https://doi.org/10.6092/issn.1825-1927/18905 ISSN 1825-1927

The normative challenges of data scraping: legal hurdles and steps forward

Jacopo Ciani Sciolla*

Abstract:

Contemporary society is characterized by the availability of immense amounts of data publicly available on the internet, both in traditional websites and social media accounts. Web scraping, a means for automatically extracting publicly available data through technical tools, is commonly used by businesses, researchers, law enforcement authorities as well as criminals to gather such data and extract value from it in several different ways. In this scenario, it is fundamental to correctly define the principles and laws applicable to publicly available data and to scraping itself, in order to define a clear boundary between legal and illegal uses. The versatility of scraping and the myriad of possible use cases reflects on the fragmentation of applicable legal instruments in different jurisdictions. This complexity requires jurists to adopt a holistic interdisciplinary approach and evaluate the implications of scraping under different legal domains like intellectual property, data protection, private law, criminal law and competition law.

Keywords:

Web scraping, internet, data ownership, public availability, EU, US, databases, copyright, privacy, data protection, artificial intelligence, human rights

1 The Data Rush

Because of its capacity to reconfigure relationships between states, subjects, and citizens, Data has become a social and political issue taking centre stage in the global policy agenda[1].

Among the many legal challenges surrounding the Big Data revolution[2], «a pivotal factor» has been considered the regulation of data ownership. Indeed, having the ability to control the circulation on the market of a more or less vast amount of data plays «a fundamental role in sustaining and developing the emergence of a European data-driven economy»"[3].

In some previous works[4], I have tried to explore to what extent data could be subject to ownership and what possibilities exist for protecting them against use by third parties. At that time, against a general scepticism about introducing new property rights on data, I warned about the risk that if property rights would have been not assigned by law, *de facto* ownership would have been allocated by the market, with a great window of opportunity for more powerful actors.

*University of Turin, Department of Law; ✉ jacopo.cianisciolla@unito.it

1. Pagallo, Ugo; Durante, Massimo. "La politica dei dati. Il governo delle nuove tecnologie tra diritto, economia e società", Mimesis, Milano, 2022 and Pagallo, Ugo. "The Politics of Data in EU Law: Will It Succeed?", Digital Society 1.3 (2022): 20.

2. A focus on these challenges is provided by Pagallo, Ugo. "The Legal Challenges of Big Data: Putting Secondary Rules First in the Field of EU Data Protection", European Data Protection Law Review 2017, 3, 1, 36-46.

3. IDC, OPEN EVIDENCE, European data market Final Report, SMART 2013/0063, 2017, available at https://ec.europa.eu/digital-single-market/en/news/final-results-european- data-market-study-measuring-size-and-trends-eu-data-economy.

4. Ciani Sciolla, Jacopo. "Property rights model v. contractual approach: how protecting non - personal data in cyberspace?", Dir. comm. Intern., 2017, 831-854 and Id., "Governing Data Trade in Intelligent Environments: A Taxonomy of Possible Regulatory Regimes Between Property and Access Rights", in *I.* Chatzigiannakis, Y. Tobe, P. Novais, O. Amft (Eds.), Intelligent Environments 2018, Workshop Proceedings of the 14th International Conference on Intelligent Environments, IOS Press, 285-297.

From that time, we had significant evidence of severe market abuses.

Cambridge Analytica made netizens concerned regarding the gathering of their online data[5]. At that time, however, there was little knowledge of how big the big-data industry actually was, and the idea was that such a scandal was possible only due to Facebook's vulnerable and leaky platform.

From the Cambridge Analytica scandal on, we have seen increased reports of mass data leaks[6], with commercial entities using software to collect substantial amounts of information from third parties' web sites in a perfect legal way or, in any case, without breaching security measures or otherwise accessing a computer without authorization.

Scraping the web, starting from accounts of the biggest and most important social media (see the Twitter example below), has become common for most actors, from researchers to law enforcement authorities, from amateurs to professionals, from philanthropists to outlaws.

Professionals are even running 'best proxy provider research tests' to determine who gives a web scraper the best access to public data online[7].

Today, web scraping is also at the core of most trainings of AI systems, particularly as regards large language models[8].

Taking a totally different stance, scraping is used for the investigation and prevention of terrorism and serious crimes around the world, e.g. for mapping, in the context of open-source intelligence techniques (OSINT)[9], online propaganda campaigns of terrorist organisations on the surface web[10].

