Technology and Regulation

Researching with Data Rights

Jef Ausloos* & Michael Veale**

right of access, GDPR, data protection, transparency, research methods

j.ausloos@uva.nl

m.veale@ucl.ac.uk

The concentration and privatization of data infrastructures has a deep impact on independent research. This article positions data rights as a useful tool in researchers' toolbox to obtain access to enclosed datasets. It does so by providing an overview of relevant data rights in the EU's General Data Protection Regulation, and describing different use cases in which they might be particularly valuable. While we believe in their potential, researching with data rights is still very much in its infancy. A number of legal, ethical and methodological issues are identified and explored. Overall, this article aims both to explain the potential utility of data rights to researchers, as well as to provide appropriate initial conceptual scaffolding for important discussions around the approach to occur.

1. The GDPR: Research Curse or Blessing?

Data protection legislation, in particular the EU General Data Protection Regulation (GDPR),[1] has been seen by some researchers as creating frustrating barriers to their work.[2] Data minimization and storage limitation restrict the extent to which large databases can be amassed for future consultation. Information requirements can limit covert or subtle collection, and sit at tension with web-scraping and research on social media. Uncertainty and anxiety in risk-averse organizations can stifle data-driven research initiatives, leaving researchers dissuaded or simply encouraging them to disregard the rules.[3] The GDPR does not appear to improve the situation much: while it recognizes the importance of scientific research through derogations from the default data protection rules, it also leaves a lot unsaid, and outsources crucial interpretative guidance to Member States.[4] If Member States do not clarify the issue in national law or regulatory guidance, the interpretative burden falls to the organizations who benefit from these exemptions. Increased public scrutiny in light of the Cambridge Analytica scandal, which involved online data collection infrastructure established by the University of Cambridge,[5] has only increased fear of infringement.

At the same time, the societal stakes for investigating the online world have never been higher, or the need more urgent.[6] Networked systems have changed individuals' experiences of the world, their enhanced and more pervasive mediating roles 'affecting the ways in which we understand our own capabilities, our relative boundedness, and the properties of the surrounding world'.[7] Many readers will not need much introduction into the 'algorithmic war-stories' unearthed in recent years that focus on the impact of these mediating systems,[8] particularly through the work of journalists, civil society and activist-minded research groups. Work by journalists such as Julia Angwin, Lauren Kirchner and Kashmir Hill has explored the way that technol-

1 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] OJ L119/1 (hereafter 'GDPR').

2 See, e.g.: Edward S Dove, 'The EU General Data Protection Regulation: Implications for International Scientific Research in the Digital Era' (2018) 46 *J Law Med Ethics* 1013; Wouter Van Atteveldt, 'Toward Open Computational Communication Science: A Practical Road Map for Reusable Data and Code' [2019] 20; Rossana Ducato, 'Data Protection, Scientific Research, and the Role of Information' (2020) 37 *Computer Law & Security Review* 105412.

3 e.g., the extensive work by David Erdos, 'Stuck in the Thicket? Social Research under the First Data Protection Principle' (2011) 19 *Int J Law Info Tech* 133; David Erdos, 'Systematically Handicapped? Social Research in the Data Protection Framework' (2011) 20 *Information & Communications Technology Law* 83; David Erdos, 'Constructing the Labyrinth: The Impact of Data Protection on the Development of "Ethical" Regulation in Social Science' (2012) 15 *Information, Communication & Society* 104.

4 GDPR, art 89(2–3).

5 https://www.theguadian.com/news/series/cambridge-analytica-files

6 Also emphasised in European Commission, 'White Paper on Artificial Intelligence - A European Approach to Excellence and Trust' (19.2.2020); European Commission, 'A European Strategy for Data' (Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions, 19.2.2020).

7 Julie E Cohen, *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice* (Yale University Press 2012).

8 Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For' (2017) 16 *Duke Law & Technology Review* 18. See also Malte Ziewitz, 'Governing Algorithms: Myth, Mess, and Methods' (2016) 41 *Science, Technology, & Human Values* 3 (on 'algorithmic drama').

* Jef Ausloos is a Postdoctoral Research Fellow at the Institute for Information Law, University of Amsterdam

** Michael Veale is a Lecturer in Digital Rights and Regulation at the Faculty of Laws, University College London

Received 3 Dec 2020, Accepted 28 Dec 2020, Published 4 Jan 2021.

ogy firms and society interact in the context of discrimination and manipulation. Research groups such as the Data Justice Lab at Cardiff University have mapped the use of technology in the public sector,[9] and research teams have made use of freedom of information rights to discover more about the way that data-driven systems are being procured.[10] Researchers from teams such as the Algorithm Auditing Research Group at Northeastern University[11] have been using bots and online scraping and analysis tools to better understand discrimination and inequality in AdTech systems,[12] and data leakage from apps to third party trackers.[13] Meanwhile, workers' collectives are mobilizing in an attempt to reclaim data from platform companies to prove their eligibility for basic employment rights, and to use as evidence in tribunals and other proceedings.[14] These efforts often run into a range of methodological, practical and legal hurdles which in many cases are easily arguable to have been preserved by powerful forces seeking to retain the secrecy that allows them to work with limited scrutiny and accountability.

We present one flipside of this (deliberately) dismal picture. Could the GDPR *enable*, rather than stifle, data-rich research? In particular, in a world where private and influential data infrastructures are coordinated by a limited number of powerful actors, might the GDPR's provisions be used as a *source* of data, rather than applying constraints on collection? We believe it can be. Researching with data rights can provide data for a range of 'digital methods', which include applying and adapting existing methods such as surveying, ethnography or text analytics to new, digitized sources of information, as well as fueling new 'natively digital' methods aimed at building understanding based on features of digital spaces (such as hyperlinking, wireless sensing, recommender systems or browsing histories) which have no clear offline analogue.[15] Potential access to such data sources is made possible through the GDPR's strengthened information provision measures, found predominantly in Articles 12 through 15, and underpinned by the overarching transparency principle in Article 5(1)(a).

2. Existing means to access enclosed data and their limits

Digital methods are plagued by the problem of 'special access', which is 'required for the study of certain natively digital objects'.[16] While much of our life is entwined with sensors and actuators,[17] this data gathered on us is not, generally, stored on our own devices. Even though the average user 'has in their pocket a device with vastly more resource than a mainframe of the 1970s by any measure', they usually end up 'using [their] devices as vastly over-specified dumb terminals.'[18] Instead, computation and data storage generally happens in rented 'cloud' infrastructure. This move to the 'cloud' is value-laden in nature, coming with a natural tendency to concentrate the power that comes from data and the constant experimental decisions made around its use in the hands of central, proprietary nodes.[19] Despite the fact that data is not, generally, considered a form of property, platforms in the informational economy have established 'de facto property arrangements' by enclosing such data using legal strategies such as terms-of-use agreements to heavily structure interactions.[20] These entities only rarely release data entirely and/or unconditionally, whether for legal, economic or technical reasons, and appear willing to fight against initiatives that would force them to do so more readily.

Lack of access has made private entities the gatekeepers of the data or infrastructure necessary for utilizing digital methods. Consequently, research that happens inside or with the blessing of these entities tends to be limited to that in the private entity's interests (notably profit and reputation), rendering it hard to impossible for outside actors (notably academia, journalists, and civil society more broadly) to perform critical parallel inquiry. Internal research undertaken for the genuine purpose of discovery, but which might impugn the firm's legitimacy, is unlikely to see the light of day.[21] Sealing off societally important data processing operations has rendered it very hard to scrutinize the practices of these entities.[22]

We identify roughly four main groups of approaches through which researchers external to these entities attempt to study them with digital methods:

- voluntary data sharing agreements (ad hoc arrangements);
- programmatic access (technical tools offered by data controllers);
- scraping and interception (independent technical tools); and

9 Lina Dencik and others, 'Data Scores as Governance: Investigating Uses of Citizen Scoring in Public Services' (Data Justice Lab, Cardiff University, 2018) https://perma.cc/39CY-H8L7 (accessed 21 August 2020); Lina Dencik and others, 'The "Golden View": Data-Driven Governance in the Scoring Society' (2019) 8 *Internet Policy Review*.

10 Marion Oswald and Jamie Grace, 'Intelligence, Policing and the Use of Algorithmic Analysis: A Freedom of Information-Based Study' (2016) 1 *Journal of Information Rights, Policy and Practice*; Robert Brauneis and Ellen P Goodman, 'Algorithmic Transparency for the Smart City' (2018) 20 *Yale Journal of Law & Technology* 103.

11 https://personalization.ccs.neu.edu

12 Michael Carl Tschantz and Anupam Datta, 'Automated Experiments on Ad Privacy Settings' (2015) *2015 Proceedings on Privacy Enhancing Technologies* 92.

13 Reuben Binns and others, 'Third Party Tracking in the Mobile Ecosystem' in *Proceedings of the 10th ACM Conference on Web Science* (WebSci '18, New York, NY, USA, ACM 2018).

14 James Farrar, 'Why Uber Must Give Its Drivers the Right to All Their Data', (*New Statesman*, 2 April 2019) https://www.newstatesman.com/america/2019/04/why-uber-must-give-its-drivers-right-all-their-data accessed 22 July 2019; 'Uber drivers demand access to their personal data' (Ekker Advocatuur, 19 July 2020) https://ekker.legal/2020/07/19/uber-drivers-demand-access-to-their-personal-data (accessed 17 August 2020).

15 Richard Rogers, *Digital Methods* (The MIT Press 2013).

16 Rogers (n 15) 15.

17 See generally Mireille Hildebrandt, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Edward Elgar Publishing 2015).

18 Jon Crowcroft and others, 'Unclouded Vision' in Marcos K Aguilera and others eds, *Distributed Computing and Networking* (Springer Berlin Heidelberg 2011) 29.

19 Seda Gürses and Joris van Hoboken, 'Privacy after the Agile Turn' in Evan Selinger and others (eds), *The Cambridge Handbook of Consumer Privacy* (Cambridge University Press 2018).

20 See Julie E Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press 2019) 44–45.

21 See e.g. Karen Hao, 'We read the paper that forced Timnit Gebru out of Google. Here's what it says' (*MIT Technology Review*. 4 December 2020) https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru (on the concerns with independence of the process surrounding the scholarly publication of a paper on bias and environmental issues in large language models co-authored by fired Google researcher Timnit Gebru).

22 See references in Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015); Nicholas Diakopoulos, 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures' (2015) 3 *Digital Journalism* 398; Muhammad Ali and others, 'Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes' [2019] arXiv:190402095 [cs], 5; European Data Protection Supervisor, 'A Preliminary Opinion on Data Protection and Scientific Research' (6 January 2020).

- data disclosure requirements (legal transparency requirements).

These approaches all have their benefits and shortcomings – discussed below – and can further be categorised along two axes, depending on (a) the relationship between researcher and data holder (collaborative v adversarial) and (b) the point of access (top-down v bottom-up) (Table 1). This last qualification is based on whether data is obtained through the entity holding the data directly, or via its users. Put briefly, *top-down* data access enables a helicopter view or overarching insights (e.g. internet platform content moderation or ad archives), but the respective data will often be very high-level, notably to safeguard users' privacy. *Bottom-up* data access enables granular insights into individuals' data (e.g. reactions to personalized media-diets), but may fail to give a global picture, may require significant technical expertise and raises legal concerns. *Collaborative* data access arrangements may be very advantageous if they work, but can create undesirable dependencies and solidify power dynamics. *Adversarial approaches* – ie independent of data holders' goodwill to release data – are therefore often the only way for researchers to obtain access to data, but come with their own set of (legal, technical, economical) challenges.

Table 1 Current approaches to data access

	Collaborative	Adversarial
Top-down	Voluntary data sharing	Data disclosure requirements
Bottom-up	Programmatic access	Scraping and interception

Against this backdrop, we believe there to be an important role for the law – democratically designed and enforceable – in framing the scope and limits of adversarial data access approaches. GDPR transparency rights show particular promise as such an adversarial, bottom-up tool for research data access (notably considering the drawbacks of scraping and interception). In order to better appreciate this, let us briefly zoom into the different approaches to data access.

2.1 Voluntary data sharing agreements

Some researchers/institutions obtain access to privately held data via 'data philanthropy' initiatives[23] and/or through amicable relationships they might entertain with the relevant actors (e.g. Facebook's *Social Science One* initiative;[24] or the UK's *Consumer Data Research Centre*[25]). Beneficial as these may be to the respective researchers, such an approach risks further solidifying existing power dynamics in academia (and the private sector). Efforts like these have also been characterized as 'corporate data philanthropy', designed to generate positive publicity rather than critical research.[26] Moreover, researchers/institutions may have several reasons for not wanting to associate with private entities as a precondition for doing research, such as fear of real or perceived loss of independence that may result from, for example, an obligatory sign-off procedure on produced findings.[27]

Threats from researchers to pull out of Facebook's *Social Science One* initiative after they were denied the data promised have only stoked scepticism about the feasibility of this ad hoc style of data access to form a basis for future digital methods.[28] Indeed, recent efforts aim to introduce more of a formal structure and regulatory oversight to data sharing arrangements, through the development of data protection codes of conduct in this area.[29]

2.2 Programmatic access

Researchers and institutions may find creative ways to re-purpose entities' existing programmatic tools, such as application programming interfaces (APIs) in order to get access to data. These allow users to access the data of themselves, others, or the environment through programmatic querying which will return machine readable data according to a given specification. There are several challenges with this approach.

APIs are generally designed with developers, not researchers, in mind, and can consequently fail to return research-grade data. API access to a stream of content may only provide a limited, non-random sample. Twitter's public APIs showed at most 1% of public tweets, and systematic biases compared to the full data-stream has cast the representativeness of reliant studies into question.[30] Such APIs in general only show public information — even then, only data that developers consider important — with available sampling and filtering commands lacking the necessary expressiveness for research.[31]

API use for research may also go against applicable Terms of Service, and researchers may therefore risk retaliatory action, such as being kicked off the platform.[32] In some jurisdictions, contract law and computer misuse law has been blurred, creating heightened legal risk as well.[33]

APIs have more recently become political tools used by platforms to exclude certain business or functionality from integration, and the interaction between developers and the changing nature of APIs has been described as 'risky territory', an 'ongoing battle' and 'hostile'.[34] Strategic changes to an API may break an entire set of business

23 See e.g., https://www.mastercardcenter.org/action/call-action-data-philanthropy.

24 Facebook, 'Facebook Launches New Initiative to Help Scholars Assess Social Media's Impact on Elections' (*Facebook Newsroom*, 9 April 2018) https://newsroom.fb.com/news/2018/04/new-elections-initiative (accessed 23 April 2019).

25 https://www.cdrc.ac.uk/about-cdrc

26 Axel Bruns, 'After the "APIcalypse": Social Media Platforms and Their Fight against Critical Scholarly Research' (2019) 22 *Information, Communication & Society* 1544, 1551.

27 Bruns (n 26) 1553. See generally the open letter regarding corporate support of research into technology and justice at https://fundingmatters.tech.

28 Camilla Hodgson, 'Facebook given Deadline to Share Data for Research', (*Financial Times,* 28 August 2019) https://www.ft.com/content/147eddec-c916-11e9-af46-b09e8bfe60c0 (accessed 11 September 2019); Social Science One, 'Public Statement from the Co-Chairs and European Advisory Committee of Social Science One' (11 December 2019) https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one (accessed 5 January 2020).

29 See arts. 40–41 GDPR. See also Mathias Vermeulen, 'The Keys to the Kingdom. Overcoming GDPR-Concerns to Unlock Access to Platform Data for Independent Researchers' (OSF Preprints 27 November 2020); 'Call for Comment on GDPR Article 40 Working Group' (EDMO, 24 Nov 2020) https://edmo.eu/2020/11/24/call-for-comment-on-gdpr-article-40-working-group (accessed 23 December 2020).

30 Andrew Yates and others, 'Effects of Sampling on Twitter Trend Detection' (2016) *Proceedings of the International Conference on Language Resources and Evaluation*.

31 Alexandra Olteanu and others, 'Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries' (2019) 2 *Front Big Data*.

32 Olteanu and others (n 31).

33 e.g., the arguments in the US case *Sandvig et al. v. Sessions*, No. 1:16-cv-01368 (D.D.C. June 29, 2016), See generally Annie Lee, 'Algorithmic Auditing and Competition Under the CFAA: The Revocation Paradigm of Interpreting Access and Authorization' (2019) 33 *Berkeley Tech L J* 1307. But see *hiQ Labs, Inc v LinkedIn Corporation* 2019 WL 4251889 (United States, Ninth Circuit).

