Court of Human Rights (ECtHR) in the case of Glukhin v. Russia, and, second, an overview of some of the various cases that have arisen, mainly in America and in Europe, following the launch of the FRT application by the US startup Clearview AI Inc. (Clearview).

Before concluding, in section 4 delves into the possible regulation of scraping in the context of AI, turning to the latest available version of the EU AI act proposal: the negotiating position of the European Parliament, where the provisions about scraping facial images are perhaps the most notable. Reference is made to the newly introduced provision banning the creation or expansion of facial recognition databases through the untargeted scrapin of facial images from the internet or CCTV footage, included among the AI practices to be prohibited, under art. 5 of the proposal. This section does also include the conclusions of this study, in which we argue that web scraping should be seen as a ubiquitous, unavoidable tool. However, rather than being passive about it, lawmakers ought to be careful to tailor new laws to fit current and foreseeable future use cases, with an eye to the most dangerous applications of current and future technologies.

2 Scraping personal data: general principles and biometric data

2.1 General principles: the need for an autonomous legal basis

In the EU, web scraping is illegal under privacy and data protection law, if personal data is gathered online, then (re)used without an appropriate legal basis. The reuse of these data without informing the data subject amounts to a violation of the principle of lawfulness, as laid down in art. 5.1, lett. (a) GDPR. Besides, to be lawful, a processing of scraped data should have a legal basis, such as consent of the data subject pursuant to art. 6.1. Such legal basis should be new, i.e. dedicated and specific to the new form of data processing.

In 2023, the Italian Data Protection Authority (DPA) issued a decision about the retrieval through scraping of websites, conservation and online publication of individuals' telephone numbers[7]. The Italian DPA first noted that the gathering of telephone numbers in view of compiling a telephone directory is only allowed through the consultation of the relevant official national

6. The lack of laws on data scraping can also be understood in the light of the principle of technological neutrality, see: Pagallo, Ugo. "The legal challenges of big data: putting secondary rules first in the field of EU data protection." *Eur. Data Prot. L. Rev.* 3 (2017): 36.

7. Garante per la protezione dei dati personali, Provv. 17 May 2023, n. 9903067.

database. That said, turning to data processing concerns, the DPA noted that the owner of the site where the leaked telephone numbers were published lacked an appropriate legal basis for processing the data it had gathered. Besides, the site lacked any contacts of the data controller, and any information on the possibility of obtaining the deletion of the data in the event of failure of the appropriate form, which was present on the website but did not appear to work.

A defence used in this Italian case and often used by scrapers is based on the public availability of data scraped online. However, public availability does not trigger the application of different rules or exceptions. This derives from the provision of art. 14.2 GDPR, covering information to be provided where personal data has not been obtained directly from the data subject. Under this norm, the controller should inform the data subject about the source from which the personal data originates, and whether it came from publicly accessible sources[8]. This clarification about the source implies that the data is still considered personal, notwithstanding its public availability.

On the same note, outside of the EU, in August 2023 twelve privacy authorities, including Canada, UK, China and Australia, issued a joint statement on data scraping and the protection of privacy, stressing the significant privacy concerns raised by indiscriminate scraping of individuals' personal information publicly available on the internet, making it clear that in most jurisdictions, the public availability of data does not imply that such data is not subject to privacy and data protection laws[9].

2.2 General principles: data protection by design and by default

Third parties' scraping of personal data can lead to a violation of another fundamental principle of the GDPR, namely the obligation to ensure data protection by design and by default, under art. 25.1 and 25.2, borne by the data controller. On this topic, in 2022 the Irish DPA has fined Meta Platforms Ireland Ltd. (Meta), the data controller of Facebook, imposing a fine of 265 mil. euros and a set of corrective measures[10]. In April 2021, following media reports highlighting that a collated dataset reported to contain personal data relating to approximately 533 million Facebook users worldwide had been made available on the web, the Irish DPA decided to start an inquiry to verify the fulfilment of Meta's obligation as data controller under the GDPR, namely on the correct application of art. 25 GDPR and on the effectiveness and integration of the technical and organisational measures implemented with respect to certain features of the Facebook social media[11]. More specifically, the investigation carried out by the Irish DPA also considered the fact that the features in Facebook thanks to which scrapers managed to grab users' data were available to be exploited by fake accounts and bots, as well as the fact that Meta had clearly identified instances of mass scraping with related bot and fake account activity in the relevant features[12].

Art. 25.1 GDPR, about data protection by design, considers the effectiveness of measures and safeguards adopted by the data controller to secure the rights of the data subject as a key element. This entails that those measures and safeguards should be designed to be robust and so that further measures may be implemented by the controller in order to correctly face any increase in risk.

In case of proceedings against the data controller following incidents such as leak of personal data, controllers should thus be able to demonstrate that the principle of data protection by design has been maintained thanks to the implemented measures[13].

Data protection by default (art. 25.2 GDPR) requires the controller to implement appropriate technical and organisational measures to ensure that only personal data necessary for each specific purpose are processed. Importantly, measures so enacted by the controller shall ensure that by default personal data are not made accessible to an indefinite number of natural persons, without the intervention of the data subject[14]. Data protection by design and by default are both expressions of the principle of accountability, at the core of the system of responsibilities set up by the GDPR.

8. See in particular Art. 14.2, lett. (f) GDPR.

9. Joint Statement on data scraping and the protection of privacy, August 24, 2023.

10. Decision of the Irish Data Protection Commission, *Meta Platforms Ireland Ltd.*, 25 November 2022, DPC Inquiry reference: IN-21-4-2.

