Towards Future-Proof, Rights-Respecting Automated Data Collection: An Examination of European Jurisprudence

Olga Kokoulina[*]

ABSTRACT

The data scraping and data collection that Application Programming Interfaces (APIs) enable are ubiquitous means of automatically and instantaneously gathering large amounts of online data. Anyone can leverage the capabilities of internet infrastructure to engage in data collection, yet the subjects of the data collection are often unaware of the full extent to which this personal data harvesting occurs. The accessibility and discreetness of large-scale automated online data collection test the overall fitness of the European data protection framework, the General Data Protection Regulation (GDPR), to provide control to individuals over their personal data while promoting innovation and ensuring legal certainty for all parties concerned.

This Article explores the extent to which the jurisprudence of the Court of Justice of the European Union (CJEU) and the guidance of data protection authorities have accommodated this technological transformation. Data controllers have been the apparent focal point of institutional responses. Absent further clarification, however, an approach focused on data controllers is likely both insufficient and unsustainable. This Article outlines areas for further clarification and additional focus—namely, the concept of "public accessibility" of online data, the notion of "reasonable expectations," the capacity of the "fairness" principle, and an increased emphasis on technological infrastructure and rights-respecting AI development.

[*] Assistant Professor, Ph.D., MSc. (Oxford), LL.M. Centre for Private Governance (CEPRI), Faculty of Law, University of Copenhagen. The Author would like to express her sincere thanks to Professors Henrik Udsen and Jens Schovsbo, and the VANDERBILT JOURNAL OF ENTERTAINMENT AND TECHNOLOGY editorial team.

TABLE OF CONTENTS

I. INTRODUCTION

Collecting information about individuals has been a ubiquitous practice throughout the history of human civilization.[1] Record curators employed different media across various cultures and periods.[2] From inscriptions on clay tablets of Mesopotamia to epigraphic texts on wood boards and papyrus sheets in Greece and Rome, maintaining records has been a consistently core component of administrative practices.[3]

1. *See, e.g.*, GUNNAR THORVALDSEN, CENSUS AND CENSUS TAKERS 3 (2017); ANDREW WHITBY, THE SUM OF THE PEOPLE: HOW THE CENSUS HAS SHAPED NATIONS, FROM THE ANCIENT WORLD TO THE MODERN AGE 3–4 (2020).

2. *See, e.g.*, Melville J. Herskovits, *Population Statistics in the Kingdom of Dahomey*, 4 HUM. BIOLOGY 252, 255 (1932) (describing the use of pebbles and sacks for record keeping in a West African kingdom); Stephen Chrisomalis, *The Origins and Co-Evolution of Literacy and Numeracy*, *in* THE CAMBRIDGE HANDBOOK OF LITERACY (David R. Olson & Nancy Torrance eds., 2009) (describing employment of notched bones, clay tokens, bundles of cotton, or wool cords as administrative mechanisms and recordkeeping tools in Mesopotamia, Mesoamerica, and China).

3. *See, e.g.*, GUILLERMO ALGAZE, ANCIENT MESOPOTAMIA AT THE DAWN OF CIVILIZATION: THE EVOLUTION OF AN URBAN LANDSCAPE 133–39 (2008) (describing clay tablet recordkeeping as

The tradition of keeping records in specialized repositories dates back to ancient times, yet it wasn't until relatively recently that such archival storage became centralized.[4] In medieval Europe, each town, guild, bishopric, or landed estate maintained its records independently.[5] The recordkeeping was primarily local and limited to registering land transactions, birth, marriage, death facts, and other legal documents.[6] The emergence of the early modern state in the late medieval and early modern periods marked the beginning of more centralized data collection efforts that culminated in the establishment of national archives and statistical offices throughout nineteenth-century Europe.[7]

The transformation from local to centralized data collection was a multifaceted process that spanned several centuries.[8] The needs of emerging nation-states, technological innovations, and the increasing complexity of governance drove this transformation.[9] Technological developments played a particular role in revolutionizing the methods and scope of data collection by enhancing the accuracy, speed, and volume of data that could be gathered and analyzed.[10] The widespread introduction of mainframe computers in the 1960s and the development of the internet three decades later further catalyzed this

---

a planning management mechanism); ANDREW N. SHERWOOD, MILORAD NIKOLIC, JOHN W. HUMPHREY & JOHN P. OLESON, GREEK AND ROMAN TECHNOLOGY: A SOURCEBOOK OF TRANSLATED GREEK AND ROMAN TEXTS 643–646 (2d ed. 2019); GEOFFREY YEO, RECORDS, INFORMATION AND DATA: EXPLORING THE ROLE OF RECORD-KEEPING IN AN INFORMATION CULTURE 1–6 (2018).

4. *See* YEO, *supra* note 3; Michel Duchein, *The History of European Archives and the Development of the Archival Profession in Europe*, 55 AM. ARCHIVIST 14, 16 (1992).

5. *See* YEO, *supra* note 3, at 12; *see generally* MICHAEL T. CLANCHY, FROM MEMORY TO WRITTEN RECORD: ENGLAND 1066–1307 (3d ed., 2013) (providing a detailed account of the development of such practices in medieval England).

6. *See* Eric Ketelaar, *Records Out and Archives In: Early Modern Cities As Creators of Records and As Communities of Archives*, 10 ARCHIVAL SCI. 201, 206 (2010).

7. *See generally* Angela Andreani, *European Renaissance Archives*, OXFORD RESEARCH ENCYCLOPEDIA OF LITERATURE (2022), https://doi.org/10.1093/acrefore/9780190201098.013.1316; NICO RANDERAAD, STATES AND STATISTICS IN THE NINETEENTH CENTURY: EUROPE BY NUMBERS (Debra Molnar, trans., 2010).

8. *See* Lars Behrisch, *Statistics and Politics in the 18th Century*, 41 HIST. SOC. RSCH. HISTORISCHE SOZIALFORSCHUNG 238, 241 (2016).

9. *See* Stuart Woolf, *Statistics and the Modern State* 31 COMPAR. STUD. IN SOC'Y AND HIST. 588, 589 (1989).

10. *See, e.g.*, Christine von Oertzen, *Machineries of Data Power: Manual Versus Mechanical Census Compilation in Nineteenth-Century Europe*, 32 DATA HISTORIES 129 (2017) (a case study of the mechanization of census compilation); MARTIN CAMPBELL-KELLY & WILLIAM ASPRAY, COMPUTER: A HISTORY OF THE INFORMATION MACHINE 19–23 (1996) (describing Herman Hollerith's contribution to the mechanization of the US census).

transformation.[11] Mainframe computers increased data processing capabilities, centralized data storage, and automated routine data collection tasks, enabling organizations—both public and private—to store and analyze data with unprecedented efficiency.[12] The internet ushered in an era of global connectivity, facilitated the instantaneous acquisition of data, and diversified the array of data types and origins available for such a collection.[13]

This transformation has elicited profound interest from a variety of actors who have been keen on harnessing this technological advancement for a wide array of purposes.[14] Cases like *Cambridge Analytica*[15] and *Clearview AI*[16] provide captivating and thought-provoking discussions on fairness of data processing and the acceptable use of collected data.[17] At the same time, they also offer telling examples

---

11. Alan F. Westin, *Databanks in a Free Society: A Summary of the Project on Computer Databases*, *in* PRIVACY: THE COLLECTION, USE, AND COMPUTERIZATION OF PERSONAL DATA 92 (1973).

12. *See, e.g.*, "Computers and Privacy" in *Privacy and the Law*, A REPORT BY JUSTICE (INTERNATIONAL COMMISSION OF JURISTS). 54–55 app. (1970) (outlining respective concerns); Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy* 53 STAN. L. REV. 1393, 1400–09 (2001).

13. ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD), THE OECD PRIVACY FRAMEWORK 86–90 (2013).

14. *See, e.g.*, Fed. Trade Comm'n, Data Brokers: A Call for Transparency and Accountability 3 (May 2014); BRUCE SCHNEIER, DATA AND GOLIATH: THE HIDDEN BATTLES TO COLLECT YOUR DATA AND CONTROL YOUR WORLD 8 (2015) (introducing modern data ecosystems and data collection patterns).

15. The *Cambridge Analytica* case refers to a significant controversy that emerged in early 2018 over the misuse of personal data by the British political consulting firm Cambridge Analytica. Having acquired the data of 50 million Facebook users and created 30 million personality profiles, the company subsequently sold the data to be used to optimize electoral outcomes. *See* Carole Cadwalladr & Emma Graham-Harrison, *Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach*, THE GUARDIAN (Mar. 17, 2018, 6:03 PM), https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election [perma.cc/3PTA-WU3T].

16. *Clearview AI* concerns the exposed activity of Clearview AI, a facial recognition software maker that accumulated an extensive collection of training data through scraping billions of photos from the internet. As was established, numerous law enforcement agencies and police departments globally had made use of the resource. Since this exposure, Clearview AI has faced allegations of breaching various privacy regulations in multiple countries, including in Europe and the United States. *See* Kashmir Hill, *The Secretive Company That Might End Privacy as We Know It*, N.Y. TIMES (Nov. 2, 2021), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html [perma.cc/YKT8-USA4].

17. *See, e.g.*, *Hearing on the Facebook/Cambridge Analytica Case—Part 2*, EUR. PARLIAMENT: COMMS. 8TH PARLIAMENTARY TERM (2014–2019) (June 25, 2018), https://www.europarl.europa.eu/committees/en/archives/8/libe/events/events-hearings [perma.cc/SZ7Q-73HS]; *Challenge Against Clearview AI in Europe,* PRIVACY INT'L, https://privacyinternational.org/legal-action/challenge-against-clearview-ai-europe [perma.cc/NJ28-Z96R] (last visited Jan. 26, 2024) (summarizing the Clearview AI investigations in Europe).

of the prevalence, ease, and often discreetness of automated collection of data that internet users perpetually generate.[18]

The collection of this purportedly public[19] data transcends several legal fields, including cybercrime,[20] intellectual property,[21] contracts, and competition law and policy.[22] Each perspective presents a distinct viewpoint and demands a particular level of engagement with technological infrastructure. For example, in cases of data extraction, automated data collection could be approached from a position of potential infringements and remedies. Appraising technological means of and obstacles to data collection then becomes a key measure for determining whether data access qualifies as "authorized"[23] or if a breach of material covenants in, for example, a website's terms of service has occurred.[24]

The discussion among European policy makers and scholars on data ownership rights could provide an alternative framework for defining publicly accessible online data.[25] As a point of departure for

18. *See, e.g.*, CONFEDERATION OF EUR. DATA PROT. ORG. WORKING GRP., GENERATIVE AI: THE DATA PROTECTION IMPLICATIONS 6, 15 (Oct. 16, 2023), https://cedpo.eu/wp-content/uploads/generative-ai-the-data-protection-implications-16-10-2023.pdf [perma.cc/6TQP-74WY] (discussing common sources of training data in generative AI).

19. *See* Michael Zimmer, *But the Data is Already Public: On the Ethics of Research in Facebook*, 12 ETHICS & INFO. TECH. 313, 315 (2010); *Joint Investigation of Clearview AI, Inc. by the Office of the Privacy Commissioner of Canada*, OFFICE OF THE PRIVACY COMM'R OF CANADA (Feb. 2, 2021), https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2021/pipeda-2021-001/ [perma.cc/BP6Y-WPP5] (noting Clearview's assertion of only collecting "publicly accessible data").

20. Could a mere practice of "data scraping" qualify as "intentional accessing to a computer without authorization or exceeding authorized access" under computer crime legislation?

21. Is it a publicly available database, which, "by reason of the selection or arrangement of their contents," constitutes the author's own intellectual creation? Is it, perhaps, a database the creation of which demands a substantial investment? What if scraping involved the use of the initial creator's names/logos?

22. What if it has the potential to promote the competition on the market by allowing small and innovative companies to get access to much needed data? *See* Case C-30/14, Ryanair Ltd v. PR Aviation BV, ECLI:EU:C:2015:10, ¶ 17 (Jan. 15, 2015).

23. *See, e.g.*, Craigslist, Inc. v. 3Taps Inc., 964 F. Supp. 2d 1178, 1182 (N.D. Cal. 2013).

24. *See, e.g.*, Benjamin L.W. Sobel, *A New Common Law of Web Scraping*, 25 LEWIS & CLARK L. REV. 147, 200 (2021).

25. *See generally Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Building a European Data Economy*, COM (2017) 9 final (Jan. 10, 2017) [hereinafter Committee Communication]; MAX PLANCK INST. FOR INNOVATION & COMPETITION, POSITION STATEMENT OF THE MAX PLANCK INSTITUTE FOR INNOVATION AND COMPETITION OF 26 APRIL 2017 ON THE EUROPEAN COMMISSION'S "PUBLIC CONSULTATION ON BUILDING THE EUROPEAN DATA ECONOMY" (Apr. 26, 2017); Andreas Wiebe, *Protection of Industrial Data – A New Property Right for the Digital Economy?*, 65 GEWERBLICHER RECHTSSCHUTZ UND URHEBERRECHT INTERNATIONALER

respective debates, the EU legal order does not generally provide for exclusive property rights with respect to "machine-generated data" despite its economic value and firms' ability to de facto control and profit from it.[26] This absence of property right protection sparked a contentious debate about the potential need for a new property right for nonpersonal data.[27] The debate focused on the proposition of an intellectual property (IP)-style right to establish data markets and guarantee equitable distribution of data benefits.[28] Although conceptually appealing, opponents of the new IP right regime extensively challenged this proposition, cautioning that such a right might stifle competition and innovation by limiting the availability of information in the public domain.[29]

The discourse provides valuable insights for the discussion on publicly accessible data, encouraging participation in normative debates about societal interests that merit protection in defining which data should be accessible to the public. By explicitly outlining policy considerations, both economic and otherwise, the discussion surrounding the limits of online data accessibility could draw on IP-fashioned discourse over incentives, risks, and protection opportunities for a variety of relevant stakeholders.[30] In this context, the technological dimension of data collection could help identify the limits of data creation and accessibility, reflecting how societal interests in

---

TEIL 877 (2016); Wolfgang Kerber, *A New (Intellectual) Property Right for Non-Personal Data? An Economic Analysis*, GEWERBLICHER RECHTSSCHUTZ UND URHEBERRECHT INTERNATIONALER, 11/2016, 989–99 (2016).

26. The scope of sui generis database right remains a contentious issue. *See, e.g.*, EUROPEAN COMMISSION, DIRECTORATE-GENERAL FOR COMMUNICATIONS NETWORKS, CONTENT & TECHNOLOGY, STUDY IN SUPPORT OF THE EVALUATION OF DIRECTIVE 96/9/EC ON THE LEGAL PROTECTION OF DATABASES (2018). The EU Commission defines machine-generated nonpersonal data as data "created without the direct intervention of a human by computer processes, applications or services, or by sensors processing information received from equipment, software or machinery, whether virtual or real." *See* Committee Communication, *supra* note 25; *see, e.g.*, JOSEF DREXL, DATA ACCESS AND CONTROL IN THE ERA OF CONNECTED DEVICES: STUDY ON BEHALF OF THE EUROPEAN CONSUMER ORGANISATION BEUC 2 (2018).

27. Herbert Zech, *A Legal Framework for a Data Economy in the European Digital Single Market: Rights to Use Data*, 11 J. OF INTELL. PROP. L. & PRAC. 460 (2016); Wolfgang Kerber, *Governance of Data: Exclusive Property vs. Access*, 47 IIC-INTERNATIONAL REV. OF INTELL. PROP. AND COMPETITION L. 759 (2016).

28. *See* Zech, *supra* note 27; Kerber, *supra* note 27.

29. Kerber, *supra* note 25 at 3 nn. 4–5 (providing a literature overview).

30. In this context, IP scholars' discussions on the concepts and limits of a public domain could be instructive as well. *See, e.g.*, Pamela Samuelson, *Mapping the Digital Public Domain: Threats and Opportunities* 66 LAW & CONTEMPORARY PROBLEMS 147 (2003).

fostering a data-centric economy interact with the motivations of individual stakeholders to generate and share data.[31]

Approaching automated collection of online data from the perspective of EU data protection law presents yet another insightful angle. The increasing ease, scale, and accessibility of automated data collection challenge the fitness of existing data protection frameworks to account for nonlinear data processing and provide for a future-proof solution.[32] As the internet has reconfigured information consumption at large, it has also shaped data accessibility in rather utilitarian terms.[33] Technologically speaking, "publicly accessible online data" commonly refers to data that can be "scraped" or "queried" by anyone interested in doing so.[34] Such automated data collection operates on a predefined set of commands and adheres to the structure of the data it is instructed to harvest, lacking an innate ability to recognize the context or sensitivity of the content it gathers.[35] The technology is incapable of distinguishing between personal and nonpersonal data, as well as between ordinary and special categories of personal data, the latter of which requires higher levels of protection due to their sensitive nature.

This Article analyzes technological developments in data collection and considers their impact in the decision-making practices of the Court of Justice of the European Union (CJEU), the Article 29 Working Party, and the European Data Protection Board (EDPB).[36] Using material scope and data controllership concepts as examples, this Article asserts that judicial responses to automated data collection are

---

31. *See*, *e.g.*, Josef Drexl, *Designing Competitive Markets for Industrial Data - Between Propertisation and Access* 16 MAX PLANCK INST. FOR INNOVATION & COMPETITION RSCH. PAPER NO. 1, 2 (2016) (discussing data ownership and data access in the data-driven environment of the Internet of Things); Ignacio N. Cofone, *The Dynamic Effect of Information Privacy Law*, 18 MINN. J. SCI. & TECH. 517, 521–22 (2017) (discussing relevant mechanisms for protecting privacy entitlements).

32. *See, e.g.*, Omer Tene, *Privacy Law's Midlife Crisis: A Critical Assessment of the Second Wave of Global Privacy Laws*, 74 OHIO ST. L.J. 1217, 1219 (2013); Tal Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 SETON HALL L. REV. 995, 995 (2016).

33. *See, e.g.*, Helen Margetts & Cosmina Dorobantu, *Computational Social Science for Public Policy*, *in* HANDBOOK OF COMPUTATIONAL SOCIAL SCIENCE FOR POLICY 3 (Eleonora Bertoni eds., 2023) (discussing opportunities and actual uses of data in computational social science). Respective research is firmly grounded on the recognition of the benefits of wide data accessibility and its use across society: given that the internet facilitates the availability of data, collection and subsequent use of such data should be advanced as a means of promoting social welfare for all.

34. *See* discussion *infra* Sections II.B.1, 2.

35. *See* discussion i*nfra* Section III.A.

36. The Article 29 Working Party (Art. 29 WP) was established by Directive 95/46/EC (Art.30) and Directive 2002/58/EC (Art.15) as an independent European advisory body on data protection and privacy. It was replaced by the European Data Protection Board upon entry into the force of the GDPR on May 25, 2018.

inconsistent and do not offer a sustainable solution for regulating data-driven innovation without undermining individual control over personal data. Part II provides a general overview of publicly accessible online data and presents the prevalent means of automating its collection. Part III incorporates technological insights into the assessment of material scope and data controllership in EU data protection laws. Part IV discusses the shortcomings of judicial responses to these developments, highlighting areas of uncertainty. Part V suggests desirable clarifications to keep the law abreast of technological advancement. Finally, Part VI offers concluding remarks.

II. DEFINING "PUBLICLY ACCESSIBLE ONLINE DATA"

Many initiatives are potentially relevant in defining the concept of "publicly accessible data." The European Union's objective of facilitating the flow of data (both personal[37] and nonpersonal[38]) has resulted in numerous legislative undertakings seeking to remove potential obstacles to the unimpeded movement of data.[39] A series of initiatives also exists to explore means of strengthening data-sharing mechanisms, enable access to and reuse of Public Sector Information across the European Union,[40] and regulate digital platforms.[41] These

37. Defined as "any information relating to an identified or identifiable natural person." Regulation (EU) 2016/679 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 (General Data Protection Regulation), 2016 O.J. (L 119)) [hereinafter GDPR] art.4(1).

38. In principle, nonpersonal data embraces two categories of data: data that does not relate to an identified or identifiable natural person (e.g., weather conditions) and data that was once personal but no longer is. *See id.* at Rec. 29 (on anonymized data).

39. *See, e.g.*, GDPR, *supra* note 37; Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a Framework for the Free Flow of Non-personal Data in the European Union, 2018 O.J. (L 303) 59.

40. *See, e.g.*, 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, 2019 O.J. (L 172) 56; Commission Implementing Regulation (EU) of 21.12.2022 Laying Down a List of Specific High-Value Datasets and the Arrangements for their Publication and Re-Use, 2023 O.J. (L 19) [hereinafter Open Data Directive]. In the EU, the regulation of public sector and publicly funded data is fundamentally guided by the principle that this data should be available for reuse, whether for commercial or noncommercial purposes. This approach aims to maximize the utility and value derived from public sector information (i.e. information held by EU Member state, regional, or local authorities), encouraging innovation and development by making such data accessible to individuals, businesses, and organizations. *See*, *e.g.*, Open Data Directive, Rec. 3, 4, 8, 10.