As a peculiar demonstration of the popularity of scraping, one might recall the provocative manifesto of "Scrapism", by the artist and researcher Sam Lavigne:

> "Scrapism, as I define it, is a counterpractice of web scraping for artistic, emotional, political, and critical ends, rather than for those of business or government. It is a process of decommodification and redatabasing, a process of eliminating artificial scarcity. At heart, it is a practice that challenges the regime of private property, with a particular focus on the ways that private property, as expressed on the web, produces and reproduces informational and material power asymmetries[11]".

This helps to understand how, in itself, the act of scraping can be seen as neutral, being based on technologies which can serve potentially any goal, both good and bad.

Web scraping is so lucrative and advantageous that it has actually spawned a industry peddling "scraping-as-a-service." One of the key findings from the new Oxylabs white paper, Alternative Data Defines Competition in the US & UK E-commerce Sectors[12], demonstrates how over 82% of e-commerce organisations are now using web scraping to gather external data to help guide their decision-making.

Instead, from the different stance of the owner of the system that is scraped, this activity will often be seen as unwanted, for various reasons, e.g. bandwidth and server overload, or loss of advertisement revenues and/or loss of control of the information content. According to 2022 State of Bot Mitigation report[13], 54% of companies lost 6% of revenue due to scraping.

5. European Parliament resolution of 25 October 2018 on the use of Facebook users' data by Cambridge Analytica and the impact on data protection (2018/2855(RSP)).

6. https://www.statista.com/statistics/290525/cyber-crime-biggest-online-data-breaches-worldwide/.

7. https://scrapingrobot.com/blog/proxy-provider/.

8. Sirisuriya, SCM De S. "Importance of Web Scraping as a Data Source for Machine Learning Algorithms-Review." 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS). IEEE, 2023.

9. See Evangelista, João Rafael Gonçalves, et al. "Systematic literature review to investigate the application of open source intelligence (OSINT) with artificial intelligence." Journal of Applied Security Research 16.3 (2021): 345-369, where OSINT is defined as "a concept that addresses the search, collection, processing, analysis, and use of information from open sources that can be legally accessed by any individual or organization".

10. Lakomy, Miron. "The virtual" Caliphate" strikes back? Mapping the Islamic State's information ecosystem on the surface web." Security Journal 36.4 (2023): 791-811.

11. Lavigne, Sam. "Scrapism: A Manifesto." Critical AI 1.1-2 (2023).

12. Oxylabs surveyed 500 UK-based and 501 US-based senior data decision-makers from UK and US ecommerce companies. See https://public-files.oxylabs.io/blog/pdf/Alternative-Data-Defines-Competition-in-the-US-and-UK-Ecommerce-Sectors.pdf

13. Kasada commissioned Atomik Research an independent market research agency to conduct a survey of 202 U.S. security and IT professionals responsible for

This clash between competing interests is at the core of controversies between website developers and scrapers.

From an ethical point of view, there is no consensus on the topic and only a few academic articles are explicitly devoted to ethical issues surrounding this practice[14]. This "grey area" is mirrored by the current legal framework as well, which can be described as particularly fragmented, given that most of the countries around the globe do not have a legislation addressing web scraping practices specifically. Rather, there exist a number of laws that can be applicable to the matter, depending on the legal system and most importantly on the circumstances of the case[15].

2 A multilayered and fragmented legal framework

Against the above mentioned framework, data scraping needs to be examined from a multi-disciplinary perspective that includes the following main legal domains of investigation.

a) The first entails to what extent scraped data are subject to third parties rights which may be enforced in order to block the unauthorised extraction. In this regard, private law (in particular tort law), criminal law, intellectual property (namely copyright, database rights) are particularly relevant.

EU Member States normally resort to copyright protection or to the *sui generis* right for the creators of databases which do not qualify for copyright, based on the EU Database Directive[16]. The *rationale* of this unique *quasi*-IP regime has been criticised by scholars[17] and has not always proven coherent and effective in view of an adequate and proportionate protection of the competing interests at stake[18]. However Article 7 of the Database Directive still provides for the closest remedy for the case in which the owner of the database can prove an extraction and/or re-utilization of a substantial part of a database.

Non-EU-countries normally tackle the issue from totally different angles[19], e.g. applying criminal law or unfair competition rules[20]. With such different approaches, the degree of protection for proprietary content on commercial web sites is not settled and may vary significantly from country to country.