34 Tania Bucher, 'Objects of Intense Feeling: The Case of the Twitter API : Computational Culture' (2013) 3 *Computational Culture: A Journal of Software Studies*; Paddy Leerssen and others, 'Platform Ad Archives: Promises and Pitfalls' (30 April 2019).

models, while privileged access can create economic advantage. Social media platform Facebook was accused by the UK House of Commons Digital, Culture, Media and Sport Committee of using API access to take 'aggressive positions' against competitor apps, taking actions leading to the failure of businesses.[35] In this context, APIs cannot be easily relied on by researchers, who may find their software rendered dysfunctional by external business logics or even a shift in functionality aimed at breaking their efforts to rigorously interrogate a system.[36] Unless many streams of their work rely on this software, researchers rarely have the time or resource to engage in this 'arms race' and maintain software in the face of sudden, unexpected and often ill-documented changes.[37] API-based research with inconvenient findings for private entities is unlikely to be sustainable.

Connectedly, and perhaps more problematically, is the fact that from a privacy and data protection point of view the use of APIs does not preclude bad faith (or at least ethically questionable) actors obtaining access to personal, or even sensitive, data. The quintessential example of this is Aleksander Kogan and Cambridge Analytica, whose Facebook add-on 'thisisyourdigitallife' harvested millions of Facebook profiles of both the users of the add-on and those users whose data they in turn had access to. The ensuing mixture of *bona fide* research and data privacy scandal has challenged the field of researchers using digital methods.[38] As a result, several APIs do not or no longer allow access to users who are not somehow connected to the requester, limiting data access to those users within the requester's 'social graph'.[39] While the dropping of access has been framed in terms of privacy and security, sceptics see it also as 'a convenient step towards disabling and evading unwanted independent, critical, public-interest scholarly scrutiny.'[40]

2.3 Scraping and interception

Researchers also rely on independent technical or methodological tools to obtain useful data otherwise sealed-off by private entities without their blessing.

Scraping tools or bots are common sources of data where APIs are restrictive or unavailable. Such an approach has some legal support in several jurisdictions with *text and data mining* exemptions in copyright laws. In some cases, such as in Japanese law, these exemptions are not restricted to actors or purposes,[41] while in other laws, such as in the UK since 2014 and in the new 2019 EU Copyright Directive, there are limitations of scope for 'non-commercial' and/or 'research' purposes.[42]

In some cases, what is of interest is how the platform, its users and its non-users behave in interaction with it. At scale, this is likely to require the use of bots or crowd workers. However, the use of both bots and crowd workers for research, particularly when bots impersonate a 'real' user or crowd workers make use of their own social profiles is, in at least some cases, legally and ethically contentious.[43] However, it also brings opportunities for co-creation of research, potentially seeing participants as co-researchers rather than research subjects.[44]

In other cases, data is trickier to obtain due to advanced enclosure techniques by firms.[45] Researchers wishing to understand what data mobile apps send and to where they send it often have to resort to monitoring users' internet traffic using a virtual private network (VPN), requiring invasive device access.[46] Approaches on the Web, which is a little more open in this regard, include browser plugins to monitor social media (WhoTargetsMe,[47] Algorithms Exposed,[48] FB-Forschung[49]), search engine (e.g. DatenSpende)[50] or general browsing activity (e.g. Robin).[51]

These approaches are more resistant to retaliatory action by the respective entities[52] or misuse by bad actors, and the active recruit-

35 See generally documents presented and published under privilege by Damian Collins MP to the Commons DCMS Committee. These documents were a selection of emails that were obtained through discovery in the US Courts in a lawsuit involving developer Six4Three and Facebook. Despite being held under seal by the San Mateo Superior Court, they were given to the UK Parliament which published them under privilege, and are available at https://www.parliament.uk/documents/commons-committees/culture-media-and-sport/Note-by-Chair-and-selected-documents-ordered-from-Six4Three.pdf. See generally Digital, Culture, Media and Sport Committee, 'Disinformation and "Fake News"' (18 February 2019).

36 e.g., Ariana Tobin Jeremy B. Merrill, 'Facebook Moves to Block Ad Transparency Tools — Including Ours' (*ProPublica*, 28 January 2019) https://www.propublica.org/article/facebook-blocks-ad-transparency-tools (accessed 17 April 2019).

37 See, on the death of Netvizz, a popular research tool for those studying Facebook, Bruns (n 26) 1549.

38 Tommaso Venturini and Richard Rogers, '"API-Based Research" or How Can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach' (2019) 0 *Digital Journalism* 1.

39 Dietmar Janetzko, 'The Role of APIs in Data Sampling from Social Media' in Luke Sloan and Anabel Quan-Haase (eds), *The SAGE Handbook of Social Media Research Methods* (SAGE Publications Ltd 2016).

40 Bruns (n 26) 1550.

41 On the Japanese text and data mining exemptions, See FutureTDM, D3.3+ Baseline Report of Policies and Barriers of TDM in Europe (Horizon 2020 Project 665940 2016) 75. This law has been recently clarified and extended by the Act of Partial Revision of the Copyright Act (Japan) 2018, which clarifies the use of copyrighted works in relation to machine learning. See generally European Alliance for Research Excellence, 'Japan Amends Its Copyright Legislation to Meet Future Demands in AI' (European Alliance for Research Excellence, 9 March 2018) http://eare.eu/japan-amends-tdm-exception-copyright (accessed 24 June 2019).

42 For the recently passed European provision, See Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance), arts 3–4; for the 2014 UK provision, See Copyright, Designs and Patents Act 1988 s 29A.

43 See generally (n 21).

44 See Alexander Halavais, 'Overcoming Terms of Service: A Proposal for Ethical Distributed Research' (2019) 22 *Information, Communication & Society* 1567, 1578.

45 See generally on data enclosure Julie E Cohen, 'Property and the Construction of the Information Economy: A Neo-Polanyian Ontology' in Leah A Lievrouw and Brian D Loader (eds), *Handbook of Digital Media and Communication* (Routledge forthcoming).

46 e.g., Abbas Razaghpanah and others, 'Haystack: In Situ Mobile Traffic Analysis in User Space' [2015] 14; Jingjing Ren and others, 'ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic' in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* (MobiSys '16, New York, NY, USA, ACM 2016); Yihang Song and Urs Hengartner, 'PrivacyGuard: A VPN-Based Platform to Detect Information Leakage on Android Devices' in *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices* (SPSM '15, New York, NY, USA, ACM 2015); Anastasia Shuba and others, 'AntMonitor: A System for On-Device Mobile Network Monitoring and Its Applications' [2016] arXiv:161104268 [cs].

47 https://whotargets.me/en

48 https://algorithms.exposed

49 https://fbforschung.de/. This tool combines a data-gathering plugin with occasional surveys with participants, enabling more in-depth information than what can merely be observed.

50 https://datenspende.algorithmwatch.org

51 Balázs Bodó and others, 'Tackling the Algorithmic Control Crisis – the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents' (2017) 19 *Yale JL & Tech* 133.

52 Though certainly not immune, as illustrated by plugins of ProPublica and WhoTargetsMe slightly being blocked by Facebook changing some of its HTML code. See generally Jeremy B Merrill (n 36); Digital, Culture, Media and Sport Committee (n 35) 64.

ment of research subjects gives an opportunity to inform them about the study and its consequences. Nevertheless, this 'reverse engineering' approach is fragile and labor intensive. Infrastructure such as operating systems or web browsers can change and be changed, disrupting these tools in the process. Because of the predominance of vertically integrated companies in the digital economy,[53] firms often control both this infrastructure and the data of research interest (e.g. Alphabet, Google and Chrome), creating issues similar to that of APIs.[54]

2.4 Data disclosure requirements

Researchers may also put their hopes in regulatory interventions (or threats of legislation) forcing more transparency. So far, legal instruments primarily focus on transparency of public sector information (i.e. freedom of information acts or the EU's Public Sector Information and Open Data Directives[55]), but new initiatives are underway to open up privately held data as well.[56] Targeted transparency and disclosure policies are familiar policy instruments in many policy areas such as the environment, health and safety.[57] Such instruments are commonly used by individuals, civil society and journalists[58] — and often designed with them in mind — but also have surprisingly high usage by commercial entities for profitable ends.[59] Some disclosures, such as curated datasets of information on disinformation, have been forced from platforms more-or-less at threat of legislation in times of political contestation.[60] Freedom of information (FoI) laws have been used to study data-driven systems already,[61] but their scope is generally limited to the public sector, and in some jurisdictions, contractors thereof.[62] Transparency obligations also exist in many safety-critical sectors, such as electronics, food, pharmaceuticals and the like, although in practice these are rarely triggered by citizens, and usually relate to access to documents through regulators, or policies concerning labelling. Data protection impact assessments (DPIAs), which may contain useful information about processing practices for researchers, are not obliged to be made public under EU data protection law, and therefore do not count amongst transparency measures covered here.[63]

A spate of new and proposed digital regulation *does*, however, include transparency reporting on digital phenomena applicable to private entities. The proposed EU Terrorist Content Regulation would have hosting service providers set out 'a meaningful explanation of the functioning of proactive measures including the use of automated tools',[64] and published annual transparency reports containing information on detection measures and statistics on takedown information.[65] The recently adopted Regulation on promoting B2B fairness and transparency, covering platforms which intermediate trade such as online e-commerce marketplaces and 'app' stores, requires providers to reveal 'the main parameters determining ranking and the reasons for the relative importance of those main parameters as opposed to other parameters', and require search engines to provide such information in 'an easily and publicly available description, drafted in plain and intelligible language' and to 'keep that description up to date'.[66] In line with its ambitious 'strategy for data',[67] the European Commission also put forward three major policy proposals at the tail end of 2020. All three – the Data Governance Act, Digital Services Act, and Digital Markets Act – place strong emphasis on transparency obligations for digital services.[68] Obligations under the proposed Digital Services Act would mandate influential 'gatekeepers' to provide data to vetted researchers investigating systemic societal risks.[69] In the run-up to the 2019 EU elections, the European Commission also managed to make a number of powerful platforms issue monthly transparency reports on a voluntary basis.[70] Inspiration might also be drawn from gender pay gap disclosure legislation increasingly common throughout the world.[71] Finally, it is also worth

53 Ian Brown and Christopher T Marsden, *Regulating Code: Good Governance and Better Regulation in the Information Age* (MIT Press 2013) xii.

54 See also Thomas Claburn, 'Google Nukes Ad-Blocker AdNauseam, Sweeps Remains out of Chrome Web Store', (*The Register*, 5 January 2017) https://www.theregister.co.uk/2017/01/05/adnauseam_expelled_from_chrome_web_store (accessed 18 June 2019).

55 Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information OJ L 345 ('PSI Directive'); from June 2021 repealed and replaced by 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 OJ L 172/56 ('Open Data Directive').

56 cf. European Commission, 'Building a European Data Economy' (10 January 2017) https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy (accessed 28 April 2018); European Commission, 'A European Strategy for Data' (Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions, 19.2.2020). For proposed regulations, see Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act) COM/2020/767 final; Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC COM/2020/825 final; Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act) COM/2020/842 final.

57 For two case studies drawing from the environmental and health context respectively, See Jef Ausloos and others, 'Operationalizing Research Access in Platform Governance What to Learn from Other Industries?' (25 June 2020). See generally: Archon Fung and others, *Full Disclosure: The Perils and Promise of Transparency* (Cambridge University Press 2007).

58 See Matt Burgess, Freedom of Information: A Practical Guide for UK Journalists (Routledge 2015).

59 See Margaret B Kwoka, 'FOIA, Inc.' (2016) 65 *Duke Law Journal* 1361.

60 See generally Amelia Acker and Joan Donovan, 'Data Craft: A Theory/Methods Package for Critical Internet Studies' (2019) 22 *Information, Communication & Society* 1590.

61 e.g., Oswald and Grace (n 10); Brauneis and Goodman (n 10).

62 The UK Information Commissioner has been active in her attempts to try to argue for contractors to fall under freedom of information law. See generally Information Commissioner's Office, Outsourcing Oversight? The Case for Reforming Access to Information Law (ICO 2019).

63 Reuben Binns, 'Data Protection Impact Assessments: A Meta-Regulatory Approach' (2017) 7 *International Data Privacy Law* 22.

64 Proposal for a Regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online (COM/2018/640 final) (hereafter Proposed Terrorist Content Regulation), art 8(1).

65 Proposed Terrorist Content Regulation art 8(3).

66 Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services (European Union 2019) Article 5.

67 European Commission, 'A European strategy for data' (n 6).

68 See references in (n 56).

69 Proposed Digital Services Act (n 56), art 31(2).

70 European Commission - DG Connect, 'Code of Practice on Disinformation' (Text, Digital Single Market, 26 September 2018) https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation (accessed 19 July 2019). Relatedly, see European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the European Democracy Action Plan' (3 December 2020).

71 The Equality Act 2010 (Gender Pay Gap Information) Regulations 2017, SI 2017/172. But note that The Financial Times caused a stir when it noted that many companies reporting their gender pay gaps under new UK legislation reported an identical mean and median: something so statistically improbably it was effectively indicative or an error, a cover-up, or both. See Billy Ehrenberg-Shannon and others, 'Cluster of UK Companies Reports Highly Improbable Gender Pay Gap', (*Financial Times*, 12 July 2017) https://www.ft.com/content/ad74ba76-d9cb-11e7-a039-c64b1c09b482 (accessed 17 June 2019).

pointing to the European Commission's ambitious data strategy, which includes the tabling of an 'enabling legislative framework for the governance of common European data spaces' by the end of 2020.[72]

As it stands under current legislation, the scope of these disclosure obligations is patchy at best. In Europe, tensions exist between FoI and privacy law,[73] which in turn limit the extent to which even public agencies can make disclosures of individual level data. Recent tensions between ICANN and European data protection regulators around the WHOIS database for website registrars have further illustrated these tensions.[74] This stands in contrast to several US cases, such as the famed *COMPAS* study into recidivism systems by *ProPublica*, where journalists used public records access to analyze a proprietary software system they accused of racial bias.[75] Replicating this method in Europe would likely run into difficulties as authorities would be unlikely to release identifiable data of convicts or ex-convicts as they did to *ProPublica* for reasons of data protection and privacy.[76]

***

These four approaches to data for digital methods all have their benefits and shortfalls. This paper does not seek to present a panacea, but it does seek to add a tool to the ever-changing toolkit. That tool is data protection transparency, in particular, the use of data rights. The rest of this paper considers legal, social, technical and ethical aspects of this proposed data source in research contexts.

3 Transparency Provisions in the GDPR

Data protection is characterized in large part by its transparency provisions. These started off as a form of general oversight over the primarily state-affiliated 'databanks' motivating early data protection law, and now are best known as tools for coping with information asymmetries that in many cases originate today's predominantly private-sector information economy.[77]

This article focusses primarily on European data protection law, and in particular the GDPR. This legal framework contains a panoply of tools, ranging from individual rights to more collectively and collaboratively-flavored provisions. Amidst this panoply, the *right to access* is explicitly highlighted in the EU Charter of Fundamental Rights. Not only should data be processed *fairly*,[78] but the Charter's Article 8(2) proclaims that everyone 'has the right of access to data which has been collected concerning him or her.'