11. Namely Facebook Search, Facebook Contact Importer, Messenger Contact Importer and Instagram Contact Importer. See point 38 of the decision.

12. See point 46 of the decision.

13. Some scholars argued that the ex post (and not ex ante) obligation of proving the effectiveness of these measures risks jeopardizing data subjects' rights. On this topic, see e.g. Veale, Michael, et al. "When data protection by design and data subject rights clash." *International Data Privacy Law* 8.2 (2018): 105-123.

14. See EDPB, Guidelines 4/2019 on Article 25 Data Protection by Design and by Default at 2: "The core of the provision is to ensure appropriate and effective data protection both by design and by default, which means that controllers should be able to demonstrate that they have the appropriate measures and safeguards in the processing to ensure that the data protection principles and the rights and freedoms of data subjects are effective".

In the case at hand, the Irish DPA identified severe risks associated with the processing, related to "fraud, impersonation and scamming". The DPA underlined that Meta, acting as controller, did not take adequate steps to assess and appraise the risk posed by scrapers and that the measures adopted by the controller when noticing their activity - such as rate limiting and bot detection measures - were insufficient[15].

The DPA thus found art. 25, paragraphs 1 and 2 infringed by Meta and issued an order to bring the data processing into compliance with the GDPR, a reprimand to Meta and two administrative fines in the amount of €150 million, and €115 million respectively.

The issues raised in this decision are mostly common to the ones underlined in the recent Joint declaration on scraping, signed by 12 privacy authorities outside of the EU, where social media companies and other website owners were warned about the privacy risks of scraping: targeted cyberattacks, identity fraud, monitoring, profiling and surveillance of individuals, unauthorized political or intelligence gathering purposes, unwanted direct marketing or spam. The joint declaration calls for social media services and websites to adopt appropriate measures to protect their users' data from scraping, to respect the legal obligations concerning the protection of the personal information published through their services. Protection should be based on multi-layered technical and procedural controls aimed to mitigate risks, such as designating a team devoted to the prevention and monitoring of scraping activities, rate limiting the number of visits by one or more suspect accounts to other accounts or to the service itself, using appropriate technical measures to detect bots (such as CAPTCHAs[16]), taking appropriate legal actions when scraping is detected, e.g. through cease and desist letters and other action to enforce terms and conditions prohibiting scraping, as well as notifying to affected individuals and competent privacy authorities, in jurisdiction where scraping may amount to a data breach.

2.3 Web scraping and biometric data

As seen above, the scraping of personal data may envisage a violation of the data subjects' rights and both third party scrapers and data controllers can be sanctioned for infringing the GDPR.

In this paragraph, an even more serious type of violation of privacy and data protection – and related principles set forth at the primary level source in the EU Charter of fundamental rights (Charter) - that can be committed through the use of scraping tools will be examined: the one concerning biometric data, more specifically facial recognition. In the EU, biometric data is defined as "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique authentification or identification of that natural person, such as facial images or dactyloscopic data"[17]. Biometric data for the purpose of uniquely identifying a natural person is included in the category of special categories of personal data, whose processing is allowed under art. 9 GDPR only under specific conditions[18], including first of all the explicit consent of the data subject (art. 9.2, lett. a[19]). Further, art. 10.2 of the law enforcement Directive sets the same requisite for processing of data by public authorities. Both instruments apply to automated processing of personal data and to manual processing that is part of a filing system[20]. However, the law enforcement directive is more specific and considered as lex specialis, thus it is applicable when public authorities process personal data for the sake of prevention, investigation, prosecution of criminal offences[21]. Scraping of documents including biometric data, such as pictures depicting an individual's facial characteristics, can be used by police authorities as a powerful tool in the framework of FRTs, and relevant applications

15. See points 102 and 140 of the decision.

16. CAPTCHA is the acronym of Completely Automated Public Turing test to tell Computers and Humans Apart. It is a program that tests whether a user is a human or a bot (or another automated program) (PC Mag, Definition of CAPTCHA). Some examples of CAPTCHAs are programs that require a user to: interpret text that is distorted, or look at a set of similar pictures and identify which of these contain a specific object.

17. See Art. 3 (13) Directive EU/2016/680 (Law Enforcement Directive); Art. 4 (14), GDPR; Art. 3 (18), Regulation (EU) 2018/1725 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC.

18. Coherently with the GDPR, the Council of Europe's modernised Convention for the Protection of Individuals with regard to the Automatic Processing of Personal Data (Convention 108+) also includes data uniquely identifying a person under the special categories of data in Art. 6.

19. Although consent is not always applicable in this case, because either EU or MS law can provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject.

20. Artt. 2(1) GDPR and art. 2 Law Enforcement Directive.

21. Recitals 11 and 12 LED and Recital 19 GDPR.

are currently being developed and used in several countries inside and outside the EU[22]. The gathering of such data can not only hinder data protection secondary legislation, also carrying specific risks for fundamental rights, such as people's dignity (art. 1 of the Charter), respect of personal life (art. 7) and protection of personal data (art. 8), with a possible impact on non-discrimination and rights of special groups (e.g. children, persons with disabilities)[23]. Art. 52.1 of the Charter sets the requirement that any limitation to the exercise of fundamental rights and freedoms must be provided for by EU or national law laying down clear and precise rules governing the scope and application of this measure and imposing safeguards so that the data subjects concerned have sufficient guarantees to effectively protect their personal data against the risk of abuse and any unlawful access or use of the processed data. Any limitation to the fundamental rights protected by the Charter must be strictly necessary and proportionate. As a matter of fact, according to the CJEU's settled case-law, derogations and limitations in relation to the protection of personal data must apply only in so far as is strictly necessary[24].

