Nadia Maccabiani
University of Brescia, Italy
nadia.maccabiani@unibs.it
ORCID: 0000-0003-1183-0854

Anna Podolska
University of Gdańsk, Poland
anna.podolska@ug.edu.pl
ORCID: 0000-0002-5380-9570

Ewelina Szatkowska
University of Gdańsk, Poland
ewelina.szatkowska@prawo.ug.edu.pl
ORCID: 0000-0001-6449-4204

https://doi.org/10.26881/gsp.2024.4.03

How Artificial Intelligence Learns: Legal Aspects of Using Data in Machine Learning

Introduction

"AI is a collection of technologies that combine data, algorithms and computing power."[1] A subset of artificial intelligence (AI) is machine learning, which uses large data sets (personal and non-personal) to find patterns and correlations from which it makes predictions and decisions. In this model (which also includes generative AI such as Chat GPT or Midjourney), AI must be properly trained. Training consists primarily of feeding the system through appropriate datasets, i.e. datasets that are sufficiently diverse, relevant, and representative (also in terms of gender, ethnicity, or age), free of errors, complete in view of the intended purpose of the system, and able to be used legally. Due to the implications of the deployment of data in the creation and use of new technologies, not only does the technical concept of data quality or the legal protection of personal data and intellectual property come into play, but also the more comprehensive concept of data justice.

Questions of data justice have been dealt with by social scientists, from the seminal work of Jeffrey Alan Johnson on open data and information justice[2] to the framework for data justice advocated by Linnet Taylor[3] and other distinct strands of

[1] White Paper on Artificial Intelligence – A European approach to excellence and trust, COM/2020/65 final. For an explication of Artificial Intelligence systems and Machine Learning models, when "a computer observes some data, builds a model based on the data, and uses the model as both a hypothesis about the world and a piece of software that can solve problems," see S. Russel, P. Norvig, *Artificial intelligence – A Modern Approach*, Hoboken 2021, p. 651.

[2] J.A. Johnson, *From open data to information justice*, "Ethics and Information Technology" 2014, vol. 16, no. 4, p. 263 *et seq*.

[3] L. Taylor, *What is data justice? The case for connecting digital rights and freedoms globally Big*, "Data & Society" 2017, vol. 4, no. 2, p. 1 *et seq*.

research.[4] Our goal is not to provide an overview of the various approaches to data justice among social scientists and philosophers. Instead, our objective is to take stock of this ongoing debate in order to highlight certain legal issues that involve aspects encompassed within the multifaceted concept of data justice. More specifically, our focus will be placed on the legal aspects that have recently been addressed by different pieces of EU legislation or EU initiatives. In this regard, the EU legislator has demonstrated an awareness of and an attempt to address concerns related to data ownership, data openness, data re-use, fair data collection and processing, data quality, and non-discrimination: all issues that are explored by researchers in the field of data justice. The EU legislator has done so through various initiatives, ranging from the individual perspective of the GDPR[5] and intellectual property provisions, to more recent and collective approaches, in the EU Strategy on Data and the EU Directive on Open Data,[6] the EU Regulation on Data Governance,[7] the Data Act,[8] the Digital Services and Market Acts,[9] and the Artificial Intelligence Act (AIA).[10] In these acts, the approach taken has predominantly followed a techno-procedural path, while a looser approach

[4] For an overview of the different approaches to data justice, see L. Dencik, J. Sanchez-Mondero, *Data Justice*, "Internet Policy Review" 2022, vol. 11, no. 1, p. 1 *et seq*.

[5] Provisions about data protection have formally evolved from a directive (Directive 95/46/CE) to the Regulation of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, OJ L 119, 4.5.2016, pp. 1–88 (Regulation EU 679/2016).

[6] A European Strategy for Data, COM(2020) 66 final, p. 1; Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast), OJ L 172, 26.6.2019, pp. 56–83, par. 13, sets out that "Public sector information or information collected, produced, reproduced, and disseminated within the exercise of a public task or a service of general interest, is an important primary material for digital content products and services and will become an even more important content resource with the development of advanced digital technologies, such as artificial intelligence, distributed ledger technologies and the internet of things. Broad, cross-border geographical coverage will also be essential in that context. Increased possibilities of re-using such information is expected, inter alia, to allow all Union businesses, including microenterprises and SMEs, as well as civil society, to exploit its potential and contribute to economic development and high-quality job creation and protection, especially for the benefit of local communities, and to important societal goals such as accountability and transparency."

[7] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), OJ L 152, 3.6.2022, pp. 1–44.

[8] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

[9] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ L 265, 12.10.2022, pp. 1–66.

[10] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj (Artificial Intelligence Act – AIA).

has been adopted in listing forbidden purposes in the context of artificial intelligence systems. However, the purposes addressed by means of different practices based on big data analytics (artificial intelligence included, and more specifically machine learning methods and foundation models) are likewise relevant for data justice, and in particular for the "instrumental" approach to data justice.[11] In light of this, our argument calls for a re-assessment of the permissible and forbidden purposes within artificial intelligence systems, in order to redefine the boundaries between the legal implications of data justice, on the one hand, and the freedom to conduct a business or the control exerted by public authorities over citizens, on the other. In addition, a comprehensive approach to data justice also underlies in-depth consideration of and respect for data ownership when it goes hand in hand with intellectual property rights, in order to strike a fair balance among openness, sharing, re-use of data, and IP.

However, we suggest that the techno-procedural approach adopted by the EU initiatives is insufficient in respect of all the multifaceted implications of data justice.

1. Establishing data justice as a legal priority

1.1. Data as a two-faced Janus: technical but legal requirements

"Data are commonly understood as measures of the world and the building blocks from which information, knowledge and value are produced. There is a long history of governments, businesses, academia and citizens producing and utilising data in order to monitor, regulate, profit from, and make sense of the world […] Data have lost none of their value, but in other respect their production and nature has been transformed through a set of disruptive innovations, including networked digital infrastructures, pervasive ubiquitous computing, cloud services and open government. Indeed, there has been a profound datification of everyday life as evermore phenomena are captured as data, and these data in turn are used to shape social and economic systems."[12] This is part of the opening address in Rob Kitchin's *The Data Revolution*. The statement clearly sets out the features and framework of a "data-driven society," nowadays, one where the classical life cycle of information is essentially dominated by gathering, processing, and extracting value from data as well as data re-use.[13] Within this framework, data lie

[11] R. Heeks, J. Renken, *Data Justice for development: What would it mean?*, Development Informatics Working Paper Series, No. 63, 2016, p. 4, remind us that "instrumental data justice means fair use of data; it therefore focuses on the outcome of use of data […]. From this perspective, there is no justice inherent to the data domain; instead justice is defined outwith that domain. For example, this would argue that there is no inherent justice or injustice about who owns data in developing countries or in development projects; concerns about justice only relate to the impact of the use of that data."

[12] R. Kitchin, *The Data Revolution – A critical Analysis of Big Data, Open Data & Data infrastructures*, London 2022, p. 3.