More recently, ensuring transparency of automated processing and profiling in particular has also become a considerable public and legislative concern. Developments in the Council of Europe illustrate this well in the (recently modernized) Convention 108[79] and earlier recommendations.[80] The modernized convention provides that each individual shall have a right 'to obtain, on request, knowledge of the reasoning underlying data processing where the results of such processing are applied to him or her'.[81] Related provisions are found in EU data protection law and French administrative law,[82] as well as in national adaptations to data protection legislation in EU member states.[83]

3.1 Flavors of data protection transparency

Transparency provisions come in many different shapes and flavors in the GDPR. Firstly, transparency provisions in the GDPR range from **overarching to concrete**. Transparency as an overarching principle informs the interpretation and application of all of the GDPR.[84] Indeed, it is listed in the first substantive provision in the GDPR, requiring any data processing operation to be lawful, fair and transparent.[85] Throughout the GDPR, more specific, concrete rights and obligations formalize how transparency should be routinely carried out.[86]

Transparency provisions have both **intrinsic and instrumental aims**.[87] The most explicit transparency provisions have a strong flavor of transparency as intrinsically important: meta-data about processing must be provided to data subjects (and often the public more broadly) upon collection,[88] upon receipt of data from a third party,[89] or upon request.[90] In other provisions, the instrumental component is more prominent, such as concerning establishing a lawful basis for processing or automated decision-making through consent;[91] in data breach notifications to data subjects;[92] in moving data to another controller;[93] and in certification mechanisms.[94]

Transparency provisions can have **different target audiences**: individual data subjects are generally considered to be the intended users of the rights to access or portability;[95] while the public at large, including

72 And which would be specifically designed to 'facilitate decisions on which data can be used, how and by whom for scientific research purposes in a manner compliant with the GDPR.' European Commission, 'A European strategy for data' (n 6) 12–13.
73 See generally Ivan Szekely, 'Freedom of Information Versus Privacy: Friends or Foes?' in Serge Gutwirth and others (eds), *Reinventing Data Protection?* (Springer Netherlands 2009).
74 See generally Stephanie E Perrin, 'The Struggle for WHOIS Privacy: Understanding the Standoff Between ICANN and the World's Data Protection Authorities' (PhD Thesis, University of Toronto 2018).
75 Julia Angwin and others, 'Machine Bias', (*ProPublica*, 23 May 2016) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Jeff Larson and Julia Angwin, 'How We Analyzed the COMPAS Recidivism Algorithm', (*ProPublica,* 23 May 2016) https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed 28 September 2018).
76 However, the EU set-up does provide a defence against recent reported uses of freedom of information law for harassment of e.g. scientists. See e.g., Claudia Polsky, 'Open Records, Shuttered Labs: Ending Political Harassment of Public University Researchers' (2019) 66 *UCLA L Rev.*
77 Jef Ausloos and Pierre Dewitte, 'Shattering One-Way Mirrors – Data Subject Access Rights in Practice' (2018) 8 *International Data Privacy Law* 4, 5–7.
78 cf. Damian Clifford and Jef Ausloos, 'Data Protection and the Role of Fairness' (2018) 37 *Yearbook of European Law* 130.
79 Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (opened for signature 10 October 2018) 228 CETS (hereafter Convention 108+), art 9(c).
80 See e.g., Council of Europe, 'Recommendation on the Protection of Individuals with Regard to Automatic Processing of Personal Data in the Context of Profiling CM/Rec(2010)13' (23 November 2010).
81 Convention 108+, art 9(c).
82 Lilian Edwards and Michael Veale, 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?' (2018) 16 *IEEE Security & Privacy* 46.
83 Gianclaudio Malgieri, 'Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards" in the National Legislations' (2019) 35 *Computer Law & Security Review.*
84 GDPR, art 5(1)(a). See generally Clifford and Ausloos (n 78).
85 GDPR, art 5(1)(a).
86 GDPR, arts 13–15.
87 See generally Ausloos and Dewitte (n 77).
88 GDPR, art 13.
89 GDPR, art 14.
90 GDPR, art 15(2–3).
91 In general for consent, GDPR, arts 4(11), 7; for automated decision-making, See GDPR, art 22(2)(c) and recital 71.
92 GDPR, art 33.
93 GDPR, art 20.
94 GDPR, art 42.
95 GDPR, arts 15, 20. But See René LP Mahieu and others, 'Collectively Exercising the Right of Access: Individual Effort, Societal Effect' (2018) 7

civil society watchdogs, often benefit through transparency obligations (often fulfilled through privacy policies or public signage).[96] Supervisory authorities are important beneficiaries of transparency, which they can obtain through a range of data controller obligations[97] as well as their own information retrieval powers.[98] Transparency provisions can also treat sensitivity with nuance and blend target audiences in doing so. For example, in the case of sensitive data in the policing context which cannot be directly released, the data subject has a right to exercise transparency provisions *through* a supervisory authority, who must verify the legality of the processing illuminated by the data they receive.[99]

Transparency provisions can **kick in either before or after data is first processed**, a topic which we will return to further below **(*ex ante* and *ex post* transparency)**. A final, related distinction distinguishes **push and pull** transparency provisions, differentiating whether the controller[100] or the target audience[101] must take the initiative before information is released. This distinction largely corresponds to transparency *obligations* versus transparency *rights*.

While these ways of categorizing GDPR transparency overlap, they help better situate the twofold goal of transparency measures in the GDPR. Transparency provisions have a protective dimension, ensuring demonstrable accountability. Yet some measures also bring an important empowerment dimension, putting control in the hands of different stakeholders, and data subjects in particular, to be more informed. Both dimensions can be considered to contribute to a common goal: redistributing power stemming from information/data asymmetries.

3.2 *Ex ante* transparency

The epicenter of transparency measures in the GDPR, as well as the most well-known and explicit, is found within Articles 13–15. The first two of these list the information that controllers—those determining the means and purposes of data processing—need to provide proactively, at their own initiative and *before* they start processing personal data.[102] In substance, Article 13 (focused on situations where personal data was obtained from individuals *directly*) and Article 14 (personal data was obtained *indirectly*) differ very little. These provisions can first and foremost be qualified as *protective* measures, forcing controllers to give proper thought to, and be upfront about, their processing operations and enabling to hold them to account later on. As such, they also serve as a useful compliance–testing tool for data protection authorities and/or other interest-groups.

Articles 13–14 also have an empowering facet to them. After all, they make data subjects — those to whom the personal data being processed relates — aware of processing taking place and as such can be seen as a *sine qua non* for empowering individuals to invoke one or more of their rights (e.g. object, erasure, portability).[103] The most important components of *ex ante* transparency relate to the scope, purposes and the lawful bases for processing, the risks involved, the retention period and how to exercise data subject rights.

3.3 *Ex post transparency*

There are two main sources of *ex post* transparency in the GDPR that can be triggered by data subjects — the right of access, commonly known as the data subject access right and the right to data portability.

3.3.1 Subject access rights

Article 15 complements *ex ante* information obligations by granting data subjects an explicit, *user-triggered right* to obtain additional information (cf. Table 1, page 15). There are two main components to this right. The first largely replicates the information that was, or should have been, provided under Articles 13–14, which is useful when the information was missed at the time or spread across multiple sources, incomplete, or not specific to the data subject's situation. In this regard, Article 15 can be qualified as an *ex post* empowerment measure and essentially gives individuals the ability to force more timely and specific transparency.[104]

The second component is more radical, at least compared to regimes that in general lack it. It demands that data controllers 'shall provide a copy of the personal data undergoing processing',[105] which explains why the right has become known as a subject access request (SAR). It is worth noting that 'processing' is an extremely broad term, meaning 'any operation or set of operations' performed on personal data.[106] Consequently, data undergoing processing is not just data actively being used, but also includes data that is being stored. Furthermore, the wide scope of personal data[107] means that opinions or comments, including those undertaken computationally or those which may be incorrect, are, *prima facie*, often going to be within the remit of the right of access.[108]

Table 2 Information Requirements under Article 15, GDPR.

Information Requirement	Art 15
Confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the **personal data**	1
Purposes of the processing	1(a)
Categories of personal data concerned	1(b)
Recipients or categories of recipients to whom the personal data **have been or will be disclosed**, in particular recipients in third countries or international organisations	1(c)
Retention period, or if that is not possible, the criteria used to determine that period	1(d)
Existence of the data subject rights to rectification, erasure, restriction of processing, and to object	1(e)
Right to lodge a complaint with a supervisory authority	1(f)
Where personal data are not collected from the data subject, any information on the **source**	1(g)

Internet Policy Review.
96 GDPR, arts 13–14.
97 e.g., GDPR, art 30(4).
98 GDPR, art 47(1).
99 Law Enforcement Directive, art 17.
100 e.g., GDPR, art 13.
101 e.g., GDPR, art 15.
102 Processing in data protection law includes collection. GDPR, art 4(2).
103 See further Ausloos and Dewitte (n 77).
104 Ausloos and Dewitte (n 77).
105 GDPR, art 15(3).
106 GDPR, art 4(2).
107 See generally Nadezhda Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10 *Law, Innovation and Technology* 40. For a view tempering the wide scope argued in that paper, See Lorenzo Dalla Corte, 'Scoping Personal Data: Towards a Nuanced Interpretation of the Material Scope of EU Data Protection Law' (2019) 10 *European Journal of Law and Technology*.
108 Case C-434/16 *Peter Nowak v Data Protection Commissioner* ECLI:EU:C:2017:994 [34].

Existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful **information about the logic involved, as well as the significance** and the envisaged consequences of such processing for the data subject	1(h)
In case of transfer to third country, information about the appropriate safeguards	2

3.3.2 Data portability

The new right to data portability offers some further promise for use in order to obtain research data. Article 20 grants data subjects the right to receive their personal data, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance.[109] Moreover, data subjects can request their personal data to be directly transferred from one controller to another where technically feasible. It is not hard to see how this provision may make the process of sharing personal data with researchers a lot smoother.[110] Indeed, in contrast to the right of access, the right to data portability actively recognizes the value and ability for data subjects to move their personal data between entities, and thus has provisions and wording that facilitate such sharing.[111] The version of the Digital Markets Act proposed by the Commission, if passed, would further strengthen data portability rights against large 'gatekeepers' by enabling them to be used continuously and in real-time.[112]

Unlike the right of access in Article 15(3), which applies to all data being processed, three important constraints limit the scope of the right to data portability:

1. It only applies to personal data that the data subject has provided to the controller, excluding for example 'inferred' and 'derived' data.[113]
2. It only applies where processing is based on 'consent' or 'necessity for the performance of a contract' as a lawful ground. This effectively exempts data processed only with one or a mixture of the four other grounds.[114] Crucially, this includes the important 'legitimate interests' ground, upon which data is gathered on an 'opt-out' or objection basis, rather than an affirmative consent basis.
3. Although not particularly restrictive for our purposes, the right to data portability only applies in situations where the respective personal data is processed 'by automated means'. Data protection also applies to physical records that meet the definition of personal data and 'which form part of a filing system or are intended to form part of a filing system'.[115] Data controllers have no obligation to digitize such data in a machine-readable format for the purposes of the right to portability, although such data remains within scope of the right of access.

3.3.3 Transparency modalities

The GDPR also lists a number of modalities to ensure transparency is effective. The key provision for this is Article 12, but some specific modalities can also be found within the respective provisions discussed above. Importantly, individuals cannot be charged a fee for claiming transparency[116] and there are strict timing requirements as well as broader conditions for the way in which transparency is provided.[117] The European Data Protection Board (EDPB)[118] has further specified that controllers should actively consider the audience's 'likely level of understanding' when accommodating transparency (e.g. appropriate level of detail, prioritizing information, format, etc.).[119] This means the controller will need to consider the context of data processing, the product/service experience, device used, nature of interactions, and so on.[120] As a result, the information obligation may also differ throughout time.[121]

Finally, it is worth keeping in mind that controllers have a duty to facilitate the exercise of data subject rights by 'implementing appropriate technical and organizational measures'[122] and only work with processors who can guarantee doing the same.[123] While the GDPR seems to imagine standard-setting and/or APIs, collaborations in complex ecosystems that facilitate data subjects' rights remain easier

109 The format should be interoperable and machine-readable, both notions being defined in EU law, cited in: Article 29 Working Party, 'Guidelines on the Right to Data Portability' (wp242, 13 December 2016)16–18. It is further specified that '[w]here no formats are in common use [...], data controllers should provide personal data using commonly used open formats (e.g.XML, JSON, CSV,...) along with useful metadata at the best possible level of granularity'.

110 Even the European Data Protection Board (previously known as Article 29 Working Party) explained how the right might be useful to learn more about music consumption by using the right with streaming services or assessing carbon footprint by using the right with loyalty cards. Article 29 Working Party, 'Guidelines on the right to data portability' (n 109) 4–5.

111 That being said, the Commission did recently state that 'as a result of its design to enable switching of service providers rather than enabling data reuse in digital ecosystems the right [to data portability] has practical limitations.' European Commission, 'A European strategy for data' (n 6) 10.

112 Proposed Digital Markets Act (n 56), art 6(1)(h).

113 The EDPB does however advocate for a broad interpretation, encompassing both 'data actively and knowingly provided by the data subject' as well as 'observed data provided by the data subject by virtue of the use of the service or the device'. Data such as search histories, browsing/location behaviour, 'raw data' collected through 'mhealth devices' (mobile health) therefore fall within the scope of the right to data portability. Article 29 Working Party, 'Guidelines on the right to data portability' (n 109) 9–11.

114 GDPR, art 6(1) lists six lawful grounds on the basis of which personal data may be processed: (a) consent; (b) necessary for the performance of a contract; (c) necessary for compliance with a legal obligation; (d) necessary to protect the data subject or another natural person's vital interests; (e) necessary for tasks carried out in the public interest, or exercise of official authority; (f) necessary for the purposes of the legitimate interests pursued by the controller or third parties, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject.

115 GDPR, art 2(1).

116 This also means that a controller cannot require you to be a paying customer as a condition to accommodate your rights. Article 29 Working Party, 'Guidelines on Transparency under Regulation 2016/679' (11 April 2018) 13. Previous empirical work has demonstrated that certain controllers effectively only enable access requests filed by people who have an account with the service and/or have bought something with the service before. See Ausloos and Dewitte (n 77) 12–13.

117 See generally Jef Ausloos and others, 'Getting Data Subject Rights Right: A Submission to the European Data Protection Board from International Data Rights Academics, to Inform Regulatory Guidance' (2020) 10 *JIPITEC*.

118 Prior to the entry into force of the GDPR, this organisation – which groups together all Member State data protection authorities – was known as the Article 29 Working Party.

119 See also Recital 60 Article 29 Working Party, 'Guidelines on transparency under Regulation 2016/679' (n 116) 11.

120 This may require running (and documenting) trials before 'going live'. See Article 29 Working Party, 'Guidelines on transparency under Regulation 2016/679' (n 116) 14.

121 cf. Article 29 Working Party, 'Guidelines on transparency under Regulation 2016/679' (n 116) 16–17.

122 GDPR, arts 12(2), 25. For a more detailed explanation on data rights modalities, See Ausloos and others (n 117).

123 GDPR, art 28.

said than done.[124]

4 Opportunities for Researching Through Data Rights

How can data rights help researchers? This will effectively depend on a variety of disciplinary, practical, legal, ethical and methodological factors. Indeed, it all starts with a research question or goal, which is to be situated in a certain (number of) discipline(s) that comes with its (their) own im-/explicit rules on valid data gathering. Next, one will need to assess what exact data is needed and what GDPR transparency measure may appropriate to capture it (cf. section 3). Researchers will also need to consider the scope of the data required, both in width (i.e. how many research subjects, if any at all, are needed to have a representative sample) and in depth (i.e. how exhaustive and/or granular does the data have to be). This scope will, in turn, inform whether research subjects are needed, and if so, how to recruit them. Researchers will also need to carefully consider an interaction strategy with data controllers (including contingency plans), which may be more or less burdensome depending on the scope, but also on the identity of the data controller.[125] Indeed, based on preliminary research (including filing access requests themselves), researchers may prepare a manual or script on how research subjects should obtain the required information and interact with data controllers.[126] Finally, researchers should also anticipate how the data they might obtain through data rights will actually be analyzed in light of the research aim. Summarized, the following seven steps may serve as a useful starting point for researchers interested in using data rights in their project:

1. Aim. What is your research goal? What purpose are you gathering data for?
2. Data. What specific data do you need to achieve said purpose?
3. Legal Approach. What GDPR transparency measure is appropriate for obtaining said data (if any)?
4. Scope. What does your (ideal) research sample look like?
5. Recruitment Strategy. Based on the scope, how to identify and recruit research participants accordingly?
6. Interaction Strategy. How will you interact with your participants and the respective data controllers?
7. Data Analysis Strategy. How will you actually gather the insights you need?

These steps remain necessarily vague, in light of the broad potential of data rights as a research method in many different disciplines. To tie it back to the many variables determining the actual usefulness of data rights for any given research project – i.e. disciplinary, practical, legal, ethical and methodological factors – the abstract workflow mentioned above will have to be given shape depending on the respective discipline(s) and research questions. There are also many *practical* factors that might influence the usefulness of data rights. Again, these will depend very much on the concrete circumstances of a given research project. Nonetheless, in order to make things more concrete, and invite readers to contemplate different use cases, this section lays out some illustrative potential and promising uses of data rights. The following section will then dig into some of the legal, ethical and methodological considerations.

For our purposes here, we identify three main categories of research (goals) as being enabled by data rights (in order of specificity):

- studying infrastructures (research into the actual infrastructures to which the respective data relates);
- studying impacts (research into how data infrastructures affect individuals, communities or society at large); and
- repurposing digital traces (research into broader questions that might be far from issues of digital rights).

4.1 Understanding infrastructures

Researchers can use data transparency rights to study digital infrastructures and practices in today's economies and society.

Studies examining data protection law in practice are one example of this. Researchers have, for example, studied the privacy policies of cloud service providers to identify common industry approaches and legal mismatches.[127] These privacy policies exist in the form they do in large part due to the GDPR's transparency provisions in Articles 13–14.[128] Other research has taken the form of exploring *how* rights are responded to by controllers, the quality of which might say something about enforcement more generally.[129]

Yet *ex post* transparency measures offer wider potential as research tools beyond studying the way the law is being interpreted and adhered to. Many use cases can be envisaged which would use specific *ex post* transparency measures to uncover substantive issues. We consider a number of them below.