It follows from the above that scraping of photos of people does not automatically amount to a processing of special categories of data, but only when photos are processed through a specific technical means, allowing the unique "identification or authentication of a natural person"[25] In accordance with the main legal principles of data protection[26], the processing of facial images must be lawful, fair and transparent, follow a specific, explicit and legitimate purpose, that is clearly defined in Member State or Union law. Besides, it should comply with the further requirements of data minimisation, accuracy, storage limitation, security and accountability. Furthermore, the obligation to carry out a Data Protection Impact Assessment (DPIA) under art. 35 GDPR should be considered applicable to FRT allowing the unique identification or authentication of individuals. In fact, the DPIA can be seen as mandatory for such use cases, in view of art. 35.3, lett. b), which states examples of mandatory DPIA, including "the processing on a large scale of special categories of data referred to in Article 9.1". Hence, one can conclude that, when developing a FRT instrument, the data controller should certainly carry out a DPIA in order to assess and be in the position to demonstrate its compliance to the GDPR, in line with the aim of this instrument[27].

3 The ruling of the ECtHR in Glukhin v. Russia and the Clearview saga

3.1 Glukhin v. Russia

In July 2023, the European Court of Human Rights (ECtHR) has issued a seminal judgment on facial recognition, the first ruling by the ECtHR on FRTs. The case was brought against the Russian Federation by a Russian national Nikolay Sergeyevich Glukhin. The proceedings concern the applicant's administrative conviction for his failure to notify the authorities of his intention to hold a (solitary) demonstration in Moscow[28]. During the investigation by the Russian police authorities, facial recognition technology was used to process the applicant's personal data and identify him. Besides, pictures of the suspect were also scraped from the internet and matched with those of the CCTVs active in Moscow.

The applicant brought the case to the ECtHR, complaining on the breach of his rights under Articles 10 (Freedom of expression) and 11 (Freedom of assembly and association) of the European Convention on Human Rights (ECHR), in the framework of the administrative offence proceedings opened against him. Concerning the processing of his personal data, Mr. Glukhin claimed the breach of his right to respect for private life, pursuant to Art. 8 of the Convention[29]. Besides, he complained about the violation

22. See e.g., for a view of the situation in England and Wales: Purshouse, Joe, and Liz Campbell. "Automated facial recognition and policing: a Bridge too far?" *Legal Studies 42.2* (2022): 209-227.; for an introduction of the EU and UK legal landscape concerning the use of FRTs, see Christakis, Theodore, et al. "Mapping the Use of Facial Recognition in Public Spaces in Europe–Part 1: A Quest for Clarity: Unpicking the'Catch-All'Term." *Available at SSRN 4110512* (2022).

23. European Union Agency for Fundamental Rights (FRA): Facial recognition technology: fundamental rights considerations in the context of law enforcement. Available online: https://fra.europa.eu/en/publication/2019/facial-recognition.

24. Joined Cases C-293/12 and C-594/12, Judgment of 8 April 2014, ECLI:EU:C:2014:238.

25. See GDPR, recital 51.

26. Art. 5 GDPR and Art. 4 LED. On the role of legal principles in the regulation of technology see: Durante, Massimo, and Luciano Floridi. "A legal principles-based framework for AI liability regulation." *The 2021 yearbook of the digital ethics lab*. Cham: Springer International Publishing, 2022. 93-112.

27. See WP29, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679.

28. *Nikolay Sergeyevich Glukhin v. Russian Federation*, App no. 11519/20 (ECtHR, 4 July 2023).

29. Pursuant to art. 8, par. 1 of the ECHR, "Everyone has the right to respect for his private and family life, his home and his correspondence".

of Art. 6 (Right to a fair trial) of the Convention, which, however, was not examined by the Court, as it concluded for the breach of Art. 8, and of art. 10 (Freedom of expression).

The Court stressed that the concept of "private life" is broad and can embrace multiple aspects of the person's physical and social identity. It is not limited to an "inner circle" in which the individual may live his or her own personal life without outside interference, but also encompasses the right to lead a "private social life"[30]. The Court ruled that the mere storing of data relating to an individual's private life amounts to an interference within the meaning of Art. 8[31]. The processing of the applicant's personal data in the framework of administrative offence proceedings against him, including the use of facial recognition technology in order to (i) identify him from the photographs and the video published on the Telegram messaging app and (ii) locate and arrest him later, envisaged an interference with his right to respect for his private life, pursuant to Art. 8 § 1 of the ECtHR. Indeed, although the Court acknowledged that there was no solid evidence about the use of FRT by the Russian police, a number of hints led the Court to conclude that they had in fact been used: firstly, Russian legislation does not require the police to make a record of their use of such technology or to notify the person concerned, hence it would have been really difficult for the applicant to prove use of FRTs. Secondly, the fact that the police were so quick - two days - in identifying the applicant as the suspect also contributed to the Court's conclusion that FRT had been used indeed[32].

Art. 8, par. 2 ECHR provides an exception to the general prohibition of interferences by public authorities with the exercise of the right to private life. The exception, raised by the Russian Federation, is applicable where explicitly provided by law, and in case of necessity, namely "in the interests of national security, public safety or the economic wellbeing of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others".