[13] This precautionary approach underlies comprehensive awareness of the risks involved in current reality, since "There has never been a state, monarchy, kingdom, empire, government, or corporation in history that has had command over such granular, immediate, varied, and detailed data about

at the centre of a circle; most current social, economic, and political processes revolve around it.[14]

As a result, data are not merely a technical component; rather, they stand at the root of increasing legal implications, not only of an individual nature but also of a collective one. From an individual standpoint, it is a question of protecting personal and, moreover, sensitive data, as well as of protecting intellectual property rights. However, if we broaden our perspective to adopt a more collective approach, it is essential to ensure data quality,[15] make data publicly available for reasons of transparency and accountability, and enable data sharing to extract further economic and socio-political value.[16] This undoubtedly implies avoiding the confinement of data within silos that can bolster oligopolistic positions. Instead, it suggests opening them to broader utilization by other public or private entities, with the aim of enhancing the delivery of public services, fostering growth, and boosting innovation.[17] In addition, this opening process necessitates ensuring the interoperability of formats and protocols governing IT systems where data are uploaded and made accessible, in order to facilitate readability and exchange with other entities.[18] However, these activities cannot violate the rights of data creators, including intellectual property rights, the essence of which is to support creativity and human creation.

It is within this context that the increasing focus of the European legislator on data processing, circulation, and the intertwined issue of data quality is paramount. Not only the well-known and settled normative discipline about the protection of personal data and intellectual property, but also the European Data Strategy document and consequent initiatives like the Open Data Directive, the Data Governance Act and the

---

subjects and objects that concern them;" thus "data has become a major object of economic, political and social investment for governing subjects," E. Ruppert, E. Isin, D. Bigo, *Data Politics*, "Big Data & Society" 2017, vol. 2, no. 5, p. 2.

14 L. Floridi, *The fourth Revolution. How the infosphere is Reshaping Human Reality*, Oxford, 2014, p. 6 *et seq.*

15 For an in-depth overview of the issues involved in data processing, beyond personal data protection, and referring rather to their accessibility, their coverage and granularity, their quality (implying how clean data are, in terms of error and how gap free; how untainted they are, in terms of bias; how consistent and complete they are, in terms of discrepancies), their veracity (referring to the authenticity, accuracy, and fidelity of the data), see R. Kitchin, *The Data Revolution*…, p. 187 *et seq*.

16 As shown by A. Ross, *The Industries of the Future*, New York 2016, p. 153, "Land was the raw material of the agricultural age. Iron is the raw material of the industrial age. Data is the raw material of the information age."

17 A European Strategy for Data…, p. 1.

18 For the concept of data integration and interoperability, see R. Kitchin, *The Data Revolution*…, pp. 196–198. The European Strategy for data (p. 6), in order to support cross-border interoperability, provides for the creation of a European data space to be accompanied by the development of sectoral data spaces in strategic areas such as manufacturing, agriculture, health, and mobility. As for the public sector, a European Interoperability Framework (EIF) – COM(2017) 134 final, was set out, as well as a Regulation (EU) 2024/903 of the European Parliament and of the Council of 13 March 2024 laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act), and a Proposal for a European Interoperability Framework for Smart Cities and Communities (EIF4SCC), were adopted by the EU.

forthcoming Data Act come into account. Further initiatives that lay down the technical requirements for data openness and data sharing by means of an interoperable structural framework that allows cross-borders data exchange[19] are also significant. In addition, the Digital Market Act provides obligations for gatekeepers in respect of fair data use, data access, and portability; likewise the Digital Services Act upholds transparency, fairness, and accountability of service providers. Finally, the draft AIA is aimed at regulating systems primarily fed with data, with a particular focus on machine learning as a subset of these AI systems, and the consequent need to implement a data governance process that aims to achieve data quality.

Briefly, in the above-mentioned pieces of legislation and initiatives the EU is trying to cope with the main legal implications of a data driven-society. They involve mutually intertwined aspects, since making data more accessible and available is a necessary "prerequisite for seizing the opportunities presented by the digital age we live in;"[20] in turn, this process needs to be supplemented by provisions about data quality and IP safeguards, with specific regard to the case of data massive deployment by machine learning systems. Consequently, throughout the above-mentioned initiatives the EU has made technical, organizational, and managerial aspects of data processing (beyond personal data being involved), legally relevant and binding.

However, the EU has not laid down strict or rigid legal requisites and requirements, since it has rather set out some principles and general clauses, delegating their specification to following technical standards.[21] Likewise, the EU has not yet adopted a specific approach to IP rights in the face of the massive deployment of free accessible online data. This path of action gives evidence of the difficult compromise pursued by the legislator: the necessity to protect fundamental rights, according to a human-centred approach, while neither stifling innovation nor hindering competition in the EU market.[22] As a consequence, technical and procedural requirements leave open a significant scope for manoeuvre to entities that deal with data; this is the reason why such requirements surely integrate necessary and pivotal aspects for the sake of data justice, but they are not enough comprehensively to safeguard this.

---

19 As stated by the European Strategy for Data…, p. 7: "The application of standard and shared compatible formats and protocols for gathering and processing data from different sources in a coherent and interoperable manner across sectors and vertical markets should be encouraged through the rolling plan for ICT standardisation and (as regards public services) a strengthened European Interoperability Framework."

20 Data Act, p. 1.

21 M. Ebers, *Standardizing AI – The Case of the European Commission's Proposal for an Artificial Intelligence Act* [in:] *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*, eds. L. Di Matteo, C. Poncibò, M. Cannarsa, Cambridge 2022, p. 330.

22 AIA, par. 1, states that the purpose of this Regulation is to promote the uptake of human centric and trustworthy artificial intelligence and to ensure a high level of protection of health, safety, fundamental rights, democracy and rule of law and the environment from harmful effects of artificial intelligence systems in the Union while supporting innovation and improving the functioning of the internal market.

1.2. Intellectual property of data – engine of innovation or a restraint on new technologies

One of the aspects of data justice is securing the ownership of data and the need to protect the intellectual property of creators whose works constitute training material for artificial intelligence. As Agnieszka Wachowska and Marcin Ręgorowicz point out, the subject of the dispute is the right of operators of "learning" systems to use publicly available data sets.[23] Obtaining one's "own" result through the mass, automatic acquisition and subsequent processing of data can lead to infringements of the rights of creators (or) other rights holders, as well as database producers, who have no real control over the use of their protected resources.

The generative AI type is particularly prone to infringement. It is trained on the basis of creative works (materials), e.g. texts, software code, or images. The system draws on the works, mixes them and then delivers (generates) works of a similar type (which can be considered collages). The final result may have varying degrees of similarity to the works used to generate it. Nevertheless, it is a direct result of the earlier human work that was copied and at the same time represents a competing work. Generative AI systems are able to flood the market very quickly with works that are substantially similar to original works; they are able to imitate style, distinctive colours, etc. From the technical point of view, such effects are achieved using a mathematical process called "diffusion."[24]

The problem of ensuring an adequate level of intellectual property protection of data is emphasized quite often by EU decision-makers,[25] but so far there have been no comprehensive legal solutions, as has been indicated by the European Parliament.[26] Recognizing that the EU needs to harmonise and to create a common European data space (an internal market for data),[27] in the field of protection of data creators the EU

---

[23] A. Wachowska, M. Ręgorowicz, *ChatGPT w praktyce – najważniejsze kwestie prawne*, https://www.traple.pl/chatgpt-w-praktyce-najwazniejsze-kwestie-prawne/ [accessed: 2023.10.13].