4.1.1 Tracking

The state of online tracking and advertising today has been both lauded for supporting online services that do not directly cost consumers money, as well as lambasted for undermining democracy, journalism and a range of fundamental rights. One thing is certain: it is a challenging area to study. Data rights provide a useful set of tools to shine further light on issues of concern.

For example, both users and researchers know little about how effective privacy protective browsers and extensions really are. While it is relatively simple to secure a device from explicitly saving tracking cookies (although that may damage Web functionality), it is very hard to disguise the unique fingerprint of a browser, particularly in the presence of advanced fingerprinting tactics utilized in modern advertising technologies.[130] Because fingerprinting does not always query

124 cf. Chris Norval and others, 'Reclaiming Data: Overcoming App Identification Barriers for Exercising Data Protection Rights' [2018] arXiv:180905369 [cs], 4.

125 As research has shown, many data controllers are often unwilling to comply in full with data access requests, unless they are repeatedly contacted. See, e.g. Ausloos and Dewitte (n 77); Jef Ausloos, 'Paul-Olivier Dehaye and the Raiders of the Lost Data' (*CITIP blog*, 10 April 2018) https://www.law.kuleuven.be/citip/blog/paul-olivier-dehaye-and-the-raiders-of-the-lost-data (accessed 23 April 2018); Mahieu and others (n 95).

126 For example, one could envisage a website or an app that makes it easier for research subjects to file access requests, follow up on them, and/or filter the personal data obtained, before it is sent to the researchers. See also section 5.3.2

127 Dimitra Kamarinou and others, 'Cloud Privacy: An Empirical Study of 20 Cloud Providers' Terms and Privacy Policies—Part I' (2016) 6 *International Data Privacy Law* 23; Jamila Venturini and others, *Terms of Service and Human Rights: An Analysis of Online Platform Contracts* (Revan 2016).

128 See section 3.2, '*Ex ante* transparency'.

129 See e.g., Ausloos and Dewitte (n 77); Mahieu and others (n 95); Janis Wong and Tristan Henderson, 'How Portable is Portable?: Exercising the GDPR's Right to Data Portability' in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (UbiComp '18, New York, NY, USA, ACM 2018); Clive Norris and others (eds), *The Unaccountable State of Surveillance: Exercising Access Rights in Europe* (Springer 2016).

130 Nick Nikiforakis and others, 'Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting' (May 2013) *2013 IEEE Symposium on Security and Privacy* 541.

a user's device directly, but observes it passively, it is unclear to what extent real protection is provided. Technologies such as re-spawning cookies, 'evercookies',[131] font and battery level fingerprinting all present methodological challenges to detect, understand and effectively and provably block.[132] The use of data rights to ascertain the data a firm actually holds on users through a separate channel could be used as means of assessing the efficacy or tracking prevention, or understanding the true nature and purpose of certain tracking practices online.[133]

The number of actors in the tracking business and the nature of their interactions with each other also considerably restricts understanding. Online advertising increasingly functions through a complex 'real-time bidding' system whereby an individual's browser, generally unbeknownst to the user, sends out personal data about them to an advertising exchange, which in turn forwards it to thousands of potential bidders. These thousands of bidders utilize the services of *data management platforms* to enrich the data received: to effectively see if your eyes are worth bidding for in relation to the adverts they are attempting to place. The UK and Belgian regulators have noted that such a system is likely not legally compliant on a number of fronts.[134] Detailed evidence is, however, scarce, due to the secrecy and complexity of these practices.

The fact that these actors often share data *server-to-server* has created a blind spot for current studies—a blind spot that data rights might help remedy. While it is possible for researchers to monitor a user's browser to observe the destination of the traffic, for example by using a VPN (local or remote) with the consent of the user,[135] data that is transmitted around the user from server-to-server cannot be observed. Researchers working in this space have to come to an unhappy compromise of either simulating these server-to-server transmissions with almost no evidence on how they actually occur in practice,[136] or to try and guess at data practices by experimenting on how users are differentially targeted further downstream.[137] If researchers were to use data rights—and, if these firms were forced to answer them truthfully and fully—information on the data, the source of the data, and potentially on the recipients could be obtained, which would help both modelling assumptions as well as enable further research questions to be answered.

Related to this, data rights may also benefit studying the intersection of user preferences and tracking infrastructures. Some researchers have been presenting users with information about tracking activities (e.g. types of data, data flows), attempting to understand the effects on decision-making by users, as well as any impacts on their ongoing formation of privacy and data control preferences.[138] To do this, they have relied on indirect methods to understand these flows, such as running an app in a virtual environment and monitoring and classifying the entities data directly flow to.[139] Yet this data is still a step removed from the tracking that has occurred to particular participants. To reflect on their own information, users would typically have to rely on tools to collect and reflect on this data,[140] such as local logging of information. Tools to give users a 'history' function on their digital activities do exist,[141] but are unwieldy to force participants to use day-to-day, and may not even log as invasively as third-party trackers currently do.[142] Insofar as these tracking infrastructures *already exist*, data rights provide an alternate means to get access to them, enabling research that takes advantage of users seeing and reflecting on tracking data that truly was captured about and relates to them.

Data rights might help economic studies too. Despite considerable interest in how online content should be funded, 'the conventional wisdom that publishers benefit too from behaviorally targeted advertising has rarely been scrutinized in academic studies'. Recent studies have indicated that when a user's cookie is available, publishers' revenue increases by only about 4%.[143] This adds to anecdotal evidence from publishers such as the New York Times that reducing tracking has increased profits in their European markets, suspected to be related to the market structure of advertising technology and the proliferation of intermediaries.[144] Data rights might help to gather datasets on which publishers, advertising technology firms, ad exchanges and other actors[145] are active in this area, and use that data to create and validate economic models which can shine light on market functioning.

In a similar vein, there has been considerable recent interest in

131 Evercookies use practices found in malware more broadly to re-establish cookies even when users or browsers attempt to purge them.

132 Gunes Acar and others, 'The Web Never Forgets: Persistent Tracking Mechanisms in the Wild' in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (CCS '14, New York, NY, USA, ACM 2014); David Fifield and Serge Egelman, 'Fingerprinting Web Users Through Font Metrics' in Rainer Böhme and Tatsuaki Okamoto eds, *Financial Cryptography and Data Security* (Lecture Notes in Computer Science, Springer Berlin Heidelberg 2015); Łukasz Olejnik and others, 'The Leaking Battery' in Joaquin Garcia-Alfaro and others eds, *Data Privacy Management, and Security Assurance* (Lecture Notes in Computer Science, Springer International Publishing 2016).

133 They would not be without their challenges of course: some tracking practices may be illegal, for example, meaning that already–infringing data controllers are unlikely to readily to openly share information.

134 Information Commissioner's Office, 'Update Report into Adtech and Real Time Bidding' (Information Commissioner's Office, 20 June 2019) https://perma.cc/X7PX-EL3L (accessed 20 June 2019); Natasha Lomas, 'IAB Europe's Ad Tracking Consent Framework Found to Fail GDPR Standard' (*TechCrunch*, 16 October 2020) https://social.techcrunch.com/2020/10/16/iab-europes-ad-tracking-consent-framework-found-to-fail-gdpr-standard (accessed 16 October 2020).

135 See e.g., Razaghpanah and others (n 46); Ren and others (n 46); Song and Hengartner (n 46); Shuba and others (n 46).

136 See e.g., Muhammad Ahmad Bashir and Christo Wilson, 'Diffusion of User Tracking Data in the Online Advertising Ecosystem' (2018) *2018 Proceedings on Privacy Enhancing Technologies* 85.

137 Tschantz and Datta (n 12).

138 See e.g., Max Van Kleek and others, 'X-Ray Refine: Supporting the Exploration and Refinement of Information Exposure Resulting from Smartphone Apps' in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18, New York, NY, USA, ACM 2018); Max Van Kleek and others, 'Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps' in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17, New York, NY, USA, ACM 2017).

139 See generally Binns and others (n 13).

140 See generally M Janic and others, 'Transparency Enhancing Tools (TETs): An Overview' (June 2013) 2013 *Third Workshop on Socio-Technical Aspects in Security and Trust* 18; P Murmann and S Fischer-Hübner, 'Tools for Achieving Usable Ex Post Transparency: A Survey' (2017) 5 *IEEE Access* 22965.

141 See e.g., Jennifer Pybus and others, 'Hacking the Social Life of Big Data' (2015) 2 *Big Data & Society* 2053951715616649.

142 Murmann and Fischer-Hübner (n 140) 22988.

143 Veronica Marotta and others, 'Online Tracking and Publishers' Revenues: An Empirical Analysis' (June 2019) *Proceedings of the Workshop on the Economics of Information Security* (WEIS 2019), 2–4 June, Boston, MA.

144 Jessica Davies, 'After GDPR, The New York Times Cut off Ad Exchanges in Europe - and Kept Growing Ad Revenue' (*Digiday*, 16 January 2019) https://digiday.com/media/gumgumtest-new-york-times-gdpr-cut-off-ad-exchanges-europe-ad-revenue (accessed 19 June 2019). See also David Beer and others, Landscape Summary: Online Targeting (Centre for Data Ethics and Innovation, HM Government 2019) 32.

145 See generally Bashir and Wilson (n 136).

competition issues around online tracking from both academics[146] and policy-makers.[147] Insofar as data rights can help with issues of accountability and provenance,[148] they may help to map the space of actors and data practices in ways which better shine light on structural power relations that matter for evidencing competition policy interventions in different jurisdictions.

4.1.2 Content moderation

For well over a decade, researchers have been investigating the freedom of expression and information implications of online copyright enforcement.[149] Considerable efforts have been put into forcing more transparency and accountability from both copyright-holders as well as the (user-generated) content platforms in taking down content.[150] More recently, growing concerns over platform power and regulatory initiatives on online content moderation have breathed new life into this work.[151] Indeed, a lot of important questions have been raised in relation to content moderation and platforms' potential political biases,[152] their role in facilitating cyber-bullying,[153] impact on inclusiveness and participation by vulnerable or minority groups,[154] and the increased privatization of the public sphere more broadly.

In mapping the available empirical literature on these issues, Keller and Leerssen make a similar distinction to those we made above separating disclosures from platforms and other direct stakeholders from independent research through, for example, APIs, secondary processing of released or scraped data, or surveys with users and other stakeholders.[155]

These methods can be lacking in depth to explore exactly the processes, data sources and reasoning automated takedowns involve. Subject access requests may be a valuable addition in researchers' toolbox, providing a legally enforceable mechanism to force platforms to be more open about their decision-processes that have affected the data subject(s) at stake. One reason for this is that any decision on (not) taking down content may significantly affect either the uploader,[156] or person(s) featuring in the actual content. In those situations, Article 15(1)h provides data subjects the right to obtain *meaningful information about the logic involved, as well as the significance and the envisaged consequences* of the respective decision(s).[157] In general, data about the uploaders' actions or account may also be considered personal data and subject to Article 15 (or 20) more broadly. Such personal data can in turn be examined for its sources, gaining a better understanding of the processing activities underlying content moderation today.

That being said, as with any of these cases, data rights are no panacea. While enabling deeper insights into certain content moderation practices, using subject access requests for mapping platform-wide trends may prove more challenging. They may, however, create new research questions and challenge commonly held assumptions about data processing for these purposes, and form an important part of a researcher's toolkit as a result.

4.2 Understanding impacts

Considerable recent concern has centered around the impact of data-driven systems, particularly in reinforcing structural disadvantage affecting marginalized communities.[158] Such systems create data infrastructures, often focused on optimization, which disregard subsets of individuals (such as those considered 'low value') or contextual and environmental factors,[159] and which may use seemingly non-sensitive data to deliberately or inadvertently make decisions based on legally protected characteristics.[160] They may perform more poorly on certain demographics, such as facial recognition or analysis systems disproportionately misclassifying or misrepresenting Black women.[161] Such systems have also been accused of using micro-targeting in an electoral context in ways unsuited for demo-

146 See e.g., Elettra Bietti and Reuben Binns, 'Acquisitions in the Third Party Tracking Industry: Competition and Data Protection Aspects' [2019] *Computer Law & Security Review*; Reuben Binns and others, 'Measuring Third-Party Tracker Power Across Web and Mobile' (2018) 18 *ACM Trans Internet Technol* 52:1.

147 See e.g., Jacques Crémer and others, 'Competition Policy for the Digital Era' (European Commission, 2019) http://ec.europa.eu/competition/publications/reports/kd0419345enn.pdf (accessed 4 April 2019).

148 See generally David Eyers and others, 'Towards Accountable Systems' (Dagstuhl Seminar 18181) [2018].

149 For a comprehensive overview, See Aleksandra Kuczerawy, *Intermediary Liability and Freedom of Expression in the EU: From Concepts to Safeguards* (Intersentia 2018); Daphne Keller and Paddy Leerssen, 'Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation' in N Persily and J Tucker (eds), *Social Media and Democracy: The State of the Field and Prospects for Reform* (CUP 2019).

150 Most notably perhaps, the early work of Wendy Seltzer and in particular the Lumen database (formerly 'Chilling Effects Clearinghouse') https://lumendatabase.org, collecting and analysing removal requests of online materials.

151 See generally Robert Gorwa, 'What is Platform Governance?' (2019) 22 *Information, Communication & Society* 854.

152 Oscar Schwartz, 'Are Google and Facebook Really Suppressing Conservative Politics?', (*The Guardian*, 4 December 2018) https://www.theguardian.com/technology/2018/dec/04/google-facebook-anti-conservative-bias-claims (accessed 1 December 2019).

153 Tijana Milosevic, 'Social Media Companies' Cyberbullying Policies' [2016] 22; Pat Strickland and Jack Dent, *Online harassment and cyber bullying* House of Commons Rep 07967 (UK House of Commons 2017).

154 Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press 2018); Stefanie Duguay and others, 'Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine' [2018] Convergence 1354856518781530; Jillian C. York and Karen Gullo, 'Offline/Online Project Highlights How the Oppression Marginalized Communities Face in the Real World Follows Them Online' (*Electronic Frontier Foundation*, 3 June 2018) https://www.eff.org/deeplinks/2018/03/offlineonline-project-highlights-how-oppression-marginalized-communities-face-real (accessed 19 July 2019).

155 Keller and Leerssen (n 149) 13–32.

156 Relatedly, it is worth referring to FairTube, an initiative set up by (semi-) professional youtubers aimed at forcing fairer and more transparent decision-making on de-monetization of YouTube content. Subject access rights played a role in this effort. René Mahieu and Jef Ausloos, 'Recognising and Enabling the Collective Dimension of the GDPR and the Right of Access' (Preprint, 2 July 2020) 29, DOI:10.31228/osf.io/b5dwm.

157 Some scholars have argued that, according to grammar found in the recitals, information will not relate to specific decisions, e.g., Sandra Wachter and others, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 76. Others have instead examined the GDPR in light of its overarching principles, arguing that specific information may, under some circumstances, be provided Andrew D Selbst and Julia Powles, 'Meaningful Information and the Right to Explanation' (2017) 7 *International Data Privacy Law* 233. No case law has definitively determined one way or another.

158 See generally Oscar H Gandy, *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage* (Routledge 2009); Tawana Petty and others, 'Our Data Bodies: Reclaiming Our Data' (*Our Data Bodies Project*, June 2018); Seeta Peña Gangadharan and Jędrzej Niklas, 'Decentering Technology in Discourse on Discrimination' (2019) 22 *Information, Communication & Society* 882.

159 Rebekah Overdorf and others, 'POTs: Protective Optimization Technologies' [2018] arXiv:180602711 [cs].

160 Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 *Calif L Rev* 671.