At first glance, the reliance on an administrative offence instead of criminal law seems at odds with the "quality of law" requirement. In order to assess the applicability of the exception, the Court examined the facts of the case, finding that the investigation of the administrative offence against the applicant had a legal basis in the domestic legislation, granting the national police the power to investigate administrative offences and to collect evidence, including that containing personal data. Besides, the laws applicable included the provision for the installation of live facial recognition CCTV cameras in the Moscow underground which were accessible to the police. However, the domestic law was found to lack provisions governing the scope and application of measures involving the use of FRTs and, importantly, of strong safeguards against the risk of abuse and arbitrariness.

Furthermore, the ECtHR pointed out that even though the aim of the interference with Mr Glukhin's rights – the prevention of crime - had been legitimate, in theory, however, the measures against the individual had been particularly intrusive, considering the facts, as Mr Glukhin had engaged in a peaceful and solitary protest, without causing any possible danger to the public or transport safety. Therefore, the Court concluded that processing of the applicant's personal data using FRTs in the framework of administrative offence proceedings cannot be regarded as "necessary in a democratic society". Thus, the ECtHR confirmed the violation of art. 8.1 ECHR.

The importance of this judgment lies in the fact that it is the first ECtHR case declaring the unlawful use of FRTs, thereby inscribing unnecessary and unproportionate gathering of facial images of an individual within prohibited policies pursuant to art. 8 ECHR[33]. The application of art. 8 of the ECHR, guaranteeing respect for private and family life, home and correspondence, seems reasonable and in line with previous case-law of the Court on the violation of biometric personal data. For instance, in Gaughran v. the United Kingdom, the Court held the disproportionality of the indefinite retention of DNA, fingerprints and facial images from all convicted adults. In particular, the ECtHR had noted the absence of key safeguards in the retention of biometrics

30. The concept of private can also include public spaces. See *López Ribalda and Others v. Spain* [GC], nos. 1874/13 and 8567/13, §§ 87-88, 17 October 2019.

31. Point 65 of the decision.

32. see *Gaughran v. the United Kingdom*, no. 45245/15, §§ 69-70, 13 February 2020, where the Court found that the storage of photographs by the police, coupled with a possibility of applying facial recognition techniques to them, constituted an interference with the right to private life.

33. Among the first scholars commenting on this landmark ruling, M. Zalnieriute notes that its significance, however important as the first precedent on this issue, might be weakened by the fact that at the time of the decision, the Russian Federation was not part of the ECHR anymore (the ECtHR confirmed its jurisdiction because the country was still part of the Convention at the time of the facts of the case. See Zalnieriute, Monika. "Glukhin v. Russia. App. No. 11519/20. Judgment." *American Journal of International Law* 117.4 (2023): 695-701. Another criticism, of more general nature, raised by M. Zalnierute concerns the Court's "procedural fetishism", already mentioned in some of her previous articles. The notion has been defined as follows, in an article about the joined cases *Big Brother Watch and Others v. United Kingdom* (*Big Brother Watch v. UK*, Apps. nos. 58170/13, 62322/14 and 24960/15) about the legality of bulk interception regimes pursuant to art. 8 and 10 ECHR: "The focus of ECtHR approach – which I call 'procedural fetishism' – is not on the substantive legality of surveillance regimes, but merely on procedural safeguards, assuming their proportionality, functionality and effectiveness". See Zalnieriute, Monika. "Big Brother Watch v. UK: Procedural Fetishism and Mass Surveillance under the ECHR.", *vervassungsblog.de* (2021).

data from convicted individuals. Indeed, the UK legislation did not include any rule about the gravity of the offence, the necessity of indefinite retention and the opportunities for review of data retention, concluding a breach of art. 8 for lack of proportionality of the domestic measures[34].

The case of Glukhin is a step forward in the direction of better protection of individuals from indiscriminate and unproportionate scraping of biometric data. This jurisprudence should inform future assessment on the legality of State policies allowing the indiscriminate processing of special categories of personal data, as well as any other exploitation of biometric data for the use of technologies that can cause a serious harm to a person's right to personal social life, and, more specifically, to digital private life[35].

In the next paragraph, we will consider an example of mass scraping of personal images, for the training phase of AI FRTs, which the rapid development of technologies makes possible for both public and private entities, certainly increasing the risks to human rights.

3.2 Clearview

The case of Clearview AI (Clearview), a US startup who used web scraping in order to create a database of billions of images to develop a facial recognition tool to be used by police authorities, shed light on the imminent risks for privacy and data protection of individuals and related human rights issues deriving from the indiscriminate and unsupervised scraping by private companies of data relating to biometric features.

Clearview's facial recognition tool is based on four key sequential steps. The first one is represented by the scraping of images of faces and associated data from publicly accessible online sources, including social media, and the storing of that information in its database. Secondly, the company creates biometric identifiers in the form of numerical representations for each image in its database. The system allows users to upload an image, to be assessed by the tool against the biometric identifiers and matched to images in the database. Finally, the system provides a list of results, containing all matching images and metadata. Clicking on any of these results, users are directed to the original source page of the image. The company decided to market its app through licensing to police authorities and to promote it by offering trial accounts to the authorities[36].

Following news articles about Clearview[37], its functioning and business model (which included inter alia the licence of its system not only to public authorities, but also to private businesses, schools and private security agencies) also social media firms reacted, by addressing cease-and-desist letters to Clearview, arguing the scraping of facial images from their sites violated their codes of conduct, privacy policies and copyright laws.

Soon after the launch of Clearview, DPAs and authorities in the sector of privacy and state Courts, especially in America and in Europe, have been dealing with the firm's FRT and its indiscriminate scraping of the web.