[24] The technique was created in 2015 by AI researchers at Stanford University. The first step is to translate the piece into its constituent elements, then small elements are removed (known as denoising) to create a lossy copy (highly compressed, similar to MP3 or JPEG files). In 2020, the technique was improved by researchers at UC Berkeley, making it possible to create better compressed training images (called hidden images). Moreover, it was discovered that hidden images can be interpolated (blended mathematically), thus creating new derivative images. The work of researchers from Munich resulted in further improvement of the process in 2022, when additional information (so-called conditioning) was introduced at the stage of the denoising process. However, this does not change the fact that the new painting is a simple consequence of copying fragments of other works. R. O'Connor, *Introduction to Diffusion Models for Machine Learning*, https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/ [accessed: 2023.10.27]; https://stablediffusionlitigation.com/ [accessed: 2023.10.26].

[25] For example: Directive (EU) 2019/1024; Regulation (EU) 2022/868; Regulation (EU) 2023/2854 (Data Act).

[26] Report on intellectual property rights for the development of artificial intelligence technologies of 2.10.2020 (2020/2015(INI)).

[27] Points 17–20 of European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies (2020/2015(INI)).

legislator currently refers to the solutions adopted in Directive (EU) 2019/790[28] and in Directive 96/9/EC.[29]

Directive 2019/790 introduced three exceptions in copyright law, i.e.: text and data mining for scientific research (Art. 3 and 4); for teaching activities (Art. 5); and for the preservation of collections by cultural heritage institutions (Art. 6). The first exception relates to the reproduction, presentation, downloading, and secondary use of all or part of a database protected by a *sui generis* right and to the use of press publications for text and data mining in connection with scientific research.[30] Article 5 covers the use of works and other protected subject matter in digital and cross-border teaching activities. The purpose of these provisions is to allow the digital use of works and other protected objects for the purpose of illustration in the context of teaching, to the extent justified solely by the non-commercial purpose to be achieved. The established exception to this concerns the exploitation of databases and works, as well as of computer programmes (reproduction). The third exception concerns the possibility of using collected works by cultural heritage institutions. Its task is to enable archiving in an appropriate amount, at any time and to the extent necessary to preserve this type of collection; however, the requirement is that works and other protected items must be permanently in the collections of a given institution.[31]

Exploiting the exceptions above, in accordance with Art. 4 section 3 of the Directive, is possible when "the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online." The regulation allows third parties to reproduce databases or works for the purpose of machine learning (whether for scientific or commercial purposes), provided that the rights holder may refuse permission in the case of commercial applications.[32]

Directive 96/9/EC takes into account the essence of "online databases" in its regulation, indicating that appropriate measures are necessary to prevent unauthorized extraction and/or re-utilization of data. Article 1(2) of this act defines the database as "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means."[33] The act states *expressis verbis* that the elements of the database may also include independent works. It is worth emphasizing the cumulative protection of databases, i.e. copyright protection and the so-called *sui generis* right. If a database constitutes a work within

---

[28] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, OJ L 130, 17.5.2019, pp. 92–125.

[29] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, pp. 20–28.

[30] E. Laskowska-Witak, *Komentarz do dyrektywy o prawach autorskich w ramach jednolitego rynku cyfrowego*, LEX/el. 2019.

[31] *Ibid*.

[32] A. Wachowska, M. Ręgorowicz, *ChatGPT w praktyce*…

[33] As indicated in recital 17, the term "database" includes: literary, artistic, musical or other collections of works or collections of other material such as texts, sound, images, numbers, facts, and data.

the meaning of copyright law, it is protected as a whole, even if the individual elements are not of a creative nature. However, *sui generis* protection is a right for the producer of a database that requires a qualitatively or quantitatively significant investment to obtain verification or presentation of its content, which is intended to protect against the extraction or re-utilization of data in whole or in significant part. Legal users of the database are only entitled to download data or re-use a non-essential part of it for any purpose, both as regards the quality and quantity of the data.

Users may not carry out acts contrary to the normal exploitation of the database or unreasonably prejudice the legitimate interests of the producer, and are obliged to respect the rights of the holder of copyright or related rights in respect of works or subject matter which constitute the contents of the database. Exceptions to the *sui generis* right are contained in Art. 9 of Directive 96/9/EC, which indicates that data may only be extracted or re-used in substantial part without the authorisation of the producer: 1) for extraction for private purposes of the contents of a non-electronic database; 2) as illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved without prejudice to the exceptions and limitations provided for in Directive (EU) 2019/790; 3) in the case of extraction and/or re-utilization for the purposes of public security or administrative or judicial procedure.[34]

It should be pointed out that a broad understanding of the concept of text and data mining, understood as an automated analytical technique for analyzing digital texts and data to generate information including, but not limited to, patterns, trends, and correlations [Art. 2(2) of Directive 2019/790], is intended to guarantee the flexibility of the definition and reduce the risk of its becoming obsolete with constant technological progress. In contrast, as Martin Kretschmer and Thomas Margoni point out, an overly broad understanding of the process of text and data mining by different users in different units of time makes the development of AI entirely dependent on exceptions to the use of data, and limiting the scope of exceptions to the right of reproduction leaves the communication of research results in a grey area. According to the authors, there is no need to establish an exception for the act of extracting information value from protected works, which is a strongly debatable position.

Despite the emphasis on creators' rights in EU legislation regarding data and technologies created based on them, and the assertion that intellectual property protection must be taken into account in the development of new technologies, an "open source" philosophy is more visible. The current legislation is not sufficiently adapted to the new conditions in which creators operate. This is expressed in broadly defined protection exceptions, as well as the legislator's narrative, which emphasizes primarily the potential resulting from access to data. Also an exception from Art. 4 section 3 of Directive 2019/790 regarding the reproduction of databases or works for the purposes of machine learning, which allows the rights holders to refuse permission

[34] Changes in *sui generis* protection are provided for by a regulation on harmonised rules on fair access to and use of data (Data Act).

when the activity has a commercial dimension, does not ensure them any protection, as in fact, in most cases, they do not even have the possibility to verify whether their protected objects are being used or in what way, as this is happening exponentially. It should be noted that infringement of intellectual property rights can already occur at the stage of programming and teaching the system, the generation of results, or during the evaluation of the right to use the obtained results.[35]

Also, the AIA will not solve the problem, although it may bring some changes to the lack of transparency. According to Art. 11, high-risk AI systems should be provided with technical (updated) documentation containing the necessary information. For example, such information could include the general characteristics, capabilities, and limitations of the system, algorithms, data, training, testing, and validation processes used, as well as documentation on the relevant risk management system.[36] This means that it will be necessary to provide the data used for AI training, but only for high-risk AI systems.