161 Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' in *Conference on Fairness, Accountability and Transparency* (FAT* 2018) (2018).

cratic society,[162] as well as manipulating individuals more generally and pervasively.[163] Concerns exist that individuals lose their ability to reflect on morally challenging tasks from pervasive use of affective (emotional) predictive systems in their ambient environments.[164] Policy-makers are also concerned about 'addiction' to devices, 'dark patterns' attempting to foster profitable but undesirable habits,[165] underpinned by systems designed to predict individuals who might be easily swayed into, for example, spending more on an app.[166]

4.2.1 Discriminatory decision-systems

The transparency provisions around machine learning in the GDPR,[167] such as Article 15(1)(h) (see Table 1) as well as access rights more generally, might be directly and indirectly useful in achieving transparency over complex, automated systems.[168] While the utility of individualized transparency has been questioned,[169] data rights could play a role in creating aggregate, societal-level transparency and accountability. Data from access rights might be used to seek inferences, data and meta-data about prediction and training data which can inform researchers around how systems function. Algorithmic 'explanations', where provided, might be compiled to shine light on the functioning of a model,[170] or compared across individuals, demographics or applications. Data rights could also help understand where models come from, which actors were involved in training and building them, and when, which is particularly salient given the rise in business models involving the trading of trained machine learning models.[171]

One example of an attempt to do just this with data protection rights can be found in the German credit scoring context. OpenSCHUFA was a campaign in Germany run by AlgorithmWatch and the Open Knowledge Foundation Deutschland attempting to reverse-engineer the main system used to determine creditworthiness of German residents. It built a data donation platform that was used by over 4,000 people to collate SCHUFA access information on the basis of data rights, in particular, asking for copies of data under the right to access that could later be analyzed. While such a campaign was a logistical success, and placed pressure on the SCHUFA, it also revealed an array of challenges in using data rights in this way, such as sampling bias, which we discuss later below.[172]

If sampling challenges can be overcome, the information that can be gathered through data rights of diverse populations might shine a light on some discrimination concerns. Studies that have tried to understand discrimination in job adverts, for example, have relied on different methods. The challenges of one of these, web scraping, have already been described.[173] Add to these the challenges of creating a credible 'data exhaust' which can be mistaken as that of a real person — a challenge which flummoxes even the intelligence services[174] — and it becomes clear that the bot approach might fast drift from the lived experience of individuals online. Others have relied on self-reported performance data from the platform itself:[175] whether such data can be trusted when there are strong incentives to make adverts look well-performing and non-discriminatory are unclear.

4.2.2 Recommenders and media exposure

A considerable deal of concern has centered around the creation of digital 'echo chambers' or 'filter bubbles' in relation to content viewed online.[176] There is limited empirical evidence to support their existence in many cases, particularly within traditional news source.[177] but the field is still poorly understood, particularly in the context of platforms working to enclose content within walled gardens.[178] Indeed, empirical work on the power of media recommender algorithms in radicalizing viewers would greatly benefit from more granular insights that access rights enable.[179] Where data about content shown, clicked on and/or viewed is stored or retained, it might prove useful for independent analysis and comparison to understand the extent of this tracking.[180]

4.3 Repurposing digital traces

Data rights can also provide data for other scientific and humanistic questions. The sensing infrastructure provided by mobile phones or 'smart' home devices have already been considered for 'citizen science' or 'community science' and 'participatory sensing'. However, these applications have typically focused on environmental factors, such as air, noise and water pollution,[181] and rely on the user send-

162 Information Commissioner's Office, Democracy Disrupted? Personal Information and Political Influence (ICO 2018).
163 Karen Yeung, '"Hypernudge": Big Data as a Mode of Regulation by Design' (2017) 20 *Information, Communication & Society* 118.
164 Sylvie Delacroix and Michael Veale, 'Smart Technologies and Our Sense of Self: Going Beyond Epistemic Counter-Profiling' in Mireille Hildebrandt and Kieron O'Hara (eds), *Life and the Law in the Era of Data-Driven Agency* (Edward Elgar 2020).
165 Forbruker Rådet, 'Deceived by Design' (27 June 2018).
166 Ronan Fahy and others, 'Data Privacy, Transparency and the Data-Driven Transformation of Games to Services' (August 2018) *2018 IEEE Games, Entertainment, Media Conference* (GEM) 1; Digital, Culture, Media and Sport Committee, 'Immersive and Addictive Technologies' (House of Commons, HC 1846, 12 September 2019).
167 See generally Edwards and Veale (n 8).
168 Edwards and Veale (n 8).
169 Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 *New Media & Society* 973.
170 Some work has recently shown that model reconstruction attacks can be heightened by the use of model explanations. e.g., Smitha Milli and others, 'Model Reconstruction from Model Explanations' in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19, New York, NY, USA, ACM 2019). Work is ongoing to understand what explanations can be used in relation to models, See further Martin Strobel, 'Aspects of Transparency in Machine Learning' in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (AAMAS '19, Richland, SC, International Foundation for Autonomous Agents and Multiagent Systems 2019).
171 Michael Veale and others, 'Algorithms that Remember: Model Inversion Attacks and Data Protection Law' (2018) 376 *Phil Trans R Soc A* 20180083.
172 See infra section 5.2.
173 See supra section 2.3.
174 Sam Jones, 'The Spy Who Liked Me: Britain's Changing Secret Service', (*Financial Times,* 29 September 2016) https://www.ft.com/content/b239dc22-855c-11e6-a29c-6e7d9515ad15 (accessed 29 April 2019).
175 Anja Lambrecht and Catherine Tucker, 'Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads' (2019) 65 *Management Science* 2966; Ali and others (n 22).
176 Frederik J Zuiderveen Borgesius and others, 'Should We Worry about Filter Bubbles?' (2016) 5 *Internet Policy Review*.
177 Zuiderveen Borgesius and others (n 176); Judith Möller and others, 'Do Not Blame It on the Algorithm: An Empirical Assessment of Multiple Recommender Systems and Their Impact on Content Diversity' [2018] *Information, Communication & Society* 1; Mario Haim and others, 'Burst of the Filter Bubble?' (2018) 6 *Digital Journalism* 330.
178 See generally Angela M Lee and Hsiang Iris Chyi, 'The Rise of Online News Aggregators: Consumption and Competition' (2015) 17 *International Journal on Media Management* 3; Paddy Leerssen, 'The Soap Box is a Black Box: Regulating Transparency in Social Media Recommender Systems' (2020) 11 *EJLT*.
179 cf. Kevin Munger and Joseph Phillips, 'A Supply and Demand Framework for YouTube Politics' [2019] 38; Rebecca Lewis, 'Alternative Influence: Broadcasting the Reactionary Right on YouTube' (18 September 2018).
180 See in this regard, Leerssen (n 178) 2.
181 Stacey Kuznetsov and Eric Paulos, 'Participatory Sensing in Public Spaces: Activating Urban Surfaces with Sensor Probes' in *Proceedings of the 8th ACM Conference on Designing Interactive Systems* (DIS '10, New York, NY, USA, ACM 2010); Prabal Dutta and others, 'Common Sense:

ing data directly rather than repurposing data collected for another purpose. Data rights could widen the scope of citizen/community science — we highlight some potential directions below.

4.3.1 Location data

Location data is one of the richest forms of data, and the rise in location aware applications has long attracted privacy concerns.[182] Because mobile phones connect so regularly to base stations, they leave a long trace of location. As users increasingly rarely turn phones off,[183] such location traces effectively extend to all times when the phone is in contact with telecoms infrastructure. As a result, telecoms data has been used by national statistical agencies and humanitarian groups alike—at times attracting considerable ethical controversy.[184] Mobile phone location data might, for example, be used to infer the type of transport someone is using,[185] socioeconomic information about them,[186] or places that they consider important,[187] among many other potential applications. But equally, with consent and with proper ethical consideration, it might be that a research subject would be happy to pass over parts of their location history to better understand some intervention or experiment they have been part of.

4.3.2 Biosensors

Commercial devices with self-monitoring sensing capabilities are becoming increasingly popular,[188] and there has been increasing interest in the medical domain in validating these consumer-grade devices to understand if their data collection has the required validity for scientific use.[189] Many devices and software are tracking physical and social characteristics of individuals, from the number of steps taken to the use of houses, vehicles, software, and even clothing. Researchers have highlighted that

> the [device] companies could allow access to more data that are collected. At present, the trackers provide users with only a subset of data that is actually collected. The companies control the output available, making the day-level summary variables the easiest to obtain. For example, despite capturing GPS and heart rate on two trackers, Fitbit currently limits the export of these full datasets. Furthermore, the resulting output is derived through proprietary algorithms that may change over time and with new features. [..] At a minimum, it would be helpful for companies to reveal what pieces of data are being used by the trackers to calculate each output measure.[190]

The role of trade secrets in this area is particularly pertinent. For example, many people use 'smart watches' to measure features of their circulation. Many research fields utilize photoplethysmography data, also known as blood volume pulse. It can be used to measure oxygen saturation, blood pressure and cardiac output, to assess autonomic function and to detect peripheral vascular disease.[191] Smart watches do not measure this directly, however: they infer it from a series of sensed measurements, often using proprietary and changing machine learning systems.[192] For researchers, this (whether in commercial or research grade) products can present challenges, as changing algorithmic systems introduce features which can be difficult to control for. For users, it might not be an issue however: they likely want the most robust and accurate measure of their heartbeat, step-count, sleep patterns or the like over time, and do not care about internal validity over the months and years of device usage.

Depending on the structure of processing, researchers interested in utilizing these sensors may be able to use transparency rights to obtain additional datasets. This might be particularly useful if and when a time comes where users are *already* using high-grade sensors in their daily lives, and research studies would work better by co-opting existing infrastructure rather than adding a further device which is not part of a user's existing routine, or may be redundant to something they already are familiar with.

4.3.3 Labor patterns

Between 1% and 5% of the EU population is estimated to have taken place in some form of paid platform work, with some countries exhibiting significantly higher rates of participation than that.[193] The growth of these markets for informal labor, such as through taxi services provided through Uber or Lyft, or workers on computers using platforms such as Amazon Mechanical Turk, has led to serious concerns, culminating in high profile legal fights, over the employment status of such individuals and the rights they possess. For example, informal work can necessarily bring a considerable amount of overhead, such as sifting through jobs online to find those which are legitimate, and being 'hypervigilant' in order to secure desirable or profitable jobs.[194] In this context, there are important factual questions, with legal ramifications, around the timings and behavior of 'gig economy' workers, such as the amount of time they are active on the app waiting for

Participatory Urban Sensing Using a Network of Handheld Air Quality Monitors' in *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems*, (ACM 11 April 2009) http://dl.acm.org/citation.cfm?id=1644038.1644095 (accessed 18 June 2019); Nicolas Maisonneuve and others, 'NoiseTube: Measuring and Mapping Noise Pollution with Mobile Phones' in Ioannis N Athanasiadis and others eds, *Information Technologies in Environmental Engineering* (Springer Berlin Heidelberg 2009).

182 AR Beresford and F Stajano, 'Location Privacy in Pervasive Computing' (2003) 2 *IEEE Pervasive Computing* 46.

183 UK regulator Ofcom report that 71% of adults claim they never turn their phones off. See Ofcom, 'A Decade of Digital Dependency' (3 May 2019) https://www.ofcom.org.uk/about-ofcom/latest/features-and-news/decade-of-digital-dependency (accessed 24 July 2019).

184 Linnet Taylor, 'No Place to Hide? The Ethics and Analytics of Tracking Mobility Using Mobile Phone Data' (2016) 34 *Environ Plan D* 319; Linnet Taylor and Dennis Broeders, 'In the Name of Development: Power, Profit and the Datafication of the Global South' (2015) 64 *Geoforum* 229.

185 Donald J Patterson and others, 'Inferring High-Level Behavior from Low-Level Sensors' in Anind K Dey and others eds, *Ubiquitous Computing 2003,* UbiComp 2003 (Lecture Notes in Computer Science, Springer Berlin Heidelberg 2003).

186 Christopher Smith-Clarke and others, 'Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14, New York, NY, USA, ACM 2014).

187 Sibren Isaacman and others, 'Identifying Important Places in People's Lives from Cellular Network Data' in Kent Lyons and others eds, *Pervasive Computing* (Lecture Notes in Computer Science, Springer Berlin Heidelberg 2011).

188 See generally Gina Neff and Dawn Nafus, *Self-Tracking* (MIT Press 2016); Deborah Lupton, 'Self-Tracking Cultures: Towards a Sociology of Personal Informatics' in *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design* (OzCHI '14, New York, NY, USA, ACM 2014).

189 e.g., Kelly R Evenson and others, 'Systematic Review of the Validity and Reliability of Consumer-Wearable Activity Trackers' (2015) 12 *International Journal of Behavioral Nutrition and Physical Activity* 159.

190 Evenson and others (n 189) 19.

191 John Allen, 'Photoplethysmography and Its Application in Clinical Physiological Measurement' (2007) 28 *Physiol Meas* R1.

192 See e.g., Empathica, 'Utilizing the PPG/BVP Signal' (*Empatica Support*, 31 March 2016) http://support.empatica.com/hc/en-us/articles/204954639-Utilizing-the-PPG-BVP-signal (accessed 19 June 2019).

193 Chris Forde and others, 'The Social Protection of Workers in the Platform Economy' (Study for the European Parliament's EMPL Committee, IP/A/EMPL/2016-11, November 2017) 38.

194 Mary L Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Houghton Mifflin Harcourt 2019).

jobs,[195] which may require data access and analysis in order to assess compensation, fairness, and even potentially an individual's legal status with regards to e.g. holiday pay, breaks workers are entitled to, or other legal rights. Data rights are already central to civil society groups, such as Worker Info Exchange,[196] but if the informal economy continues to increase in density and complexity, more advanced, collective use of digital rights to gather data to understand exploitation, labor patterns, and the changing nature of work may be required.[197]

5 Considerations for Researching with Data Rights

While the opportunities seem promising, the research use of data rights is made difficult by several nuanced limitations. In this section we delineate some of the most important limitations, categorized as legal, methodological and ethical considerations.

5.1 Legal considerations

Even though the right of access is grounded in both the principles of the GDPR and Article 8 of the Charter, there are still legal questions as to its utility in the research context. Some of these issues have clearer answers in guidance and case law than others do. In this section, we group and tackle some of the major issues, misconceptions and open questions around the use of access rights in the contexts discussed earlier.

5.1.1 Motivation of the request

Prima facie, it might appear that a data controller could seek to refuse a request because enabling research was not a stated purpose of the GDPR. Yet case law and regulatory guidance falls behind the view that GDPR rights are *intent-agnostic*. Access rights have commonly been used in relation to highly specific pieces of information, often as part of disputes that might be related to issues of criminal,[198] employment,[199] financial,[200] fiscal,[201] immigration,[202] trust[203] or defamation proceedings.[204] These types of cases can create, in the words of AG Bobek, 'certain intellectual unease as to the reasonable use and function of data protection rules'.[205]

As European data protection has traditionally had a close connection with the right to privacy, one might argue that it is especially aimed at safeguarding the respective individual's interests. If such a view were taken, data protection transparency measures to gather research data might then appear to misuse/retrofit a legal device for unintended purposes, calling its legal enforceability into question. Yet there is an argument to be made that this type of usage is aligned *extremely well* with data protection's primary, historical purpose of regulating data infrastructures underlying society (from large, centralized data mainframes to the complex ecosystem today) rather than (just) supporting individually-focused privacy. The GDPR's legal toolbox that gives some level of control over personal data and/or how it is processed, and the use of these tools is arguably envisaged to be used by a range of stakeholders, including regulators, academics, journalists, artists and civil society organizations, not just by individual data subjects for purely individualistic purposes. As such, the GDPR's transparency measures, as a general tool with many potential uses for promoting oversight and agency, can only be intent agnostic: it is up to these stakeholders to use them flexibly as part of governance, self-determination and oversight.[206]

Indeed, the right of access is an explicit part of the fundamental right to data protection in the Charter, and courts and regulators have held that a 'privacy' motive is not required for its use. In *YS and others* for example, the Court of Justice made no reference to fact that the claimants were seeking to use the right of access in order to support litigation as evidence that their use of rights should fail. National case law has been supportive of this approach too. For example, both English[207] and Dutch[208] courts in recent years have reached a clear consensus that access requests are purpose-blind, and the guidance of the Information Commissioner's Office[209] and Autoriteit Persoonsgegevens[210] is in alignment with this. It is worth noting that some restrictions on motivation of access rights exist at national level to prevent data subjects from being coerced into making them.[211]

Especially insofar as research aims to shed light on the use of personal data in contemporary infrastructures, research uses of data rights seem not just possible within this intent-agnostic regime, but a prime example of an empowerment mechanism working on the side of data subjects.