41. European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on the Digital Services Act: Improving the Functioning of the Single Market, 2020 O.J. (C 404) 2; 2021 O.J. (C 404) 31; *see also* Andrej Savin, *The EU Digital Services Act: Towards a More Responsible Internet*, 24 J. INTERNET L. 15, 21 (2021).

regulatory interventions undeniably inform and shape the contours of data accessibility (or lack thereof) by defining the type of data they seek to govern[42] and the objective they aim to achieve by proposing respective rules.[43]

This Article aims to deepen these discussions by focusing on a particular set of data often framed as publicly accessible online data. While the notion of publicly accessible online data itself seems rather intuitive, its actual contours are not particularly clear.[44] Publicly accessible online data is, technically, not an exact antipode of "offline data." Similarly, defining it as the opposite of "publicly restricted online data" is also marred with challenges, including determining the best method to comprehensively demarcate these restrictions.[45] Thus, a preliminary reflection on the specifics of publicly accessible online data proves instructive.

A. On a Baseline of "Public Accessibility"

In painting the initial contours of publicly accessible online data, a reflection on what public access should exclude helps elucidate these contours. At the outset, the general architecture and availability of data on the internet and the common classifications of the data that data analytics generate are important to consider.

As for the taxonomy of collected data, online data collection is an indispensable step in conducting big data research.[46] Essentially, the internet has significantly augmented research capacities: it has grown to be both an unprecedented data repository and a key tool for

42 *See, e.g.*, Directive (EU) 2019/1024 O.J. (L 172) 56.

43. *See, e.g.*, Regulation (EU) 2022/868 of the European Parliament and of the Council on European Data Governance and Amending Regulation (EU) 2018/1724 (Data Governance Act) (in particular, Art. 4 on a general prohibition of exclusive arrangements related to the reuse of data held by public sector bodies), 2019 O.J. (L 152) 1, 48 (requiring nonexclusivity); 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), 2022 O.J. (L 277) 1, 31.

44. *See* Zimmer, *supra* note 19, at 315; Danah Boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO., COMMC'N & SOC'Y 662, 672 (2012).

45. The endeavour would likely involve a systemic analysis of considering paywall conditions, firewall restrictions, compliance with anti-piracy regulations, and the general presence of internet filtering.

46. Defined as "the practice of combining huge volumes of diversely sourced information and analysing them, using more sophisticated algorithms to inform decisions." *See* European Data Protection Supervisor, *Opinion 7/2015, Meeting the Challenges of Big Data* 7 (Nov. 19, 2015).

conducting a wide range of inquiries.[47] Its "self-documenting and self-archiving" capabilities have resulted in an incredibly detailed data treasure trove of social interactions, daily transactions, and corresponding information created therefrom.[48]

The internet has equally been a formative and transformative tool itself. From virtual ethnography and online experiments to recruitment and interaction with research subjects, the internet has expanded, democratized, nuanced, and challenged traditional research methods.[49] Both quantitative and qualitative research have witnessed a rise in innovative approaches applied to a variety of research areas, and commercial as well as public entities have been pivotal forces in advancing this trend.[50] However, the degree and scale of performed data analysis have been generally uneven due to the varied amount of data, skills, and other resources available to researchers.[51]

Regarding commercial entities in particular, the data researchers generate in pursuit of predicting, for instance, consumer behavior, is often discussed in general terms of "observed" and "inferred" data.[52] A company's data mining activity typically includes

47. For an excellent introduction to computational research, *see generally* NIGEL G. FIELDING, GRANT BLANK & RAYMOND M. LEE, THE SAGE HANDBOOK OF ONLINE RESEARCH METHODS (2016).

48. Howard T. Welser, Marc Smith, Danyel Fisher & Eric Gleave, *Distilling Digital Traces: Computational Social Science Approaches to Studying the Internet*, *in* THE SAGE HANDBOOK OF ONLINE RESEARCH METHODS 117 (2008).

49. *Id.* at 121.

50. *See, e.g.*, *id.* By expanding access to big data, enabling new methods of data collection and analysis, and facilitating collaboration and communication among researchers and participants, the internet paved a way to exploring topics that were traditionally out of reach for research enquiry. Studies on the estimated impact of education policies on exam performance in all schools in England using public databases, exploration of social conventions of marginalized online communities, network analysis on digital platforms are some of the examples of innovative approaches in this context. *See* FIELDING ET AL., *supra* note 47) (providing examples and discussions on ethical, technical, and legal considerations of respective research); *see* DATA SCIENCE AND SOCIAL RESEARCH: EPISTEMOLOGY, METHODS, TECHNOLOGY AND APPLICATIONS (Carlo Natale Lauro, Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia & Marina Marino, eds., 2017) (providing an introduction on data science in general—defined broadly as a multidisciplinary approach grounded in statistical and computer science methods and supplemented by expertise of various domains).

51. *See* Ralph Schroeder & Jamie Halsall, *Big Data Business Models: Challenges and Opportunities*, 2 COGENT SOC. SCI. 1, 2 (2016).

52. *See* ARTICLE 29 DATA PROT. WORKING PARTY, GUIDELINES ON THE RIGHT TO DATA PORTABILITY 8, 10 (Apr. 5, 2017); WORLD ECONOMIC FORUM: RETHINKING PERSONAL DATA: A NEW LENS FOR STRENGTHENING TRUST 16 (2014), https://www3.weforum.org/docs/WEF_RethinkingPersonalData_ANewLens_Report_2014.pdf [perma.cc/U5Y2-XDMK]; EUR. DATA PROT. BD., GUIDELINES 8/2020 ON THE TARGETING OF SOCIAL MEDIA USERS 13 (2021), https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202008_onthetargetingofsocialmediausers_en.pdf [perma.cc/8Q3Z-GST5].

data obtained through recording users' interactions with services or devices ("observed" data).[53] Social media plug-ins or tracking pixels, GPS location, and financial transactions can serve as sources of this data.[54] Then, during the computational analysis of collected data, companies usually obtain what is commonly referred to as "inferred" data: predictions about consumer behavior and other issues of interest.[55] In this context, both "observed" and "inferred" data could qualify as "online data" in its traditional sense as data originating from or connected to the internet. However, this "observed" or "inferred" data is not easily confined to the category of publicly accessible data. The relevant corporate entities typically limit the dissemination of "observed" and "inferred" types to their organizational systems.[56] Moreover, for a variety of reasons, from IP rights to users' settings, this data is typically not available to the public at large.

This is not to say, however, that such data cannot become available and accessible online for reasons that are occasionally beyond the control of the data holding entity. The information on fairly large datasets "leak[ing] online" is not uncommon;[57] some data that is classified or otherwise excluded from public access is routinely available on a "Darknet," a subset of the deep web only accessible through special software.[58] However, these examples are rather extreme: while the data could be technically "online," it naturally belongs to a deeper content

53. EUR. DATA PROT. BD., GUIDELINES 8/2020, *supra* note 52.

54. *Id.*

55. *Id.* at 14.

56. *Id.* at 10. Organizational knowledge encompasses a wide range of information such as the organization's strategies, best practices, methods, and processes of working and decision-making algorithms. In the context of internal data management in particular, organizational knowledge often includes algorithms used for extracting inferences and a body of data collected therewith. Trade secrets and contractual arrangements are among the widely employed means of protecting such knowledge. *See*, *e.g.*, Tanya Aplin, Alfred Radauer, Martin A. Bader & Nicola Searle, *The Role of EU Trade Secrets Law in the Data Economy: An Empirical Analysis*, 54 ICC INT'L REV. INTELL. PROP. & COMPETITION L. 826 (2023).

57. *See* Tony Romm, *Facebook Says a New Bug Allowed Apps to Access Private Photos of Up To 6.8 Million Users*, WASH. POST (Dec. 14, 2018, 1:24 PM), https://www.washingtonpost.com/technology/2018/12/14/facebook-says-newbug-allowed-apps-access-private-photos-up-million-users [perma.cc/D5V5-N3YV]; Rachna Khaira, *Rs 500, 10 Minutes, and You Have Access to Billion Aadhaar Details*, THE TRIBUNE (Jan. 5, 2018, 1:58 PM), http://www.tribuneindia.com/news/nation/rs-500-10-minutes-and-you-haveaccess-to-billion-aadhaar-details/523361.html [perma.cc/64ML-F947]; Aaron Holmes, *533 Million Facebook Users' Phone Numbers and Personal Data Have Been Leaked Online*, BUS. INSIDER (Apr. 3, 2021), https://www.businessinsider.com/stolen-data *have been* leaked-online [perma.cc/GWX4-TMNQ].

58. *See Israel: Police Looking at Chareidim In Theft of Population Database*, YESHIVA WORLD, JERUSALEM (Oct. 24, 2011, 10:49 AM), https://www.theyeshivaworld.com/news/headlines-breakingstories/106550/israel-police-looking-at-chareidim-in-theft-of-populationdatabase.html [perma.cc/MXT8-X6FT].

layer of the internet that is not easily accessible for the typical internet user. Put differently, it seems rather intuitive that public access should only embrace the visible part of the internet (commonly referred to as the "surface web"), while leaving out the "hidden" part of the internet (i.e., the "deep web").[59]

B. Public Accessibility as a Technological Construct: Automating Collection

Beyond this first contour, there are some necessary line variations that are prudent to consider. Accepting that internally generated and stored commercial data is not publicly accessible online data does not equate to saying that the remainder of the online data is within easy reach for the public. While online data might be "accessible" in principle, the specific terms of this access are actually a complex interaction between legal and technical aspects.[60]

At the outset, online data has no innate or acquired capacity to signal its originators' individual preferences for its dissemination and use.[61] There is no universally accepted standard of the "Privacy Commons" license attached to each set of data bits.[62] Nor is there a widely adopted data nutrition or genealogy database or registry against which it would be possible to run a check as to the origin of data and its originators' preferences for its use.[63] At present, operational "control

59. For a simple explanation of different content layers, *see generally* Paul McFedries, *The Language of the Dark Web*, IEEE SPECTRUM (Sept. 20, 2017), https://spectrum.ieee.org/the-language-of-the-dark-web [perma.cc/P6Z3-9Q3L]. On crawling that part of the web, *see generally* Abdullah Alharbi, Mohd Faizan, Wael Alosaimi, Hashem Alyami, Alka Agrawal, Rajeev Kumar & Raees Ahmad Khan, *Exploring the Topological Properties of the Tor Dark Web*, 9 IEEE ACCESS 21746, 21746 (2021).

60. A few of these aspects are contractual restrictions, IPR considerations, personal data protection, security measures, exclusion codes, Creative Commons, and Wikipedia communities.

61. *See* Jonathan Zittrain, *Privacy 2.0*, 2008 U. CHI. LEGAL F. 65, 106 (discussing data genealogy).

62. *Id.* Privacy Commons licenses are analogous to Creative Commons licenses for creative works. Creative Commons licenses are a set of copyright licenses that enable creators to grant certain permissions to others regarding the use of their work while still retaining some rights. Creative Commons licenses provide a flexible and standardized way for creators to share their creative works with the public by allowing them to specify how their works can be used, modified, and shared. *See, e.g.*, Lawrence Lessig, *The Creative Commons,* 65 MONT. L. REV. 1 (2004).

63. *But see, e.g.*, *Empowering Data Scientists and Policymakers with Practical Tools to Improve AI Outcomes*, THE DATA NUTRITION PROJECT, https://datanutrition.org/ [perma.cc/5TR8-532A] (last visited Mar. 4, 2023) (showcasing a data nutrition project). A possible registry could have included, for example, information on the original purpose of the processing and employed legal basis for such processing, recipients of data, period of storage, preferences as to further use

pods" for personal information allowing users to moderate data sharing and access do not yet exist.[64] However, some code-backed means for regulation are in operation.[65] Data is essentially embedded within the structure and architecture of the internet. This means that to access and gather the data, one is naturally compelled to follow the innate logic of the technological settings. For example, it is possible to manually retrieve data from a Wikipedia page, record users' reviews of a movie on IMDb, or take notes of someone's Facebook profile data. However, the non-automated export of this data is a tedious, time-consuming, and costly enterprise. As a manual endeavor, the scale of this work largely hinges on the availability of human resources and the ability to scroll, click, copy, and paste. While theoretically possible, launching a large-scale and error-free data collection and analysis in practice presents a practical challenge.

Recognizing the limitations of manual data extraction, automation thus becomes increasingly imperative to streamline the process of data collection. In this context, publicly accessible online data generally pertain to information that is amenable to being "scraped"[66] or "queried."[67] Different technological pathways for accessing data entail diverse legal implications and qualifications under the established decisional court practice in the European Union. Therefore, it is essential to delve into technological attributes in the following sections.

---

of the data in statistical research, or preferences as to the use of data in international data transfer. *See* Zittrain, *supra* note 61, at 110.

64. *See* Jonathan Weber, *Tim Berners-Lee Seeks to Reinvent Internet to Leave User in Control*, BUS. DAY (Jan. 12, 2021, 4:20 PM), https://www.businesslive.co.za/bd/world/2021-01-12-tim-berners-lee-seeks-to-reinvent-internet-to-leave-user-in-control [perma.cc/3LV5-5KGG]. In the context of Tim Berners-Lee's Solid project, PODs stand for "Personal Online Data Stores." Essentially conceived as individualized data repositories where users can store various types of personal information, PODs are an essential component of the framework of a more user-centric and decentralized web architecture where individuals are meant to have greater agency, autonomy over their digital presence and more control over their personal information. *See What is Solid?*, SOLID, https://solid.mit.edu/ (last visited Mar. 25, 2024) (describing the project).

65. For a spectrum of debates on the capacity of the software to regulate and shape one's behavior, *see generally* Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, 76 TEX. L. REV. 553, 555 (1998); LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE 7 (2009); Tim Wu, *When Code Isn't Law*, 89 VA. L. REV. 679, 684 (2003).

66. "Scraped" information is understood generally as information that has been extracted or harvested from websites or online sources using automated tools or scripts. *See* discussion *infra* Section II.B.2.

67. "Queried" information refers to information that is retrieved using a "query," which is a structured command or statement in a database query language that describes specific conditions and criteria for data collection. *See* discussion *infra* Section II.B.1.

1. Programmatic Access Through API

Despite its crucial and fundamental role in enabling software to communicate, the concept of an Application Programming Interface (API) has not been a frequent figure in legal debates.[68] As a set of rules, protocols, and tools allowing for communication and interaction between different components of software system, API has long existed in the technical universe, but has rarely transcended into a nontechnological space.[69] Over time, though, issues like intellectual property protection of protocols[70] and the API's role in enabling competition[71] opened a path for integrating API into legal debates and discussions. The following discussion presents just a handful of remarks on API's technical dimensions.

Firstly, APIs permeate digital reality. Several types of APIs might be routinely at work providing myriad functionalities that device users often take for granted.[72] For instance, to share photos using a social networking mobile application, at least three APIs are involved: a hardware API (to access photos on a phone camera), a library API (to transpose colors), and a web API (to send them to a server over the internet).[73] A web API, by contrast, is a web interface: it essentially acts like a user's interface but is meant for and used by the software instead.[74] Thus, whenever one uses a phone app, the user typically interacts with it through the phone's screen (the "user interface"). An API enables the same interaction, but at the level of software: one application communicates to another through the set of definitions and protocols that make up its interface.

Secondly, in much of its design, an API is a "layer of abstraction": it hides specifics of the underlying service.[75] A metaphor is helpful to

68. *See, e.g.*, Catalina Goanta, Thales Bertaglia & Adriana Iamnitchi, *The Case for a Legal Compliance API for the Enforcement of the EU's Digital Services Act on Social Media Platforms*, 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 1341, 1348 (2022).

69. *See, e.g.*, Saša Milić, *APIs: The Digital Glue*, MEDIUM (Oct. 4, 2020), https://medium.com/api3/apis-the-digital-glue-7ac87566e773 [perma.cc/SL5M-U8S2] (providing a historical account of APIs); *What Is an API (application programming interface)?*, IBM, https://www.ibm.com/topics/api (last visited Mar. 23, 2023).

70. *See, e.g.*, Google, LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1201 (2021) (noting that this characteristic is by no means neglectable).

71. Chris Riley, *Unpacking Interoperability in Competition*, 5 J. CYBER POL'Y 94, 99 (2020).

72. A few examples are Communication APIs, Cloud IPs, Payment Gateway APIs, and Social Media APIs.

73. ARNAUD LAURET, THE DESIGN OF WEB APIS 4 (2019).

74. *Id.*

75. Sudeshna Roy, *What Is API Integration? (The Complete Guide)*, KNIT (Aug. 16, 2023), https://www.getknit.dev/blog/what-is-api-integration-the-complete-guide [perma.cc/PY9P-7SWG].

explain this concept. An API is akin to a restaurant waiter taking an order;[76] though customers will typically not know how exactly their meals will be prepared (e.g., what recipe will be used, how many people will prepare it, etc.), a waiter ensures that a cooking staff accepts and executes their orders. So, the waiter (API), a de facto representative of the restaurant (an API provider), mediates the relationship between the customers (e.g., mobile phone applications) and the staff of the restaurant (e.g., a server). This mediation and decoupling results in a situation where the customers and the service providers know little to nothing about each other. By design, the restaurant staff enjoy much flexibility and autonomy in executing orders.

The same goes for API providers in a virtual world, whether they consist of a news outlet,[77] social microblogging service,[78] media streaming service,[79] online payment system,[80] or an approaching three-billion-user[81] social networking platform.[82] API providers retain significant control over their data offerings in terms of who has what and how.[83] Consequently, data access procedures are more robust,[84] content of respective inquiries is typically released in a structured format, and the overall process of data collection is more stable and predictable.[85]

2. Web Scraping

The use of an API is not the only way to access online data. In

76. LAURET, *supra* note 73, at 11.

77. *Documentation: All You Need to Know About the API Is Here*, THE GUARDIAN, https://open-platform.theguardian.com/documentation/ [perma.cc/J7L4-ZPHY] (last visited Mar. 4, 2023).

78. *Twitter API v2*, TWITTER: DEV. PLATFORM, https://developer.twitter.com/en/docs/twitter-api [perma.cc/RH85-7D36] (last visited Mar. 4, 2023).

79. *Web API*, SPOTIFY: FOR DEVS., https://developer.spotify.com/documentation/web-api/reference/ [perma.cc/F696-C5NR] (last visited Mar. 4, 2023).

80. *Get Started with PayPal REST APIs*, PAYPAL DEV., https://developer.paypal.com/docs/api/overview/ [perma.cc/3YFC-T2MX] (last visited Mar. 4, 2023).

81. *Facebook Users in the World*, INTERNET WORLD STATS (2021), https://www.internetworldstats.com/facebook.htm [perma.cc/FTR8-XUNY] (listing the global Facebook user population at 2,803,147,884).

82. *Meta for Developers*, META, https://developers.facebook.com/tools/explorer/ [perma.cc/QJ89-6UBA] (last visited Mar. 4, 2023).

83. *See also* Oscar Borgogno & Giuseppe Colangelo, *Data Sharing and Interoperability: Fostering Innovation and Competition Through APIs*, 35 COMPUT. L & SEC. REV. 1, 10 (2019).

84. API data access procedures are standardized across many computer languages, so the data collection process can be replicated in various software environments as well. *See* SIMON MUNZERT, CHRISTIAN RUBBA, PETER MEIßNER & DOMINIC NYHUIS, AUTOMATED DATA COLLECTION WITH R: A PRACTICAL GUIDE TO WEB SCRAPING AND TEXT MINING 277 (2014).

85. *See, e.g.*, *id.*

fact, under certain circumstances, it might not be a feasible or beneficial means to collect data at all. For example, some websites might not provide a public API.[86] In other instances, companies may moderate API use by implementing a tiered pricing model, where the least restrictive tier could be cost prohibitive for some users and the free tier prohibitively limits the features available to users.[87] It also could be the case that an API does not provide access to particular data that is published on a website.[88]

To advance with data collection in these cases, one would need to explore alternative approaches. This process, called "web scraping," involves "the construction of an agent to download, parse, and organize data from the web in an automated manner."[89] The web, as a massive network of resources, relies on a multitude of protocols or "rules" that ensure its operation.[90] For example, the lingua franca of the web is the Hypertext Transfer Protocol (HTTP) that enables communication between a web client (a web browser) and a web server (e.g., a computer that deals with the HTTP requests).[91] A HyperText Markup Language (HTML) typically encodes the messages users exchange and communicate.[92] This language instructs web browsers to present the underlying data in a certain way, such as displaying headlines, links, tables, and other content. End users, accessing the web typically

86. *See, e.g.*, OPENAPIHUB, *Major Types of API – Public API, Private API & Partner API* (Jan. 10, 2022), https://blog.openapihub.com/en-us/3-major-types-of-api-public-api-private-api-partner-api/.

87. Some examples of tiered pricing models are as follows: Google Maps offers a free tier with limited usage, while access to more extensive features and higher usage limits requires upgrading to a paid plan. *Pricing That Scales to Fit Your Needs*, GOOGLE MAPS PLATFORM, https://mapsplatform.google.com/pricing/ (last visited Apr. 13, 2024). Amazon Web Services (AWS) offers a range of cloud computing solutions, encompassing APIs for storage, computing, and databases. *Amazon API Gateway Pricing*, AWS, https://aws.amazon.com/api-gateway/pricing/ (last visited Apr. 13, 2024). AWS implements a tiered pricing structure, with varying pricing tiers contingent upon usage levels and service attributes. *Id.*

88. For instance, the API may not grant access to seller-specific metrics or detailed product reviews, both of which are readily available on the website. Also, features such as real-time pricing or inventory levels may be unavailable through the API due to, for example, data latency.