Internal terms of services binding on users of individual portals also do not support the rights of creators. When posting songs on popular platforms such as Google, YouTube, or X, it is worth knowing that users are granting a license to these entities. For example, under Google Terms of Service, the license allows users to host, reproduce, distribute, transmit, and use the content, for example to save it on Google's systems and make it accessible from anywhere; it permits changing user content, e.g. by reformulating or translating it, as well as sublicensing these rights, among others, to develop new technologies and services for Google.[37]

An even broader scope of licences was adopted in the X terms of service of 29 September 2023.[38] License (with the right to sublicense) includes the right "to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods now known or later developed (for clarity, these rights include, for example, curating, transforming, and translating). This license authorizes us to make your content available to the rest of the world and to let others do the same. You agree that this license includes the right for us to provide, promote, and improve the Services and to make Content submitted to or through the Services available to other companies, organizations, or individuals for the syndication, broadcast, distribution, repost, promotion, or publication of such Content on other media and services, subject to our terms and conditions for such Content use." These internal provisions indicate that when using social media, we should be aware that our content will serve as training (validation, testing) material in the process of machine learning.[39]

---

35 A. Wachowska, M. Ręgorowicz, *ChatGPT w praktyce…*

36 See: Annex IV, AIA.

37 Google Terms of Service of 5 January 2022, https://policies.google.com/terms?hl=en&fg=1 [accessed: 2023.10.26].

38 https://twitter.com/en/tos [accessed: 2023.10.25].

39 For example: the case of Copilot v Microsoft and GitHub concerned the use of data published on the social network. Repositories owned by Copilot were exploited by the defendants to train

These problems are not merely abstract in nature, as is shown by lawsuits concerning the intellectual property of training data. One of the first case, Getty Images v. Stability AI, is pending in a federal court in Delaware. The lawsuit concerns the copying of over 12 million photos along with captions and metadata. Damage was estimated at $150,000 for each work, which could mean a total of $1.8 trillion.[40] A similar claim was filed before the High Court of Justice in London.[41] In another case, artists Sarah Andersen, Kelly McKernan, and Karla Ortiz sued Midjourney, Stable Diffusion, and DeviantArt.[42] In turn, OpenAI, the creator of ChatGPT, was sued by a group of writers and journalists, including: Michael Chabon, David Henry Hwang, Matthew Klam, Rachel Louise Snyder, and Ayelet Waldman. They claim that their works were copied without their consent and used to teach the generator how respond to commands entered by people.[43] Satirist Sarah Silverman accused OpenAI of unlawfully generating book summaries.[44] *The New York Times*, in turn, is considering accusing the chatbot of plagiarism.[45] The results are difficult to predict. Companies working on artificial intelligence indicate that the use of data protected by copyright is possible under the principles of fair use provided for in US law; another difficulty is proving that the work was actually used by AI.[46]

The lack of appropriate legislative solutions has prompted protests from the arts community. On 9 August 2023, leading global news and publishing organisations (among them: Agence France Presse, European Pressphoto Agency, the European Publishers' Council, the National Press Photographers Association, the National Writers

---

generative AI. According to GitHub, users who publish their code on the platform have agreed to the viewing, usage, indexing, and analysis of public code. For this reason the owners of the portal are entitled to use, including for commercial purposes, the published data. In this case, it involved the creation of codes that were very similar to or even duplicated user codes. It is true that Copilot published under an open source licence, but the claimant considers that the scope of use, including copying of published data, does not fall within the licence granted. T. Claburn, *Microsoft and GitHub are still trying to derail Copilot code copyright legal fight*, https://www.theregister.com/2023/07/01/microsoft_github_copilot/ [accessed: 2023.10.21].

40 M. O'Brien, *Photo giant Getty took a leading AI image-maker to court. Now it's also embracing the technology*, https://apnews.com/article/getty-images-artificial-intelligence-ai-image-generator-stable-diffusion-a98eeaaeb2bf13c5e8874ceb6a8ce196 [accessed: 2023.10.26].

41 Getty Images, *Statement of 17 January 2023*, https://newsroom.gettyimages.com/en/getty-images/getty-images-statement [accessed: 2023.10.26].

42 J. Vincent, *AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit*, https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart [accessed: 2023.10.24].

43 Ch. DiFeliciantonio, *Authors Michael Chabon, David Henry Hwang sue OpenAI over copyright concerns*, https://www.sfchronicle.com/tech/article/michael-chabon-open-ai-lawsuit-copyright-18360019.php [accessed: 2023.10.20].

44 Z. Small, *Sarah Silverman Sues OpenAI and Meta Over Copyright Infringement*, https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html [accessed: 2023.10.26].

45 B. Allyn, *'New York Times' considers legal action against OpenAI as copyright tensions swirl*, https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl [accessed: 2023.10.25].

46 Skeptical about the chances of winning, see: N. Senkowska, *"Trening" sztucznej inteligencji: co z prawami twórców dzieł, na których ćwiczy*, "Rzeczpospolita", 13.07.2023.

Union, The Associated Press, The Authors Guild) presented an open letter calling for copyright protection to be taken into account in the development of generative AI models. In particular, they demanded disclosure of the training sets used to create generative AI models.[47]

Taking into account the increasing number of lawsuits and objections from the press, OpenAI has announced that website operators have the option to block the content published there for Chat GPT.[48] The ability to block content, especially through media portals, seems to be most desirable. Many entities, including Reuters, Getty Images, *The Guardian*, *The New York Times*, *The Chicago Tribune*, CNN, Australia's ABC, *The Canberra Times*, *The Newcastle Herald*, and other content providers have already banned Chat GPT from using the content they generate.[49]

These examples show that the lack of adequate protection for creators will affect the quality of the data, that is – in turn, as is shown in par. 2.1 – of paramount importance for the sake of data justice. There is a risk that if reliable/quality data (e.g. valuable press releases) are blocked for AI systems, these systems will be fed with datasets containing a significant number of errors (e.g. fake news).

Intellectual property rights over data, as an element of data justice, need to be balanced according to a more instrumental approach than what is currently available. Protection in this area should focus, above all, on the transparency of training data sources or planned methods of use (applied not only to high-risk AI systems), but also in order to protect human creativity (in three dimensions: training, processing, and producing works competing with the original) should be based on the need to obtain an author's consent. First, the owners of new solutions currently do not publicly disclose information about the origin of the data used.[50] Second, both the regulations

---

[47] https://drive.google.com/file/d/1jONWdRbwbS50hd1-x4fDvSyARJMCgRTY/view [accessed: 2023.10.23], see also https://www.publishers.org.uk/global-summit-on-ai-the-importance-of-intellectual-property-to-the-success-of-safe-artificial-intelligence/ [accessed: 2023.10.28].

[48] New privacy policies from Google and the Meta-owned platforms introduce the possibility for users to block user-generated content. However, the works collected so far remain in the database. D. Milmo, *The Guardian blocks ChatGPT owner OpenAI from trawling its content*, https://www.theguardian.com/technology/2023/sep/01/the-guardian-blocks-chatgpt-owner-openai-from-trawling-its-content [accessed: 2023.10.28]. At the same time, the system has been improved, so that the chat is based on current content and not on data posted on the Internet until 2021. Data blocking is also possible by using appropriate plug-ins, https://pl.wordpress.org/plugins/block-chat-gpt-via-robots-txt/ [accessed: 2023.10.28]. Another solution is the Nightshade tool, which disrupts training data. Here again, new technologies are ahead of the law, E. David, *Artists can use a data poisoning tool to confuse DALL-E and corrupt AI scraping*, https://www.theverge.com/2023/10/25/23931592/generative-ai-art-poison-midjourney [accessed: 2023.10.28].