On this layer of analysis, the very nature of the data subject to extraction may trigger the application of other relevant legal domains, such as data protection and trade secrecy law.

Data protection authorities in Europe and other regions are particularly active. A very recent example lies in the Italian DPA's public announcement of an investigation into current practices of scraping of personal data on the web as a means for training AI algorithms. The purported aim of this investigation is that of verifying the adoption of security measures by website owners, in a way which should adequately prevent the massive scraping of personal data by third parties[21].

---

mitigating bots. All organizations surveyed currently have anti-bot solutions in place. Fieldwork took place between August 18th and August 29th of 2022. The report is available at the following link https://get.kasada.io/hubfs/Reports%20and%20Datasheets/2022%20State%20of%20Bot%20Mitigation%20-%20Final%20Report.pdf.

14. See e.g. Gold, Zachary, and Mark Latonero. "Robots welcome: Ethical and legal considerations for web crawling and scraping." Wash. JL Tech. & Arts 13 (2017): 275. Albeit this scarcity of literature about the ethics of scraping, lessons may be learned from works building on the ethics of copying. See e.g. Hick, D. H.; Schmücker, R. (Eds.). The Aesthetics and Ethics of Copying. London: Bloomsbury Publishing, 2016; Pagallo, U. The Troubles with Digital Copies: A Short KM Phenomenology. In: Digital Rights Management: Concepts, Methodologies, Tools, and Applications; Hershey: IGI Global, 2013: 1379-394.

15. Ciani Sciolla, Jacopo and Pagallo. Ugo. "Anatomy of web data scraping: ethics, standards, and the troubles of the law" forthcoming.

16. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

17. Derclaye, Estelle and Husovec, Martin, "Sui Generis Database Protection 2.0: Judicial and Legislative Reforms", EIPR 44 (2021): 323.

18. A clear example of the contradictions of this regime and its interpretation by the EUCJ is represented by the ruling in Innoweb (Judgment of 19 December 2013, C-202/12, ECLI:EU:C:2013:850), where the Court The Court defined the actions of metasearch engines as "nearly parasitic", and later national rulings which did not follow the principles set forth by the top EU court. For instance, years later, the Italian Supreme Court (Cass. Civ. 18 December 2018 n 2290 and 2289, Ryanair v. Viaggiare), the Italian Supreme Court stated that scraping online databases is not automatically illegal, safe for possible infringements of IP rights.

19. Ciani Sciolla, Jacopo and Pagallo. cit.

20. In the U.S., the Computer Fraud and Abuse Act ("CCFAA"), which punishes "whoever ... intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains ... information".

21. Document no. 9952078, 22 November 2023.

Trade secrecy law has proven, especially in the US, to be a valuable alternative for protecting a database, which total amount of quotes, differently from those specifically searched by the users and showed as a result of consultation, are not public available[22].

Even if no rights are granted to the database maker, this may seek some contractual protection, through the terms of service governing the interaction between the platform and the users. In this regard, private law is again relevant to our purposes, to the extent to which it establishes the grounds of validity and effectiveness of websites' terms and conditions towards its human and non-human visitors[23].

Another area of legal interest focuses on the scraping activities that do not encroach on the data holder's rights, because of exceptions or limitations to such rights or due the nature of the data, urging to keep them as open as possible. The former is the case of copyright law that, upon certain stringent conditions, allows text and data mining. It is also the case of the Digital Services Act[24], whose article 40(12) provides researchers with a legal defence to internet service providers' attempts to shut down access to publicly available data[25]. The latter refers to the so called open data, which, according to the public sector information directive, as a rule should not be restricted in their public availability.

Another limitation to the data holder's rights may arise from antitrust law, which may prevent from using such rights as a tool for restricting competition and access to the market[26].

No single special issue can at present compound and account for all the envisaged and listed issues and questions, precisely because of the all-pervading dimension and in-built functioning of the data scraping operations and applications in our information societies. In this perspective, the aim of this special issue is to raise at least some critical and challenging issues concerning the main fields of investigation.

Let us briefly introduce them.

3 Critical and challenges issues

The articles included in this special issue regard different facets of the matter under discussion, looking at the legal infrastructure applicable and to the existing case-law under two main different perspectives.

a) the first concerns the legal boundaries to text and data mining and the *discrimen* between lawful and unlawful data scraping activities;

b) the latter pertains the use of the powerful tools of scraping for law enforcement purposes and the somewhat difficult relationship between them.