5.1.2 Infringement of the rights and freedoms of others

Controllers might (partially) fend off access and portability requests when they can establish that accommodating them would 'adversely

195 *Uber BV v Aslam* [2018] EWCA Civ 2748 at [100]; 'Uber drivers demand access to their personal data' (n 14).
196 Farrar (n 14). See further https://workerinfoexchange.org and Mahieu and Ausloos (n 157) 8–10; 29; 'Uber drivers demand access to their personal data' (n 14).
197 The European legislator already took a step in this direction with: Directive (EU) 2019/1152 of the European Parliament and of the Council of 20 June 2019 on transparent and predictable working conditions in the European Union of 2019 EP, CONSIL 32019L1152, EP, CONSIL (European Union EP, CONSIL 2019).
198 *Kololo v Commissioner of Police for the Metropolis* [2015] EWHC 600 (QB). *Lin & Anor v Commissioner of Police for the Metropolis* [2015] EWHC 2484 (QB).
199 *Ittihadieh v 5-11 Cheyne Gardens RTM Company Ltd & Ors* [2017] EWCA Civ 121.
200 *Rechtbank Zwolle-Lelystad 103434 / HA RK 04-215 9 maart 2005*; Parket bij de Hoge Raad 9 Nov 2018.
201 Amélie Lachapelle and Elise Degrave, 'Le Droit d'accès Du Contribuable à Ses Données à Caractère Personnel et La Lutte Contre La Fraude Fiscale, Revue Générale Du Contentieux Fiscal, 2014, 5, p. 322-335' [2014].
202 Joined Cases C-141/12 and C-372/12 *YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel v M and S* ECLI:EU:C:2014:2081.
203 *Dawson-Damer & Ors v Taylor Wessing LLP* [2017] EWCA Civ 74.
204 *Rudd v Bridle & Anor* [2019] EWHC 893 (QB).
205 Case C-13/16 *Valsts policijas Rīgas reiona pārvaldes Kārtības policijas pārvalde v Rīgas pašvaldības SIA 'Rīgas satiksme'* ECLI:EU:C:2017:43, Opinion of AG Bobek, para 93.
206 See also: René Mahieu and Jef Ausloos, 'Harnessing the Collective Potential of GDPR Access Rights: Towards an Ecology of Transparency' [2020] *Internet Policy Review*.
207 e.g., *Dawson-Damer & Ors v Taylor Wessing LLP* [2017] EWCA Civ 74 at [104]–[108]; *B v The General Medical Council* [2019] EWCA Civ 1497 at [79] ('the general position is that the rights of subject access to personal data [...] are not dependent on appropriate motivation on the part of the requester') and case law cited therein.
208 *Parket bij de Hoge Raad, 9 November 2018* ECLI:NL:PHR:2018:1273, at para 3.37.
209 See generally Information Commissioner's Office, 'Subject Access Code of Practice' (9 June 2017) 47.
210 Autoriteit Persoonsgegevens, *Recht op inzage* (Netherlands Autoriteit Persoonsgegevens).
211 e.g., Data Protection Act 2018 (United Kingdom) s 184, which, albeit not relevant to researching through data rights, creates offences for employers or providers of contracts to make arrangements conditional on the production of 'relevant records' obtained by use of a SAR. See generally (in relation to the previous regime) Alexander de Gaye and Sabba Mahmood, 'Enforced Subject Access—is It Finally the End?' (2014) 15 *Privacy and Data Protection* 10; Information Commissioner's Office, Enforced Subject Access (Section 56) (ICO 2015).

affect the rights and freedoms of others'.[212] Understanding of this clause by the EDPB has centered on two rights which might be balanced against information rights in the GDPR — the right to privacy and trade secrets/intellectual property rights.[213]

Privacy of third parties. The rights to privacy and data protection of third parties is one of the most important roles of this provision, and likely its most common use. It is common that personal data relates to more than one natural person — messages, notes about one person made by another, ratings and reputation systems, or shared 'smart' devices, for example. This is not a *carte blanche* to refuse data provision, however. The European Court of Human Rights (ECtHR) has held that an access rights regime would be in breach of Article 8 of the Convention if there was no independent authority to determine if access had to be granted if an individual to whom data also relate failed to provide or withheld consent.[214]

The European Court of Human Rights has also weighed in on the argument that the inclusion of some personal data in a document renders it ineligible for release. In *Társaság a Szabadságjogokért v Hungary,*[215] an NGO attempted to access a complaint to the Constitutional Court submitted by a member of parliament. The Government of Hungary denied this request, arguing that the complaint contained the personal data of the member, and consequently was ineligible for release. The Court found it 'quite implausible that any reference to the private life of the MP [...] could be discerned from his constitutional complaint', and noted that it would be 'fatal for freedom of expression in the sphere of politics if public figures could censor the press and public debate in the name of their personality rights, alleging that their opinions on public matters are related to their person and therefore constitute private data which cannot be disclosed without consent'.[216] It found a violation of Article 10 (freedom of expression), in relation to the freedom to 'receive and impart information and ideas without interference by public authority and regardless of frontiers'.

This emerging regime appears favorable to the use of data rights in research, particularly if ethical reviews are undertaken to carefully consider third party privacy interests.[217]

Intellectual property of the controller. The EDPB anticipated that controllers will invoke this clause in relation to an adverse effect on *their* rights and freedoms. [218] A clear example would be where a trade secret or IP argument is forwarded by the controller.[219] Yet, as counselled in the recitals to the GDPR, 'the result of those considerations should not be a refusal to provide all information to the data subject.'[220] How this would play out in the situation where access requests *en masse* might threaten intellectual property in a different way is unclear. It is worth noting however that it would be very difficult for a data controller to accurately pre-empt the fact that data rights were being used in that way. Indeed, from the CJEU's case law on copyright protection, it can be derived that the mere potentiality of an IP breach will not generally be sufficient to impinge on the right to data protection in Article 8 of the Charter (which includes a right of access as mentioned before).[221]

Freedom to conduct a business. It could also be envisaged that a company claims that its 'freedom to conduct a business'[222] has been adversely affected. Yet the freedom to conduct a business is not an absolute right, but must be considered in relation to its societal function.[223] Restrictions of and interferences with this freedom are possible in cases where they correspond to an objective of general interest pursued by the Union, and respect the 'actual substance' of the freedom.[224] Furthermore, the Court has upheld that the tentative wording of Article 16,[225] which differs from that of other rights and freedoms in Title II of the Charter, reflects a broader leeway to restrict this freedom than they would have otherwise.[226]

Indeed, cases where the Court has held a measure in breach of Article 16 are rare, and even in these cases have only been in breach when read closely with EU secondary legislation.[227] In *Scarlet Extended*, the Court held that the installation of 'a complicated, costly, permanent computer system at [the company's] own expense' (to monitor internet traffic) would be a 'serious infringement' of the freedom to conduct a business in Article 16 of the Charter.[228] This was upheld in *SABAM v Netlog*.[229] However, it is important to consider the broader context in both cases, where the freedom to conduct a business aligned with the respective service providers' rights to data protection and freedom of expression (resp Articles 8 and 11 Charter). Moreover, in the latter case the Court relied specifically on the explicit language of the IPR Enforcement Directive to this effect, which forbids intellectual property enforcement measures that are 'unnecessarily complicated or costly'.[230] No comparable language or provision exists

212 GDPR, arts 15(4), 20(4).
213 See Article 29 Working Party, 'Guidelines on the Right to Data Portability (n 109)9–10 (mentioning only these two areas as examples of an issue to be considered as part of Article 20[4]).
214 *Gaskin v United Kingdom* [1990] EHRR 36. The United Kingdom for example entered a balancing test into the law as a result, stating there is no obligation to provide data 'to the extent that doing so would involve disclosing information relating to another individual who can be identified from the information' unless that third individual has consented to the release or it is reasonable to do so (determined with regard to the type of information, duties of confidentiality which might exist, attempts to seek consent, capabilities to give consent or express refusal of consent). See Data Protection Act 2018 (United Kingdom) sch 2 para 16.
215 *Társaság a Szabadságjogokért v Hungary App* no 37374/05 (ECtHR 2009).
216 *Társaság a Szabadságjogokért v Hungary* (n 215) para 37.
217 Research purposes benefit from a more lenient approach in the GDPR, See GDPR, art 89. See generally Miranda Mourby and others, 'Governance of Academic Research Data under the GDPR—Lessons from the UK' (2019) 9 *International Data Privacy Law* 192. See also Article 29 Working Party, 'Guidelines on the right to data portability' (n 109) 12.
218 See Article 29 Working Party, 'Guidelines on the right to data portability' (n 109) 9–10 (mentioning only these two areas as examples of an issue to be considered as part of Article 20[4]).
219 Gianclaudio Malgieri, 'Trade Secrets v Personal Data: A Possible Solution for Balancing Rights' (2016) 6 *International Data Privacy Law* 102; Jef Ausloos, *The Right to Erasure in EU Data Protection Law. From Individual Right to Effective Protection* (Oxford University Press 2020) 343–48.
220 GDPR, recital 63. Similarly, See Caroline Colin and Yves Poullet, 'Du Consommateur et de Sa Protection Face à de Nouvelles Applications Des Technologies de l'information: Risques et Opportunités' (2010) 2010/3 *DCCR* 94, 117; Malgieri (n 219) 103.
221 See case law references in Ausloos (n 219) 345.
222 Charter, art 16.
223 Joined Cases C-184/02 and C-223/02 *Spain and Finland v Parliament and Council* ECLI:EU:C:2004:497; Ausloos (n 219) 335–43.
224 Case C-554/10 *Deutsches Weintor eG v Land Rheinland-Pfalz* ECLI:EU:C:2012:526, para 54.
225 'The freedom to conduct a business in accordance with Union law and national laws and practices is recognised.'
226 Case C-283/11 *Sky Österreich GmbH v Österreichischer Rundfunk* ECLI:EU:C:2013:28, para 47.
227 See generally Peter Oliver, 'What Purpose Does Article 16 of the Charter Serve?' in Ulf Bernitz and others (eds), *General Principles of EU Law and European Private Law* (Kluwer Law International 2013).
228 Case C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL* (SABAM) ECLI:EU:C:2011:771.
229 Case C-360/10 *Belgische Vereniging van Auteurs Componisten en Uitgevers CVBA (SABAM) v Netlog NV* ECLI:EU:C:2012:85.
230 Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights (Text with EEA relevance) OJ L 157/45 ('IPR Enforcement Directive'), art 3(1).

in the GDPR. Even in cases where significant costs are placed upon a business, such as in *Denise McDonagh v Ryanair Ltd*, where the airline's duty to provide care after the eruption of the Icelandic volcano Eyjafjallajökull left passengers stranded, the existence of articles in secondary legislation that could be understood to reconcile fundamental rights (in this case, freedom to conduct a business and the right to property with the right to consumer protection) led the Court to rule no breach of the right to conduct a business had occurred.[231]

The GDPR has many provisions designed to respect (or enable Member States to navigate) the balancing between, among other fundamental rights and freedoms, Articles 8 and 16 of the Charter, such as Article 12 (on transparency modalities), Article 14(5) (on situations where information obligations can be avoided or relaxed) and Article 23(1) ('Restrictions'). Furthermore, for information-intensive companies, the marginal cost of providing information to each individual once a compliant infrastructure is established is very low (compared to, for example, flight compensation). Indeed, it does not generally require the establishing of any new modalities of communication, as information-intensive companies already have data and computational infrastructures, as well as log-in accounts and/or email, which can be used to this end.[232] Consequently, in agreement with many scholars,[233] we do not see much chance of the freedom to conduct a business as standing in the way of the use of data rights, including in research situations.

5.1.3 Abuse of rights?

The idea that an access right could, in certain situations, construe an abuse of rights was considered by Advocate General (AG) Kokott in her opinion in *Nowak*.[234] Abuse of rights is, however, 'rarely used, or at least not successfully',[235] usually implicated in politically charged, high level issues concerning freedom of expression or freedom of association, often when pitted against values of the defense of democracy. Yet, as AG Kokott noted, the risk of abuse of rights which was present in the 1995 Data Protection Directive is 'resolved' in the GDPR by the considerations of the rights and freedoms of others (see section 5.1.2).[236] We agree, noting further that the intent-agnostic nature of the right to access under the GDPR makes abuse more difficult to construe.[237]

5.1.4 Data, not documents

It is important to note that accommodating the right of access does not necessarily require sharing an exact copy of the data on the servers (or in the manual filing system) of the data controller in question. Both the CJEU and national courts have affirmed that a SAR is not a right to access whole documents, for example to provide context, but the right to the personal information contained within.[238] Such data could be extracted and provided in a variety of forms, and need not be in the original format. Indeed, there are times where that original format might actually be undesirable, such as if it is proprietary in nature, requiring the data subject to have specific software or expertise to examine it. No cases have been ruled on or are pending in the CJEU relating to the new right to data portability, but we can safely assume that that right, too, does not provide access to documents. As a result, there will be research designs that are better suited to freedom of information legislation,[239] or access to environmental information legislation,[240] which both can provide documentation for matters within their respective scopes. In many cases however, data controllers may find it easier to provide documents, and as such while it cannot be relied upon, data rights may be useful in studies where the original context is crucial for understanding.

5.1.5 Lack of consistency and machine readability

The 2012 GDPR proposal had a role for the European Commission, through implementing acts, of specifying a standard for the format of SAR responses in different sectors.[241] This aspect of the GDPR was a casualty of the intense, half-decade political battle over the text. The result is that access (and portability) rights do not have a common standard or format which data subjects can expect. This, in turn, makes it hard to build tools which are data controller agnostic, and which are reliable enough not to break if a data controller decides to switch the form of response they provide.[242] While codes of conduct and certification mechanisms under the GDPR may yet provide a means to help standardize this area, [243] we are still to see one on access or portability rights take concrete shape[244] — although a plethora of third parties seeking to sell back-end software to data controllers with the promise of consolidation and automation have emerged.[245]

Obtaining machine-readable data is crucial for research.[246]

231 Case C-12/11 *Denise McDonagh v Ryanair Ltd* ECLI:EU:C:2013:43 at [59]–[65].

232 ccf. Case C-649/17 *Bundesverband der Verbraucherzentralen und Verbraucherverbände—Verbraucherzentrale Bundesverband eV v Amazon EU Sàrl* ECLI:EU:C:2019:576, where the Court emphasised with reference to the freedom to conduct a business that a firm should not be obliged to establish a phone line for the purposes of communication with consumers where alternative means of direct and effective communication have been established.

233 See Hielke Hijmans, 'The European Union as a Constitutional Guardian of Internet Privacy and Data Protection' (PhD Thesis, University of Amsterdam 2016) 196, 216–17, 258; Ausloos (n 219) 333–49.

234 Case C-434/16 *Peter Nowak v Data Protection Commissioner* ECLI:EU:C:2017:582, Opinion of AG Kokott, paras 42–50.

235 Lorna Woods, 'Abuse of Rights' in Steve Peers and others (eds), *The EU Charter of Fundamental Rights : A Commentary* (Hart Publishing 2014) 1545.

236 Case C-434/16 *Peter Nowak v Data Protection Commissioner* ECLI:EU:C:2017:582, Opinion of AG Kokott, para 48.

237 See section 5.1.1.

238 Joined Cases Joined Cases C-141/12 and C-372/12 *YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel v M and S* [2019] ECLI:EU:C:2014:2081 [48].; *Dunn v Durham County Council* [2012] EWCA Civ 1654, [2013] 2 All ER 213 at 16; *Ittihadieh v 5-11 Cheyne Gardens RTM Company Ltd & Ors* [2017] EWCA Civ 121 at 93; *Rudd v Bridle* [2019] EWHC 893 (QB); *Rechtbank Noord-Holland (24 May 2019)* ECLI:NL:RBNHO:2019:4283; Parket bij de Hoge Raad 9 Nov 2018.

239 e.g., Freedom of Information Act 2000 (United Kingdom).

240 e.g., implementations of the UN/ECE Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters ('Aarhus Convention') such as transpositions of Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC OJ L 41/56.

241 Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM/2012/011 final - 2012/0011 (COD)), Art 15.

242 This is comparable to the politics of APIs and programmatic access, See generally section 2.2 above.

243 GDPR, arts 40, 42.

244 Though it is worth pointing to a recent initiative by the European Digital Media Observatory (EDMO), which is setting up a working group in order to develop a code of conduct on 'Access to Data Held by Digital Platforms for the Purposes of Social Scientific Research'. See notably: 'Call for Comment on GDPR Article 40 Working Group' (n 29); Vermeulen (n 29). There are also some self-regulatory initiatives, none of which really seem to have gained a lot of traction, most notably the 'Data Transfer Project' (with among its contributors: Apple, Facebook, Google, Microsoft and Twitter). See https://datatransferproject.dev

245 The IAPP compiled a list of such providers, accessible at https://iapp.org/resources/article/privacy-tech-vendor-report.

246 cf. European Commission, 'A European strategy for data' (n 6) 10.