[49] A. Bogle, *New York Times, CNN and Australia's ABC block OpenAI's GPTBot web crawler from accessing content*, https://www.theguardian.com/technology/2023/aug/25/new-york-times-cnn-and-abc-block-openais-gptbot-web-crawler-from-scraping-content [accessed: 2023.10.20]; see also B. Haring, *BBC Will Block ChatGPT AI from Scraping Its Content*, https://deadline.com/2023/10/bbc-will-block-chatgpt-from-scraping-its-content-1235566868/ [accessed: 2023.10.14].

[50] The data often comes directly from social media, websites, or databases, including those created by non-profit projects such as LAION (Large-scale Artificial Intelligence Open Network). It provides free access to various types of databases, making a reservation that "this large-scale dataset is non-

of social networking sites, where works may come from, and artificial intelligence systems assume a presumption of consent, which significantly weakens the position of the artist. The adopted techno-procedural model is not complete enough, and the existing regulations introduce extensive exceptions that leave authors practically without protection. It is also interesting that in relation to new technologies, the protection of authors is at a weaker level than in relation to traditional forms of use of works (despite the greater potential for threats). However, an approach that guarantees the same protection regardless of the tool used is justified under the scheme of data justice.

2. Navigating legal significance: unravelling the journey towards data justice

2.1. A multifaceted concept of data justice

Data justice is a concept originally developed within the realm of the social sciences; however, in the context of digital society nowadays, it also deserves closer examination within the domain of legal studies.[51] There is no universally agreed upon and established definition of what data justice is or, consequently, how to address data (in) justice.[52] Data justice is a multifaceted concept that encompasses different aspects: the existing inequalities reflected and multiplied by data and relevant discriminations; the ways of gathering and processing data; the purposes addressed by the deployment of data; new digital rights; the "politics" of data[53] with the implied asymmetries of data power and "ownership" among private bodies and between private bodies and public authorities; and a society forged by data and for data.[54]

However, beyond these different facets, it is possible to isolate some aspects of legal relevance that deserve major discussion among legal scholars. First, the focus of data justice is on groups, in addition to individuals, and it extends beyond any personal data involvement. Second, open data policies surely align with social justice principles, because of their "democratic" approach to data sharing. But they hold a reverse side:

curated. It was built for research purposes to enable testing model training on larger scale for broad researcher and other interested communities, and is not meant for any real-world production or application."

51 L. Dencik, J. Sanchez-Mondero, *Data*…, p. 3, remind us that "To speak of data justice is thus to recognise not only how data, its collection and use, increasingly impacts on society, but also that datafication is enabled by particular forms of political and economic organisation that advance a normative vision of how social issues should be understood and resolved. That is, data is both a matter *in* and *of* justice; datafication embodies not only processes and outcomes of (in)justice, but also its own justifications."

52 R. Kitchin, *The Data Revolution*…, p. 287, observes that "there is no shared common understanding of the moral principles of social justice – and by association data justice – and how to achieve it."

53 E. Ruppert, E. Isin, D. Bigo, *Data Politics*…, p. 3.

54 L. Dencik, J. Sanchez-Mondero, *Data*…, p. 3.

the exacerbation of existing inequalities, since data often reflect deeply ingrained socio-cultural biases and discrimination;[55] the risk of further discrimination stemming from the repurposed and broader deployment of inferred (anonymous) data, originally collected for specific groups of people (e.g., unwell, elderly, or disabled individuals)[56] and for specific purposes concerning these groups; and the infringement of IP rights. Third, data justice upholds a procedural approach and participatory rights to ensure data sharing, data quality, and non-discriminatory practices. Considering this context, the data justice approach can also be deemed supportive of the establishment of new digital rights, which can be enforced either individually or collectively.[57] This means that the legal system endows citizens or groups with appropriate legal tools to defend their claims related to fair data processing, but, in doing so, it also charges them with the responsibility to consider possible cases of data (in)justice.

For our limited purposes, it is worth recalling the methodology implemented by the EU legislator to cope with data quality, data and IP protection, data security, the underlying risks of discrimination, and the objectives of making data more open and available for re-use. In this respect, the methodological pattern followed by the EU is mainly built on a techno-procedural-driven approach.[58] Thus, some organisational steps are required to be embedded in the technology itself (i.e. by design and by default);[59] furthermore, some procedural fulfilments are listed to set out a governance process addressed with tackling the question of transparency and data quality (i.e. Art. 10, AIA).

---

55 As stressed by the Opinion of the European Economic and Social Committee on Artificial intelligence, *The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society*, in OJ 2017/C 288/01, par. 3.5. "There is a general tendency to believe that data is by definition objective; however, this is a misconception. Data, may be biased, may reflect cultural, gender and other prejudices and preferences."

56 As stressed by L. Taylor, *What is data justice*…, p. 2, "the greatest burden of dataveillance (surveillance using digital methods) has always been borne by the poor […] Beyond socio-economic status, gender, ethnicity and place of origin also help to determine which databases we are part of, how those systems use our data and the kinds of influence they can have over us."

57 B. Custers, *New digital rights: Imagining additional fundamental rights for the digital era*, "Computer Law and Security Review" 2022, vol. 44, pp. 9–10, refers (among others) to the right to change your mind, the right to start over with a clean (digital) slate, and the right to expiry dates for data.

58 In this respect, digital constitutionalism rests on safeguards addressed to protect fundamental rights and democratic values, thus making private digital bodies accountable by means of procedural fulfilments, transparency, and due process in order to limit their discretionary margin of manoeuvre and mitigate the risks stemming from their practices: see G. De Gregorio, *Digital Constitutionalism in Europe – Reframing Rights and Powers in the Algorithmic Society*, Cambridge 2022, p. 312 *et seq.*; O. Pollicino, *Judicial Protection of Fundamental Rights on the Internet. A Road Towards Digital Constitutionalism?*, Oxford 2021.

59 L.A. Bygrave, *Hardwiring Privacy* [in:] *The Oxford Handbook of Law, Regulation, and Technology*, eds. R. Brownsword, E. Scotford, K. Yeung, Oxford, 2018, p. 755, recalls that "with such embedment, the automated processes of the architecture will help automate legal norms, thus making the latter largely self-executing."

2.2. From data quality to data justice

We know that data are not neutral for various reasons. On the one hand, they involve human choices that may be questionable; they imply certain methodologies over others, the collection of certain data over others, and different modes of gathering, selecting, measuring, and analysing data. On the other hand, in today's society, data represent power, not only in terms of economic revenues as viewed through the lens of antitrust and competition law, but also because they enable a profound understanding of citizens' habits and preferences, facilitating profiling, predictions, and subsequently, personalized decision-making,[60] according to the settled relationship between knowledge and power, well described by Michel Foucault,[61] and by the more recent "surveillance capitalism" theory.[62]

The scale of this data power can be proven in terms of numbers, i.e. by the global market size gained in recent years by intelligent data processing and the pace of its increasing growth rate estimated for coming years;[63] but it can also be proven by the influence of data and information on the exercise of fundamental rights and freedoms like the freedom of expression and information (i.e. disinformation, misinformation) as well as on democratic processes (i.e. the Cambridge Analytica case with reference to the US elections and the UK referendum).