Under the first domain, the issue collects papers analysing the implications of web scraping both from a private and criminal law perspective.

The private law perspective is provided by "*Web Scraping: A Private Law Perspective*", authored by Alessandra Quarta and Michael W. Monterossi. The paper points out how current case law is mainly grounded on the distinction between restricted and non-restricted websites as key factor for assessing whether data scraping falls within the fair competition rules. The authors evoke the famous Ryanair cases, where the damages claimed by the airline were linked to the sales drop of travel packages, complementary services or advertising revenues, allegedly caused by the diversion of online traffic in favour of the scraping online travel agencies. Quarta and Monterossi question this case law based on the evaluation of websites' technical protective

22. Geoffrey Xiao, "Data Misappropriation: A Trade Secret Cause of Action for Data Scraping and a New Paradigm for Database Protection", 24 Colum. Sci. & Tech. L. Rev. 125 (2022).

23. Maurizio Borghi and Stavroula Karapapa, "Contractual restrictions on lawful use of information: sole-source databases protected by the back door?", E.I.P.R. 2015, 37(8), 505-514.

24. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance).

25. This provision has been called the 'CrowdTangle Provision', and aims to facilitate access to already public data by protecting data scraping and, potentially, requiring platforms to facilitate APIs and other forms of access to this data. See Paddy, Leerssen. "Platform research access in Article 31 of the Digital Services Act: Sword without a shield?" Available at https://verfassungsblog.de/power-dsa-dma-14/ (2021). See also Husovec, Martin. "How to Facilitate Data Access under the Digital Services Act." Available at SSRN 4452940 (2023).

26. Hans Graux. "Sharing Data (Anti-)Competitively. Will European data holders need to change their ways under the proposed new data legislation?" Luxembourg: Publications Office of the European Union, 2022.

measures (TPMs), arguing that their circumvention should not be held enough to infringe unfair competition rules. Instead, Courts should attribute a higher value to the reuse of data, rather than their mere extraction, with a view to evaluate deceptive and parasitic behaviours as well as the related competitive harm suffered by the scraped website. Another valuable proposition in this contribution lies in the idea that the mere reliance on websites' protection measures is at odds with the EU current policy on non-personal data[27], which aims at untangling and fostering their circulation and reuse. The same is valid for the EU Data Act[28], whose provisions, *inter alia*, will force companies to share raw, non-personal data on their IoT products with other economic actors.

"*Web scraping and AI training in the Directive 790/19*" by Chiara Gallese explores the text and data mining exception in the Copyright in the Digital Single Market Directive 2019/770, ("CDSM Directive")[29] and its availability as a safe harbour for scraping data to train AI generative models. Based on the ambiguous wording of Article 4 CDSM, the author proposes an interesting and provoking opinion according to which generative models training would not be included in the definition of text and data mining adopted by the EU legislator. Consequently, in order to train generative models for commercial purposes, companies need to seek permission from each copyright holder before scraping the web or reusing materials that were originally scraped for research, teaching, or cultural purposes.

The criminal law perspective is provided by Rosa Maria Vadalà's contribution on "*Criminal law for data scraping. The compatibility with the value of data analytics in the modern age"*. The contribution examines existing criminal offences under Italian criminal law related to acts of unlawful scraping. The *fil rouge* is the principle of *extrema ratio* as a ground rule of Italian criminal law, *i.e.* the idea that in a liberal democracy the intervention of criminal law should be seen only as a last resort. Thus, the *extrema ratio* principle is presented as a filter to avoid the risk of over criminalization of scraping activities as well as the duplication of civil and criminal sanctions.

The second research area addressed by this issue, concerning the scraping of data by law enforcement authorities and its impact on fundamental rights, is explored by the following articles.

"*Data collection via web scraping: privacy and facial recognition after Clearview*" by Fabrizio Lala reviews privacy and data protection laws and principles applicable to the scraping of biometric collected by facial recognition technologies ("FRTs") under the lens of two major judicial cases. First, the ECtHR judgement in Glukhin v. Russia, ruling that the right to private life (art. 8 ECHR) does not protect only the "inner circle" in which the individual live his or her own personal life without outside interference, but it also encompasses the right to lead a "private social life"[30]. Second, the Clearview AI case, gathering billions of pictures of people posted online for building a database to be exploited by law enforcement authorities[31], which led to several decisions by national DPAs as well as a EU Parliament resolution on the use of AI by the police and judicial authorities in criminal matters[32]. The author reflects on the echoes of these cases in some of the Parliament proposed amendments to the EU AI Act proposal[33], seeking to forbid the introduction in the EU of "AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage"[34].