Among the first cases is the one brought in 2020 to Court in the state of Illinois, US, by ACLU, which was settled by the parties in 2022[38]. The main provision of the settlement agreement restricts Clearview's practices in Illinois, barring the company from selling access to its database to any entity in Illinois for five years, including state and local police. Besides, under the settlement, Clearview accepted a permanent ban from making its faceprint database available to most businesses and other private entities across the United States[39]. In addition, Clearview agreed to allow Illinois residents to have their facial data blocked from

34. See Amankwaa, Aaron, and Carole McCartney. "Gaughran vs the UK and public acceptability of forensic biometrics retention." *Science & Justice* 60.3 (2020): 204-205 and Tuazon, Oliver M. "Universal forensic DNA databases: acceptable or illegal under the European Court of Human Rights regime?" *Journal of Law and the Biosciences* 8.1 (2021): lsab022.

35. Argren, Rigmor. "Using the European Convention on Human Rights to Shield Citizens from Harmful Datafication." *Kristoffersson, Proceedings from first annual FIRE conference (Uppsala: iustus, 2023)*. 2023.

36. For a detailed analysis on the right to digital private life, as shaped by the jurisprudence of the ECtHR, see Dauvergne, Peter. "The corporate politics of facial recognition." *Identified, Tracked, and Profiled*. Edward Elgar Publishing, 2022. 59-68.

37. See in particular Hill, Kashmir. "The secretive company that might end privacy as we know it." *Ethics of Data and Analytics*. Auerbach Publications, 2022. 170-177. First appeared: New York Times, 18 January 2020, available at: https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

38. Circuit Court of Cook County, Illinois, May 28, 2020, No. 9337839.

39. The settlement is available at: https://www.aclu.org/legal-document/exhibit-2-signed-settlement-agreement?redirect=exhibit-2-signed-settlement-agreement.

Clearview's database[40].

This settlement was possible thanks to the peculiarity of Illinois' legal system in this area, namely by the Biometric Information Privacy Act (BIPA), passed in 2008[41]. The law prohibits private actors from gathering individuals' biometric information, unless the data subject is informed in writing of the kind of data collected and on the specific purpose and time of the processing and, importantly, the person concerned signs a written consent to that processing. BIPA also prohibits the sale of consumers' biometric data and allows them to take to Court companies who violate the terms of the law[42]. Indeed, Illinois can be seen as an exception in the US privacy laws landscape, where consumer privacy legislation has surfaced in recent years, especially in some states such as California, Connecticut, Colorado, Utah and Virginia. For instance, California has enacted the California Consumer Privacy Act (CCPA) and the California Privacy Rights Act (CPRA), which regulate the collection of consumer personal data and their sharing with third parties. However, to this day, there is no comprehensive US privacy federal law, although legislative proposals have been put forward in recent years[43].

Another major difference with the European experience is the lack of statute adequately protecting publicly available personal information, hence the attitude of Clearview's stance towards the cease and desist letters it has received from Google, LinkedIn and other companies to request the stop of scraping of their platforms, claiming that it has a First Amendment right to access publicly available information[44].

In North America, Clearview also entered the Canadian market. Soon thereafter, in February 2021, the Office of the Privacy Commissioner of Canada (OPC) issued the findings of a joint investigation with other national and provincial competent authorities: The Commission d'accès à l'information du Québec (CAI), the Information and Privacy Commissioner for British Columbia (OIPC BC) and the Information and Privacy Commissioner of Alberta (OIPC AB), to verify whether collection, use and disclosure of personal information by means of Clearview's facial recognition tool complied with Canadian federal and provincial privacy laws applicable to the private sector[45]. The aim of the investigation was twofold: on the one hand, verifying if the company obtained consent to collect, use and disclose personal data, as prescribed by the Canadian privacy act and the provincial acts[46], and, on the other, if Clearview collected, used and disclosed such personal data for an appropriate purpose. Besides, the CAI also sought to determine whether the US startup had reported to CAI the creation of a database of biometric characteristics or measurements, pursuant to the law of Quebec.

The decision makes clear that by no means the scraping of publicly available information is exempted from privacy consent requirements[47]. The decision illustrates that the collection of images was carried out for inappropriate purposes, gathering information of people mostly unrelated to any crimes, being furthermore unreasonable and indiscriminate since based on the data found on publicly accessible websites. This implied a clear risk of significant harm to the individuals concerned, totally unaware of the collection and processing of their biometric facial arrays. The intention of the Canadian DPAs was to order or recommend the company to cease offering their facial recognition services to clients in Canada, cease the collection, use and disclosure of images and biometric facial arrays collected from individuals in Canada and delete images and biometrics collected from individuals in Canada. However, before the end of the investigation, Clearview had already withdrawn from the Canadian market.

40. At the time of writing, Illinois residents can request the blocking of their data from Clearview's database using an opt out web form, available at: https://edelson.com/clearviewoptout/.

41. Biometric Information Privacy Act (BIPA), 740 Ill Comp. Stat. 14/1 (2008).

42. Hunt-Blackwell, Sarah. "You Have the Right to Remain Private: Safeguarding Biometric Identifiers in Civil and Criminal Contexts." *Tul. J. Tech. & Intell. Prop.* 24 (2022): 205.

43. See Zhang, W. "Comprehensive Federal Privacy Law Still Pending", NAT'L L. REV. (Jan. 22, 2020). Available at: https://www.natlawreview.com/article/comprehensive-federal-privacy-law-still-pending [perma.cc/N9EH-27MT]; Jordan Yallen, Comment, "Untangling the Privacy Law Web: Why the California Consumer Privacy Act Furthers the Need for Federal Preemptive Legislation", 53 *LOY. L.A. L. REV.* 787 (2020); "US House lawmakers keep federal privacy legislation top of mind", *IAPP*, 1 March 2023.