Machine-readable data is not the same as digital data. For example, a PDF containing tabular data is designed to be printed rather than read and processed by a computer, and as such are not generally marked-up in such a way which makes automatic processing easy.[247] Machine-readable has been defined in EU law as 'a file format structured so that software applications can easily identify, recognize and extract specific data, including individual statements of fact, and their internal structure.'[248]

A teleological reading of the GDPR would require controllers to share personal data from the right to access in a consistent, machine-readable format unless great effort was involved. This effort is unlikely to be required in the context of online services, which given their automated nature already hold the data in such forms, as both consistency and machine readability are key to their business models — and to take more effort to obstructively destroy such properties would be highly questionable in light of the overarching data protection principle of fairness.[249] Even more so considering the European Commission's more recent push for 'stricter requirements on interfaces for real-time data access and making machine-readable formats compulsory for data from certain products and services'.[250]

5.1.6 (Re)identifying data subjects

Article 11(1) explains that controllers do not have to retain personal data *only* for the ability to potentially accommodate data subject rights at a later stage. Put differently, the requirement to accommodate data subject rights does not prevent them from anonymizing their datasets. Be that as it may, data subjects still have the possibility to provide the controller with additional information so as to (re-) identify their data in anonymized data-sets.[251] In practice however, this may lead to a frustrating back-and-forth between data subject and controller, where the data controller appears to have designed systems that are deliberately challenging to reidentify data subjects within.[252] In particular, the data controller may argue that the data, while clearly falling within the GDPR's scope (with high re-identification potential and in practice used to target or single out data subjects), may not be re-identifiable to the very high reliability needed to ensure that data not relating to an individual is delivered to them by mistake.[253] This is an argument Apple makes to refuse accommodating access requests with regard to the voice-data gathered in relation to its Siri-service.[254] Such arguments will generally be insufficient to block access requests entirely however.[255]

5.1.7 'Disproportionate effort?'

Some data controllers have read into data protection law the existence of a 'disproportionate effort' exemption which would exempt them from fulfilling an access request.[256] Such an exemption does not appear to exist in the GDPR, although it did in some transpositions of the now defunct 1995 Data Protection Directive.[257] Complaints around this are ongoing and it seems likely that more clarity will be forthcoming. Indeed, even if the increasing complexity of data processing ecosystems may render it hard to accommodate the core transparency requirements,[258] it does not exonerate controllers. To the contrary, Recital 58 highlights transparency is even more important in complex situations involving many actors.[259] When the controller processes a large quantity of personal data, Recital 63 does permit the controller to request the data subject to specify the information or processing activities to which the request relates.

One related provision that *does* exist in the GDPR is the ability to refuse a request if the nature of that request is 'manifestly unfounded or excessive, in particular because of their repetitive character'. Where this is done, 'the controller shall bear the burden of demonstrating the manifestly unfounded or excessive character of the request.'[260] This provision relates to the character of the request itself, rather than the character of the burden of fulfilling that request.

The EDPB have noted that for information society services such as large social media firms which specialize in automated data processing, 'there should be very few cases where the data controller would be able to justify a refusal to deliver the requested information, even regarding multiple data portability requests.'[261] They also note that the cost of building the infrastructure to comply with these requests is irrelevant to the notion of 'excessive' requests. In particular, they state that 'the overall system implementation costs should neither be charged to the data subjects, nor be used to justify a refusal to answer portability requests.'[262] Under these conditions, it appears that there are limited general reasons to refuse a data subject access request or portability request on effort grounds.[263]

5.1.8 National exemptions

It should also be noted that Article 23 grants Member States (and the EU legislator) the ability to install specific exemptions to the rights of access/portability in their national and/or sector-specific laws.[264] While most of the situations in which such exemptions can be pre-

247 The EDPB has stated that PDFs are unlikely to meet portability requirements, also noting that the requirements of portability must be interpreted in the context of the intention of the portability requirement, which the recitals (68) note is to promote interoperability. See Article 29 Working Party, 'Guidelines on the right to data portability' (n 109) 18. See also Ausloos and others (n 117) 286–87.

248 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] OJ L 172/56, art 2(13).

249 See generally Clifford and Ausloos (n 78).

250 European Commission, 'A European strategy for data' (n 6) 20.

251 GDPR, art 11(2).

252 This is further detailed in: Michael Veale and others, 'When Data Protection by Design and Data Subject Rights Clash' (2018) 8 *International Data Privacy Law* 4; Ausloos (n 125). For mobile app-specific considerations, See Norval and others (n 124).

253 On the security implications of data rights, See Andrew Cormack, 'Is the Subject Access Right Now Too Great a Threat to Privacy?' (2016) 2 *European Data Protection Law Review* 15; Coline Boniface and others, 'Security Analysis of Subject Access Request Procedures How to Authenticate Data Subjects Safely When They Request for Their Data' (*2019 - Annual Privacy Forum*, 13 June 2019) https://hal.inria.fr/hal-02072302/document (accessed 4 April 2019).

254 Veale and others (n 252).

255 Ausloos and others (n 117) 308–09.

256 Ausloos (n 125).

257 e.g., Data Protection Acts 1998, 2003 (Ireland) s 4(9) (repealed).

258 René Mahieu and others, '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' (2019) 10 *JIPITEC*.

259 Article 29 Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (Guidelines, 6 February 2018) 25.

260 GDPR, art 12(5).

261 Article 29 Working Party, 'Guidelines on the right to data portability' (n 109) 15.

262 Article 29 Working Party, 'Guidelines on the right to data portability' (n 109) 15.

263 See generally Ausloos and others (n 117).

264 For a review, See Access Now, One Year Under the EU GDPR: An Implementation Progress Report (Access Now 2019).

scribed relate to specific contexts[265] and are subject to conditions,[266] there is a catch-all included that makes it hard to anticipate the level of derogations to access/portability rights. Especially because this catch-all—enabling EU or Member State laws to restrict data subject rights in order to safeguard 'the rights and freedoms of others'—may be deployed in any kind of legislation (so not just the GDPR implementation laws). For example, while the seminal *Nowak* case in 2017 highlighted that data protection subject access rights applied to exam scripts,[267] this jurisprudence had limited direct applicability within the United Kingdom, which had an exemption for exam scripts being subject to access requests since 1998, replicated in the new law of 2018.[268]

In any case, such exemptions or derogations ought to be formulated and interpreted restrictively and narrowly. Hence, it is fair to say that the default position should be that data subject rights *are* applicable, *unless* the controller can clearly establish the applicability of a (national) exemption or derogation.[269] Such derogations are subject to potential challenge on the grounds of data protection principles and the fundamental right to data protection more generally.

5.2 Ethical considerations

While not easily split from other concerns, there are several ethical challenges that are distinctly applicable to data rights in research.

5.2.1 Who are the research subjects?

One ethical argument against the use of data rights in research is that it places a heavy burden on infrastructures that can prevent them from carrying out their normal function. A relevant question is whether data controllers (and their staff) would then be research subjects in the context of such a study.

Useful analogies can be found in studies of the peer review system. A 1982 study considered the rejection of duplicate papers by fictitious less-prestigious authors by selective American psychology journals.[270] They submitted 12 papers that journals had already accepted, authored by researchers from prestigious American psychology departments, but changed the names on the papers to fictious ones to see whether the prestige of the authors biased the reviewers' responses. A different 1987 study investigated whether social work journals' editorial processes were biased in favor of studies showing interventions to be effective, sending 146 submissions to test this hypothesis.[271] Both works were published by journals only trepidatiously and in an unusual manner. Despite referees' reservations about both the rigor of both studies, the journals that published these pieces (*Behavioral and Brain Sciences* and *Science, Technology and Human Values* respectively) did so only alongside commentaries (5 and 55 (short form) commentaries respectively) on relevant methodological and ethical issues.[272] In later years, the issue of studies into peer review was reignited by the 'Sokal affair', where a paper designed to be non-sense was submitted by Alan Sokal, a physics professor into a post-modern cultural studies journal and accepted, and follow-up events that have become known as 'Sokal Squared'.[273]

Scholars considering the ethical implications of these types of studies have questioned the 'social overhead of social research',[274] asking whether the 'costs of studying and correcting an injustice consume so many resources that they create new injustices, or create a net social loss [..] if too many people designed [peer review bias testing experiments], they would simply clog the peer review machinery altogether and bring the system to its knees.'[275] Parallel concerns have been raised in relation to issues of 'survey fatigue',[276] that 'indiscriminate use of surveys may be undercutting their effectiveness as a data collection approach by creating survey fatigue and lowering response rates',[277] particularly among student populations.[278] Others have considered that perhaps the journal editors and peer reviewers should have consented in line with widely accepted norms of research ethics. '[S]cientists do have rights,' one commentator noted, 'and [those] rights are not less than those guaranteed other human subjects'.[279] Others yet consider it important to weigh the stress on the system with the need to scrutinize gatekeepers of power and prestige.[280]

A key question to take away and analyze from this is whether formal processes of research ethics should be engaged simply because individuals are burdened as a result of the research. It is not clear in the case of data rights that simply because a human is involved in the fulfilment of a statutory obligation that the research should be treated as 'human subject' research.

There are some jurisdictions that have exempted studies concerning data rights from ethical review on the basis that disclosures mandated by legislation already have processes of custodianship associated with them and built into their respective regimes. Canada's three

265 For example, national security; defence; public security; prevention, investigation, detection or prosecution of criminal offences. See GDPR, art 23(1).

266 GDPR, art 23(2).

267 Case C-434/16 *Peter Nowak v Data Protection Commissioner* ECLI:EU:C:2017:994 (Nowak).

268 Data Protection Act 1998 (United Kingdom) sch 7 para 9 (repealed); Data Protection Act 2018 (United Kingdom) sch 2 para 25.

269 While certainly an interesting and much needed exercise, mapping the different implementations of Article 23 across EU Member States, even when only focusing on GDPR implementation laws, far reaches beyond the scope of this paper.

270 Douglas P Peters and Stephen J Ceci, 'Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again' (1982) 5 *Behavioral and Brain Sciences* 187.

271 William M Epstein, 'Confirmational Response Bias Among Social Work Journals' (1990) 15 *Science, Technology, & Human Values* 9.

272 Susan E Cozzens, 'Editorial' (1990) 15 *Science, Technology, & Human Values* 5.

273 Issues that arose in the Peters and Ceci and the Epstein studies also returned in subsequent peer-review 'hoax' studies, such as the so-called Sokal Affair, where the mathematician Alan Sokal sought to test his belief that the journal Social Text would accept an article that did not make sense, but supported the editors' ideological views. Despite the Sokal Affair reaching higher peaks of notoriety than either Peters and Ceci's or Epstein's controversies, Sokal submitted only a single paper, and therefore it is the parallel with the two studies above that is the most interesting for our purposes. See generally Stephen Hilgartner, 'The Sokal Affair in Context' (1997) 22 *Science, Technology, & Human Values* 506. On the later hoaxes, See Yascha Mounk, 'What an Audacious Hoax Reveals About Academia', (*The Atlantic*, 10 May 2018) https://www.theatlantic.com/ideas/archive/2018/10/new-sokal-hoax/572212 (accessed 30 November 2019).

274 See generally the special issue commencing with Joan E Sieber, 'Whose Ethics? On the Perils and Dilemmas of Studying Powerful Persons' (1983) 9 *SASP Newsletter* 1.

275 Mary Clark, 'Comments from the Side Lines' (1983) 9 *SASP Newsletter* 10, 11.

276 Stephen R Porter and others, 'Multiple Surveys of Students and Survey Fatigue' (2004) 2004 *New Directions for Institutional Research* 63.

277 Curtis A Olson, 'Survey Burden, Response Rates, and the Tragedy of the Commons' (2014) 34 *Journal of Continuing Education in the Health Professions* 93, 93.

278 Stephen R Porter, 'Survey Research Policies: An Emerging Issue for Higher Education' (2005) 2005 *New Directions for Institutional Research* 5, 8.

279 Michael J Mahoney, 'Bias, Controversy, and Abuse in the Study of the Scientific Publication System' (1990) 15 *Science, Technology, & Human Values* 50, 53.

280 Rachelle D Hollander, 'Journals Have Obligations, Too: Commentary on "Confirmational Response Bias"' (1990) 15 *Science, Technology, & Human Values* 46; Mahoney (n 279) 53.

federal research agencies note in their statement on ethical conduct for human-subject research that

> [r]esearch that relies exclusively on information that is publicly available, or made accessible through legislation or regulation, does not require REB [Research Ethics Board] review. Exemption from REB review for research involving information that is legally accessible to the public is based on the presence of a legally designated custodian/steward who protects its privacy and proprietary interests (e.g., an access to information and privacy coordinator or a guardian of Canadian census data).281

It is worth considering freedom of information (FoI) rights as a parallel case. A recent paper by Walby and Luscombe makes three core arguments in favor of not subjecting FoI-based research to ethical review.[282] Firstly, they claim that FoI already involves a bureaucratic vetting process, and only results in data being officially published by governments and redacted as appropriate with respect to national legislation. To extend ethical review to FoI-based research could be considered a form of unwarranted 'ethics creep',[283] where researchers become subject to restrictions on the use of *secondary* data. Data protection rights too have such built-in exemptions. Secondly, they use an analogy to the legal notion of *double jeopardy* to argue that researchers should not be subject to both the process of the 'quasi-ethical' exemptions in FoI law *and* university procedure. Thirdly, they argue that research ethics processes cannot infringe on a citizenship right: universities should not block a researcher's right to know, which in some cases (like New Zealand) is even constitutional in nature. They note a university refusing to push a right to know to its limit additionally could be accused of not carrying out its duty as a knowledge-seeking institution. The fundamental rights nature of access rights in EU law make this additionally convincing in the case of data protection.

Yet there is a significant difference between freedom of information and data protection transparency rights: the former are supposedly subject independent, whereas the latter are most certainly not. This creates ethical challenges that are more unique to data subject rights, to which we now turn.

5.2.2 Privacy of research subjects

Unlike the case argued above for data controllers, in many cases, those undertaking transparency requests — the data subjects themselves — should be treated as human subjects.

If the researcher themselves is gathering information (e.g. that relates to them) with data rights, fewer ethical considerations around the data subject are relevant. For example, a researcher may only need a single response per data controller to answer their research question. They may also be fabricating data subjects, such as through simulating web or app behavior to study tracking. Yet the researcher undertaking data requests alone does not mean that there are no ethical issues relating to data subjects. A helpful parallel is autoethnography — a qualitative research method that uses a researcher's autobiographical experiences as primary data to analyze and interpret the sociocultural meanings of such experiences.[284] In autoethnography, while the subject of the study is ostensibly the researcher, many other individuals are implicated through the stories being told and analyzed.[285] Where data relates to more than one person, these privacy issues may require ethical considerations that cannot be resolved by the data subject–researcher alone .[286]

However, in many of the scenarios illustrated above,[287] we have envisaged recruiting participants to carry out data rights where one of the aims is to contribute to the research project in question. This raises several issues.

While one of the tenets of research ethics is informed consent, information asymmetries in data rights use cases make this challenging.[288] The research team will not always be able to foresee the content or categories of personal data returned to the data subject, posing two main challenges.

The first is that the data subject might discover something that distresses them. There seems little need to pre-emptively protect subjects from dismal revelations about, for example, the sheer extent of online tracking, or reflection on their own experiences through data more generally.[289] Indeed, a call for participation could be structured to make the aim of triggering such experiences clear. However, data often inadvertently relate to more than one person,[290] and may reveal sensitive information that, for example, could create rifts and divisions between families and friends.

The second challenge is that the returned data might be so complex, or rich with potential inferences, that the individual themselves is unable to accurately appraise the sensitivity of what it is they are handing over to researchers. Individuals participating in citizen or participatory science projects do express privacy concerns, but a tendency to focus on 'openness, sharing, and the personal and collective benefits that motivate and accompany participation' can mask these and limit the attention paid to them by coordinating researchers.[291] This is problematic because even 'dull' seeming data framed as part of a significant collective good, such as smart meters in the context of climate change, can be extremely revealing of individuals' lifestyle and preferences.[292] Practices around genetic research indicate some of the challenges when individuals provide extremely potent data about themselves to third parties.[293] Yet in these cases, what genetic

281 Canadian Institutes of Health Research and others, Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (2014) 16.
282 Kevin Walby and Alex Luscombe, 'Ethics Review and Freedom of Information Requests in Qualitative Research' (2018) 14 *Research Ethics* 1.
283 Kevin D Haggerty, 'Ethics Creep: Governing Social Science Research in the Name of Ethics' (2004) 27 *Qualitative Sociology* 391.
284 Heewon Chang, 'Autoethnography in Health Research: Growing Pains?' (2016) 26 *Qual Health Res* 443, 444.
285 See generally on the ethics of autoethnography Martin Tolich, 'A Critique of Current Practice: Ten Foundational Guidelines for Autoethnographers' (2010) 20 *Qualitative Health Research* 1599; Anita Gibbs, 'Ethical Issues When Undertaking Autoethnographic Research with Families' in *The SAGE Handbook of Qualitative Research Ethics* (SAGE Publications Ltd 2018).
286 See generally on the entangled nature of privacy Solon Barocas and Karen Levy, 'Privacy Dependencies' [2019] 95 *Washington Law Review* 555.
287 See supra section 4.
288 See on the overlap with data protection law: European Data Protection Supervisor (n 22) 18 et seq.
289 See generally Petr Slovák and others, 'Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection' in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17, New York, NY, USA, ACM 2017).
290 This is referred to as a bycatch by Barocas and Levy (n 286).
291 Anne Bowser and others, 'Accounting for Privacy in Citizen Science: Ethical Research in a Context of Openness' in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17, New York, NY, USA, ACM 2017) 2134.
292 Ian Brown, 'Britain's Smart Meter Programme: A Case Study in Privacy by Design' (2014) 28 *International Review of Law, Computers & Technology* 172; Michael Veale, Data Management and Use: Case Studies of Technologies and Governance (The Royal Society and the British Academy 2017).
293 See generally, concerning the list of findings that researchers should report in the United States by way of a voluntary code, Sarah S Kalia

diagnosis can potentially do, can be communicated to research subjects better than what a (personal) dataset of unknown variables of unknown extent might reveal.