89. SEPPE VANDEN BROUCKE & BART BAESENS, PRACTICAL WEB SCRAPING FOR DATA SCIENCE (1st ed. 2018).

90. For a high-level introduction to the internet layered model, *see, e.g.*, Kevin Werbach, *A Layered Model for Internet Policy*, 1 J. ON TELECOMM. & HIGH TECH. L. 37 (2002).

91. MUNZERT ET AL., *supra* note 84, at ch. 5 (providing a primer on relevant protocols).

92. *See Hypertext Markup Language - 2.0*, INTERNET ENGINEERING TASK FORCE (Nov. 1995), https://www.ietf.org/rfc/rfc1866.txt [https://perma.cc/HY8W-KJMV] (describing the protocol).

through a web browser, do not see the HTML document itself.[93] They are instead greeted with its visually appealing interpretation: a web layout full of images, video, animation, styling, and other content.[94] In web scraping, however, this human-friendly representation is not that helpful; rather, one needs to access the HTML source and its elements.[95]

Typically, the steps involved include loading the HTML into a chosen programming environment and applying algorithms to then extract necessary information.[96] Various data retrieval situations may arise, such as dealing with dynamic webpages, complex structures beyond HTML tables, or extracting plain text. Users need to adjust their web scraping approaches to achieve efficient results.[97] Several technologies and programming languages are available and suitable for web scraping tasks—such as Python, Ruby, Node.js, R, and PHP[98]—and users can opt for building a web scraper by themselves, using approaches that require basic or advanced coding skills, or purchasing a ready-made commercial solution.[99]

---

93. Brian O'Grady, *What is HTML? An Introduction*, CODE INSTITUTE, https://codeinstitute.net/global/blog/what-is-html-and-why-should-i-learn-it/ [https://perma.cc/26UY-W9PW] (last visited Apr. 18, 2024).

94. One can access the HTML source code of a web page by right-clicking on the page under consideration and choosing a "view page source" option.

95. *See* Werbach, *supra* note 90, at 49–50.

96. E.g., by traversing the HTML tree and extracting values of some tags, or by employing natural language processing (NLP) techniques.

97. Users could do so by employing various tools and techniques commonly used in web scraping and data extraction such as XPath (a query language used to identify and extract specific data from the structure of a webpage), JSON parsers (a data interchange format commonly used for transmitting data between a web server and a client), Selenium (a tool used for interacting with dynamic webpages), and regular expressions (tools often used to extract data from unstructured or semi-structured text). *See* IBM, *XPath overview*, https://www.ibm.com/docs/en/app-connect/11.0.0?topic=xpath-overview (Mar. 27, 2024); JSON FORMATTER, *JSON Parser Online*, https://jsonformatter.org/json-parser (last visited Apr. 18, 2024); Davis David, *Scraping Dynamic Websites with Python*, BRIGHT DATA, https://brightdata.com/blog/how-tos/scrape-dynamic-websites-python#:~:text=Selenium%20is%20an%20open%20source,or%20tasks%20on%20dynamic%20websites (last visited Apr. 18, 2024); AWS, *Extracting string fragments using a regular expression*, https://docs.aws.amazon.com/glue/latest/dg/transforms-regex-extractor.html [https://perma.cc/ECR9-UWLE] (last visited Apr. 18, 2024).

98. Each of these programming languages has its strengths and weaknesses, and the choice of the language often depends on the specific requirements of the project and a collection of external tools, packages, and resources accessible to developers.

99. *See, e.g.*, JAY M. PATEL, GETTING STRUCTURED DATA FROM THE INTERNET RUNNING WEB CRAWLERS/SCRAPERS ON A BIG DATA PRODUCTION SCALE (2020) (providing a general overview of web scraping approaches); Parsehub Home Page, https://www.parsehub.com/ [https://perma.cc/B6BZ-G9ER] (last visited Apr. 18, 2024) (a free web scraping tool); Scrapy Home Page, https://scrapy.org/ [https://perma.cc/QGJ3-G2HL] (last visited Apr. 18, 2024) (an open source web scraping Python library and a community of contributors); Web Scraper Home Page, https://webscraper.io/ [https://perma.cc/H3MF-BZ3J] (last visted Apr. 18, 2024) (a Google Chrome browser extension).

C. Public Accessibility as a Technological Construct: Points of Control

Various industries and academic researchers alike routinely employ API and web scraping for large-scale data collection. Data collection is a core constituent for businesses focusing on branding and competition,[100] web technology monitoring,[101] and "people analytics."[102] Data collection is also an auxiliary yet indispensable tool for covering the gamut of theoretical and practical applications. For instance, collecting sentiment data from social media can enable or bolster surveillance in financial markets, which StockPulse's "Emotional Data Intelligence," a system that analyzes emotions to make financial decisions, demonstrates.[103] In a different context, malicious actors can use such data investigating political orientations, as the now-defunct *Cambridge Analytica* illustrates.[104] Additionally, automated large-scale data collection can support information campaigns, akin to the web scraping Statistics Canada undertook during the COVID-19 pandemic.[105] Yet, irrespective of these varied goals, each method of data collection comes with its own set of limitations. These include technical challenges related to exerting control and restricting access.

---

100. One example of a business focused on branding and competition is JungleScout, a tool used to optimize product listings, manage and improve customer feedback, and manage customer communication. *See, e.g.*, *Find the Keywords That Count*, JUNGLESCOUT, https://www.junglescout.com/features/keyword-scout/ [perma.cc/A8B8-DYG7] (last visited Feb. 17, 2024).

101. One example of a web technology monitoring service is BuiltWith, a software tool designed to provide users with comprehensive insights into the technological infrastructure behind websites. *See, e.g.*, *Find out What Websites are Built With*, BUILTWITH, https://builtwith.com/ [perma.cc/A3P9-HS2Q] (last visited Feb. 17, 2024).

102. One example of a "people analytics" provider is HiQ Labs, a service that specializes in workforce analytics and human capital management by generating insights about employees and potential candidates. *See, e.g.*, hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1187 (9th Cir. 2022).

103. *See How Does Social Media Influence Financial Markets? NASDAQ Is Deploying Emotional Data Intelligence By Stockpulse to Find Out*, STOCKPULSE, https://stockpulse.ai/press-article/how-does-social-media-influence-financial-markets-nasdaq-is-deploying-emotional-data-intelligence-by-stockpulse-to-find-out/ [perma.cc/R4SH-WLVY] (last visited Feb. 17, 2024) ("Emotional Data Intelligence" by Stockpulse is a framework that revolves around analyzing sentiments and emotions expressed on social media and other digital platforms to inform financial market
decisions).

104. *See* Katie Harbath & Collier Fernekes, *History of the Cambridge Analytica Controversy*, BIPARTISAN POL'Y CTR. (Mar. 16, 2023), https://bipartisanpolicy.org/blog/cambridge-analytica-controversy/ [perma.cc/J48N-ZV6E].

105. *Web Scraping During the COVID-19 Pandemic*, STATS. CAN. (Feb. 12, 2024), https://www.statcan.gc.ca/en/our-data/where/web-scraping/covid-19 [perma.cc/R6MP-XQQE].

Commonly, the use of an API is a starting point for data collection inquiry.[106] As this Article discussed earlier, the API provider usually exercises principal control over conditions for data collection practices.[107] Thus, to gain API access, the API provider might require one to apply for a special account, like a developer account.[108]

Furthermore, the process of sending requests often includes a requirement for authentication, like acquiring OAuth credentials,[109] which allows the server to verify the identity of the party requesting the data and regulate the interaction of the third party with a data provider.[110]

Thus, authenticated APIs act as checkpoints defining access conditions like volume, variety, quality, and velocity of requested data. Various factors can influence an API provider's decision as to what and how data becomes accessible for automated data collection including the privacy preferences of its users and the specific types of data involved.[111]

For example, X (formerly Twitter) sets forth rate limits restricting the number of data requests an app can make to a given API

106. *See, e.g.*, Stine Lomborg & Anja Bechmann, *Using APIs for Data Collection on Social Media*, 30(4) THE INFO. SOC'Y 256 (2014) (discussing methodological challenges of API-enabled research).

107. *See* discussion *infra* Section II.B.1.

108. *See Tap Into What's Happening to Build What's Next*, X DEV. PLATFORM, https://developer.twitter.com/en [perma.cc/NV4N-4V68] (last visited Feb. 17, 2024). Sometimes, there might be an extra requirement to submit the account for approval by the platform (e.g. X developer account) or undergo a verification process (e.g. Facebook verification practice). *See Developer Policy Support*, X DEV. PLATFORM, https://developer.twitter.com/en/support/x-api/policy [perma.cc/EFK9-3DW3] (last visited Feb. 17, 2024); Stephanie Curran, *Developer Platform Will Now Require Business Verification for Advanced Access,* META (Feb. 1, 2023), https://developers.facebook.com/blog/post/2023/02/01/developer-platform-requiring-business-verification-for-advanced-access/ [perma.cc/J7GG-MPAG].

109. OAuth credentials are typically issued by a server with which the API is associated. For the following legal assessment, however, OAuth credentials could be considered as expression of control by the API provider over data processing. *See* Google Identity, *Using OAuth 2.0 to Access Google APIs*, GOOGLE (Jan. 10, 2024), https://developers.google.com/identity/protocols/oauth2#:~:text=Google%20APIs%20use%20the%20OAuth,from%20the%20Google%20API%20Console [perma.cc/2K7R-NK2S].

110. *See* MATTHEW A. RUSSELL & MIKHAIL KLASSEN, MINING THE SOCIAL WEB: DATA MINING FACEBOOK, TWITTER, LINKEDIN, INSTAGRAM, GITHUB, AND MORE app. at B (3rd ed. 2019).

111. *Cf.*, *e.g.*, considerations internalized within API access in cases of Facebook, LinkedIn, United Nations, Walters Art Museum, and ProPublica Congress.

within a specified duration.[112] Similar restrictions could be defined as "quotas per day" or could be connected to the use of an IP address.[113]

As far as restrictions on the type of data available for collection are concerned, the evolution of the Facebook API offers a telling cautionary tale. Facebook has routinely positioned itself as a tool to "give people the power to build community and bring the world closer together."[114] Not surprisingly, due to its dependence on advertising income, the data the platform collects is of great value and contributes the necessary information to construct a "social interest graph,"[115] exposing the connections between people and their interests.[116]

The development of the platform's API has mirrored the ongoing struggle regarding access to this data and the associated knowledge. The initial period (2006–2015) featured a relatively nonrestrictive API environment that allowed for querying a wide range of objects, including people, photos, pages, events, and connections.[117] The subsequent period (2015 onward), however, has been anything but similar. For a number of years, Facebook has been gradually imposing limits on data access through a succession of APIs updates.[118] A series of public inquiries into the platform's data practices and the app's access permissions has ultimately expedited Facebook's API policies' self-review and update progression .[119] Most recently, the Facebook and

---

112. *Documentation: Rate Limits*, X DEV. PLATFORM, https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits [perma.cc/UX8R-PRCW] (last visited Mar. 4, 2023).

113. *Limits and Quotas on API Requests*, GOOGLE ANALYTICS, https://developers.google.com/analytics/devguides/config/mgmt/v3/limits-quotas [perma.cc/NGX8-32KQ] (last visited Mar. 4, 2023).

114. FAQs, META: META INV. RELS., https://investor.fb.com/resources/default.aspx [perma.cc/J2JP-AWJX] (last visited Mar. 4, 2023).

115. *See, e.g.*, Bernie Hogan, *Social Media Giveth, Social Media Taketh Away: Facebook, Friendships, and APIs*, 12 INT'L J. COMMC'N 592, 593 (2018).

116. *See id.* In the context of the social graph, "interests" denote a wide range of topics, activities, preferences that users express or engage with. *See, e.g.*, Deepak Nayal, *Why Is Interest Graph More Interesting Than Social Graph?*, MEDIUM (July 23, 2011), https://medium.com/@dnayal/why-is-interest-graph-more-interesting-than-social-graph-59d5e105d567 [perma.cc/KTK4-QCKA].

117. *See, e.g.*, Jesse Weaver & Paul Tarjan, *Facebook Linked Data via the Graph API*, 4 SEMANTIC WEB J. 245, 245 (2013).

118. *See, e.g.*, Mike Schroepfer, *An Update on Our Plans to Restrict Data Access on Facebook*, META (Apr. 4, 2018), https://about.fb.com/news/2018/04/restricting-data-access/ [perma.cc/JT4L-43YU].

119. *See, e.g.*, OFF. OF THE DATA PROT. COMM'R OF IRELAND, REPORT OF AUDIT: FACEBOOK IRELAND LTD 81 (Dec. 21, 2011), https://www.pdpjournals.com/docs/87980.pdf [perma.cc/LHX4-T88P] hereinafter Report of the Audit on Facebook; ATLE ÅRNES, JØRGEN SKORSTAD & LARS-HENRIK PAARUP MICHELSEN, SOCIAL NETWORK SERVICES AND PRIVACY: CASE STUDY OF FACEBOOK 5 (Apr. 15, 2011) https://www.datatilsynet.no/globalassets/

Cambridge Analytica scandal has become a poster child for data misuse that has captured public attention and prompted several legal and political actions globally.[120] Regulators' concerns mainly pertained to the normative implications of Facebook's API configurations enabling third-party app data access.[121] Consequently, the present iteration of Facebook's API is considerably more restricted, suggesting a shift away from prioritizing developers to protecting users, and narrowing the scope of research that external inquirers can conduct.[122]

The evolution of Facebook's API serves as a prominent instance showcasing the substantial influence API providers hold in determining the terms and conditions for data collection practices. Operating as gatekeepers, API providers play a crucial role in shaping technological settings of data accessibility within their data repositories. Consequently, recipients of online data collected via APIs should be wary of the resulting limitations of the API provider's decision in terms of data access. In addition to the API provider's control over modifications to access conditions, it also has the power to revoke data access unilaterally and without any advance warning.[123]

The case of web scraping is notably different in terms of the amount of control it offers developers and the fewer technological barriers to data collection it offers. As this Article has explained, web scraping implies the employment of an agent to download, parse, and

---

global/english/11_00643_5_parti_rapport_facebook_2011.pdf [perma.cc/TJ53-KFC5]. As evident from the Report of the Audit on Facebook, the process of API policies' revision is mostly conducted internally by Facebook following the assessment of security, business impact, cost, and the developers' experience.

120. *See, e.g.*, Cadwalladr & Graham-Harrison, *supra* note 15. There have been multiple legislative inquiries and regulatory investigations. *See, e.g.*, *Facebook, Social Media Privacy, and the Use and Abuse of Data: Testimony Before the Committee on the Judiciary and Committee on Commerce, Science and Transportation* (2018) (statement of Mark Zuckerberg, Chief Executive Officer of Facebook) https://www.commerce.senate.gov/2018/4/facebook-social-media-privacy-and-the-use-and-abuse-of-data [https://perma.cc/RBA9-FUQ4]; House of Commons. Digital, Culture, Media and Sport Committee. *Disinformation and 'fake news': Final Report*, PARLIAMENTARY COPYRIGHT HOUSE OF COMMONS (2019) https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf; FEDERAL TRADE COMMISSION, *FTC Imposes $5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook*, (July 24, 2019), https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions-facebook [https://perma.cc/Y2C6-AANM] (describing a series of the FTC enforcement actions).

121. Hogan, *supra* note 115.

122. *See, e.g.*, Deen Freelon, *Computational Research in the Post-API Age,* 35 4 POL. COMMC'N 665, 665 (2018); Axel Bruns, *After the 'APIcalypse': Social Media Platforms and Their Fight Against Critical Scholarly Research*, 22 INFO., COMMC'N & SOC'Y 1544, 1544 (2019).

123. *See API Terms and Conditions*, ZIM, https://www.zim.com/help/api-terms-and-conditions [perma.cc/5MKG-P7YC] (last visited Feb. 21, 2024).

arrange data in an automated manner.[124] In Google search results,[125] for instance, the information displayed is routinely obtained through the use of "web crawlers," which "locate and sweep up the content of web pages methodically and automatically."[126] The main virtual obstacle that can restrict this type of crawling is the use of "exclusion codes," commonly implemented in the form of a "robots.txt" file.[127] Using this robots.txt file, website owners can specify their preferences regarding which sections of their site should be crawled and by which entities.[128] However, although the use of the robots.txt file and tags like "noindex" or "noarchive" is a common and accepted practice in the industry, it does not enjoy legal enforcement.[129] Thus, from a purely technical perspective, while the data hosts' use of exclusion codes might discourage data access in some instances, it cannot effectively prevent such access, provided that the website remains publicly accessible.[130]

The internet environment is both an invaluable source of data and a key enabler of its collection. Consequently, any automated data collection practices demand careful attention to the sheer amount of available online data and digital infrastructure used to host it. Technically speaking, the settings of API access or "exclusion codes" could restrict the automated process of data collection. When online data is accessed via APIs, data hosts generally have the ability to monitor and control the data collection process using technical measures such as authenticating protocols.[131] Web scraping, by

124. MUNZERT ET AL., *supra* note 84, at 216.

125. It has to be noted, however, that for the sake of simplicity, this discussion does not distinguish between "bots" ("spiders," "crawlers," "wanderers") and "agents." For a more nuanced approach to possible definitions and taxonomy, *see* Stan Franklin & Art Graesser, *Is it an Agent, or Just a Program?: A Taxonomy for Autonomous Agents* (Intelligent Agents III: Agent Theories, Architectures, and Languages, ECAI '96 Workshop, 1996).

126. Case C-131/12, Google Spain, S.L. v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317 (May 13, 2014).

127. *See* Martijn Koster, *A Standard for Robot Exclusion*, ROBOTSTXT, https://www.robotstxt.org/orig.html [perma.cc/YE2U-RR8L] (last visited Jan. 28, 2024).

128. Zittrain, *supra* note 61, at 102; *see How to Write and Submit a Robots.txt File*, GOOGLE SEARCH CENTRAL, https://developers.google.com/search/docs/crawling-indexing/robots/create-robots-txt [https://perma.cc/TZ6W-TQ84] (last visited Aug. 4, 2023).

129. Robot.txt played a particular role in establishing Bidder's Edge's unauthorized access to eBay's website. *See* eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1071–72 (N.D. Cal. 2000). This particular venue could be further explored in enhancing individuals' control over their personal data. *See infra* Part V.

130. *See* Patel, *supra* note 99, at 371–93 (discussing ethics and legality of web scraping); Alfred Ng & Steven Musil, *Clearview AI Hit with Cease-and-Desist from Google, Facebook over Facial Recognition Collection*, CNET (Feb. 5, 2020, 6:10 PM), https://www.cnet.com/news/privacy/clearview-ai-hit-with-cease-and-desist-from-google-over-facial-recognition-collection/ [perma.cc/3NVV-KH6U].

131. Google Identity, *supra* note 109; RUSSELL & KLASSEN, *supra* note 110.

contrast, does not provide comparable technical levers of control for data hosts. Although embedding "exclusion codes" can deter some scraping entities from accessing the data, as long as the websites in question remain accessible for browsing to the public at large, online data collection persists to some extent.

Thus, the technological dimension of automated online data collection casts the net of "public accessibility" widely. In technical terms, "publicly accessible online data" implies data that could be "scraped" or "queried" by anyone interested in doing so. While the use of APIs might be a more controlled and robust method of collecting online data, web scraping essentially provides an alternative route for collecting the same data, void of the insurmountable technical restrictions that data hosts impose.

III. DATA PROTECTION DIMENSION OF LARGE-SCALE AUTOMATED DATA COLLECTION

As mentioned in the Introduction, data accessibility has been a central focus in discussions surrounding cybercrime, intellectual property, contractual matters, and issues related to competition law.[132] It has a distinct ethical dimension as well.[133] Thus, the question of what online data ought to be considered publicly accessible has risen against the backdrop of fundamental issues of research transparency, reproducibility, and the responsibility of researchers and institutional review boards.[134]

The following reflection, however, focuses on data protection law. It locates technological insights of large-scale data collection within a legal domain, focusing in particular on relevant CJEU case law

132. *See*, *e.g.*, Tess Macapinlac, *The Legality of Web Scraping: A Proposal,* 71 FED. COMMC'NS. L.J. 399, 401 (2018); Borgogno & Colangelo, *supra* note 83, at 1, 3; Benjamin L.W. Sobel, *A New Common Law of Web Scraping*, 25 LEWIS & CLARK L. REV. 147, 152 (2021).

133. *See, e.g.*, Mike Thelwall & David Stuart, *Web Crawling Ethics Revisited: Cost, Privacy, and Denial of Service,* 57(13) J. AM. SOC'Y FOR INFO. SCI. & TECH., 1771, 1771 (2006); ANNETTE MARKHAM & ELIZABETH BUCHANAN, ETHICAL DECISION-MAKING AND INTERNET RESEARCH*:* RECOMMENDATIONS FROM THE AOIR ETHICS WORKING COMMITTEE (VERSION 2.0) 2 (2012), https://aoir.org/reports/ethics2.pdf [perma.cc/KVG4-8KYR] (last visited Jan. 26, 2024); Russell Brewer, Bryce Westlake, Tahlia Hart & Omar Arauza*, The Ethics of Web Crawling and Web Scraping in Cybercrime Research: Navigating Issues of Consent, Privacy, and Other Potential Harms Associated with Automated Data Collection, in* RESEARCHING CYBERCRIMES: METHODOLOGIES, ETHICS, AND CRITICAL APPROACHES 435 (Anita Lavorgna & Thomas J. Holt eds., 2021).