In addition to these asymmetries of power, further risks emerge when datasets and their analytics are built upon biased data;[64] or even if data do not contain prejudices and are not inherently discriminatory, they can be deployed in ways that yield discriminatory outcomes (the Aadhaar system in India is telling);[65] or in any case deployed in order to re-shape human behaviour according to the will of the data controller, impinging upon people's free will. In short, data have become a "political and social practice" and so they share with it stereotypes, gaps, prejudices, and biases.[66]

---

60 J. van Dijck, *Datification, dataism and dataveillance: Big data between scientific paradigm and ideology*, "Surveillance & Society" 2014, vol. 12, no. 2, p. 197 *et seq.*; G. De Gregorio, *Digital Constitutionalism in Europe*…, describes the shift that has occurred within the framework of the current algorithmic society – from the freedom to conduct a business to real private digital bodies that exercise functions traditionally offered by public authorities.

61 M. Foucault, *La naissance de la biopolitique. Course au Collège de France (1978–1979)*, Paris 2004.

62 S. Zuboff, *The Age of Surveillance Capitalism – The Fight for a Human Future at the New Frontier of Power*, London 2019.

63 See *Intelligent Document Processing Market Size, Share & Trends Analysis Report*, 2023, https://www.grandviewresearch.com/industry-analysis/intelligent-document-processing-market-report [accessed: 2023.10.28].

64 G.A. Grasso, *GDPR Feasibility and Algorithmic Non-Statutory Discrimination*, Naples 2022, p. 10, underlines that "the presence of bias […] leads to systematic errors that influence judgement and decisions. These distortions or false representations of reality may also affect computer systems, which consistently and unfairly discriminate against certain individuals or group of individuals in favour of others, denying opportunity or generating unwanted results for unreasonable or inappropriate reasons."

65 For a description of the discriminatory background and consequences of this for India's biometric database, see L. Taylor, *What is data justice*…, pp. 4–5.

66 E. Ruppert, E. Isin, D. Bigo, *Data Politics*…, p. 3, state that "the production of data is a social and often political practice that mobilizes agents who are not only objects of data (about whom data

Data sharing assumes that the data made available are of good quality to prevent discrimination resulting from big data analytics.[67] In this last respect (avoiding discrimination), not only are data gathering procedures and data quality relevant, but also the way they are processed by different algorithms, the algorithm models chosen, their uses, and the final aims addressed.[68] In respect of the mentioned risks, not only are individual rights in the foreground, but also the interests of entire groups: "Big data and associated analytics are radically transforming how people are treated collectively."[69] This holds true, not only when decisions are taken and they produce legally binding effects on people, such as the case of artificial intelligence systems used in order to allocate social allowances or benefits, select workers, candidates or students, or implement predictive justice etc.[70] This is also true when the insight and knowledge gained through artificial intelligence systems about the habits, behaviours, and cognitive patterns of groups are employed in a softer but equally striking way, by nudging people's freedom of will and conduct.[71] In this respect, power (either public or private) becomes "intimate and efficient. It knows us. It learns from us."[72] In other instances, this in-depth knowledge about people, gives rise to a real "dataveillance" that entails an enforced disciplinary control over people, such as is the case in China.[73]

This highlights the necessity for the integration of the existing procedural approach with more stringent legislative intervention in defining which uses should be prohibited when big data analytics come into play, particularly through artificial intelligence systems. There are two primary reasons for this. First, a procedural approach

---

is produced) but that they are also subjects of data (those whose engagement drives how data is produced). Our question thus shifts to social practices and agents. Data does not happen through unstructured social practices but through structured and structuring fields in and through which various agents and their interests generate forms of expertise, interpretation, concepts, and methods that collectively function as fields of power and knowledge."

67 Cases of algorithmic run by data that have brought about discriminatory outcomes are too well-known to be described; it suffices to recall the Loomis case in the US, that of Syri in the Netherlands, or Amazon's automated recruitment system.

68 European Union Agency for Fundamental Rights – Report, *#BigData: Discrimination in Data Supported Decision – Making*, Luxembourg 2018, p. 3.

69 R. Kitchin, *The Data Revolution*…, p. 214.

70 For an overview of the different uses of artificial intelligence systems and the consequent legal implications, see B. Custers, E. Fosch-Villaronga, *Law and Artificial Intelligence – Regulating AI and Applying AI in Legal Practice*, The Hague 2022.

71 For the nudge theory see R.H. Thaler, C. Sunstein, *Improving decisions about Health, Wealth and Happiness*, New York 2009.

72 J. Cheney-Lippold, *We are Data: Algorithms and the Making of Our Digital Selves*, New York 2017, p. 107.

73 In this respect the testimony offered by the investigative journalist Geoffrey Cain before the U.S. Senate Committee on the Judiciary Subcommittee on Human Rights and the Law on 10 June 2023 is telling: "The Chinese Communist Party (CCP) has engineered a vast AI-powered surveillance system literally called 'Sky Net'. It runs AI-powered 'alarms' that notify the police and intelligence services when someone unfurls a banner, when a foreign journalist is traveling to certain parts of the country, and when someone from an ethnic minority is present. The government accuses entire groups, such as Muslim Uyghurs, of posing a terrorist threat, and relentlessly persecutes them with the use of AI tools."

is unable to eliminate the root-causes of inequalities deeply embedded in cultural and socio-economic structures. Second, while procedures can help to ensure transparent, fair, and accountable data processing in accordance with standards and best practices, they do not tackle the nature of the objectives pursued by the massive deployment of data.[74] In this respect, as set out by research developed by the Global Partnership on Artificial Intelligence,[75] data justice encompasses not only a focus on data openness, data sharing, data governance, data quality, and transparent and non-discriminatory algorithms, but also a focus on the targeted objectives, when operators, whether public or private, run algorithms based on data and deliver assessments, forecasts, or decisions. Therefore, it is not only data themselves that are likely to cause harm, but even in the event that data are not biased, the aims pursued, too, could potentially threaten fundamental rights and freedoms.

It is certainly a complex challenge for legislators to address all potential data uses that either cause or are likely to cause significant harm to people; moreover this is so because the legislator is requested to take into account competing interests of businesses and lobbies that often push in different directions. This has been proven by the recent "Joint industry call for a risk-based AI Act that truly fosters innovation": it contends that "the list of prohibited AI systems would create unnecessary red tape and legal uncertainty."[76]

Consequently, it seems that the approach to the objectives served by practices involving artificial intelligence, as adopted by the EU legislator, especially with regard to data-fed systems, requires further discussion and insight. This is primarily due to the current uncertainty surrounding the scope of the objectives that certain artificial intelligence systems can enable, especially with regard to their ability to cause harm to people or be a likely cause of harm (above all, foundation models and general-purpose AI systems: Art. 1, par. 1, AIA.[77] Hence, a more comprehensive evaluation and adjustment are required concerning the boundaries between unacceptable goals pursed by the deployment of big data (as currently outlined in Art. 5 of the AIA) and high-risk systems (as currently outlined in Art. 6 of the AIA). Therefore, it is the scope of Art. 5 that requires a more in-depth evaluation. This is to place greater emphasis on

[74] C. D'Ignazio, L.F. Klein, *Data feminism*, Cambridge 2020, p. 60, denounce the insufficiency of a procedural and data governance approach when addressing data ethicists' position. M. Veale, F. Zuiderveen Borgesius, *Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach*, "Computer Law Review International" 2021, no. 4, underline the inadequacy of the EU Draft Artificial Intelligence Act, and equally denounce the shortcomings of the prohibitions listed in Art. 5 of the proposal.