Flavia Giglio's contribution "*Moderation of illegal content and social media scraping. Privacy and data protection constraints in the processing of publicly available data by law enforcement authorities*" investigates the use of current scraping possibilities in social medias by law enforcement authorities as a form of Open Source Intelligence ("OSINT") to the purpose of identifying and removing illicit content online. The article includes a careful evaluation of the impact of scraping and retention of data by LEAs on the exercise of fundamental rights, with a focus on the right to freedom of expression. The author stresses the difficulties

27. See e.g. European Commission Communication on A European strategy for data, Brussels, 19.2.2020, COM(2020) 66 final.

28. Proposal for a Regulation on harmonised rules on fair access to and use of data (Data Act), Brussels, 23.2.2022, 68 final, 2022/0047 (COD).

29. Directive (EU) 2019/770 of the European Parliament and of the Council of 20 May 2019 on certain aspects concerning contracts for the supply of digital content and digital services.

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. 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.

32. 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)).

33. 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

34. Text of the proposal adopted by the EU Parliament on 14 June 2023, Amendment 225, establishing new art. 5, par. 1, point d b (new).

to strike a right balance between the protection of publicly available data collected from social media contents and the need to prevent or fight politically motivated crimes and concludes that a further reflection is needed to avoid that the legitimate objective to tackle illegal or harmful content online would degenerate in practices of mass surveillance.

Elisabetta Stringhi's paper, "*Hallucinating (or poorly fed) LLMs?*", delves into one of the major problems concerning the reliability of scraping tools for public interest purposes: the accuracy of data. As the author argues, scraping is all but a selective tool. Hence, the data it allows to gather include all sorts of information, including outdated data. As Stringhi explains, the current situation is alarming, in view of the scale and magnitude of LLMs' output generation. To overcome or limit the underlying risks, the author suggests, *inter alia*, that future research should aim to avoid inaccuracies by design, through the implementation of measures to timely detect design errors, also envisaging the use of synthetic data as a means for reducing potential harms deriving from inaccurate datasets.

4 Future research lines

Some aspects of these contributions are interlinked and show a common research direction. While all the papers agree on the central role of data scraping and the importance of its regulation for the future of our society, they basically focus on the following research question: is data scraping lawful based on existing legislation? which loopholes may be detected in the normative framework? Which gaps should be addressed in order to allow a safe use of scraping technologies for public interest's purposes, such as security, safeguarding individuals fundamental rights.

In this analysis, they bring to light that two normative macro regions exist (EU vs US), which tackle the scraping phenomenon building upon different values and ground norms[35]. The gap concerns the understanding of the notion of "public availability of data" and its possible (legal) appropriability by scrapers. In the EU, public availability does not imply that any appropriation is legal. On the contrary, in the US, the public availability of data normally does not preclude its appropriability, based on the strong role played by the First Amendment.

The research line mentioned above is valuable in that it, commendably, pushes towards *ad hoc* initiatives at the international level with a view to narrow the existing gaps which still separates the macro regions. However, the editor aims to take advantage of these few lines for suggesting a prospective agenda for future research in this field.

To this purpose, I shall take the cue from the recommendations recently issues by some Data protection authorities, inviting social media platforms, other websites and individuals to adopt measures against unlawful data scraping[36].

Scraping has become a hot topic in 2023, even in mainstream media, with the likes of Twitter and Redditt publicly taking stance and issuing specific measures to prevent, fight or limit the possible negative effects of scrapers[37]. The recent lawsuit that X Corp., formerly Twitter, filed against a non-profit seeks tens of millions of dollars in damages based on the reputational harm allegedly caused by the publication of a report carried out by scraping public information on Twitter and arguing that Twitter allowed hate speech and misinformation contents to remain up on the platform, even if they violated Twitter's own policies[38].

In my view, this recommendation may lead in the long run to a rush towards finding new ways to fight back against AI companies that use public available data to train their models without the data holder's permission. This even if they engage in clearly public-interest-focused research or investigations.