134. *See, e.g.*, Jessica Vitak, Nicholas Proferes, Katie Shilton & Zahra Ashktorab, *Ethics Regulation in Social Computing Research: Examining the Role of Institutional Review Boards*, 12 J. EMPIRICAL RSCH. ON HUM. RSCH. ETHICS 372, 372 (2017).

and the guidance of Article 29WP/EDPB.[135] Admittedly, the issue of large-scale data collection transcends multiple institutions of data protection law.[136] It also serves as a litmus test for exploring and defining the evolving boundaries among data protection, other fundamental rights, and competition law in Europe.

Nevertheless, this Article's focus on the material scope and data controllership is sufficient to elucidate some fundamental challenges that the technology affordances pose. As discussed in Section III.A below on material scope, recent judicial interpretation in automated data collection cases effectively suggests the application of the strictest form of the personal data protection regime nearly by default.[137] Automated data collection implies no meaningful methodology of distinguishing between personal versus nonpersonal and ordinary versus special category data.[138] In this context, a broad and multipronged concept of "personal data" loses its ability to serve as a dynamic boundary, thus frustrating the logic underlying a higher level of protection for particularly sensitive data.[139] Recent judicial interpretation of data controllership shows inconsistency and uncertainty.[140] The data controller is a principal bearer of responsibility for compliance with data protection obligations and has an ultimate duty to facilitate the exercise of an extensive array of data subjects' rights.[141] A commitment to "ensure a high level of protection" of data

135. Article 29 Working Party (Art. 29 WP) was established by Directive 95/46/EC (Art.30) and Directive 2002/58/EC (Art.15) as an independent European advisory body on data protection and privacy. It was replaced by the European Data Protection Board upon entry into force the GDPR (May 25, 2018). *Legacy: Art. 29 Working Party*, EUR. DATA PROT. BD. https://www.edpb.europa.eu/about-edpb/who-we-are/legacy-art-29-working-party_en [perma.cc/ES29-NR6U] (last visited Mar. 18, 2024).

136. *See, e.g.*, JERSEY OFFICE OF THE INFO. COMM'R: JOINT STATEMENT ON DATA SCRAPING AND THE PROTECTION OF PRIVACY (Aug. 24, 2023), https://jerseyoic.org/media/f0jnyjix/gpa-iewg-data-scraping-joint-statement-august-2023.pdf [perma.cc/C4SN-LJLH].

137. *See* Case C-252/21, Meta v. Bundeskartellamt, ECLI:EU:C:2023:537, ¶ 69 (2023).

138. *See* Michèle Finck & Frank Pallas, *They Who Must Not Be Identified—Distinguishing Personal from Non-Personal Data Under the GDPR*, 10 INT'L DATA PRIV. L. 11, 11 (2020).

139. On a special regime of protection of sensitive data, *see* Christopher Kuner & Ludmila Georgieva, *Processing of Special Categories of Personal Data*, *in* THE EU GENERAL DATA PROTECTION REGULATION: A COMMENTARY 79, 80 (Christopher Kuner, Lee A. Bygrave & Christopher Docksey eds., 2020).

140. *See* Rene Mahieu, Joris Van Hoboken & Hadi Asghari, *Responsibility for Data Protection in a Networked World – On the Question of the Controller, 'Effective and Complete Protection' and Its Application to Data Access Rights in Europe*, 10 J. INTELL. PROP., INFO. TECH. & ELEC. COM. L. 39, 49 (2019); Michèle Finck, *Cobwebs of Control: The Two Imaginations of the Data Controller in EU Law,* 11 INT'L DATA PRIV. L. 333, 337 (2021).

141. GDPR, *supra* note 37, art. 5(2).

subjects[142] largely drives CJEU jurisprudence, which has promoted a broad and expansive interpretation of data controllership.[143] Despite its commendable intention, however, it is far from certain that the CJEU's established practice actually fulfils its commitment. The fundamental challenge of the online environment is a plurality of entities potentially capable of large-scale data processing. The established CJEU decisional framework, nevertheless, has in large part evolved around large players like Meta (previously Facebook) and Google.[144] As will be shown below, when confronted with the fundamentals of prevalent data collection models and data sharing on the internet, the court has approached them piecemeal. Thus, based on the emergent case law, the assumption of joint controllership in API-enabled data collection would be a safe compliance strategy. In web scraping cases, the court has produced a series of judgments around the search engine function, carefully carving out pathways for ensuring the rights of data subjects.[145] Recognizing the importance of a search engine's role in providing information access and exercising freedom of expression, the court has not addressed the questions of transformational search capabilities in a "future-proof" manner.[146] An exclusive focus on existing search models leaves significant room for uncertainty regarding how it should address

142. Case C-2010/16, Wirtschaftsakademie Schleswig-Holstein, ECLI:EU:C:2018:388, ¶¶ 26–28 (June 5, 2018).

143. Katerina Tassi & Ruth Boardman, *The CJEU Rules on the Liability of Controllers*, IAPP (Jan. 4, 2024), https://iapp.org/news/a/the-cjeu-rules-on-the-liability-of-controllers/#:~:text=The%20CJEU%20reaffirmed%20the%20broad,purposes%20and%20means%20of%20processing [perma.cc/C6AE-E6X4] (last visited Feb. 21, 2024).

144. It is worth noting, however, that the practice of the national Data Protection authorities features a variety of market players beyond the two named. *Id.* Article 3 of the GDPR outlines the territorial scope of the Regulation based on two key criteria: the "establishment" criterion specified in Article 3(1), and the "targeting" criterion outlined in Article 3(2). *See* GDPR, *supra* note 37, arts. 3(1), 3(2). An establishment refers to "the effective and real exercise of activities through stable arrangements" as defined in Recital 22 of the GDPR. *See* GDPR, *supra* note 37, rec. 22. The absence of an establishment under Art. 3(1) of the GDPR within the EU does not automatically exclude processing activities by a data controller or processor established in a third country from the GDPR's scope. Under a "targeting criterion," (Art. 3(2) of the GDPR) such processing will still be governed by the GDPR when and if it relates to the "offering of goods or services" or the "monitoring of behaviour" of individuals located in the Union. *See* EUR. DATA PROT. BD., GUIDELINES ON THE TERRITORIAL SCOPE OF THE GDPR 03/2018 (2020).

145. Jure Globocnik, *The Right to Be Forgotten is Taking Shape: CJEU Judgments in* GC and Others *(C-136/17) and* Google v CNIL *(C-507/17)*, 69 GRUR INT'L, 380, 380 (2020).

146. *See, e.g.*, Emily Bender, *Large Language Models on the Web: Anticipating the Challenge*, DIGWATCH (Oct. 12, 2023, 1:30 AM), https://dig.watch/event/internet-governance-forum-2023/large-language-models-on-the-web-anticipating-the-challenge-igf-2023-ws-217 [perma.cc/9LM5-CUQY] (session report on a challenge of large language processing models embedded in a search).

more integrated search models, predominantly voice or image-based inquires, and large language models (LLMs) like ChatGPT.

These issues are fundamental, and, as the following will demonstrate, illustrative of the relative preparedness of the EU data protection framework as a relevant and enforceable legal instrument for addressing an increasingly immersive online environment of data subjects.

A. GDPR Material Scope: On Personal Data and Processing

The right to privacy and data protection are integral components of two distinct yet interconnected systems aimed at safeguarding human rights in Europe. The first system pertains to the European Convention on Human Rights (ECHR),[147] which serves as an international agreement the European Court of Human Rights (ECtHR) enforces and interprets. The second system is founded on the jurisprudence of the CJEU, which guarantees the protection of fundamental human rights within the EU.[148] The relationship between the CJEU and ECtHR is complex, but for present purposes, it suffices to say that ECtHR jurisprudence invariably serves as a source of inspiration for the CJEU.[149]

The CJEU initially recognized fundamental human rights rights within the system of the general principles of EU law, closely mirroring the system of protection afforded under the ECHR regime, other international instruments, and constitutional traditions common to the Member States.[150] Subsequent incorporation of the EU Charter of

147. Convention for the Protection of Human Rights and Fundamental Freedoms, European Convention on Human Rights, Nov. 4, 1950, Europ.T.S. No. 5, 213 U.N.T.S. 221.

148. *See* Consolidated Version of the Treaty on European Union O.J. (C 326) [hereinafter TEU] art. 2 (acknowledging that The European Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights). *See id.* art. 19 (mandating that the Court of Justice of the EU shall ensure the interpretation and application of the EU Treaties).

149. *See, e.g.*, Koen Lenaerts, President, Court of Justice of the European Union, *The ECHR and the CJEU: Creating Synergies in the Field of Fundamental Rights Protection* (Jan. 26, 2018) (describing the role that the European Convention of Human Rights, as interpreted by the ECtHR, has had and continues to have on the EU legal order); Juliane Kokott & Christoph Sobotta, *The distinction between privacy and data protection in the jurisprudence of the CJEU and the ECtHR*, 3 INT'L DATA PRIVACY L. 222 (2013) (providing an introduction to distinct frameworks and scope of protection of privacy and data protection of individuals in Europe).

150. This jurisprudential approach of recognizing the particular subset of human rights as "fundamental" is particularly evident in the following landmark cases: Case 29/69, *Stauder v. Ulm*, 1969 E.C.R. 419 (where the Court of Justice explicitly acknowledged the importance of fundamental rights as enshrined in the general principles of Community law and protected by the

Fundamental Rights of the European Union[151] into primary EU law[152] further strengthened protection of these rights by formally codifying them as "fundamental rights" and increasing their visibility.[153]

Both the ECHR and the EU Charter contain clauses on the protection of privacy. Article 8 of the Convention, and similarly Article 7 of the Charter, affirm that everyone has the right to respect for his or her private and family life, home, and communications.[154] At the same time, only the EU Charter explicitly refers to the right to data protection as an active entitlement, which comprises various measures that grant individuals both preemptive and retrospective protection of their rights.[155] These measures of the data protection regime encompass

---

Court); Case 11/70, *Internationale Handelsgesellschaft mbH v. Einfuhr- und Vorratsstelle fuër Getreide und Futtermittel*, 1970 E.C.R. 1125 para. 4 ("…[r]espect for fundamental rights forms an integral part of the general principles of law protected by the Court of Justice. The protection of such rights, whilst inspired by the constitutional traditions common to the Member States, must be ensured within the framework of the structure and objectives of the community"); Case 4/73, *Nold KG v. Commission*, 1974 E.C.R. 491 para. 13 ("As the court has already stated, fundamental rights form an integral part of the general principles of law, the observance of which it ensures. In safeguarding these rights, the Court is bound to draw inspiration from constitutional traditions common to the Member States, and it cannot therefore uphold measures which are incompatible with fundamental rights recognized and protected by the constitutions of those states. Similarly, international treaties for the protection of human rights on which the Member States have collaborated or of which they are signatories, can supply guidelines which should be followed within the framework of community law."). *See* GLORIA GONZÁLEZ FUSTER, THE EMERGENCE OF PERSONAL DATA PROTECTION AS A FUNDAMENTAL RIGHT OF THE EU (2014) (describing fundamental rights to privacy and data protection in particular); HIELKE HIJMANS, THE EUROPEAN UNION AS GUARDIAN OF INTERNET PRIVACY: THE STORY OF ARTICLE 16 TFEU (2016); ORLA LYNSKEY, THE FOUNDATIONS OF EU DATA PROTECTION LAW (2015).

151. Charter of Fundamental Rights of the European Union, 2012 O.J. C 326/391 (hereinafter EU Charter).

152. The charter was formally proclaimed in Nice in December 2000 by the European Parliament, the Council of the European Union and the Commission; It became legally binding with the entry into force of the Treaty of Lisbon in December 2009. Treaty of Lisbon amended Art. 6(1) of the TEU to include that "the Union recognises the rights, freedoms and principles set out in the Charter of Fundamental Rights of the European Union of 7 December 2000, as adapted at Strasbourg, on 12 December 2007, which shall have the same legal value as the Treaties."

153. Preamble of the EU Charter.

154. Art. 8 of the ECHR; Art. 7 of the EU Charter.

155. EU Charter art. 8. There is a substantial scholarship focused on the interplay between the two rights (right to privacy and data protection). *See, e.g.*, Orla Lynskey, *Deconstructing Data Protection: The 'Added-Value' of a Right to Data Protection in the EU Legal Order*, 63 INT'L & COMPAR. L. Q. 569 (2014); Lee A. Bygrave, *Privacy and Data Protection in an International Perspective*, 56 SCANDINAVIAN STUDS. IN L. 165 (2010); Gloria González Fuster & Raphaël Gellert, *The Fundamental Right of Data Protection in the European Union: In Search of an Uncharted Right*, 26 INT'L REV. OF L., COMPUTS. & TECH. 73 (2012).

principles,[156] rights of data subjects,[157] and supervision obligations,[158] which have been further developed in secondary EU law[159] such as the GDPR, as well as through the case law of both national data protection authorities and the CJEU.

The GDPR—a comprehensive data protection legislation that the European Union implemented in May 2018—only applies when "processing" concerns "personal data."[160] The decisional practice of the CJEU and national data protection authorities mandates a broad interpretation of both concepts.[161] Personal data comprises "any information relating to an identified or identifiable natural person."[162] Such information can be available in a wide variety of forms (e.g., sound, image, or binary code), including both objective (individual's health test result) as well as subjective (individual's assessment by an employee) statements, and concerns all kinds of activities of an individual (e.g., family as well as professional life details).[163] Pseudonymized data—information on individuals who are indirectly identifiable—are considered personal data. Anonymous data—information that does not relate to an identified or identifiable natural person—fall outside the scope of data protection legislation.[164] As

156. "[D]ata must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law." EU Charter art. 8(2).

157. "Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified." EU Charter, art. 8(2).

158. "Compliance with these rules shall be subject to control by an independent authority." EU Charter, art. 8(3).

159. EU law distinguishes primary sources of law (e.g., constituent EU Treaties such as TEU and TFEU and their protocols and the EU Charter of Fundamental Rights and the general principles established by the CJEU) and secondary sources of law (legislative sources such as regulations, directives, and decisions, and non-legislative sources such as implementing acts). *See* Paul Craig & Gráinne de Búrca, EU LAW: TEXT, CASES, AND MATERIALS (2020) (describing the foundations of EU Constitutional law). *See id.* ch. 5 (describing EU instruments and the hierarchy of norms).

160. GDPR, *supra* note 37, art. 2.

161. *See, e.g.*, ARTICLE 29 DATA PROT. WORKING PARTY, OPINION 1/2010 ON THE CONCEPTS OF "CONTROLLER" AND "PROCESSOR" 3 (2010), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp169_en.pdf [perma.cc/QX86-QL3Q] [hereinafter ARTICLE 29, OPINION 1/2010]; ARTICLE 29 DATA PROT. WORKING PARTY, OPINION 4/2007 ON THE CONCEPT OF PERSONAL DATA 6 (June 20, 2007), https://www.pdp.ie/docs/1030.pdf [perma.cc/ZA5U-8NC8] [hereinafter ARTICLE 29, OPINION 4/2007]; Case C-582/14, Patrick Breyer v. Bundesrepublik Deutschland, ECLI:EU:C:2016:779, ¶ 56 (Oct. 19, 2016); Case C-434/16, Nowak v. Data Prot. Comm'r, ECLI:EU:C:2017:994, ¶ 62 (Dec. 20, 1017) (on "personal data"); Case C-101/01, Bodil Lindqvist, ECLI:EU:C:2003:596, ¶ 95 (Nov. 6, 2003); Case C-131/12, Google Spain SL. v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317, ¶ 95 (May 13, 2014); Case C-25/17, Jehovan Todistajat, ECLI:EU:C:2018:57 ¶ 75 (July 10, 2018) (on "data processing").

162. GDPR, *supra* note 37, art. 4(1).

163. *See* ARTICLE 29, OPINION 1/2010, *supra* note 161, at 6–7; Case C-434/16, Nowak ¶ 34.

164. GDPR, *supra* note 37, rec. 26, art. 4(5).

advanced data analysis methods and hardware improve concurrently with an increase in the availability of data points, the limits and defensive reach of anonymization are becoming increasingly apparent. Ultimately, technological advancements coupled with the EU policy commitment to effective and complete protection of individuals have led to a high threshold for truly anonymized data and an ever-expanding scope of "personal data."[165]

As this Article has discussed previously, automated data collection concerns both personal and nonpersonal data.[166] For example, IMDb, a free, user-generated source of production details of over 400,000 movies, short videos, and video games, proves instructive.[167] The website contains nonpersonal information, such as genre, release dates, and technical specifications of media.[168] However, it also includes information that falls under the definition of "personal data" in Article 4(1) of the GDPR.[169] Some of these data points are information related to identified or identifiable natural persons by virtue of being "about" an individual, like usernames and email addresses.[170] In instances of a user's movie preferences and reviews, such information might be personal data owing to its ability "to evaluate, treat in a certain way, or influence the status or behaviour of an individual" or create "an impact on a certain person's rights and interests."[171] As expounded in Article 29WP guidance, these are the cases where attributes of the "purpose" of information use and the "result" of such use become instrumental in

165. *See* Luc Rocher, Julien M. Hendrickx & Yves Alexandre de Montjoye, *Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models*, 10 NATURE COMMC'NS 1, 1 (2019); Khaled El Emam & Cecilia Alvarez, *A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques*, 5 INT'L Data PRIV. L. 73, 86 (2015); Nadezhda Purtova, *The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law*, 10 L., INNOVATION & TECH. 40, 42 (2018); Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI*, 2 COLUM. BUS. L. REV. 494, 495 (2019).

166. *See* ARTICLE 29, OPINION 4/2007, *supra* note 161, at 16.

167. *See IMDb Statistics*, IMDB: PRESS ROOM (Dec. 2023), https://www.imdb.com/press-room/stats/ [perma.cc/JQ84-UEU8] https://www.imdb.com/pressroom/stats/; Mo Saraee, S. White & J. Eccleston, *A Data Mining Approach to Analysis and Prediction of Movie Ratings, in* DATA MINING V. (A. Zanasi, N. F. F. Ebecken & C. A. Brebbia, eds., 2004) (big data research example).

168. *See* Saraee et al., *supra* note 167. Nonpersonal data is defined here as "data that does not relate to an identified or identifiable natural person" and thus falls outside of the material scope of the GDPR. *See* GDPR *supra* note 37, art. 2, 4(1). Essentially, nonpersonal data embraces two categories of data: data that does not relate to an identified or identifiable natural person (e.g., weather conditions) and data that was once personal, but no longer is. GDPR, *supra* note 37, rec. 26 (on anonymized data); *see, e.g.*, Finck & Pallas, *supra* note 138, at 13.

169. *See* GDPR, *supra* note 37, art. 4(1).

170. *See* ARTICLE 29, OPINION 4/2007, *supra* note 161, at 9.

171. *Id.* at 11.

qualifying certain information as "personal data."[172] Given one of IMDb's business commitments to provide users with personalized recommendations,[173] it is reasonable to propose that IMDb might take users' movie preferences and reviews to draw certain inferences, moderate users' access to the site, and otherwise shape the interaction of users with the platform. Such data use carries an undeniable potential to bring certain information into the realm of "personal data," even if and when the relevant information is not about individuals per se.[174]

Some web pages might also contain information that enjoys a special, higher protection regime due to its sensitive nature.[175] This personal data might explicitly relate to, for example, individuals' race, ethnicity, or political opinions.[176] Some personal data might also fall under this heightened protection regime owing to its capacity to "reveal" sensitive information about an individual.[177] For example, images of users typically qualify as personal data under the GDPR since they usually allow for the identification of the individuals depicted.[178] However, images could also reveal other information about individuals that might indicate "racial or ethnic origin," political opinions, religious or philosophical beliefs, health condition, or sexual orientation.[179] In other words, the context, background, and particular details of images might be decisive in bringing this personal data under the label of "special categories of personal data."[180]

---

172. *Id.* at 8.

173. *What to Watch FAQ*, IMDB https://help.imdb.com/article/imdb/discover-watch/what-to-watch-faq/GPZ2RSPB3CPVL86Z?ref_=helpsect_pro_3_8# [perma.cc/CY5R-ZCLF] (last visited Jan. 26, 2024).

174. *See* Case C-434/16, Nowak v. Data Prot. Comm'r, ECLI:EU:C:2017:994 ¶ 34 (Dec. 20, 2017) (as an illustration of the CJEU approach). Importantly, such information might also be qualified as "personal data" related to other individuals, especially in cases where the recommendation system has an embedded social media component. Telling examples are Letterboxd, Voteflix, StampSocial, and similar services.

175. *See* GDPR, *supra* note 37, art. 9(1).

176. *Id.*

177. *See* ARTICLE 29 DATA PROT. WORKING PARTY, ADVICE PAPER ON SPECIAL CATEGORIES OF DATA ("SENSITIVE DATA") 6 (2011), https://ec.europa.eu/justice/article-29/documentation/other-document/files/2011/2011_04_20_letter_artwp_mme_le_bail_directive_9546ec_annex1_en.pdf [perma.cc/PQF2-JDCE] [hereainfter ARTICLE 29, ADVICE PAPER].