44. Parks, A.M. "Unfair Collection: reclaiming control of publicly available personal information from data scrapers." *Michigan Law Review*, 120, 5.; Xiao, G. "Bad Bots: Regulating the Scraping of Public Personal Information", 34 *HARV. J.L. & TECH.* 702 (2021).

45. Joint investigation of Clearview AI, Inc. by the Office of the Privacy Commissioner of Canada, the Commission d'accès à l'information du Québec, the Information and Privacy Commissioner for British Columbia, and the Information Privacy Commissioner of Alberta.

46. Canada's Personal Information Protection and Electronic Documents Act (PIPEDA), Quebec's Act Respecting the Protection of Personal Information in the Private Sector (Quebec's Private Sector Act), and Act to Establish a Legal Framework for Information Technology (LCCJTI), British Columbia's Personal Information Protection Act (PIPA BC), and Alberta's Personal Information Protection Act (PIPA AB).

47. Confédération des syndicats nationaux, CAI 1009621-S et 1009629-S, decision by C. Chassigneux, November 12, 2019 [in French]. See also Quebec's Private Sector Act, section 13.

It is interesting to note that the Privacy Commissioner of Canada also underlined concerns on the general level of accuracy of FRTs, in particular with respect to certain demographics. It is indeed a fact noted by several scholars that while false negatives (i.e. the failure to identify an individual whose face is recorded in the reference database are an issue limited to the individual however harsh in terms of privacy compliance and human rights this may be), false positive, (i.e. matching faces that actually belong to two different individuals), presents compelling risks of harm to individuals, particularly when facial recognition is used in the context of law enforcement[48].

Another interesting aspect of the document lies in the importance given to the contractual terms of some of the scraped services, such as Google, Facebook, Twitter, YouTube and LinkedIn, who had sent cease and desist letters to Clearview to stop collecting data in violation to their terms. In Canada, this aspect was considered as further evidence of the unlawfulness of the processing.

Not long after launching operations in North America, the company also tried to enter foreign markets, such as the UK, Australia and EU Member States, like Italy, Sweden and Greece.

The UK and Australian DPA launched a joint investigation, in accordance with the UK Data Protection Act 2018 and the Australian Privacy Act[49]. The Information Commissioner's Office (ICO) issued a fine of more than £ 7.5m, finding that Clearview had violated UK data protection laws by failing to use the information of people in the UK in a fair and transparent manner, given the lack of information provided to individuals, failing to establish a lawful reason for collecting personal data, to put in place a process to limit data retention to a limited time, and also failing to meet the higher data protection standards required for biometric data. Besides, ironically, the ICO decision reports that, when individuals asked information on their possible inclusion in the database, Clearview had them provide further images of them to verify it[50].

Another Clearview case that is worth mentioning is the one by the Italian DPA, the "Garante per la protezione dei dati personali". On 20 February 2022, the Garante imposed a ban on any further collection through web scraping techniques of images and the relevant metadata concerning persons in the Italian territory and on further processing of the standard and biometric data handled by the company via its facial recognition system and regarding persons in the Italian territory. The DPA also ordered the erasure of all the data processed by the facial recognition system with regard to persons in the Italian territory, subject to the obligation to timely reply to such requests for the exercise of the rights under articles 15 to 22 GDPR, as may have been received from data subjects in accordance to art. 12(3) GDPR, ordering the company to designate a representative in the territory of the European Union. In addition, Clearview received a 20 million euros fine[51].

As already seen above, in the EU, the processing of biometric data for the purpose of uniquely identifying a natural person, in principle prohibited, can be carried out under specific conditions and on a limited number of grounds, the main one being for reasons of substantial public interest. Processing must take place on the basis of EU or national law, subject to the requirements of proportionality, respect for the essence of the right to data protection and appropriate safeguards. Since any processing of biometric data for the purpose of uniquely identifying a natural person would relate to an exception to a prohibition laid down in EU law, it would be subject to the Charter, which, however addressed to EU institutions and to the Member States when implementing Union law, (Art. 51(1)) may also have a horizontal effect, thus affecting relations between private parties too[52].

3.3 Final remarks: towards the EU ACT

At the time of writing this contribution, the discussion of the EU AI Act proposal[53] has entered the final stage before its adoption, the phase known as trilogue, the informal tripartite meetings between the three institutions involved in the EU legislative process

48. See e.g. Raposo, Vera Lúcia. "When facial recognition does not 'recognise': erroneous identifications and resulting liabilities." *AI & SOCIETY* (2023): 1-13.

49. The joint investigation was also based on the Global Cross Border Enforcement Cooperation Arrangement and on the MOU between the ICO and the OAIC.

50. ICO, 26 May 2022, Clearview AI Inc. Also the Australian Commissioner issued a decision against Clearview, ordering Clearview AI Inc. to destroy and not collect any more images of individuals in Australia. See Whitfield-Meehan, Sarah. "Privacy: Biometric recognition technology and the'Clearview AI'decision." LSJ: Law Society Journal 86 (2022): 85-87.

51. Garante per la protezione dei dati personali. Ordinanza ingiunzione nei confronti di Clearview AI - 10 febbraio 2022 [9751362].

52. Frantziou, Eleni. "The Horizontal Effect of the Charter: Towards an Understanding of Horizontality as a Structural Constitutional Principle." *Cambridge Yearbook of European Legal Studies* 22 (2020): 208-232.