Indeed, even if, in compliance with the opt-out clause in the Digital Single Market Directive (art. 4), many AI companies allow copyright holders to opt out of having their works used to train AI models, a growing number of stakeholders raised concerns over the effectiveness of this solution. First, this puts the onus on content creators to actively protect their IP, rather than requiring

35. See e.g. Petkova, Bilyana. "Privacy as Europe's first Amendment." European Law Journal 25.2 (2019): 140-154 and Xiao, Geoffrey. "Bad bots: regulating the scraping of public personal information." Harv. JL & Tech. 34 (2020): 701.

36. Joint statement on data scraping and the protection of privacy released by the office of the Australian Information Commissioner (OAIC) and 11 of its international data protection and privacy counterparts, members of the GPA's International Enforcement Cooperation Working Group (IEWG). See https://www.oaic.gov.au/newsroom/global-expectations-of-social-media-platforms-and-other-sites-to-safeguard-against-unlawful-data-scraping.

37. Frenkel, Sheera, and Stuart A. Thompson. "'Not for Machines to Harvest': Data Revolts Break Out Against AI." International New York Times, July 16, 2023.

38. Sheera Frenkel and Ryan Mac, Twitter Sues Nonprofit That Tracks Hate Speech, The New York Times, July 31, 2023.

the AI developers to secure the IP to the work prior to using it. Then, without clear guidelines for standardized machine readable reservations, the opt-out mechanism is unlikely to work in practice[39].

Against this framework, movements are growing with the aim to help tip the power balance back from AI companies towards data holders, especially artists, by creating powerful deterrent against disrespecting privacy, copyright or intellectual property of data holders[40].

For instance, Glaze allows artists to "mask" their own personal style to prevent it from being scraped by AI companies. By changing the pixels of images in ways invisible to the human eye, the tool manipulates machine-learning models to interpret the image as something different from what it actually shows[41].

Another tool, called Nightshade, lets artists add invisible changes to the pixels in their art before they upload it online so that if it's scraped into an AI training set, it can cause the resulting model to break in chaotic and unpredictable ways. Using it to "poison" training data could damage future iterations of image-generating AI models, such as DALL-E, Midjourney, and Stable Diffusion, by rendering some of their outputs useless[42].

Both these techniques exploit the security vulnerability in generative AI models[43], arising from the fact that they are trained on vast amounts of data that have been hoovered from the internet.

The more people use these features, the more biased data can be scraped into the model, the more damage the techniques will cause.

At the same time, AI companies such as OpenAI, Meta, Google, and Stability AI are facing a slew of lawsuits from artists who claim that their copyrighted material and personal information was scraped without consent or compensation[44].

This trend may generate in the long run a shift from open to more closed and "entry limited" digital systems, as data scraping becomes more accessible and threatens traditional business models.

These forecasts urge as well to move our research lines from what can AI companies do with publicly available data to what data holders are entitled to do in order to lawfully prevent their data from being scraped. Is the adoption of technological protection measures (TPMs) freely available to prevent or dissuading AI from scraping protected items? At this stage, only copyright law takes care of regulating such aspect[45] and this means that the protection of TPMs does not apply to non-copyrightable items.

At the same time, another research line should focus on AI model security and robustness. User's trust on AI model and the consequent commercial success strongly depends on whether robust defences against "poisoning attacks" shall be available.

5 Bibliography

M. Borghi and S. Karapapa, *Contractual restrictions on lawful use of information: sole-source databases protected by the back door?*, EIPR 2015, 37(8), 505-514.

J. Ciani Sciolla, *Property rights model v. contractual approach: how protecting non - personal data in cyberspace?*, Dir. comm. Intern., 2017, 831-854

39. See Keller, Paul and Warso, Zusanna. "Defining best practices for opting out of ML training", Open Future policy Brief no. 5, 29 September 2023, https://openfuture.eu/wp-content/uploads/2023/09/Best-_practices_for_optout_ML_training.pdf; Nielbock, Leander. "Defining best practices for opting out of ML training – time o act." September 29, 2023, https://communia-association.org/2023/09/29/defining-best-practices-for-opting-out-of-ml-training-time-to-act/.

40. Heikkilä, Melissa. "This new data poisoning tool lets artists fight back against generative AI." MIT Technology Review, October 23, 2023, https://www-technologyreview-com.cdn.ampproject.org/c/s/www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/amp/.