178. Case C-212/13, František Ryneš v. Úřad pro Ochranu Osobních Údajů, ECLI:EU:C:2014:2428 ¶ 22 (Dec. 11, 2014).

179. *See id.* Examples of these images include photos during participation in a political campaign, photos of individuals wearing religious clothing and symbols, and photos featuring individuals with disabilities. *See id.*

180. *See* Catherine Jasserand, *Legal Nature of Biometric Data: From 'Generic' Personal Data to Sensitive Data*, 2 EUR. DATA PROT. L. REV. 297, 311 (2016) (on contextual and purposeful definition regarding photographs and images).

The GDPR, like its predecessor Directive 95/46/EC, affords special categories of personal data a higher level of protection.[181] As stated in recital 51 of the GDPR, this approach is necessary as this type of personal data is, by its nature, "particularly sensitive in relation to fundamental rights and freedoms"[182] and the "context of their processing could create significant risks to the fundamental rights and freedoms."[183] In other words, a multitiered data protection system in the European Union is grounded on the assumption that misuse of some personal data could have more severe consequences on the individual's fundamental rights.[184] These consequences might have, for example, discriminatory effect[185] or lead to financial loss and damage to reputation.[186]

The principal rule while dealing with special category data is that its processing is prohibited unless exceptions under Article 9(2)

181. *See* Council Directive 95/46, art. 8, 1995 O.J. (L 281) 31, 40–41 (EC); GDPR, art. 9, 2016 O.J. (L 119) 1, 38 (EU).

182. GDPR, recital 51, 2016 O.J. (L 119) 1, 10 (EU); EU Charter, 2012 O.J. C 326/391 (listing rights). The EU Charter is a primary law of the European Union. *See The European Union's Primary Law*, EUR-LEX (Nov. 12, 2022), https://eur-lex.europa.eu/EN/legal-content/summary/the-european-union-s-primary-law.html [perma.cc/H4Y3-22AS]. It contains a wide array of rights that has received protection in established case law of the CJEU, the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) and constitutional traditions of the EU Member States. *See* EU Charter, 2012 O.J. C 326/391; Case 29/69, Stauder v. Ulm, 1969 E.C.R. 419, 422 (first mention of "fundamental rights"); Allan Rosas, *The Court of Justice of the European Union: A Human Rights Institution?*, 14 J. HUM. RTS. PRAC. 204, 205 (2022); Israel Butler, *The EU Charter of Fundamental Rights: What Can It Do?*, OPEN SOC'Y FOUNDS. (Feb. 2013), https://www.opensocietyfoundations.org/uploads/48675542-b638-4f98-8984-24b65adc6d65/eu-charter-fundamental-rights-20130221.pdf [perma.cc/K8RE-CNH3]. Privacy and data protection are recognized as fundamental rights in Art.7–8 of the EU Charter. EU Charter, art. 7–8, 2012 O.J. C 326/391, at 397. A helpful resource on application of the Chapter is THE EU CHARTER OF FUNDAMENTAL RIGHTS: A COMMENTARY (Steve Peers, Tamara Hervey, Jeff Kenner & Angela Ward eds., 2d ed. 2021).

183. GDPR, recital 51, 2016 O.J. (L 119) 1, 10 (EU).

184. ARTICLE 29, ADVICE PAPER, *supra* note 177, at 4.

185. GDPR, recital 71, 2016 O.J. (L 119) 1, 14 (EU).

186. GUIDELINES ON PERSONAL DATA BREACH NOTIFICATION UNDER REGULATION 2016/679, DATA PROTECTION WORKING PARTY 23 ("…[n]otification of a breach is required unless it is unlikely to result in a risk to the rights and freedoms of individuals, and the key trigger requiring communication of a breach to data subjects is where it is likely to result in a high risk to the rights and freedoms of individuals. This risk exists when the breach may lead to physical, material or non-material damage for the individuals whose data have been breached. Examples of such damage are discrimination, identity theft or fraud, financial loss and damage to reputation. When the breach involves personal data that reveals racial or ethnic origin, political opinion, religion or philosophical beliefs, or trade union membership, or includes genetic data, data concerning health or data concerning sex life, or criminal convictions and offences or related security measures, such damage should be considered likely to occur.").

apply.[187] The list of data categories is exhaustive, though the EU Member States have room to maintain or introduce further conditions, including limitations on the processing of genetic data, biometric data, or data concerning health.[188]

In practice, interpreting the scope of the special categories of personal data might be challenging, particularly in the context of automated large-scale data collection. For example, it could be difficult to define concepts such as "philosophical beliefs" or "political opinions."[189] Furthermore, ongoing advancements in computational capabilities, combined with increasing levels of interconnectivity, result in an irreversible growth in complementary data sources that nefarious actors could leverage to deduce sensitive information about individuals.[190]

The nature and techniques of large-scale automated data collection render issues of complementary data sources additionally problematic. As *Cambridge Analytica* illustrates, social media websites allow their users to indicate their preferences, which can range from "liking" a page of a political party, to marking their attendance at locations that could reflect their religious beliefs.[191] Users also upload countless images, the context, background, and details of which could expose particularly sensitive personal information, such as health information. As explained above, some of this information should not fall under the definition of "publicly available."[192] The combination and synthesis of information during data mining could encompass data such as exact GPS locations or financial transactions. Companies often consider this data "proprietary" and do not share it indiscriminatory

187. *See* GDPR, recital 51–56, 2016 O.J. (L 119) 1, 10–11 (EU); *id.* art. 9, at 38 (derogations include explicit consent of the data subject (Art.9(2(a)), data processing necessary for the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject in the field of employment and social security and social protection law (Art.9(2(b)), and protection of the vital interests of the data subject or of another natural person where the data subject is physically or legally incapable of giving consent (Art.9(2(c))).

188. *Id.* art. 9, at 39.

189. *See* ARTICLE 29, ADVICE PAPER, *supra* note 177, at 8, 10.

190. *See, e.g.*, Peiyu Liu, Shouling Ji, Lirong Fu, Kangjie Lu, Xuhong Zhang, Jingchang Qin, Wenhai Wang & Wenzhi Chen, *How IoT Re-Using Threatens Your Sensitive Data: Exploring the User-Data Disposal in Used IoT Devices*, 2023 *IEEE SYMP. ON SEC. & PRIV* 3365, 3365 (2023).

191. For example, users actively indicate their presence at these locations through the platform's features such as the "check-in" option available on Facebook. *See, e.g.*, Carole Cadwalladr, *The Great British Brexit Robbery: How Our Democracy Was Hijacked*, THE GUARDIAN (May 7, 2017, 4:00 PM), https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy [perma.cc/YH3Y-7G5N]; EUR. DATA PROT. BD., GUIDELINES 8/2020, *supra* note 52, at 32; ARTICLE 29, ADVICE PAPER, *supra* note 177, at 8 (referencing a belief in climate change as a philosophical belief).

192. *See supra* Part II.

through API access points or make it available for web scraping.[193] Apart from this inferred or observed information, however, a wealth of other potentially sensitive information remains within the ambit of public accessibility.

The technological dimension of automated data collection suggests that the process itself might neither be capable of accommodating the transformation from personal data to a special category of data nor distinguishing between personal and nonpersonal data. The nature of automated data collection, particularly through web scraping, inherently lacks the capability for nuanced contextual assessment.[194] Therefore, collection practices often fail to meaningfully distinguish between personal and nonpersonal data, or to identify when ordinary personal data transition into a more sensitive category.[195]

Likewise, though access through APIs would, in principle, confer more control over the type of collected data, it still does not capture the transition of ordinary personal data into special category data.[196] As for nonpersonal and personal data collection, the ultimate qualification largely depends on the system of permissions (or restrictions) the data host embeds into the design of an API. For example, a "metasearch travel engine" platform allows public access to price information through an API, while at the same time restricting such access to personal information that it collects, like the IP addresses of the users who consulted its website.[197] Yet Facebook's API, as discussed above, provided for wide API-enabled access to a great variety of personal information, including potentially "special category" data containing personal information (e.g., users' names, email addresses, and messages) until such access became much more restrictive following *Cambridge Analytica*.[198]

---

193. *See, e.g.*, EUR. Data PROT. BD., GUIDELINES 8/2020, *supra* note 52, at 32; Case C-252/21, Meta v. Bundeskartellamt, ECLI:EU:C:2023:537, ¶ 64 (July 4, 2023).

194. By focusing mostly on elements like HTML source, headlines, links, and tables, *see* EJ Stanley, *What is Data Scraping and How to Use It: A Complete Guide*, FORTRA https://www.fortra.com/resources/guides/what-is-data-scraping-and-how-use-it [perma.cc/2YCH-4GTF] (last visited Feb. 22, 2024); Zarsky, *supra* note 32, at 1012–15.

195. *See, e.g.*, Zarsky, *supra* note 32, at 1012–15 (discussing issue in the context of big data at large).

196. *See id.*; Borgogno & Colangelo, *supra* note 83, at 9–10.

197. *See* Introduction, SKYSCANNER, https://developers.skyscanner.net/docs/intro [perma.cc/XX4K-Q5YW] https://skyscanner.github.io/slate/#api-documentation (last visited Feb. 22, 2024); *Skyscanner Privacy Policy*, SKYSCANNER (Jan. 3, 2024), https://www.skyscanner.com/media/privacy-policy? [perma.cc/AJV7-NSGS].

198. Data Protection Act 1998: Monetary Penalty Notice from the Info. Comm'rs Off., to Cathay Pacific Airways Ltd. 9–10 (2018), https://ico.org.uk/media/action-weve-taken/mpns/2617314/cathay-pacific-mpn-20200210.pdf [perma.cc/82TQ-TT5L] [hereinafter Monetary Penalty Notice].

Thus, the EU data protection framework assumes a practically binary perspective on the world of data, where only nonpersonal data is placed outside of its reach.[199] It also stipulates a general prohibition on processing of certain special categories of personal data, unless Article 9 GDPR exceptions cover such processing.[200] In contrast with this neat framework, however, the reality operates along a spectrum where the data's "identifiability," as well as its determination as particularly sensitive, are often relative and highly contextual.[201] Against this background, automated large-scale data collection offers a telling example of the technological development that challenges the ability of a data controller to engage in an assessment of the nature, scope, and risks of data processing, as well as, more crucially, to meaningfully comply.

On multiple occasions, the CJEU has ruled on different aspects of the matter of data controller compliance.[202] A prime example is a CJEU preliminary ruling on the case involving Google Spain and Google Inc.[203] The original complaint, lodged by a Spanish national, Mr. Costeja González, with the Spanish data protection authority (AEPD), concerned the request to remove or conceal the personal data about proceedings for the recovery of social security debts from the search engine results.[204] As the case advanced through the national court system and culminated in a preliminary ruling proceeding,[205] the CJEU

199. Zarsky, *supra* note 32, at 1012–13.

200. GDPR, recital 51–56, 2016 O.J. (L 119) 1, 10–11 (EU); *id.* art. 9, at 38.

201. Sophie Stalla-Bourdillon & Alison Knight, *Anonymous Data v. Personal Data—A False Debate: An EU Perspective on Anonymisation, Pseudonymisation and Personal Data*, 34 WIS. INT'L L.J. 284, 301 (2016); Karen McCullagh, *Data Sensitivity: Proposals for Resolving the Conundrum*, 2 J. INT'L COM. L. & TECH. 190, 190–201 (2007).

202. *See, e.g.*, Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317, ¶¶ 72, 98 (May 13, 2014).

203. *See id.*

204. *Id.* at ¶¶ 14–15. Initially, the information was published in a printed version of the newspaper in early 1998. *Id.* at ¶ 14. It appeared in search engine results after the newspaper was digitalized and subsequently indexed by the search engine. *Id.* at ¶¶ 14–15, 19.

205. The preliminary ruling procedure is a mechanism of cooperation between the CJEU and national courts with the view of ensuring the effective and uniform application of the EU law. Court of Justice of the European Union, Recommendations to National Courts and Tribunals in Relation to the Initiation of Preliminary Ruling Proceedings, 2019 O.J. (C 380) 1, 2. Under Article 267 of the TFEU, the CJEU has jurisdiction to give preliminary rulings on the interpretation of Union law and on the validity of acts adopted by the institutions, bodies, offices or agencies of the Union. Consolidated Version of the Treaty on the Functioning of the European Union art. 267, 2012 O.J. (C 326) 47, 164 [hereinafter TFEU]. For more on the role, decision-making and cooperation mechanisms under 267 TFEU, *see* Court of Justice of the European Union, Recommendations to National Courts and Tribunals in Relation to the Initiation of Preliminary Ruling Proceedings, 2019 O.J. (C 380) 1.

had to consider the qualification of web scraping in terms of "data processing."[206]

First, the Advocate General[207] (AG) suggested that the fact some data have the quality of "personal data" but might not be indistinguishable as such for the web search provider does not change the qualification of its activity as "data processing."[208] The Advocate General also pointed out that, as the personal data in the source web pages appear "in a certain sense random," no technical or operational differences may exist for the functions of the search engine concerning targeting all web pages accessible on the internet.[209] In issuing its findings, the CJEU largely agreed with this assessment.[210] The court acknowledged that search engines undertake operations that are explicit examples of data processing, emphasizing that its finding was not affected even when the search engine performs the same operations in respect of other types of data.[211] The argument that the search engine does not distinguish between other types of data and personal data similarly did not alter the finding of the court,.[212]

The CJEU expanded this logic even further in a 2023 case involving Meta.[213] Emphasizing a significant risk to the fundamental rights and freedoms in cases of processing a special category of personal data, the CJEU underscored a general prohibition against such processing unless any of the exceptions of Article 9(2) GDPR applied.[214] As the CJEU pointed out, this fundamental prohibition existed

206. *See* Case C-131/12, *Google Spain* at ¶¶ 21–31.

207. The institute of Advocate Generals (AG) was established by art. 252 of the TFEU. TFEU, *supra* note 205 art. 252. The provision stipulates that "[i]t shall be the duty of the Advocate-General, acting with complete impartiality and independence, to make, in open court, reasoned submissions on cases which, in accordance with the Statute of the Court of Justice of the European Union, require his involvement." *Id.* The CJEU is assisted by eight AGs who are appointed for a six-year term. *Id.* art. 252–53. In proceedings before the CJEU, AGs typically "frame" the case and legal arguments used; summarizes and systematizes available case law on the matter; and submits a non-binding opinion ahead of the delivery of the judgment by the CJEU. *See*, *e.g.*, Michal Bobek, *A Fourth in the Court: Why Are There Advocates-General in the Court of Justice?*, 14 CAMBRIDGE Y.B. EUR. LEGAL STUDS. 529 (2012) (discussing history and role of the AG).

208. Opinion of Advocate General Jääskinen ¶ 72, Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (June 25, 2013), https://curia.europa.eu/juris/document/document.jsf?text=&docid=138782&doclang=EN [perma.cc/S7L7-9653].

209. *Id.*

210. *See* Case C-131/12, *Google Spain* at ¶ 28.

211. *Id.*

212. *See id.* at ¶¶ 28–29.

213. *See* Case C-252/21, Meta v. Bundeskartellamt, ECLI:EU:C:2023:537, ¶¶ 64–70, 89 (July 4, 2023).

214. *Id.* at ¶¶ 64–67.

"independent of whether the information revealed is correct and of whether the controller is acting with the aim of obtaining information that falls within one of the special categories."[215] Since the prohibition applies regardless of the stated purpose of the processing, the mere prerequisite of collection of data might qualify as a special category and thus trigger the heightened regimes of protection.[216] In the *Meta* case, the court addressed a situation where data are collected "*en bloc* without it being possible to separate the data items from each other at the time of collection."[217] The court highlighted that such a scenario is subject to the processing regime for special categories of data under Article 9 of the GDPR, which attaches "if [the data] contain[] at least one sensitive data item and none of the derogations in Article 9(2) of that regulation applies."[218]

Thus, confronted with cases of automated data collection where no technological means of distinguishing between personal versus nonpersonal and ordinary versus special categories of personal data exists, the court not only confirmed the wide reach of the EU data protection law in principle, but also upheld the two-tier protection regime. Consequently, the CJEU effectively extended a stricter requirement of compliance on a whole dataset, as long as it included "one sensitive data item."[219]

B. On Data Controllers

A data controller is a natural or legal person, public authority, agency, or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data.[220] As data protection provisions strive to provide for "effective and complete protection" of data subjects,[221] they place the ultimate responsibility of compliance on the party that actually exercises control over data processing.[222] As the Article 29 Working Party clarified and the European Data Protection Board subsequently confirmed, one might infer this control from a variety of attributes, the assessments of which

215. *Id.* at ¶ 69.

216. *Id.* at ¶ 70.

217. *Id.* at ¶ 89.

218. *Id.*

219. *See id.*

220. GDPR, art. 4, 2016 O.J. (L 119) 1, 33 (EU).

221. Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317, ¶ 34 (May 13, 2014); Case C-25/17, Jehovan todistajat, ECLI:EU:C:2018:551 ¶ 66 (July 10, 2018).

222. ARTICLE 29, OPINION 1/2010, *supra* note 161, at 9.

should be executed in a factual rather than a formal analysis.[223] Thus, the capacity to exert actual influence over data processing might be explicitly laid down in legal provisions. It might also stem from implicit competence accompanying the assumed roles in the relationships from other legal contexts, such as labor and contract law. The overarching principle, however, is that control should be effective rather than nominal. Courts should check the existence of such a control against the factual circumstances of the case.[224]

In carrying out the respective inquiry, one needs to address two central issues: first, who determines the "why" and the "how" of the data processing at stake; and second, if any particular circumstances suggest joint controllership. Both questions essentially require the identification of the party acting as a data controller. As the GDPR explicitly mentions the possibility of joint controllership in its definition of a "data controller," the starting analytical point for both questions is fundamentally identical.[225] In cases of joint controllership, however, the GDPR imposes a specific requirement to make arrangements among joint controllers regarding their respective responsibilities for compliance with GDPR obligations.[226]

As this Article will discuss below, the decisional practice of the CJEU and guidelines of the national data protection authorities promote a particularly pragmatic and policy-driven view of data collection controllership. The "practical assumptions" stemming from the relevant court decisional practice are different in cases of the API-enabled and web scraping data collection methods.[227] The assessment

223. *See id.* at 8–9; EUR. DATA PROT. BD., GUIDELINES 07/2020 ON THE CONCEPTS OF CONTROLLER AND PROCESSOR IN THE GDPR 10–11 (2020), https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202007_controllerprocessor_en.pdf [perma.cc/JD9Z-XQ3T] [hereinafter EUR. DATA PROT. BD., GUIDELINES 07/2020].

224. *See* Opinion of Advocate General Mengozzi ¶ 68, Case C-25/17, Jehovan todistajat (Feb. 1, 2018), https://curia.europa.eu/juris/document/document.jsf?docid=198949&doclang=en [perma.cc/6RBK-UGSA] ("For the purposes of determining the 'controller' within the meaning of Directive 95/46, I am inclined to consider . . . that excessive formalism would make it easy to circumvent the provisions of Directive 95/46 and that, consequently, it is necessary to rely upon a more factual than formal analysis . . . ."). The factual analysis implies the establishment of data controllership based on the ability of the entity to "exert actual influence over the data processing," "by virtue of an exercise of decision-making power" in concrete circumstances of the case. *See* EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 9–12.

225. *See* GDPR, art. 4, 2016 O.J. (L 119) 1, 33 (EU).

226. *Id.* art. 26 rec. 79.

227. *See* discussion *infra* Sections III.B.1, III.B.3; ARTICLE 29, OPINION 1/2010, *supra* note 161, at 9 ("[T]he need to ensure effectiveness requires that a pragmatic approach is taken with a view to ensure predictability with regard to control. In this perspective, rules of thumb and practical presumptions are needed to guide and simplify the application of data protection law.

of the technological attributes in the former instance suggests a strong indication of the joint controllership. The technological element in the latter case has been predominantly internalized in constructing a specific regime of compliance in selected cases of the search engines only. It is important to recognize that while both demonstrated approaches to data controllership ultimately aim to enhance compliance, their effectiveness, impact and precedential value differ. The first approach, which advocates for joint responsibility, can enhance accountability and data protection, yet it may also dilute responsibility and create a false sense of data subjects' control over data processing. The narrow focus of the second approach, on the other hand, might restrict its applicability and effectiveness across broader contexts.

1. Programmatic Access and Controllership

Programmatic access to and extraction of data through APIs encompass a wide range of decision-making configurations underlying the data processing operation. Since API providers hold a principal, gatekeeping position shaping and moderating "accessibility" of data, they commonly assume the role of a data controller, exercising decision-making power over the purpose and means of the data processing.[228] When it comes to the entity using the API access to the collected data, however, the required assessment becomes less straightforward.

In principle, the use of a common data-processing system or infrastructure does not necessarily imply joint controllership.[229] However, following available CJEU case law and EDPB Guidelines, it is reasonable to assume that joint controllership is present in an overwhelming number of programmatic data access cases merely by virtue of the technological configuration.[230]

Joint controllership covers a wide variety of cases, from instances where two or more parties decide on the "purposes and means" of data processing together to cases where parties' decisions on purposes and means of data processing are complementary yet

---

This calls for an interpretation of the Directive ensuring that the "determining body" can be easily and clearly identified in most situations, by reference to those - legal and/or factual - circumstances from which factual influence normally can be inferred, unless other elements indicate the contrary.").