[75] *Advancing Data Justice Research and Practice – An Interim Guide for Policymakers for the 2022 AI UK event*, https://gpai.ai/projects/data-governance/data-justice/ [accessed: 2023.10.14].

[76] https://ccianet.org/library/joint-industry-call-for-a-risk-based-ai-act-that-truly-fosters-innovation/ [accessed: 2023.10.14].

[77] See also: Art. 28b Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD), https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html [accessed: 2023.10.26].

the idea that data justice is a multifaceted concept that encompasses not only data, their quality, and their ways of processing (as is done by Art. 6 and following articles), but also the objectives pursued through their processing.

In this light, an "instrumental justice"[78] approach needs to be better implemented and deepened by the legislator. Indeed, this seems to be the effort that the EU legislator has tried to undertake: it is worth noting that during the legislative process for the approval of the EU's draft AIA, the European Parliament expanded the list of artificial intelligence systems prohibited within the EU.[79] This expansion included: the prohibition of *ex-post* remote biometric identification systems, in addition to real-time remote biometric identification systems; the deployment of sensitive characteristics for biometric categorisation, predictive policing, and emotion recognition; and indiscriminate scraping of biometric data from social media or CCTV footage to create facial recognition databases. Likewise, the opinion delivered on the draft AIA by the European Economic and Social Committee recommended expanding the lists of the AI systems banned from the EU.[80]

Caution proves especially beneficial when certain applications are built on uncertain theoretical foundations, as was the case with the Basic Emotion Theory (BET).[81] Concerning this the AIA has resulted in a ban on the practice of biometric categorization and the associated emotion recognition systems, whereas in the original European Commission proposal, these practices were neither forbidden nor classified as high-risk systems.[82] In a similar way, and as a relevant instance, the current Annex III provisions on high-risk AI systems encompass systems intended to be used for determining access to certain essential public services or activities (education and vocational training, recruitment: i.e. Annex III, point 3 and point 4), or monitoring the

---

78 According to the perspective described by R. Heeks, J. Renken, *Data Justice for development*…, p. 4, "concerns about justice only relate to the impact of the use of that data." See also quotation n. 4.

79 Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – P9 TA(2023)0236.

80 In its opinion (INT/940, par. 4.8), the EESC called for "a ban on use of AI for automated biometric recognition in publicly and privately accessible spaces (such as recognition of faces, gait, voice and other biometric features), except for authentication purposes in specific circumstances (for example to provide access to security sensitive spaces): a ban on use of AI for automated recognition of human behavioural signals in publicly and privately accessible spaces; a ban on AI systems using biometrics to categorise individuals into clusters based on ethnicity, gender, political or sexual orientation or other grounds on which discrimination is prohibited under Art. 21 of the Charter; a ban on the use of AI to infer emotions, behaviour, intent or traits of a natural person, except for very specific cases, such as some health purposes, where patient emotion recognition is important."

81 The Basic Emotion Theory (BET), developed by psychologist Paul Ekman in the 1960s, suggests that it is possible to understand people's emotions based on their facial expressions. The psychologist also argued that his theory had universal applicability because the expressions are the same for all human beings. However, over the years, various studies have demonstrated the invalidity of BET, since how a human being manifests his/her emotions changes according to different socio-cultural environments.

82 https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236EN.html [accessed: 2023.10.14].

behaviour of students (i.e. Annex III, point 3). In this respect, their inclusion in the list of the forbidden AI systems should better comply with the concept of "instrumental" data justice, since these purposes (monitoring students or determining access to education and job) have the potential to infringe people's fundamental rights (access to education or work) or people's free will and behaviour (especially when monitoring tools are employed towards students at certain stages of their development); or, in any case, because similar tools could foster blackmail practices.

An instrumental justice approach could also fit the purpose of IP protection, because its focusing on the goals addressed by the deployment of data can help to strike a fair balance between conflicting interests and the consequent definition of the adequate exceptions in copyright law.

The objection to a wider list of prohibited AI practices is based on concerns about stifling innovation and harming competition as a consequence of a too rigid approach. Nevertheless, it is also true that such an approach introduces a criterion of certainty that holds value at this "state of the (uncertain) art" for the cost-benefit assessment usually practiced by market operators.

Consequently, a precautionary approach that re-assesses certain purposes addressed by data deployment, more specifically by means of artificial intelligence practices, moving them away from the existing high-risk artificial intelligence systems list in order to integrate them into the list of banned artificial intelligence practices in the EU, is considered valuable for two primary reasons of "instrumental" data justice: on one hand, the significance of the interests involved (fundamental rights and freedoms) and the uncertainty (at the state of the art) of the scope and probability of potential harms; on the other, the enhancement of legal certainty that implies a rebalance of the boundaries between interests involved in the multifaceted concept of data justice and those involved in freedom to conduct business or control of public authorities over citizens.

Conclusions

The pieces of EU legislation and EU initiatives seem to comply with (part of) the data justice approach. They foster data openness and data sharing (Data Governance Act and Data Act); they provide for new digital rights (i.e. transparency, free portability of data for users and third parties, submission of complaints; see here: Data Market Act and Data Services Act, Artificial Intelligence Act); they focus on technical and procedural fulfilments and a consequent data governance framework aimed to safeguard data quality (primarily, the Artificial Intelligence proposal). All this gives evidence of the effort made by the legislator in safeguarding the interests of individuals and groups, while promoting innovation and acknowledging the fluid and relative nature of data quality, which depends on specific objectives and is consequently challenging to define in precise legal terms.

However, some substantial safeguards are loosely defined. In this respect, what makes the substance of data quality is mainly delegated to private and technical standards; transparency, fairness and accountability in data processing are enacted by procedural obligations; prohibited uses remain confined in a limited list; and copyright exceptions are ill-balanced in reference to the purposes that they address.

At this time, it is crucial to emphasize one key issue: such an approach complies with only a part of the issues involved in data justice. Thus, a more comprehensive approach to data justice calls for a more incisive intervention by public authorities.

The legal implications at stake, especially in relation to social justice outcomes, are too significant to be overcome by and to be limited to procedures and rights of access or addressing complaints in order to safeguard individuals or groups. Thus, these legal implications should not be reduced to mere technical and procedural requirements to be implemented by operators, overseen and enforced by supervisory authorities, or to new enforceable digital rights.