41. See https://glaze.cs.uchicago.edu/.

42. Masterson, Victoria. "What is Nightshade – the new tool allowing artists to 'poison' AI models?." World Economic Forum, November 14, 2023, https://www.weforum.org/agenda/2023/11/nightshade-generative-ai-poison/.

43. David, George et al. "Safeguarding Generative Artificial Intelligence (AI) with cybersecurity measures. Risk insights and building blocks for secure Generative AI solutions." Deloitte, September 2023, https://www2.deloitte.com/content/dam/Deloitte/in/Documents/risk/in-ra-safeguarding-generative-artificial-intelligence-noexp.pdf.

44. De Vynck, Gerrit. "AI learned from their work. Now they want compensation." The Washington Post, July 16, 2023, https://www.washingtonpost.com/technology/2023/07/16/ai-programs-training-lawsuits-fair-use/.

45. Council Directive 2001/29/EC, Art. 6.

J. Ciani Sciolla, *Governing Data Trade in Intelligent Environments: A Taxonomy of Possible Regulatory Regimes Between Property and Access Rights*, in I. Chatzigiannakis, Y. Tobe, P. Novais, O. Amft (Eds.), Intelligent Environments 2018, Workshop Proceedings of the 14th International Conference on Intelligent Environments, IOS Press, 285-297.

E. Derclaye, and M. Husovec. *Sui generis database protection 2.0: judicial and legislative reforms,* EIPR 44 (2021): 323.

J. R. G. Evangelista, et al., *Systematic literature review to investigate the application of open source intelligence (OSINT) with artificial intelligence,* Journal of Applied Security Research 16.3 (2021): 345-369

Z. Gold, and M. Latonero, *Robots welcome: Ethical and legal considerations for web crawling and scraping,* Wash. JL Tech. & Arts 13 (2017): 275.

H. Graux, *Sharing Data (Anti-)Competitively. Will European data holders need to change their ways under the proposed new data legislation?*, Luxembourg: Publications Office of the European Union, 2022.

D. H. Hick and R. Schmücker, (Eds.), *The Aesthetics and Ethics of Copying*, London: Bloomsbury Publishing, 2016.

K. Hill, *The secretive company that might end privacy as we know it,* Ethics of Data and Analytics. Auerbach Publications, 2022. 170-177.

M. Husovec, *How to Facilitate Data Access under the Digital Services Act,* Available at SSRN 4452940 (2023).

IDC, OPEN EVIDENCE, *European data market Final Report*, SMART 2013/0063, 2017.

M. Lakomy, *The virtual" Caliphate" strikes back? Mapping the Islamic State's information ecosystem on the surface web,* Security Journal 36.4 (2023): 791-811.

S. Lavigne, *Scrapism: A Manifesto*, Critical AI 1.1-2 (2023).

SCM. Sirisuriya, De S. *Importance of Web Scraping as a Data Source for Machine Learning Algorithms-Review*, 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS). IEEE, 2023.

U. Pagallo, *Algo-rhythms and the beat of the legal drum,* Philosophy & Technology 31.4 (2018): 507-524.

U. Pagallo, *The Legal Challenges of Big Data: Putting Secondary Rules First in the Field of EU Data Protection*, European Data Protection Law Review 3.1 (2017): 36-46.

U. Pagallo, and J. Ciani Sciolla, *Anatomy of web data scraping: ethics, standards, and the troubles of the law*, forthcoming.

U. Pagallo, *The Politics of Data in EU Law: Will It Succeed?* Digital Society 1.3 (2022): 20.

U. Pagallo, *The Troubles with Digital Copies: A Short KM Phenomenology*, In: Digital Rights Management: Concepts, Methodologies, Tools, and Applications; Hershey: IGI Global, 2013: 1379-394.

U. Pagallo and M. Durante (Eds.), *La politica dei dati. Il governo delle nuove tecnologie tra diritto, economia e società*, Mimesis, Milano, 2022.

B. Petkova, *Privacy as Europe's first Amendment*, European Law Journal 25.2 (2019): 140-154.

G. Xiao, *Bad Bots: Regulating the Scraping of Public Personal Information*, 34 Harv. J. L. & Tech. 702, 2021.

G. Xiao, *Data Misappropriation: A Trade Secret Cause of Action for Data Scraping and a New Paradigm for Database Protection*, 24 Colum. Sci. & Tech. L. Rev. 125, 2022.