228. *See, e.g.*, Case C-210/16, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v.Wirtschaftsakademie Schleswig Holstein GmbH, ECLI:EU:C:2018:388, ¶ 30 (June 5, 2018).

229. EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 20.

230. *See, e.g.*, Case C-131/12, Wirtschaftsakademie Schleswig Holstein, ECLI:EU:C:2018:388 at ¶¶ 38–39.

distinct.[231] The former category of joint participation—"common" decision-making—comprises a group of cases where the "common intention" of the parties is evident, meaning the parties determine purposes and means of data processing together. That could be, for example, where entity A provides API access to entity B for carrying out a joint research project or a marketing undertaking. Given that these joint endeavors potentially include a myriad of data processing operations, joint controllership is present only with regard to those operations where parties collectively make decisions on purposes and means.[232] That is to say, if the entity B decides to use the received data for an auxiliary parallel project without any coordination with or participation by entity A (the API provider), entity B becomes the sole data controller with regard to this particular data-processing operation.[233]

Another category of joint participation, converging decision-making, draws on CJEU rulings regarding joint controllership.[234] It practically aims at capturing more nonlinear, "converging," and interdependent data-processing patterns common in a modern networked ecosystem.[235] The classification of joint controllership in these cases hinges on the presence of an "inextricable link," or the "inseparability" of data processing activities. In other words, it encompasses instances where two or more parties engage in data processing practices that are so connected that they would not be possible without both parties' participation.[236] Furthermore, a party's decisions concerning this data processing must be complementary and necessary to the effect of "having a tangible impact on the determination of the purposes and means."[237]

For parties to implement a required appraisal following the EDPB Guidelines, they must analyze their data-processing activities, breaking them down into individual operations to determine which

231. EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 18.

232. *See* Case C-40/17, Fashion ID GmbH & Co. KG v. Verbraucherzentrale NRW eV, ECLI:EU:2019:629 ¶ 74 (July 29, 2019) ("By contrast, and without prejudice to any civil liability provided for in national law in this respect, that natural or legal person cannot be considered to be a controller, within the meaning of that provision, in the context of operations that precede or are subsequent in the overall chain of processing for which that person does not determine either the purposes or the means.").

233. *See* EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 21–22.

234. *See, e.g.*, Case C-131/12, *Wirtschaftsakademie Schleswig Holstein*, ECLI:EU:C:2018:388 at ¶¶ 35–39; Case C-40/17, *Fashion ID* ¶¶ 76–78; Case C-25/17, Jehovan todistajat, ECLI:EU:C:2018:551 ¶¶ 69–75 (July 10, 2018).

235. *See, e.g.*, Mahieu, Hoboken & Asghari, *supra* note 140, at 87.

236. EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 18.

237. *Id.*

involve another party's participation.[238] Then, the party must examine whether respective decisions are "convergent" on a point of determining the purpose and means of the processing.[239] The EDPB Guidelines neither elaborate on the precise level and scope of the required convergency of decision-making, nor thoroughly discuss the various factors that would indicate decision-making complementarity.[240] The EDPB Guidelines merely refer to a nub of the "tangible impact" on the determination of the purposes and means.[241] They also point out that such purposes could be qualified as linked or complementary when "a mutual benefit arising from the same processing operation" characterizes them.[242]

Following CJEU case law and EDPB Guidelines, an API-enabled data extraction strongly suggests an inseparable nature of data processing. As this Article previously discussed, an API provider exercises a principal control over the manner of data collection.[243] It defines both the objectives of data processing and the manner of obtaining said objectives.[244] It can also unliterally change these rules.[245]

However, the entity receiving data might still have a certain margin for decision-making. For example, it can choose a particular developer's status, such as accessing data using an "in-house" or a "third-party" API solution.[246] Additionally, it has the flexibility to select APIs offering different functionalities, such as exclusively retrieving data or integrating data manipulation and social media capabilities.

---

238. *See id.* at 10, 14 (2021), (discussing the assignment of responsibility according to stages of data processing); Case C-40/17, *Fashion ID* ¶ 72 ("[T]he processing of personal data may consist in one or a number of operations, each of which relates to one of the different stages that the processing of personal data may involve."); *Id.* at ¶ 74 ("[A] natural or legal person cannot be considered to be a controller, within the meaning of that provision, in the context of operations that precede or are subsequent in the overall chain of processing for which that person does not determine either the purposes or the means."); Case C-683/21, Nacionalinis visuomenės sveikatos centras prie Sveikatos apsaugos ministerijos v. Valstybinė duomenų apsaugos inspekcija, ECLI:EU:C:2023:949 ¶ 42 (Dec. 5, 2023).

239. EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 18, 48.

240. *See generally id.* at 19. Thus, in most recent case involving joint controllers, the CJEU merely reiterated the EDPB Guidelines on the point. *See id.*; Case C-683/21, *Nacionalinis Visuomenės Sveikatos* ¶ 42.

230. *See* EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 18.

242. *Id.* at 19. It is important to note, however, that the EDPB Guidelines 07/2020 point out that a "mutual benefit" attribute is not a decisive, but rather an indicative attribute. *Id.*

243. *Id.* at 13.

244. *Id.*

245. Hogan, *supra* note 115, at 594.

246. *See* Case C-210/16, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v.Wirtschaftsakademie Schleswig Holstein GmbH, ECLI:EU:C:2018:388 ¶¶ 30–44 (June 5, 2018).

Also, the data-receiving entity using API-enabled access ultimately has a "decisive influence" over the *general terms* of data retrieval from a particular data repository. For example, entities using API access can select specific configurations, such as the number of data requests, thereby setting key variables that influence the overall data flow and the availability of data for other entities.

This ability to define such parameters exemplifies control over data processing as underscored in CJEU case law concerning joint controllership. A notable case illustrating this point is *Wirtschaftsakademie Schleswig-Holstein* ruling where a German company was involved in providing educational services via a Facebook fan page, thus raising questions about joint controllership responsibilities[247] Upon examining the role each party played in processing the data of Facebook users, the CJEU concluded that the case actually involved two data controllers.[248] It deemed Facebook a data controller by virtue of "primarily determining the purposes and means of processing the personal data of users of Facebook and persons visiting the fan pages hosted on Facebook."[249] At the same time, the CJEU classified *Wirtschaftsakademie Schleswig-Holstein* as a data controller owing to the organization's ability to "define parameters" of the overall data collection occurring under its purview.[250] Defining parameters entailed exerting influence over data processing by "targeting specific audience" and pursuing the objectives of "managing and promoting its activities" on Facebook.[251] Since the respective decisions had a decisive impact on Facebook statistics of visits to the page, the CJEU concluded that the German company contributed to the processing of personal data in the capacity of a joint controller rather than on an individual basis.[252]

In a similar case, *Fashion ID*, a German online clothing retailer had embedded the Facebook "like" social plugin on its website.[253] According to the CJEU, Fashion ID acted as a joint controller with regard to the processing of personal data of Facebook users through collecting and transmitting their data to the Fashion ID website.[254] As the CJEU pointed out, such data processing would have not happened

247. *Id.* at ¶ 2.

248. *See id.* at ¶¶ 31–39.

249. *Id.* at ¶ 30.

250. *Id.* at ¶¶ 30, 36.

251. *Id.* at ¶ 39.

252. *See* Case C-210/16, Wirtschaftsakademie, Judgment ¶¶ 36, 38, 39.

253. Case C-40/17, Fashion ID GmbH & Co.KG v. Verbraucherzentrale NRW eV, ECLI:EU:C:2018:1039, Opinion ¶ 1 (July 29, 2019).

254. *See id.* ¶¶ 66–70.

without Fashion ID installing the Facebook social login in the first instance.[255] Thus, the *Fashion ID* decision to embed third-party content that result in data collection was complementary and necessary to the effect of "having a tangible impact on the determination of the purposes and means" of data processing.[256]

Drawing on these judgments, one can assume joint-controller status by merely "enabling" an act of personal data collection and subsequently exercising rather limited influence over the parameters of such collection.[257] In principle, the entity receiving data in cases of API-enabled data collection will almost always meet this rather low threshold. By engaging in data collection practices, it initiates an act of data processing that would have not occurred otherwise. When the entity opts for certain configurations of API access, it exerts enough influence to have some form of control over how the data processing operation proceeds.

Lastly, regarding the "complementarity" of data processing purposes, the "mutual benefit"—which is not a determinative but rather an indicative attribute of joint controllership, according to the EDPB Guidelines[258]—could be presumed in a large subset of API cases by default.

As the CJEU expounded in *Fashion ID*, a "mutual benefit" could take form of a "benefit from the commercial advantage."[259] In the case at hand, Fashion ID embedded the Facebook "Like" button in order to benefit from increased publicity for its goods.[260] The data-processing operations were performed "in the economic interests of both Fashion ID and Facebook Ireland, for whom the fact that it can use those data for its own commercial purposes is the consideration for the benefit to Fashion ID."[261] The examination of the "mutual benefit" limb in the *Fashion ID* case offers a telling illustration of an overall economic

255. *See id.* ¶ 78.

256. *See* EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 18; Case C-683/21, Nacionalinis visuomenės sveikatos centras prie Sveikatos apsaugos ministerijos v. Valstybinė duomenų apsaugos inspekcija, ECLI:EU:C:2023:949, Judgment ¶ 43 (Dec. 5, 2023); EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 20.

257. *See* Finck & Pallas *supra* note 138, at 3 (2021).

258. *See* EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 19. EDPB Guidelines 07/2020 do not offer an extensive elucidation of the meaning of "mutual benefit" in the context of joint controllership. The sole clarification available is provided through the observation that "the mere existence of a mutual benefit (for ex. commercial) arising from a processing activity does not give rise to joint controllership." *See id.* at 20.

259. *See* Case C-40/17, Fashion ID GmbH & Co. KG v. Verbraucherzentrale NRW eV, ECLI:EU:C:2019:629, Judgment ¶ 80 (July 29, 2019).

260. *See id.*

261. *Id.*

reality: utilizing APIs for automated data collection is a central business strategy decision for many platforms and web services.[262] Essentially, APIs promote interoperability and allow for outsourcing software development.[263] Thus, API-enabled data collection, in principle, extends the functionality and increases the versatility of platforms and web services, in turn greatly benefiting API providers.[264]

Following the analytical framework of CJEU case law and EDPB Guidelines on data controllership, API-enabled data collection exhibit a strong indication of joint controllership, primarily owing to the technical configuration of respective data processing operations. In effect, a great number of the entities should recognize themselves as joint controllers if and when they access data through APIs. Given that automated data collection often occurs inconspicuously,[265] it remains

262. *See, e.g.*, DANIEL JACOBSON, GREG BRAIL & DAN WOODS, APIS: A STRATEGY GUIDE 11–36 (2012).

263. *See* Robert Bodle, *Regimes of Sharing: Open APIs, Interoperability, and Facebook*, 14 INFO., COMMC'N & SOC'Y 320 (2011); Borgogno & Colangelo, *supra* note 83, at 1–4.

264. *See* Taina Bucher, *Objects of Intense Feeling: The Case of the Twitter API*, COMPUTATIONAL CULTURE (2013), http://computationalculture.net/objects-of-intense-feeling-the-case-of-the-twitter-api/ [perma.cc/UQE2-6MSK].

265. *See, e.g.*, Monetary Penalty Notice, *supra* note 198, at 10 ("[t]o the extent that the App had access to the identity of those who had exchanged Facebook messages with a user of the App, or to the content of such messages, the individuals who had exchanged such messages with users of the APP: were not informed that the App was given access to such information; and were not asked to consent to such access"). It has to be noted in this regard that the duty to inform data subjects of such an access in the European Union is also regulated by Article 5(3) of Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 concerning the processing of personal data and the protection of privacy in the electronic communications sector. Council Directive 2009/136/EC, 2009 O.J. (L 337) 28 (EC). The Directive is widely known for establishing the requirement of presenting cookie notice and obtaining consent for data subjects in cases of information access. The attempts of revisiting the Directive have been ongoing for years, with the EU Commission Proposal for a new Regulation on Privacy and Electronic Communications (ePR) dating back to 2017. As the proposal hit stalemate, several legislative acts and initiatives have emerged to address some of the contentious issues in the meantime. For example, the European Data Protection Board recently published the Guidelines 2/2023 on Technical Scope of Art. 5(3) of ePrivacy Directive and announced the public consultation on the matter (closed Jan. 18, 2024). The Guidelines intend to "remove ambiguities related to the application of Art. 5(3) to tracking tools". The EDPB also published the Guidelines 03/2022 on Dark patterns in social media platform interfaces (Feb. 14, 2023) tackling the issue of "unintended, unwilling and potentially harmful decisions" taken by data subjects in regards of their personal data. The problem of transparency, information duty obligation and users' choice online has been widely debated in the EU and led to a number of legislative outputs. *See* Inge Graef, *The EU Regulatory Patchwork for Dark Patterns: An Illustration of an Inframarginal Revolution in European Law?*, *in* Ramsi A. Woodcock, TOWARD REVOLUTION: MARKETS AS WEALTH DISTRIBUTERS (forthcoming 2023).

unclear whether this approach yields the desired outcome of "effective and complete protection of data subjects."[266]

2. Web Scraping and Data Controllership

As this Article has discussed, web scraping generally denotes the employment of an agent to download, parse, and arrange data in an automated manner.[267] In the case of Google, terms such as "crawling," "indexing," and "listing" capture these processes.[268] Web users typically perceive the sequence of these steps as a seamlessly consolidated and momentary action.[269] However, the examination of data processing operations behind this "momentary action" encompasses several key EU rulings on data controllership.[270]

As discussed above, the *Google Spain* case provides a telling example of the CJEU approach to web scraping at large.[271] While both the Advocate General and CJEU agreed that the processes comprising search engine activity amounted to "data processing,"[272] they clearly disagreed on how that finding expressed itself in terms of data controllership. As the Advocate General argued:

> "[T]he general scheme of [Data Protection] Directive, most language versions and the individual obligations it imposes on the controller are based on the idea of responsibility of the controller over the personal data processed in the sense that the controller is aware of the existence of a certain defined category of information amounting to personal data and the controller processes th[ese] data with some intention which relates to their processing as personal data."[273]

The AG then emphasized that the search engine could not distinguish personal data from other data in the course of crawling, had no control over or relationship with the content of third-party source web pages, and did not "in law or in fact" fulfill the obligations of a data

---

266. *See* Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), Judgement, ECLI:EU:C:2014:317 ¶ 34 (May 13, 2014); Case C-210/16, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v. Wirtschaftsakademie Schleswig Holstein GmbH, ECLI:EU:C:2018:388 ¶¶ 28 (June 5, 2018).

267. *See* JERSEY OFFICE OF THE INFO. COMM'R, *supra* note 136.

268. *See, e.g.*, U.S. Patent No. 6,285,999 B1 (filed Jan. 9, 1998).

269. *See How Google Search Works*, GOOGLE https://www.google.com/search/howsearchworks/how-search-works/ [perma.cc/A6PS-RXEA] (last visited Feb. 4, 2023).

270. *See, e.g.*, Case C-131/12, *Google Spain*, Judgement ¶ 1; Case C-136/17, GC, AF, BH, ED v. Commission Nationale de L'informatique et des Libertés (*GC and Others*), ECLI:EU:C:2019:773, Judgement ¶ 1 (Sept. 24, 2019); Case C-460/20, TU & RE v. Google, ECLI:EU:C:2022:962, Judgment ¶ 1 (Dec. 8, 2022).

271. *See* Case C-131/12, *Google Spain*, Judgement ¶ 18.

272. *See* Opinion of Advocate General Jääskinen, *supra* note 208, ¶ 83.

273. *Id.* ¶ 82.

controller in relation to the personal data on source web pages.[274] From these observations, the AG concluded that the internet search engine service provider cannot be defined as a data controller, at least in terms of its crawling function.[275] However, according to the AG, that does not then mean that the search engine is not a data controller for the purposes of other relevant processes.[276] In the view of the AG, these processes primarily relate to the search engine's index.[277] Ahead of the CJEU issuing its guidance, the AG proposed that Google exercises its influence in establishing an "information location tool." In effect, Google decides how to actually structure its index, including[278] whether certain search results are blocked.[279] Thus, the AG appears to have drawn a distinct line between the initial step of automated data collection and Google's subsequent operations, including establishing and presenting a directory of web pages accessible on the internet.[280] According to the AG's opinion, only at the latter stage does Google assume the role of a data controller by exercising control over the data processing in question.[281]

However, appealing to the definition of a "data controller" and its principal objective to provide for "effective and complete protection of data subjects," the CJEU ruled that Google is a "data controller."[282] The court did not meaningfully engage with the AG's argument that the interpretation of the Directive should be based on "a rule of reason, in other words, the principle of proportionality."[283] Neither did it expand on its own statement that the search engine must ensure, "within the framework of its responsibilities, powers and capabilities," that all

274. *See id.* ¶¶ 86–89.

275. *See id.* ¶¶ 84, 86, 89.

276. *See id.* ¶¶ 91–93.

277. *See id.* ¶ 91.

278. *See id.*; *see, e.g.*, *Indexing Pages to be Included in Search Results*, GOOGLE https://support.google.com/programmable-search/answer/4513925?hl=en https://support.google.com/programmable-search/answer/4513925?hl=en [perma.cc/KL75-ECUJ] (last visited Jan. 28, 2024) (in Google's words, the Google index is akin to an index in a library, with a difference being that the listing concerns all of the webpages instead of books).

279. Opinion of Advocate General Jääskinen, *supra* note 208, ¶ 91.

280. *See id.* The AG seems to connect this transition—from Google-as-intermediary to Google-as-data controller—with a transformation from processing a mere computer code that might contain personal data, to processing data on identified or identifiable natural person in some semantically relevant way.

281. *See id.* ¶¶ 9, 93 (as clarification, the AG also points out that the control is also evident in a decision of Google not to comply with the exclusion codes).

282. *See* Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317, Judgement, ¶ 33–34 (May 13, 2014).

283. *See* Opinion of Advocate General Jääskinen, *supra* note 208, ¶¶ 30, 79, 88.

processing complies with provisions of the Directive.[284] The court did emphasize, however, the important role the internet and search engines play in modern society.[285] The processing at stake, according to the CJEU, is "liable to affect significantly the fundamental rights to privacy and to the protection of personal data": it enables any internet user to obtain through the list of results a structured overview of the information relating to individuals.[286] This information "potentially concerns a vast number of aspects of [an individual's] private life" and would have not been available as easily without the search engine.[287] The court also affirmed that although some website publishers have the option of making particular information unavailable for Google's processing, this option does not absolve the search engine from its obligations as a data controller.[288]

Thus, while the AG Opinion and CJEU concurred on qualifying the search engine activity as "data processing," they differed in their assessment of which party assumes the role of data controller.[289] One of the reasons for such a divergence seems to lie in the distinct vantage points from which the AG and the court reviewed the case. The AG's perspective was based on a rather pragmatic objective of avoiding the expansion of the data protection obligations to the extent they might challenge the very lawfulness of the search engine functioning.[290] Despite the underlying concern, however, it is critical to note a certain level of artificiality in the AG's approach of dissecting Google's search operations to pinpoint a starting point of data controllership. It appears that the respective decisions as to how to structure indexing and whether to comply with exclusion codes are typically "inextricably linked" with the very process of accessing webpages to collect data. One can thus contend that the actual decisions on the issues are made when a developer first writes the code, predating the software's launch.

Applying the AG Opinion's reasoning, the hypothetical operator of a camera system would not act as a data controller regarding the camera's data collection until he watched the recording and assessed the automatic processing of that personal data.[291] However, the

284. Case C-131/12, *Google Spain*, Judgement ¶ 83.

285. Opinion of Advocate General Jääskinen, *supra* note 208, ¶ 36.

286. Case C-131/12, *Google Spain*, Judgement ¶ 80.

287. *Id.*

288. *See* Opinion of Advocate General Jääskinen, *supra* note 208, ¶¶ 30, 40.

289. *Cf. id.* ¶ 89; Case C-131/12, *Google Spain*, Judgement ¶¶ 33, 34.

290. *See* Opinion of Advocate General Jääskinen, *supra* note 208, ¶¶ 89, 90.

291. *Compare id. with* Case C-212/13, František Ryneš v. Úřad pro Ochranu Osobních Údajů, ECLI:EU:C:2014:2428, Judgment ¶¶ 33–35 (Dec. 11, 2014). *Ryneš* concerned applicability

operator of the camera system installed it in the first place, decided on its positioning, the timing of recording, etc.[292] Thus, affixing the status of a data controller to the operator's inspection of the video material would, in effect, place the camera's initial collection of the data outside of the scope of data protection law altogether.[293] Against this background, the approach of the CJEU in recognizing Google as a data controller throughout the entire data lifecycle appears to be a strategy of avoiding granting blanket immunity exactly in those cases.[294]

Following the *Google Spain* reasoning, it is possible to advance that the entity behind the web scraping will most likely act as a data controller throughout the entirety of the data collection.

3. On What Makes Google Google

Search engine activity concerns the totality of data website publishers and content providers upload to the web. For the purposes of data protection law, these parties should be deemed the primary decision-makers and data controllers concerning what data ultimately become available online. Both the AG and CJEU duly acknowledge this existence of distinct data processing activities website publishers and search engines perform.[295]

As the CJEU pointed out in the *Lindqvist* case, loading personal data on an internet page constitutes data processing that results in information becoming available for viewing "by an indefinite number of people living in many places at almost any time."[296] Furthermore,

---

of the EU Data Protection Directive to individuals using surveillance cameras for personal security purposes. While examining the data controllership status, the CJEU did not delineate the timeline of the data processing activity to attribute the status of a data controller to a specific event, such as Mr. Rynes' decision to review the recording. *See id.*

292. *See* Case C-212/13, *Ryneš*, Judgment ¶ 13 (such as decisions to direct recording to the entrance of the house, the public footpath, and the entrance to the house opposite).