In the light of the previous arguments, data justice needs to receive legal enshrinement in order to encompass all the multifaceted aspects described. Consequently, it should not be confined solely to personal data protection or to the broader collective concerns about which data are collected and how they are collected and processed. Nor should it be restricted to questions of data ownership, data openness, data sharing, relevant and underlying procedural and governance issues, or new digital rights. It should also address the purposes for which such data are utilized.

Literature

Allyn B., *'New York Times' considers legal action against OpenAI as copyright tensions swirl*, https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl.

Bogle A., *New York Times, CNN and Australia's ABC block OpenAI's GPTBot web crawler from accessing content*, https://www.theguardian.com/technology/2023/aug/25/new-york-times-cnn-and-abc-block-openais-gptbot-web-crawler-from-scraping-content.

Bygrave L.A., *Hardwiring Privacy* [in:] *The Oxford Handbook of Law, Regulation, and Technology*, eds. R. Brownsword, E. Scotford, K. Yeung, Oxford 2018.

Cheney-Lippold J., *We are Data: Algorithms and the Making of Our Digital Selves*, New York 2017.

Claburn T., *Microsoft and GitHub are still trying to derail Copilot code copyright legal fight*, https://www.theregister.com/2023/07/01/microsoft_github_copilot/.

Custers B., *New digital rights: Imagining additional fundamental rights for the digital era*, "Computer Law and Security Review" 2022, vol. 44.

Custers B., Fosch-Villaronga E., *Law and Artificial Intelligence – Regulating AI and Applying AI in Legal Practice*, The Hague 2022.

David E., *Artists can use a data poisoning tool to confuse DALL-E and corrupt AI scraping*, https://www.theverge.com/2023/10/25/23931592/generative-ai-art-poison-midjourney.

De Gregorio G., *Digital Constitutionalism in Europe – Reframing Rights and Powers in the Algorithmic Society*, Cambridge 2022.

Dencik L., Sanchez-Mondero J., *Data Justice*, "Internet Policy Review" 2022, vol. 11, no. 1.

DiFeliciantonio Ch., *Authors Michael Chabon, David Henry Hwang sue OpenAI over copyright concerns*, https://www.sfchronicle.com/tech/article/michael-chabon-open-ai-lawsuit-copyright-18360019.php.

D'Ignazio C., Klein L.F., *Data feminism*, Cambridge 2020.

van Dijck J., *Datification, dataism and dataveillance: Big data between scientific paradigm and ideology*, "Surveillance & Society" 2014, vol. 12, no. 2.

Ebers M., *Standardizing AI – The Case of the European Commission's Proposal for an Artificial Intelligence Act* [in:] *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*, eds. L. Di Matteo, C. Poncibò, M. Cannarsa, Cambridge 2022.

Floridi L., *The fourth Revolution. How the infosphere is Reshaping Human Reality*, Oxford 2014.

Foucault M., *La naissance de la biopolitique. Course au Collège de France (1978–1979)*, Paris 2004.

Getty Images, *Statement of 17 January 2023*, https://newsroom.gettyimages.com/en/getty-images/getty-images-statement.

Grasso G.A., *GDPR Feasibility and Algorithmic Non-Statutory Discrimination*, Naples 2022.

Haring B., *BBC Will Block ChatGPT AI from Scraping Its Content*, https://deadline.com/2023/10/bbc-will-block-chatgpt-from-scraping-its-content-1235566868/.

Heeks R., Renken J., *Data Justice for development: What would it mean?*, Development Informatics Working Paper Series, No. 63, 2016.

Johnson J.A., *From open data to information justice*, "Ethics and Information Technology" 2014, vol. 16, no. 4.

Kitchin R., *The Data Revolution – A critical Analysis of Big Data, Open Data & Data infrastructures*, London 2022.

Laskowska-Witak E., *Komentarz do dyrektywy o prawach autorskich w ramach jednolitego rynku cyfrowego*, LEX/el. 2019.

Milmo D., *The Guardian blocks ChatGPT owner OpenAI from trawling its content*, https://www.theguardian.com/technology/2023/sep/01/the-guardian-blocks-chatgpt-owner-openai-from-trawling-its-content.

O'Brien M., *Photo giant Getty took a leading AI image-maker to court. Now it's also embracing the technology*, https://apnews.com/article/getty-images-artificial-intelligence-ai-image-generator-stable-diffusion-a98eeaaeb2bf13c5e8874ceb6a8ce196.

O'Connor R., *Introduction to Diffusion Models for Machine Learning*, https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/.

Pollicino O., *Judicial Protection of Fundamental Rights on the Internet. A Road Towards Digital Constitutionalism?*, Oxford 2021.

Ross A., *The Industries of the Future*, New York 2016.

Ruppert E., Isin E., Bigo D., *Data Politics*, "Big Data & Society" 2017, vol. 2, no. 5.

Russel S., Norvig P., *Artificial intelligence – A Modern Approach*, Hoboken 2021.

Senkowska N., *"Trening" sztucznej inteligencji: co z prawami twórców dzieł, na których ćwiczy*, "Rzeczpospolita", 13.07.2023.

Small Z., *Sarah Silverman Sues OpenAI and Meta Over Copyright Infringement*, https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html.

Taylor L., *What is data justice? The case for connecting digital rights and freedoms globally*, "Big Data & Society" 2017, vol. 4, no. 2.

Thaler R.H., Sunstein C., *Improving decisions about Health, Wealth and Happiness*, New York 2009.

Wachowska A., Ręgorowicz M., *ChatGPT w praktyce – najważniejsze kwestie prawne*, https://www.traple.pl/chatgpt-w-praktyce-najwazniejsze-kwestie-prawne/.

Veale M., Zuiderveen Borgesius F., *Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach*, "Computer Law Review International" 2021, no. 4.

Vincent J., *AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit*, https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart.

Zuboff S., *The Age of Surveillance Capitalism – The Fight for a Human Future at the New Frontier of Power*, London 2019.

Summary

Nadia Maccabiani, Anna Podolska, Ewelina Szatkowska

How Artificial Intelligence Learns. Legal Aspects of Using Data in Machine Learning

Recalling the debate around data justice in order to highlight which parts of this multifaceted concept have been endowed with legal relevance by EU legislation or initiatives, the paper argues that the EU should implement a more "instrumental" approach to data justice. This perspective emphasizes a stronger focus on the purposes addressed by the deployment of data within AI systems.

Keywords: data justice, artificial intelligence, intellectual property, data quality.

Streszczenie

Nadia Maccabiani, Anna Podolska, Ewelina Szatkowska

Jak uczy się sztuczna inteligencja. Prawne aspekty wykorzystywania danych w uczeniu maszynowym

Tocząca się debata na temat sprawiedliwości danych daje możliwość wskazania, które elementy tej wielowymiarowej koncepcji zostały odzwierciedlone w prawodawstwie oraz inicjatywach UE. W artykule argumentuje się, że UE powinna wdrożyć bardziej „instrumentalne" podejście do sprawiedliwości danych. Perspektywa ta podkreśla konieczność silniejszego skupienia się na celach, którym ma służyć wykorzystanie danych w systemach AI.

Słowa kluczowe: sprawiedliwość danych, sztuczna inteligencja, własność intelektualna, jakość danych.