53. 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. {SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}, Brussels, 21.4.2021 COM(2021) 206 final 2021/0106 (COD), 3.5 Fundamental rights, p. 11.

to reach provisional agreement on legislative files, under the EU ordinary legislative procedure. Indeed, following the draft regulation proposal by the EU Commission in April 2021, the Council has adopted its common position ("general approach"[54]) on the Act on 6 December 2022, where, inter alia, the scope of the legislation is clarified, explicitly excluding national security, defence and military purpose.

In June 2023, the Parliament has agreed on its negotiating position, which includes substantial modifications to the Commission proposal.

As regards the topic of this study, there are several notable novelties, as a particular form of scraping, a topic which was not even mentioned in the draft proposed by the Commission, has been now introduced in Art. 5 of the Act, regarding prohibited artificial intelligence practices. As a consequence, in this version of the text, scraping is also present in some of the new recitals the Parliament has suggested to introduce.

Let us start this brief analysis by illustrating the new version of Art. 5 of the act. Amendment 225 (set to introduce new point d b to art. 5.1 of the draft) by the Parliament introduces among the prohibited AI practices "the placing on the market, putting into service or use of AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage". The reason for this amendment is explained by one of the new recitals suggested by the EU's law making body, illustrating that "the indiscriminate and untargeted scraping of biometric data from social media or CCTV footage to create or expand facial recognition databases add to the feeling of mass surveillance and can lead to gross violations of fundamental rights, including the right to privacy. The use of AI systems with this intended purpose should therefore be prohibited.". This is coherent with the backbone of the AI Act proposal, i.e. the proportional risk-based approach, which categorizes AI activities according to different levels of risk[55]. In practice, the legislative text forbids those AI applications whose related risks appear too high (Art. 5), while establishing safeguards for high-risk activities, posing significant risks to the health and safety or fundamental rights of individuals (Art. 6 et seq.).

The Parliament amended the list of AI systems prohibited in the EU, clearly expressing its intention to ban the use of biometric identification systems, not only for real-time use, as proposed by the Commission, but instead for both real-time and ex-post use. The only exceptions to this suggested ban to the use of biometrics concern cases of severe crime and pre-judicial authorisation for ex-post use. Furthermore, the Parliament would like to ban all biometric categorisation systems using sensitive characteristics (such as gender, race, ethnicity, citizenship status, religion, political orientation), predictive policing systems (based on profiling, location or past criminal behaviour), emotion recognition systems (used in law enforcement, border management, workplace, and educational institutions) and, as already mentioned, systems using indiscriminate scraping of biometric data from social media or CCTV footage to create facial recognition databases.

The reference to the recent Clearview saga seems clear in the Parliament's position. This does not come as a surprise, considering the reaction of the body to the news about the app's facial recognition technology. Soon after the publication of the New York Times article unfolding Clearview practices in the field of facial recognition, Members of the Parliament (MEPs) had addressed written questions to the Commission, to know if and how the system was being used in EU countries and what where the implications for EU rules on privacy, data protection and data processing and EU-US bilateral agreements on privacy[56]. Some MEPs also wrote a letter to the EDPB to express their concerns and ask for guidance. Perhaps even more importantly, the Parliament issued its first official position about the risks entailed by systems such as the one of the US FRT startup only a few months after the Commission's AI proposal. Indeed, in October 2021, The EU law making body presented a resolution on Artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters, where, inter alia, it expressed great concern "over the use of private facial recognition databases by law enforcement actors and intelligence services, such as Clearview AI, a database of more than three billion pictures that have been collected illegally from social networks and other parts of the internet, including from EU citizens", inviting Member States to oblige law enforcement actors to disclose the use of Clearview or equivalent technologies, recalling the opinion of the EDPB, that the use of a service such as that of Clearview by law enforcement authorities in the EU would "likely not be consistent with the EU data protection regime"[57] and calling for a

54. Interinstitutional File: 2021/0106(COD).

55. On the risk-based approach, see e.g., Mahler, Tobias. "Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act Proposal." *Nordic Yearbook of Law and Informatics* (2021).

56. Questions for written answer to the Commission E-000507/2020 and E-000491/2020.

57. As noted by the EDPB, in the letter of 10 June 2020, in reply to MEPs.

ban on the use of private facial recognition databases in law enforcement.[58]

In turn, this resolution followed the Joint opinion on the AI Act by the EDPB and the EDPS, published in spring 2021, where the two bodies called for a "general ban on any use of AI for an automated recognition of human features in publicly accessible spaces", and for a ban on categorization, emotion analysis and scraping the internet to create facial image databases[59]. The EDPB later reaffirmed the need for a ban of use cases of facial recognition technologies, which pose unacceptably high risks to individuals and society ("red lines"), considering that processing of personal data in a law enforcement context relying on a database "populated by collection of personal data on a mass scale and in an indiscriminate way, e.g. by"scraping" photographs and facial pictures accessible online, in particular those made available via social networks, would, as such, not meet the strict necessity requirement provided for by Union law."[60]. Strict necessity is an aspect of the broader principle of proportionality[61].

As seen above, the creation of facial recognition databases through the untargeted scraping of facial images can lead to the violation of fundamental rights protected by the Charter. In this sense, the amendments proposed by the European Parliament are coherent with the purported goal of the AI Act, as exposed in the Commission proposal, i.e. guaranteeing a high level of protection for fundamental rights and addressing risks through a risk-based approach[62], enhancing and promoting the protection of the fundamental rights protected by the Charter, with specific regard to the right to human dignity (Art. 1), respect for private life and protection of personal data (Articles 7 and 8), nondiscrimination (Art. 21) and equality between women and men (Art. 23).