293. *See* ARTICLE 29, OPINION 1/2010, *supra* note 161, at 9 (2010) (which in itself contradicts the EU data protection approach to a data controller as a "functional concept," the aim of which is to allocate responsibilities); EUR. DATA PROT. BD., GUIDELINES 07/2020, *supra* note 223, at 9 (concepts of controller and processor).

294. However, that is not to say that the CJEU position and reasoning are void of shortcomings. *See, e.g.*, Giovanni Sartor, *Search Engines as Controllers: Inconvenient Implications of a Questionable Classification*: Case C-131/12 Google v. Agencia Española de Protectión de Datos (AEPD) and Mario Costeja Gonzalez, *Judgment of 13 May 2014*, 21 MAASTRICHT J. OF EUR. & COMPAR. L. 570 (2014).

295. *See* Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317, Judgement ¶ 35 (May 13, 2014); Opinion of Advocate General Jääskinen, *supra* note 208, ¶ 39; *see also* Case C-460/20, TU & RE v. Google, ECLI:EU:C:2022:962, Judgment ¶ 50, (Dec. 8, 2022).

296. Case C-101/01, Bodil Lindqvist, ECLI:EU:C:2003:596, Judgment ¶¶ 25, 58 (Nov. 6, 2003).

depending on the status of the content provider, type of information, and underlying motivation behind making information public, this act may constitute a manifestation of the freedom of expression the EU Charter of Fundamental Rights protects.[297]

Turning back to automated data collection, the ways in which data had been made public in the first place were not a primary concern of the *Google Spain* ruling.[298] Furthermore, despite the AG Opinion's invitation to closely consider competing rights of freedom of expression and information as well as the right to conduct a business, the CJEU did not follow that path.[299] Instead, the CJEU limited its holding to a mere mention of the "interest of the general public [...] in having access to the information" while considering the scope of the data subject's right to require the search engine to delist personal information from its index.[300] The court primarily focused on the effect Google's activity had on providing for increased accessibility, interconnectedness, and ubiquity of data.[301] It was against this backdrop that the court mandated that search engines must ensure, "within the framework of its responsibilities, powers and capabilities,"[302] that their data processing activities comply with the provisions of data protection law.[303]

Considering this ruling in terms of data accessibility and collection, it is worth highlighting the vagueness of the framework of "responsibilities, powers and capabilities" for tailoring liability of the

297. *See id.* ¶ 35; Opinion of Advocate General Jääskinen, *supra* note 208, ¶¶ 103, 129. In the European Union, the right to freedom of expression is protected in Article 11 of the Charter, which essentially corresponds to Article 10 of the European Convention of Human Rights (ECHR). *See* EUR. UNION AGENCY FOR FUNDAMENTAL RTS. & COUNCIL OF EUR., HANDBOOK ON EUROPEAN DATA PROTECTION LAW 54–69 (2018).

298. *See* Case C-131/12, *Google Spain*, Judgement ¶ 62 (the questions submitted for a preliminary ruling posited on a premise that publication of information was lawful).

299. *See* Opinion of Advocate General Jääskinen, *supra* note 208, ¶ 120.

300. *See* Case C-131/12, *Google Spain*, Judgement ¶ 97 (note also the court's choice in presenting the dilemma to mention "interest" rather than the right to freedom of expression and information).

301. *See* Orla Lynskey, *Grappling with "Data Power": Normative Nudges from Data Protection and Privacy*, 20 THEORETICAL INQUIRIES L. 190, 204–05 (2019).

302. *See* Case C-131/12, *Google Spain*, Judgement ¶¶ 38, 83. The notion of "responsibilities, powers and capabilities" introduced by the CJEU in *Google Spain* (paras. 38, 83) has been widely debated in the scholarship. Despite the regular recourse of the CJEU to the framework in its data protection cases, there is significant ambiguity as to the scope and normative implications of the concept. *See, e.g.*, Hielke Hijmans, *Right to Have Links Removed: Evidence of Effective Data Protection*, 21 MAASTRICHT J. EUR. & COMPAR. L. 559 (2014); Miquel Peguera, *The Shaky Ground of the Right to be Delisted*, 18 VAND. J. ENT. & TECH. L. 507 (2015); Mahieu, Hoboken & Asghari, *supra* note 140, at 40.

303. *See* Case C-131/12, *Google Spain*, Judgement ¶¶ 38, 83.

web scraping entity.[304] It remains unclear whether Google has a heightened responsibility due to its role in rendering the data content providers and publishers upload "even more available."[305] Additionally, ongoing uncertainty persists regarding the precise relationship between "responsibilities, powers[,] and capabilities," and the size[306] of Google and its approach to users' privacy.[307] More generally, the method of appraising the "responsibilities, powers and capabilities" framework with regard to web scraping entities other than Google continues to be a matter of uncertainty. In particular, it remains unclear how to apply this framework in cases of commercial databases with large-scale search functionalities,[308] search engines embedded in social network services,[309] or LLMs with internet browsing capabilities.[310]

Currently, there are no comprehensive answers to these questions, despite several recent CJEU judgments that aimed to provide more clarity.[311] For instance, the *GC, AF, BH, ED v. CNIL* case (*GC and Others*) has offered some, albeit limited, guidance on this matter.[312] In this case, the CJEU explored the framework of "responsibilities, powers and capabilities" of search engine operators (SEOs) specifically in the context of dereferencing links to third-party

304. *See id.*

305. *See, e.g.*, Mykola Makhortykh, Aleksandra Urman & Roberto Ulloa, *How Search Engines Disseminate Information About COVID-19 and Why They Should Do Better*, 1 HARV. KENNEDY SCH. MISINFORMATION REV. 3 (2020) (research on effect of search engines on information dissemination in times of the public health crisis).

306. *See Search Engine Market Share Worldwide*, STATCOUNTER, https://gs.statcounter.com/search-engine-market-share [perma.cc/YG4Z-Y3AV] (last visited Mar. 18, 2023) (According to the Statcounter, Google accounts for 93.18 percent of search queries, followed by Bing (2.87 percent) and Yandex (1.02 percent)); *see also* Lynskey, *supra* note 301, at 191.

307. *Cf.* DuckDuckGo, which is an internet search engine with a pronounced focus on users' privacy. *About DuckDuckGo*, DUCKDUCKGO, https://duckduckgo.com/about [perma.cc/HW7M-YASN] (last visited Mar. 4, 2023).

308. *See* Christopher Kuner, *The Court of Justice of the EU Judgment on Data Protection and Internet Search Engines: Current Issues and Future Challenges* 8 (London Sch. Econ. & Pol. Sci., Law, Soc'y and Econ., Working Paper 3, 2015).

309. *See* David Lindsay, *The 'Right to be Forgotten' in European Data Protection Law*, in EMERGING CHALLENGES IN PRIVACY LAW: COMPARATIVE PERSPECTIVES 295 (Normann Witzleb, David Lindsay, Moira Paterson & Sharon Rodrick eds., 2014).

310. *See* Danny Goodwin, *Google Gemini is Here – and It's Already Being Tested in Search*, SEARCH ENGINE LAND (Dec. 6, 2023, 11:16 AM), https://searchengineland.com/google-testing-gemini-search-435516 [perma.cc/X89V-W3U7] (last visited Jan. 29, 2024).

311. *See* Case C-136/17, *GC and Others*, Judgment, ¶ 37.

312. *See id.*

websites containing sensitive personal data.[313] As the CJEU pointed out, the prohibition and restrictions for the special category of personal data apply "to every kind of processing of the special categories of data referred to in those provisions and to all controllers carrying out such processing," with no derogation for SEOs.[314] Furthermore, even if such a derogation existed, it would have to run counter to the very objective of the rule to provide for enhanced protection of sensitive data.[315]

However, the court also stressed that the specific features of the SEO's processing could have an effect on the extent of the operator's responsibility and obligations.[316] As the court underscored, the SEO's responsibility stems from referencing and displaying links that might contain sensitive information rather than because the sensitive information "appear[s] on a web page published by a third party."[317] It follows that the responsibility can apply only "because of the referencing"[318] "and therefore by means of an *ex post facto* verification under the supervision of the competent national authorities, on the basis of a request by the data subject."[319] Thus, the CJEU confirmed that no exemption from compliance with data protection law exists for search engines per se.[320] This is the case even when ascertaining the legality of processing sensitive data *ex ante* is not feasible.[321] However, in effect, the judgment proposed a particular *modus operandi* providing for a de facto assumption of the lawfulness of processing in the absence of a data subject's successful claim for de-referencing.[322]

313. *See* GC and Others, Judgment, ¶ 24–25 (the information under consideration contained a satirical photomontage of a former politician; a reference to an applicant as public relations officer of the Church of Scientology; judicial investigation relating to an applicant and a reference to an applicant found guilty of sexual assaults on children).

314. *See id.* ¶¶ 42–43.

315. *See id.* ¶ 44.

316. *See id.* ¶ 45.

317. *Id.* ¶ 46; *See* Case C-136/17, GC, AF, BH, ED v. Commission Nationale de L'informatique et des Libertés (GC and Others), ECLI:EU:C:2019:773, Opinion of Advocate General M. Szpunar ¶ 55 (Jan. 10, 2019), https://curia.europa.eu/juris/document/document_print.jsf?mode=req&pageIndex=0&doc-cid=209686&part=1&doclang=EN&text=&dir=&occ=first&cid=2646504 [perma.cc/3YV3-SB5V].

318. Case C-136/17, *GC and Others*, Judgment ¶ 46 (referencing C-131/12, *Google Spain* ¶ 80 and its potential to "significantly affect the data subject's fundamental rights to privacy and to the protection of the personal data relating to him").

319. *Id.* ¶¶ 56, 47 (citing Opinion of Advocate General M. Szpunar, *supra* note 317).

320. *Id.* ¶ 45.

321. *Id.* ¶ 48.

322. *Id.* ¶ 47.

The findings were further confirmed in the CJEU's recent ruling in *RE v. Google*.[323] The case concerned a request to dereference links to content containing allegedly inaccurate claims as well as the request to de-reference applicants' photographs in the form of preview images ("thumbnails").[324] By affirming distinct layers of data processing, the CJEU pointed to a "significant" and "additional" effect of the search engine activity on the fundamental rights to privacy and protection of personal data.[325] The court also emphasized that "referencing" the page facilitates access to information with respect to individuals and ultimately may play a "decisive role" in the dissemination of such information.[326]

The case essentially sought guidance on two separate balancing exercises. The first concerned the burden of proof in cases of dereferencing in general, and the second related to the scope of the delisting obligation with respect to the thumbnails.[327]

Considering the former weighing-up exercise of balancing conflicting fundamental rights,[328] the CJEU specified that data subjects shall exert "reasonably . . .required" efforts to establish manifest inaccuracy of the data they seek to remove from the search results.[329] Concurrently, the CJEU acknowledged the risks of imposing an excessive burden on the search engine.[330] In this regard, the ruling was a natural extension of *GC and Others*, affirming a special *ex post* verification regime for the search engine.[331] However, the CJEU, in stark contrast to its first ruling on the matter nearly a decade earlier,

323. *See* Case C-460/20, TU & RE v. Google, ECLI:EU:C:2022:962, Judgment ¶¶ 49–53, 56–58 (Dec. 8, 2022).

324. *Id.* ¶¶ 19–20.

325. *Id.* ¶¶ 50–52.

326. *See e.g.*, Opinion of Advocate General Pitruzzella ¶ 15, Case C-460/20, TU v. Google, ECLI:EU:C:2022:962 (Dec. 8, 2022), https://curia.europa.eu/juris/document/document.jsf?docid=257515&doclang=EN [perma.cc/2GQW-EU3F]; Case C-460/20, TU & RE v. Google, ECLI:EU:C:2022:962, Judgment ¶¶ 50, 52, 93 (Dec. 8, 2022).

327. Case C-460/20, *TU v. Google*, Judgment ¶¶ 48, 89.

328. GDPR, Rec. 4: "The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality." For an account of the principle of proportionality as a means of balancing competing rights and interests *See* Gráinne de Búrca, *The Principle of Proportionality and its Application in EC Law,* 13 YEARBOOK OF EUROPEAN LAW 105 (1993); Tor-Inge Harbo, *The Function of the Proportionality Principle in EU Law,* 16 European Law Journal 158 (2010); Jan H. Jans, *Proportionality Revisited,* 27 LEGAL ISSUES OF ECONOMIC INTEGRATION 239 (2000).

329. *Id.* ¶ 68.

330. *Id.* ¶¶ 73–74.

331. *Id.* ¶ 53; *see* Case C-136/17, GC, AF, BH, ED v. Commission Nationale de L'informatique et des Libertés (*GC and Others*), ECLI:EU:C:2019:773 ¶ 47 (Sept. 24, 2019).

spent considerably more time discussing the search engine's responsibility in categories of "reasonableness," "excessiveness," and "risk" of a deterrent effect on the exercise of freedom of expression and information.[332]

The second weighing-up exercise pertained to the delisting obligation with respect to the preview images of individuals.[333] Essentially, it invited the court to further engage with the issues of search engine functioning, online content creation, and data accessibility.[334] The court established that displays of photographs of data subjects constituted a "particularly significant interference with their rights to private life and individuals' personal data" [335]

Drawing on a European Court of Human Rights (ECtHR) ruling in a case on publication of photos in magazines nearly twenty years ago,[336] the CJEU pointed to a particular quality of photographs to act as one of the central attributes of one's personality by revealing a person's unique characteristics and distinguishing that person from others.[337] The reference to the ECtHR case concerning traditional photography and print media served as a timely reminder of the stark contrast in technological settings. While the ECtHR ruling took place in a context where the dissemination of printed photographs was relatively controlled and limited in scope,[338] the more contemporaneous CJEU ruling occurred against a cultural backdrop of overwhelmingly ubiquitous connectivity and the ever-evolving phenomenon of the online

---

332. Case C-460/20, *TU v. Google*, Judgment ¶ 71.

333. *Id.* ¶ 89. Preview images of individuals typically refer to small, scaled-down versions of images that depict individuals. These images are often used as previews or thumbnails in various contexts, such as search engine results, social media platforms, or online galleries. They provide a glimpse or preview of the full-size image and help users quickly identify and select the content they are interested in viewing.

334. Opinion of Advocate General Pitruzzella, *supra* note 326, ¶¶ 2–4.

335. Case C-460/20, *TU v. Google*, Judgment ¶ 94.

336. Von Hannover v. Germany (No. 2), 2012-II Eur. Ct. H.R. ¶¶ 95, 98, 103. The case concerned the publication of a series of photos of Princess Caroline of Monaco by tabloid magazines. *Id.* ¶¶ 16, 20. The photos were taken without the knowledge of the Princess and captured scenes of her daily life. *Id.* ¶ 121. The ECtHR concluded that in the case at hand the publication of photos did not contribute to a debate of general interest and the Princess's right to privacy and family life (enshrined in Article 8 of the European Convention on Human Rights) was violated. *See id.* ¶¶ 117, 124.

337. Case C-460/20, *TU v. Google*, Judgment ¶ 95.

338. At the time of the ECtHR judgment, the internet had not yet become the most dominant channel for information distribution. *See Individuals using the Internet (% of population)*, WORLD BANK, https://data.worldbank.org/indicator/IT.NET.USER.ZS [https://perma.cc/RXH5-DFTD] (last visited Feb. 4, 2023) (featuring only a handful of European countries boasting the connectivity rate over 50 percent of population).

search.[339] It was against this background that the CJEU delivered its judgment on delisting image search results.[340]

The question presented related to the "informative value" of thumbnails as search results.[341] The display of thumbnails is a decontextualization exercise where images appear separately from the text of the original publications.[342] At the same time, by virtue of the link they contain, thumbnails are naturally connected to the internet pages from which they originate.[343] Therefore, in practice, a request to dereference thumbnails naturally raises a broader question of how much, if at all, their original context matters. More generally, however, a concept of "informative value" highlights the very functioning of the search, especially in the field of graphic content.[344] As the referring court, AG, and CJEU all concurred, displaying thumbnails constitutes "autonomous" search engine processing.[345] This processing is distinct from that of the original webpage publisher; its underlying grounds for data processing might be different, and so too could be the consequences of such processing for the data subject.[346] The CJEU did not elaborate much further on normative consequences of such a distinction in terms of "responsibilities, powers and capabilities."[347] However, it noted in dicta that image results might contribute to a particularly intense interference with fundamental rights "owing to the aggregation, in a search by name, of all information concerning the data subject which is found on the internet."[348] Furthermore, the court particularly stressed the potential of photographs—as a nonverbal means of communication—to generate increased user interest, while

339. *See* Olaf Kopp, *Google MUM Update: What Can SEOs Expect in the Future?*, SEARCH ENGINE LAB (Apr. 15, 2022, 6:00 AM), https://searchengineland.com/google-mum-update-seo-future-383551 [perma.cc/W8FU-33G].

340. *See, e.g.*, Opinion of Advocate General Pitruzzella, *supra* note 326, ¶¶ 15, 28.

341. Case C-460/20, *TU v. Google*, Judgment ¶ 89.

342. The original webpages might also be removed, like was in the case under consideration. *See id.* ¶ 19.

343. *See, e.g.*, *id.* ¶ 36.

344. *See, e.g.*, Opinion of Advocate General Pitruzzella, *supra* note 326, ¶¶ 56–58.

345. Summary of the Request for a Preliminary Ruling ¶¶ 20, 25, Case C-460/20, *TU v. Google*, ECLI:EU:C:2022:962 (Dec. 8, 2022); Opinion of Advocate General Pitruzzella, *supra* note 326, ¶ 57.

346. Case C-460/20, *TU v. Google*, Judgment ¶ 102.

347. *Id.* ¶¶ 51, 53.

348. *Id.* ¶ 104.

simultaneously remaining subject to multiple interpretations by virtue of being disconnected from the original publication.[349]

The court has continued to construct and detail a data protection application regarding a prevalent model of the search engine: from *Google Spain*'s early example of an early case of linking one's name to the electronic version of newspaper to *RE v. Google*'s more recent example of court recognition of the informative value of search preview images.[350] Against a decade-long timeline, a series of rulings on Google's dereferencing is a testament to the persistent challenge that internet technology poses for regulators and the judiciary.

Now, nearly two decades later, it remains questionable whether the GDPR adequately accommodates the "search engine privacy" phenomenon.[351] Initially envisioned as a technology that could execute extracting and structuring information from the World Wide Web,[352] search engines have gradually acquired distinctly governance-related attributes as well.[353] The EU Google cases challenged the CJEU to address this dual function in a coherent manner. Aiming to ensure effective data protection, the CJEU largely refrained from meaningful engagement with the technological attributes of web scraping. Yet, it dedicated considerable attention to the principal role Google assumes in shaping access to online information at large. Placing a pronounced focus on the "display" of information, the CJEU crafted a particular ex post verification regime to acknowledge the lawfulness of Google's processing while simultaneously devising a mechanism to enforce

349. *Id.* ¶ 100 ("The publication of photographs as a non-verbal means of communication is likely to have a stronger impact on internet users than text publications. Photographs are, as such, an important means of attracting internet users' attention and may encourage an interest in accessing the articles they illustrate . . . .").

350. *Id.* ¶ 89; Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317, Judgement ¶¶ 14, 80, 97 (May 13, 2014).

351. *See, e.g.*, ARTICLE 29 DATA PROT. WORKING PARTY, OPINION 1/2008 ON DATA PROTECTION ISSUES RELATED TO SEARCH ENGINES 4, 12–13 (2008), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2008/wp148_en.pdf [perma.cc/V9DX-D83R] (early discussions on privacy and search engines); Omer Tene, *Privacy: The New Generations*, 1 INT'L DATA PRIV. L. 15, 21–22 (2011).

352. *See* Sergey Brin, *Extracting Patterns and Relations From the World Wide Web*, *in* THE WORLD WIDE WEB AND DATABASES 172, 180 (Paolo Atzeni, Alberto O. Mendelzon & Giansalvatore Mecca eds., 1999).

353. *See, e.g.*, Eszter Hargittai, *The Social, Political, Economic, and Cultural Dimensions of Search Engines: An Introduction*, 12 J. COMPUT.-MEDIATED COMMC'N 769, 771, 774 (2007); Lucas D. Introna & Helen Nissenbaum, *Shaping the Web: Why the Politics of Search Engines Matters*, 16 INFO. SOC'Y 169, 171–74 (2000). An emerging strand of scholarship provides an organization-centric perspective on digital platforms. *See, e.g.*, Liang Chen, Tony W. Tong, Shaoqin Tang & Nianchen Han, *Governance and Design of Digital Platforms: A Review and Future Research Directions on a Meta-Organization*, 48 J. OF MGMT. 147 (2022).

individual rights.[354] The framework of "responsibilities, powers and capabilities" of a search engine appears particularly connected with Google's special role as an "intermediary" aiding access to the original hosts of data.