In view of the above, the position expressed by the European Parliament seems in line with the legislation applicable to the processing of biometric data and with the position of the EDPB. As concerns the topic of this contribution, it seems that it would be reasonable to clearly exclude indiscriminate forms of scraping of biometric data. This would be in line with the established legal and human rights principles. Moreover, a clear exclusion would be wise, considering the maturity reached by FRTs, which could allow anyone to scrape the internet for facial images and build up a database for facial recognition, with clear (though not yet fully explored) risks for human rights.

4 Bibliography

A. Amankwaa, and C. McCartney, *Gaughran vs the UK and public acceptability of forensic biometrics retention,* in "Science & Justice" 60.3 (2020): 204-205.

R. Argren, *Using the European Convention on Human Rights to Shield Citizens from Harmful Datafication*, in "Kristoffersson, Proceedings from first annual FIRE conference" (Uppsala: iustus, 2023). 2023.

T. Christakis, *Mapping the use of facial recognition in public spaces in Europe*, *Part 1* in "Report of the AI-Regulation chair" (AI-Regulation.Com) COM), MIAI, 2022.

P. Dauvergne, *The corporate politics of facial recognition*, in *Identified, Tracked and Profiled*, Elgar, 2022.

M. Durante, *Computational power: the impact of ICT on law, society and knowledge*, Routledge, 2021.

M. Durante, and L. Floridi, *A legal principles-based framework for AI liability regulation,* in *The 2021 yearbook of the digital ethics lab.* Cham: Springer International Publishing, 2022. 93-112.

E. Frantziou, *The Horizontal Effect of the Charter,* in *Cambridge Yearbook of European Legal Studies*, Vol 22, 2020.

58. Artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters. European Parliament resolution of 6 October 2021 on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters, (2020/2016(INI)).

59. EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 18 June 2021.

60. EDPB Guideline 05/2022 on the use of facial recognition technology in the area of law enforcement, 12 May 2022, pqrr. 103-104.

61. The content of the strict necessity test has been clarified by the CJEU in the *Schwartz* case (C-291/12 *Schwarz v Stadt Bochum* ECLI: EU: C:2013:670, para 46), dealing with the processing of fingerprints data: "*in assessing whether such processing is necessary, the legislature is obliged, inter alia, to examine whether it is possible to envisage measures which will interfere less with rights recognised by Articles 7 and 8 of the Charter but will still contribute effectively to the objectives of the European Union rules in question*".

62. The risk approach of the AI Act follows the direction already undertaken by the GDPR, which presents data controller accountability as a "meta-principle", see: Paseri, Ludovica, Sébastien Varrette, and Pascal Bouvry. "Protection of Personal Data in High Performance Computing Platform for Scientific Research Purposes." *Annual Privacy Forum*. Cham: Springer International Publishing, 2021.

S. Hunt-Blackwell, *You Have the Right to Remain Private: Safeguarding Biometric Identifiers in Civil and Criminal Contexts*, in "Tul. J. Tech. & Intell. Prop." 24 (2022): 205.

T. Mahler, *Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act Proposal,* in *Nordic Yearbook of Law and Informatics*, 2022.

U. Pagallo, *The legal challenges of big data: putting secondary rules first in the field of EU data protection*, in "Eur. Data Prot. L. Rev." 3 (2017): 36.

A. M. Parks, *Unfair Collection: reclaiming control of publicly available personal information from data scrapers,* in "Michigan Law Review", 120, 5.

L. Paseri, S. Varrette, and P. Bouvry, *Protection of Personal Data in High Performance Computing Platform for Scientific Research Purposes*, *Annual Privacy Forum*. Cham: Springer International Publishing, 2021.

O. M. Tuazon, *Universal forensic DNA databases: acceptable or illegal under the European Court of Human Rights regime?* In "Journal of Law and the Biosciences" 8.1 (2021): lsab022.

J. Purshouse; L. Campbell, *Automated facial recognition and policing: A Bridge too far?* Cambridge University Press, 27 August 2021.

V. L. Raposo, *When facial recognition does not 'recognise': erroneous identifications and resulting liabilities*, in "AI & Soc", 2023.

M. Veale, R. Binns, and J. Ausloos, *When data protection by design and data subject rights clash*, in "International Data Privacy Law" 8.2 (2018): 105-123.

S. Whitfield-Meehan, *Privacy: Biometric recognition technology and the 'Clearview AI' decision*, in "LSJ: Law Society Journal" 86 (2022): 85-87.

G. Xiao, *Bad Bots: Regulating the Scraping of Public Personal Information*, 34 in "HARV. J.L. & TECH." 702, 2021.

J. Yallen, *Untangling the Privacy Law Web: Why the California Consumer Privacy Act Furthers the Need for Federal Preemptive Legislation*, 53 in "LOY. L.A. L. REV." 787, 2020.

M. Zalnieriute, *Big Brother Watch v. UK: Procedural Fetishism and Mass Surveillance under the ECHR*, in "vervassungsblog.de" (2021).

M. Zalnieriute, *Glukhin v. Russia. App. No. 11519/20. Judgment,* in "American Journal of International Law" 117.4 (2023): 695-701.

W. Zhang, *Comprehensive Federal Privacy Law Still Pending*, in "NAT'L L. REV.", 2020.

B. Zhao, *Web scraping*, in L. A. Schintler, C. L. McNeely (Eds.), *Encyclopedia of Big Data*, Springer, 2017.