This "intermediary" role, however, becomes less clear in a context where the displayed information might appear more removed and disconnected from its original source,[355] thus rendering the underlying activity into a more autonomous data processing exercise. *TU v. RE*, the case delisting thumbnail requests, evinces this phenomenon. Atomizing and dissecting layers and dimensions of search engine functioning is an ongoing exercise for the court. It has yet to elaborate upon and clarify the framework not only with respect to the emerging search capabilities like ChatGPT, but also regarding other entities employing large-scale data collection through web scraping.

IV. TOWARD "FUTURE-PROOFING" AUTOMATED DATA COLLECTION PRACTICES

CJEU case law and data protection authorities' guidance presently do not include any specific exceptions acknowledging the effect of the automated data collection enabling the unobtrusive assembly of large-scale datasets. As this Article has discussed, data collection technology does not differentiate between personal data versus nonpersonal as well as "ordinary" personal data versus special category data.

Yet, the internal logic of data protection laws, which specify distinct layers of protection depending on the type of data at issue, seems to largely dismiss inability of automated collection to distinguish among these aforementioned data categorizations.[356] The fundamental mismatch between technology and law persists in CJEU case law and guidance from data protection authorities.[357] Reiterating their commitment to ensure a high level of personal privacy protection, such institutions' pertinent decisional practices are firmly grounded in the

354. Case C-136/17, GC and Others, Judgment, ¶ 47. *See also* EUR. DATA PROT. BD., GUIDELINES 5/2019 ON THE CRITERIA OF THE RIGHT TO BE FORGOTTEN IN THE SEARCH ENGINES CASES UNDER THE GDPR (PART 1) ¶¶ 1, 53 (2020), https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201905_rtbfsearchengines_afterpublicconsultation_en.pdf [perma.cc/X3H9-M4H9].

355. This is demonstrated by Case C-460/20, *TU v. Google* regarding de-referencing thumbnails. Judgment, ¶¶ 19–20.

356. *See* GDPR, *supra* note 37, art. 6, 9.

357. As discussed, examples of assessment of the material scope and data controllership evince. *See supra* Section III.A.

continuous expansion of the concepts of "personal data" and "data controllership."[358]

This formalism then nearly triggers by default the strictest form of the personal data protection regime; the scale and automation of the data collection process imply a high probability that a certain part of the collected data would potentially "reveal" sensitive information, bringing the whole set of collected data into a special category entitled to heightened protection.[359] The application of the heightened form of data protection[360] subsequently increases respective compliance costs[361] and potentially undermines the effectiveness and credible threat of enforcement.[362] Furthermore, depending on the mode and particular settings of automated data collection—web scraping or API-enabled—the responsibility for data protection compliance lies with either a single controller or joint controllers.

The CJEU's established decisional practice on the matter is, however, inconclusive and demands further clarification. First, it is uncertain how courts should apply Google case law[363] to other actors using data scraping.[364] Equally unclear is how courts should apply the framework of "responsibilities, powers and capabilities" to distinct dimensions and modes of online searches employing distinct search queries.[365] In other words, the outcome of the balancing exercise between data protection and freedom of information, as illustrated through various discussed Google cases, does not seem to elucidate a

358. *See, e.g.*, GDPR, *supra* note 37, art. 2, 4, § 1; Mahieu, Hoboken & Asghari, *supra* note 140, at 40; Finck, *supra* note 140, at 335; Finck & Pallas, *supra* note 138, at 4–5.

359. *See* Case C-252/21, Meta v. Bundeskartellamt, ECLI:EU:C:2023:537, Judgment ¶¶ 51, 69 (July 4, 2023).

360. In a form of, for example, additional requirements concerning lawful grounds for processing, *see* GDPR, *supra* note 37, art. 9, certain limitations as to the scope of the automated individual decision-making, *see id.* art. 22, §§ 1, 4, and higher level of administrative fines, *see id.* art. 83, § 5.

361. *See, e.g.*, Milda Macenaite, *The "Riskification" of European Data Protection Law Through a Two-Fold Shift*, 8 EUR. J. RISK REGUL. 506, 511–12, 515 (2017) (discussing a risk-based approach to compliance).

362. *See, e.g.*, Purtova, *supra* note 165, at 32.

363. *See* discussion *supra* Section II.B.3.

364. Debates around web scraping for the purpose of training Generative AI solutions are evolving. *See, e.g*, Blake Brittain, *Google says data-scraping lawsuit would take 'sledgehammer' to generative AI* (Reuters, Oct. 17, 2023, 12:54 PM CDT) https://www.reuters.com/legal/litigation/google-says-data-scraping-lawsuit-would-take-sledgehammer-generative-ai-2023-10-17/.

365. *See, e.g.*, David Pierce, *The AI takeover of Google Search starts now*, THE VERGE (May 10, 2023, 12:51 PM CDT), https://www.theverge.com/2023/5/10/23717120/google-search-ai-results-generated-experience-io [https://perma.cc/3RUZ-EJNA] (discussing search models' development).

universal formula.[366]Rather, the framework is not readily applicable to different types of information displays or different business models and applications such as LLM-based chatbots. Moreover, the exact boundaries and application of the joint controllership should be better calibrated to account for the actual level of control entities exercise.[367] Absent these clarifications, it is unclear how to rebut a nearly omnipresent assumption of joint controllership in programmatic data collection cases.[368]

Perhaps more fundamentally, whether joint controllership actually serves the purpose of strengthening the personal data protection interminably proves a questionable prospect. As the *Cambridge Analytica* scandal demonstrates, data stewardship issues extend well beyond the exclusive context of data protection.[369] A data platform acting as one of the data processing parties that defines responsibilities for other joint controllers could be a way to exert pressure on proper avenues and modalities of exercising individual data subjects' rights. The corollary, however, is complex arrangements, augmentation of power imbalances among data controllers, and reduced enforcement and protection of data subjects' rights.[370]

The commitment to ensure the highest possible level of data protection in the European Union stands on par with the EU ambition to enable big data analytics and benefit from data-driven innovation. The recent regulatory EU examples, taken together, provide a compelling public declaration to unleash the untapped potential of

366. In this context, the Guidelines 5/2019 on the criteria of the Right to be Forgotten in search engine cases under the GDPR provide yet another telling illustration of the challenge involved in reconciling data protection with freedom of speech. This balancing act is described in the Guidelines through several examples of the potential clash between respective interests showcases the difficulty of evaluating the impact of de-listing both on individual rights and societal values. *See generally* GUIDELINES 5/2019 ON THE CRITERIA OF THE RIGHT TO BE FORGOTTEN IN THE SEARCH ENGINES CASES UNDER THE GDPR, EUROPEAN DATA PROTECTION BOARD (July 7, 2020) https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201905_rtbfsearchengines_afterpublicconsultation_en.pdf.

367. *See* Mahieu, Hoboken & Asghari, *supra* note 140.

368. It is worth noting that a very recent ruling from the CJEU concerning the AdTech industry confirmed its ongoing approach of extending joint controllership. *See generally* Case C-604/22, AB Europe, ECLI:EU:C:2024:214. While underscoring the CJEU's commitment to ensure comprehensive data protection across entities involved in processing activities, the decision does not provide any clear and definite guidelines as to when shared control should not be assumed. *Id.*

369. *See* ORG. FOR ECON. COOP. & DEV., DATA STEWARDSHIP, ACCESS, SHARING AND CONTROL: A GOING DIGITAL III MODULE SYNTHESIS REPORT 23, (2023), https://one.oecd.org/document/DSTI/CDEP(2022)6/FINAL/en/pdf [perma.cc/J98X-G8LB].

370. Mahieu, Hoboken & Asghari, *supra* note 140, at 58, 59; Finck, *supra* note 140, at 334, 341.

"unused" and closely kept data through devising means of data sharing and curtailing gatekeeping capabilities.[371]

Online data is a foundational unit of myriad data-centered undertakings. Clearview AI amassed a vast database of three billion data points to create what the company described as a "bias-free algorithm" aimed at assisting law enforcement in solving crimes, such as financial fraud, human trafficking, and crimes against children.[372] The utilization of online data has also facilitated significant advancements in a number of scientific and public health initiatives. Projects around sequencing the human genome, creating databases of phenotypic trait data and health data based on online data not only broaden the scope of research capabilities but also democratize access to scientific data and more collaborative and inclusive research settings.[373]

Yet, the case of online collection for big data analytics is also special. Unlike big data analytics grounded on "observed" and "inferred" data, the online data collection path is typically available without a direct prior relationship established between a data subject and a data controller. In principle, practically anyone with the ability to automate the process of collection through the internet's infrastructure can harness online data. The discreteness of collection methods, plurality of potential data controllers, and diversity of potential data controllers pose particular risks to meaningful data protection enforcement, thus further undermining individuals' control over their personal data. Expanding notions and boundaries of "personal data," "special category data," and "data controllership" as a strategy in this case might not only be unsustainable but also could undermine the regulatory relevance and power of the GDPR holistically.

The data protection framework contains several means intended to steer the process of data collection. These means include defining legal grounds,[374] conducting a data protection impact assessment

371. *See, e.g.*, *Commission Proposal for a Regulation on Harmonised Rules on Fair Access to and Use of Data (Data Act)*, at 1, COM (2022) 68 final (Feb. 2, 2022); 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) 2022 O.J. (L 265) ¶¶ 6, 7.

372. HOAN TON-THAT https://hoantonthat.com/ [perma.cc/3Z2K-8BQZ] (last visited Feb. 4, 2023) (personal page of the CEO and co-founder of the Clearview AI); Hill, supra note 16.

373. *See, e.g.*, PERS. GENOME PROJECT: GLOB. NETWORK, https://www.person-algenomes.org/gb#about [perma.cc/ZN4B-Q5BR] (last visited Feb. 4, 2023). The project aims at enhancing the understanding of how genetics and environmental factors contribute to human traits. It also supports the development of personalized medicine by using principles of open data sharing.

374. *See, e.g.*, GDPR, *supra* note 37, art. 6, 9.

(DPIA),[375] and meeting requirements of data protection by design and by default[376] to ensure the overall compliance with the GDPR. Means of rebalancing a control asymmetry between a data controller and data subjects exist through, for example, the institute of data subjects' rights. However, the online environment and online data collection in principle test the overall ability of the GDPR to resort to this risk-based regulation as a proportionate and adaptive strategy. The ultimate effect of sustaining the status quo could become apparent in inconsistent and selective enforcement, a lack of legal certainty,[377] and a misleading appearance of compliance.

Against this background, an increased focus on detailing and further elaborating upon issues of material scope and controllership in automated data collection appears both timely and warranted.[378] Moreover, it is crucial to examine how alternative existing safeguards, viewed from the perspective of a data subject and while the data collection is actively occurring, could act as reinforcements of control.

There are several potential methods to accomplish such an examination. For example, national data protection authorities could further explore and advance the concept of "public accessibility" of online data in data protection terms. Unlike laws elsewhere,[379] the European data protection law does not contain a definition of "publicly available online data." However, it does contain a handful of terms that could define its metes and bounds.[380] For example, clarifying the scope and area of application of personal data the data subject renders

375. *Id.* art. 35.

376. *Id.* art. 25.

377. *See id.* rec. 7.

378. *See, e.g.*, *Global expectations of social media platforms and other sites to safeguard against unlawful data scraping*, THE OFFICE OF THE AUSTRALIAN INFORMATION COMMISSIONER (Aug. 24, 2023) https://www.oaic.gov.au/newsroom/global-expectations-of-social-media-platforms-and-other-sites-to-safeguard-against-unlawful-data-scraping [https://perma.cc/ERX6-4ZYJ].

379. *Cf.* Rossijskaja Federacija Federal′nyj Zakon O Personal′nyh Dannyh [Federal Law of the Russian Federation on Personal Data], Sobranie Zakonodatel'stva Rossiĭskoĭ Federatsii [SZ RF] [Russian Federation Collection of Legislation] 2006, No. 152-FZ, Art. 6(10) (defining "personal data made publicly accessible by data subjects" as "personal data the access to which was provided by a data subject himself or at his request"). It has to be noted, however, that the term has disappeared from the text of the law due to "Amendments to the Federal Law on Personal Data" No.519-ФЗ dated Dec. 30, 2020. Irina P. Golovanova, Russia; *Amendments to the Federal Law on Personal Data Takes Effect*, NAT'L L. REV. (Mar. 3, 2021), https://www.natlawreview.com/article/russia-amendments-to-federal-law-personal-data-takes-effect [perma.cc/MJ8S-ZDNC]. Instead, a new category of "personal data permitted for dissemination by the data subject" was introduced. *Id.*

380. *See, e.g.*, GDPR, *supra* note 37, art. 14, § 2(f) (information obligation); *id.* art. 9, § 2(e) (exception to a prohibition on processing sensitive data in case where such data was "manifestly made public by the data subject").

"manifestly public" could be an initial step in that direction.[381] Establishing clear parameters for personal data that individuals publicly disclose has the potential to enhance clarity, ensure uniform enforcement, and empower individuals in overseeing their data protection choices at least with regard to particularly sensitive online data. Additionally, this approach could further the European Union's objective of encouraging data altruism and establishing a Common European Health Data Space.[382]

Another potentially helpful path could be to delve into the notion of "reasonable expectations" and the capacity of the "fairness" principle to accommodate and articulate some of the fundamental conventions of automated data collection.[383] Some of the GDPR balancing assessments already incorporate the concept of "reasonable expectations," such as those concerning lawful processing.[384] Furthermore, an analysis of reasonable expectations was prominently featured in a recent CJEU case on Meta data processing[385] and in some national data protection cases.[386] Apart from the interpretation of fairness in the context of data subjects' reasonable expectations, it is also possible to explore the connection of the "fairness" principle with the closely associated principles of lawfulness and transparency of data processing.[387] Transparency in particular is a challenging requirement to meet in the context of automated data collection. Exploring non-transparency of

381. Currently, there are insufficient clarifications regarding when personal data is "made manifestly public." *See, e.g.*, Edward S. Dove & Jiahong Chen, *What Does it Mean for a Data Subject to Make their Personal Data 'Manifestly Public'? An Analysis of GDPR Article 9(2)(e)*, 11 INT'L DATA PRIV. L. 107, 108 (2021).

382. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A European Strategy for Data*, COM (2020) 66 final (Feb. 19, 2020); Data Governance Act *supra* note 43. The EU initiative of creating Common European Data Spaces relates to a data governance framework that aims at facilitating the use of health data for research, personalized medicine and evidence-based policy-making. *See Common European Data Spaces*, EUROPEAN COMMISSION, https://digital-strategy.ec.europa.eu/en/policies/data-spaces [https://perma.cc/2N7S-T8QY].

383. *See* EUR. DATA PROT. BD., GUIDELINES 8/2020, *supra* note 52, at 10, 48. Discussions on a "household" exemption might be also instrumental, especially in the context of social media services. *See, e.g.*, ARTICLE 29 DATA PROT. WORKING PARTY, OPINION 05/2009 ON ONLINE SOCIAL NETWORKING 5 (2009), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2009/wp163_en.pdf [perma.cc/68DB-R7RW].

384. GDPR, *supra* note 37, rec. 47 (legitimate interests).

385. Case C-252/21, Meta v. Bundeskartellamt, ECLI:EU:C:2023:537, Judgment, ¶ 47, July 4, 2023.

386. *See, e.g.*, Norwegian Data Protection Authority, Administrative Fine – Grindr LLC 23, 47 (Dec. 13, 2021), https://www.datatilsynet.no/contentassets/8ad827efefcb489ab1c7ba129609edb5/administrative-fine---grindr-llc.pdf [perma.cc/MZC4-CJ4F] (regarding the data sharing practices of the Grindr LLC).

387. *See* GDPR, *supra* note 37, art. 5, § 1.

data processing as an attribute of undermined fairness of processing therefore could be a promising path. Furthermore, research proposals to interpret "fairness" based on the insights of digital ethics and computational research practice might not only be theoretically insightful but also practically instrumental in bringing more control to individuals.[388]

Ultimately, placing greater emphasis on the technological infrastructure of the internet and the implementation of ethical AI safeguards could serve to "future-proof" the data protection framework indirectly. The European Union's adoption of the EU Artificial Intelligence Act (EU AI Act) marks a significant step in regulating AI technologies within the region and beyond.[389] While the EU AI Act does not explicitly target large-scale data collection as its primary focus, it regulates the development, deployment, and use of artificial intelligence systems, which often include systems trained on a large corpus of data or involved in large data collection.

Under the EU AI Act, high-risk AI systems are subject to stricter regulations due to their potential to cause significant harm or infringe upon fundamental rights.[390] These high-risk systems include, for example, those used in critical infrastructure, such as transportation and energy, as well as those involved in areas like healthcare and law enforcement.[391] The AI Act imposes stricter regulations on such high-risk AI systems, requiring rigorous conformity assessments before they can be placed on the market or used in the EU.[392] These assessments evaluate aspects such as data quality, technical robustness, and compliance with safety and transparency requirements.[393] Additionally, high-risk AI systems must meet specific transparency and documentation obligations to ensure accountability and facilitate oversight.[394]

---

388. *See, e.g.*, Damian Clifford & Jeff Ausloos, *Data Protection and the Role of Fairness*, 37 Y.B. EUR. L. 130, 131, 186 (2018).

389. European Parliament Press Release, Artificial Intelligence Act: MEPs adopt landmark law (Mar. 13, 2024), https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law [https://perma.cc/8WXB-22ZY].

390. *See Explanatory Memorandum to 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*, COM (2021) 206 final (Apr. 21, 2021).

391. *See* AI Act, Annex III, Art. 6(2).

392. *See, e.g.*, AI Act, Art. 6(1)(b), 16, 19.

393. AI Act, ch. 2.

394. AI Act, art. 17.

In the context of the AI systems involved in large-scale data collection, the upcoming framework has a significant potential to improve the transparency of AI solutions in data collection processes. This increased transparency is expected to empower data subjects by providing them with better insights into the mechanics of data collection and offering greater control over exercising their data protection rights. Simultaneously, conformity and risk assessment exercises this governance framework facilitates should enable entities engaged in automated data collection to more effectively identify and address potential risks related to data privacy. The AI Act and the GDPR application should ideally thus synergistically result in a robust foundation for promoting trustworthy and responsible AI innovation while safeguarding individuals' rights and freedoms regarding data protection.

Finally, the engagement with internet technological architecture such as "robot.txt exclusion codes" for guiding data collection processes could also provide for needed synergy between law and technology by defining and distributing liability. Additionally, it could be helpful to reflect on a potential connection between the online information presentation and its effect on the feasibility of the required contextual assessment of personal data under the GDPR.[395] The recent Google case concerning thumbnail delisting requests, *TU and RE*, offers a promising starting point in this regard.[396] Further reflection on the presentation and compilation of information in the context of user interface might prove valuable in updating and enhancing the contextual assessment of special category data.

Automated online data collection shows no sign of fading into oblivion.[397] As technological advancements progress, it becomes even more evident that online data will remain one of the cornerstones of continuous data-driven transformations. It is essential that the legal framework provides a future-proof solution to accommodate technological advancement. The solution must be sufficiently flexible to allow and encourage the type of innovation that society regards as

395. Regularly held calls for evidence within the framework of reports evaluating the application of the General Data Protection Regulation (GDPR) present a promising forum for such reflection and stakeholders engagement. *See, e.g.*, REPORT ON THE GENERAL DATA PROTECTION REGULATION, EUROPEAN COMMISSION, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14054-Report-on-the-General-Data-Protection-Regulation_en [https://perma.cc/8S3E-5A5Q].

396. *See* Case C-460/20, TU v. Google, ECLI:EU:C:2022:962, Judgment ¶ 2 (Dec. 8, 2022).

397. *See, e.g.*, Alistair Barr & Adam Rogers, *Death by LLM: Stack Overflow's Decline, and Its Plan to Survive, Shows the Future of Free Online Data in an AI World*, BUS. INSIDER (Aug. 3, 2023, 4:00 AM), https://www.businessinsider.com/stack-overflow-crisis-future-of-online-data-ai-world-2023-7 [perma.cc/D9VY-ZVK4] (on use of online data as a training data).

beneficial. At the same time, however, it must ensure that parties do not achieve innovation at the expense of individual rights and freedoms. As this Article has discussed, the EU data protection framework strives to ensure a high level of personal data protection. Notwithstanding this aim, based on a broad interpretation of central terms and figures of data protection law, it remains unclear whether such laws possess longevity. From the modest outlook of today, it seems that the current approach yields more uncertainty and concerns about the removed control over one's personal data than it provides for the much-needed flexibility and resilience to greet a bright tomorrow.

V. CONCLUSION

Automated online data collection frequently takes place discreetly, often without users actively participating or even being aware of the process. Through the analysis of material scope and data controllership concepts, this Article demonstrates that the prevalent EU legal solutions lack consistency and fail to provide a sustainable framework for regulating data-driven innovation while preserving individual control over personal data. As such, there is a pressing need for further exploration and development of regulatory approaches to address the challenges posed by evolving data collection technologies.

Given this context, a heightened emphasis on delineating and elaborating on the contours of accessible public data seems both timely and justified. To ensure that online data-driven innovation does not materialize at the expense of personal data control, it is critical to proactively explore potential remedies and legal responses. These responses could involve a range of actions, from enhancing clarity through enforcement authorities and courts' decision-making practices to industry-led initiatives such as heightened accountability and compliance-focused endeavors enabled by the simultaneous application of the GDPR and the AI Act.