In both these cases, the genetic analogy lends an important structural finding that may ameliorate concerns. This field has emphasized 'a duty on the part of a research investigator to consider what incidental and secondary results might occur from genomic testing, to create a plan for the possible return of results to participants, and to inform research participants of that plan before the tests are conducted'.[294] In data rights, a similar plan should be made clear. In the case where only a small subset of data would ever be needed and analyzed, a strict plan should be made to discard the rest as soon as possible, either before it leaves the research subject's control, or as soon as possible after if separation is technically challenging. If the aim is for the research subject to explore the data themselves, researchers should be aware of the potential for findings about e.g. others in the datasets that may concern or alarm the researcher and prepare the data subject accordingly. Particular care should be taken if the researchers are to ask open-ended questions of potentially large and unknown datasets provided by research subjects, and situations where researchers do this on their own without supervision or guidance from research subjects may be best avoided unless there is a very clear and justified reason to do so.

Apart from this, researchers should of course also comply with relevant legal protections, including the GDPR, that are aimed at safeguarding research subjects' privacy. This holds particularly true for key data protection principles such as 'purpose limitation', 'data minimisation', 'storage limitation', 'integrity and confidentiality'.[295] As emphasized by the EDPB, 'the principles of necessity and proportionality are essential' and it will not be sufficient for researchers to simply claim that the processing of personal data is 'necessary for the purposes of scientific research'.[296] Important here is that 'informed consent' in research ethics should be distinguished from research/data subjects consenting to the processing of their personal data (consent being one out of six grounds for rendering the processing of personal data lawful).[297] Indeed, in some situations there might be a clear imbalance between data subjects and the controller/researcher (e.g. because of the scale of the research project and/or how invested the research/data subject may be), which would challenge the GDPR-requirement for consent to be *freely given*.[298] Put briefly, to the extent researchers plan on receiving personal data of their participants, they will have to give due regard to data protection law. In this regard, it is worth referring to the European Commission's plans to propose a data governance legal framework that would also include rules to 'facilitate decisions on which data can be used, how and by whom for scientific research purposes in a manner compliant with the GDPR.'[299]

5.2.3 Risk of retribution

A last risk, which does not have considerable legal support but which may nevertheless pose ethical risks for data/research subjects is the possibility of some retribution by a data controller. Prior research into data rights has highlighted the tendency of some data controllers, for example, to respond to access requests as if they were erasure requests, presumably to avoid regulatory burden of troublesome data subjects.[300] This indicates a security risk that is posed to data subjects in relation to the availability of the data in the systems they are being asked to query. Deleting data before access has been provided may be considered a violation of the GDPR (notably the fairness, lawfulness and integrity principles in Article 5(1)), subject to considerable fines and even criminal prosecution in some countries.[301] However, in some cases it is also possible that the data controller or their agents are personally known to the research subject: for example, in the case of previous employers or medical practitioners. Considerations must be given to the social repercussions of requesting research subjects to use rights against controllers such as these.

5.2.4 Relationship to enforcement action

As data controllers are often responding to data rights in ways that do not seem compliant with the law,[302] researchers may feel they should work with research subjects to author complaints to data protection authorities to ensure the law is properly upheld. Given the overburdened and under-resourced nature of many authorities,[303] we feel this move should be supported in general as researchers will often be very well placed to explain breaches in detail and clarify important technical issues to the regulators. However, this does raise a challenge when research subjects are involved, as while a complaint seems like a simple form, in many jurisdictions it can open a legal process with the research subject as a party. While the research subject should not be put under any legal liability as a result, there is a small possibility they may be asked to eventually be party or intervenor to a legal case that could occur, such as an appeal against the decision of a supervisory authority. If this is undertaken, the potential role of the research subject going forward should be made clear, and while researchers may wish to provide the means and support for a research subject to complain, they should emphasize that this aspect should be considered an activity independent of the research project.

5.2.5 Broader ethical issues

None of this is to suggest that research questions themselves cannot bring ethical issues that are not well characterised by privacy concerns. A mass data access campaign to access and utilise biometric data to create facial recognition systems, for example, can bring ethical questions regardless of individual data subjects' consent. These are out of scope of this paper, which focusses on issues more specific to researching through data rights.

and others, 'Recommendations for Reporting of Secondary Findings in Clinical Exome and Genome Sequencing, 2016 Update (ACMG SF v2.0): A Policy Statement of the American College of Medical Genetics and Genomics' (02 2017) 19 *Genet Med* 249.

294 Christine Weiner, 'Anticipate and Communicate: Ethical Management of Incidental and Secondary Findings in the Clinical, Research, and Direct-to-Consumer Contexts (December 2013 Report of the Presidential Commission for the Study of Bioethical Issues)' (2014) 180 *Am J Epidemiol* 562.

295 Article 5 GDPR.

296 European Data Protection Supervisor (n 22) 11–12, 16.

297 Article 6(1) GDPR

298 European Data Protection Supervisor (n 22) 18.

299 European Commission, 'A European strategy for data' (n 6) 12–13.

300 Ausloos and Dewitte (n 77).

301 In the UK, for example, it is considered a criminal offence 'to alter, deface, block, erase, destroy or conceal information with the intention of preventing disclosure of all or part of the information that the person making the request would have been entitled to receive'. Data Protection Act 2018 s 173(3).

302 Ausloos and Dewitte (n 77); Mahieu and others (n 95); Janis Wong and Tristan Henderson, 'The Right to Data Portability in Practice: Exploring the Implications of the Technologically Neutral GDPR' [2019] *International Data Privacy Law*.

303 See generally European Data Protection Board, 'First Overview on the Implementation of the GDPR and the Roles and Means of the National Supervisory Authorities' (Report presented to the European Parliament's Civil Liberties, Justice and Home Affairs Committee (LIBE), 26 February 2019).

5.3 Methodological considerations

5.3.1 Integrity of research

Certain uses of data rights might struggle for methodological validity when assessed in a strictly quantitative frame. In particular, some scholars advance a quantitative approach as a general template for conducting research with inferential, empirical validity in both quantitative and qualitative projects.[304] One characteristic result of this logic is the advice that increasing the number of records (assuming they are sampled in a random manner) will increase inferential leverage. For *ex post* data rights especially, this can be challenging, as uptake of the use of rights in a particular study might be limited, both in a general sense and among specific subgroups. According to a classic quantitative view, this might mean that the sample may be insufficiently large or representative to draw generalizable statistical conclusions from.

These problems mainly arise, however, when we confuse data rights and their potential with 'Big Data' research. The logic of research over large datasets made available through the digital economy, such as scraped Web data or global search patterns,[305] is that even data not collected for a particular purpose might reveal important societal phenomena due to the number of subjects and the richness of collected data. As *ex post* data rights require manual effort, they are not akin to this type of research, but more akin to citizen or participatory science. This field has well-known effort and participation biases, such as oversampling on weekends[306] or in certain areas[307] which researchers actively work to compensate.[308]

This indicates that data rights are more useful when certain characteristics of a research program are met. Studies that are considering small, well defined populations are apt for data rights. If participants were always going to be enlisted and worked with directly, and perhaps compensated for their time, then many of the biases simply reduce down to the classic representativeness challenges in fields such as psychology. If an attempt is made to generalize from a small sample to the world, significant challenges exist, such as capturing phenomenon as they manifest in easily accessed 'convenience samples' of participants,[309] such as students on campus,[310] which may differ from the world more generally. However, if the aim is to study exactly that type of student, this poses little problem.

If there is pre-existing reason to believe that a phenomenon will be homogeneous across populations, then data rights may also be appropriate. If the aim is, for example, to study how web tracking systems work online, these remain the same between individuals, although the websites sampled and technologies (such as tracker blockers) used may differ. In this situation, researchers are a gateway into a homogenous phenomenon, such as policy or infrastructure. Where this becomes challenging is where the aspect of infrastructure observed is heavily contingent on the data subject, as German credit scoring reverse-engineering effort OpenSCHUFA discussed above[311] found when it was unable to study issues such as discrimination due to a bias in white, male volunteers. OpenSCHUFA reflected that they 'were not able to get the attention of demographic groups that are probably most affected by poor SCHUFA scores' and as a result it was difficult to make generalizable conclusions, or understand all parts of the system.[312]

Statistical and methodological challenges around data rights must also be seen in the context of the pitfalls and biases in 'Big Data' research about the digital economy[313] — and data rights can potentially help provide alternative datasets as a check on these biases for the same types of phenomena — for example, for focusing on obtaining data about certain difficult to identify populations and communities that may be underserved or underrepresented in 'Big Data' held either by firms or obtained through other methods by external researchers.

5.3.2 Interactional considerations

Data requests can be made directly by the data subject or indirectly by an individual or organization mandated by a data subject. The latter option, however, can present difficulties as data controllers are concerned around releasing data to individuals pretending to be the data subject.[314] Individuals having been given demonstrable power of attorney are unlikely in practice to see problems of authentication,[315] but other agents, such as researchers, may be refused or requested for specific information to aid verification which only the data subject can provide. The data may also be provided to the data subject for sending on further to the third party again, necessitating a significant back-and-forth. We leave detailed legal analysis of mandating data rights to third parties to future work, but note that this is a challenging area, and in the absence of clear judicial clarification, it seems unlikely that controllers will adopt a consistent approach broadly necessary for research.

If rights are not to be delegated to a third party, it will be up to data subjects to interact with the data controller and obtain the necessary data, and to make all or relevant portions of that data available for research. This is easier said than done, as interaction with these controllers can take many different forms along a spectrum of collaborative to adversarial. In some cases, adversarial approaches

304 e.g., Gary King and others, *Designing Social Inquiry* (Princeton University Press 1994).

305 e.g., Shihao Yang and others, 'Accurate Estimation of Influenza Epidemics Using Google Search Data via ARGO' (2015) 112 *Proceedings of the National Academy of Sciences* 14473.

306 Jason R Courter and others, 'Weekend Bias in Citizen Science Data Reporting: Implications for Phenology Studies' (2013) 57 *Int J Biometeorol* 715.

307 Yexiang Xue and others, 'Avicaching: A Two Stage Game for Bias Reduction in Citizen Science' in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (AAMAS '16, Richland, SC, International Foundation for Autonomous Agents and Multiagent Systems 2016).

308 e.g., Chankyung Pak and others, 'Auditing Algorithms With Donated Data: Methods for Poor Scientists?' (*ICA*, Virtual, 20–[26]May 2020).

309 Robert A Peterson and Dwight R Merunka, 'Convenience Samples of College Students and Research Reproducibility' (2014) 67 *Journal of Business Research* 1035.

310 e.g., Patricia M Greenfield, 'Sociodemographic Differences Within Countries Produce Variable Cultural Values' (2014) 45 *Journal of Cross-Cultural Psychology* 37 (arguing that the difference between student populations from different socioeconomic backgrounds can be larger than cultural differences between countries); Paul HP Hanel and Katia C Vione, 'Do Student Samples Provide an Accurate Estimate of the General Public?' (2016) 11 *PLoS One* (arguing that different student populations significantly differ from the general public in ways that are difficult to explain).

311 See supra section 4.2.1.

312 'OpenSCHUFA' (OpenSchufa, no date) https://openschufa.de/english (accessed 24 June 2019).

313 Olteanu and others (n 31).

314 See generally Coline Boniface and others, 'Security Analysis of Subject Access Request Procedures How to Authenticate Data Subjects Safely When They Request for Their Data' [2019] *Annual Privacy Forum*, Jun 2019, Rome, Italy; Cormack (n 255).

315 See eg Information Commissioner's Office, 'Right of Access' (Guide to the GDPR, 12 August 2019) https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-of-access (accessed 1 December 2019).

may be required as data controllers are unwilling to provide data they are required to by law. It may be possible for the research team to pre-empt and avoid these adversarial encounters by testing the process for relevant controllers before the research begins, allowing identification of any hurdles, the enlistment of the local data protection authority if required,[316] and the creation of both a tailored request and pre-built responses that are suitable for the particular issues and views of the controller in question. Researchers will have to consider participants' skills when crowd-sourcing data-gathering using the GDPR. This can be dealt with to some extent, by providing explanations, personal or technical assistance and tools.[317]

In some cases however, the research project may have to be postponed while enforcement or legal action can be carried out.[318] On the more collaborative side of the spectrum, one could imagine company and researchers agreeing to include a specific tag in participants' access requests so that they are prioritized and/or responded to in a predefined format. Researchers may also simply rely on available 'download my data' functionalities already offered by many online services, which currently only generally provide a fraction of eligible data, but which may be suitable for the research question.

Information may be provided in a variety of ways, such as files through secure drop facilities, as email attachments with or without passwords, or on physical media (particularly for data outside of the digital economy such as CCTV footage). The research team must prepare for these different formats and create a secure, suitable and ideally easy-to-use way for data subjects to grant access to this data. There may be an important role for the researchers to carry out an initial request to create more bespoke guidance of what to expect from a data controller. Relatedly, the research team should also make efforts to ensure that data subjects are not storing this data in insecure ways, and advise them on the correct storage or disposal if appropriate.

6 Conclusion

The concentration and privatization of data infrastructures, turns (mainly big technology) companies into *de facto* gatekeepers of research agendas. Independent researchers have developed a wide variety of approaches in order to pierce through enclosed datasets, each with their benefits and drawbacks. This article outlines a fairly new approach to add to researchers' toolset for obtaining relevant research data (Section 3). Compared to other tools, data rights under the GDPR have the advantage of being potent (legally enforceable) and enabling access to very fine-grained data (Section 4). That being said, they also raise a number of (legal, ethical and methodological) issues whose significance will vary depending on the actual research projects (Section 5).

Given the multi-faceted nature of using data rights outlined throughout this paper, it is not possible to outline a detailed procedure or plan that would fit each potential research project. That said, the seven steps identified at the start of section 4 may serve as a useful starting point for researchers interested in using data rights in their project.[319] The research team should reflect upon the process in the context of methodological, ethical, legal and data security and protection challenges described in Section 5. Such analysis will depend in large part on national and local processes specific to different countries, university systems or funders. Methodological issues will be largely discipline-specific, and cross-cutting guidance cannot be easily provided linking this specific data collection approach to the broad and welcome array of potential analysis techniques.

In conclusion, using data rights requires a triangle of expertise – domain, technical and legal – the constellation of which may vary from one research project to another. Any research project will of course rely on adequate *domain expertise* relating to the actual research questions. Data rights in particular require a minimum level of *legal expertise* to properly identify the opportunities and limitations, as well as manage the interaction strategy. Finally, *technical expertise* may be necessary in order to understand and process the data received.

***

Researching with data rights is still at a very early stage. Our aim with this article was both to explain the potential utility of data rights to researchers, as well as to provide appropriate initial conceptual scaffolding for important discussions around the approach to occur. We do not claim to have exhausted either the possibilities or the challenges of using the transparency provisions in data protection law for research, and offer only a non-exhaustive tour through some of the issues and questions that might arise. Data rights may not be the right tool for every job, but there are many investigations of data and power in particular that remain open. Data protection is a flexible instrument designed to address asymmetries of informational power, and we believe researchers should be at the forefront of finding new ways to use that flexibility for societally critical knowledge generation.

Acknowledgements

This article underwent countless alterations and is the product of many discussions. We would like to thank everyone that has provided us with their feedback in many different forms. We are particularly grateful for the engaged audience at TILTing Perspectives (Tilburg, 2019), as well as the rich conversation, led by Joshua Kroll, at the the Privacy Law Scholars Conference where an earlier version of this paper was workshopped (Berkeley, June 2019). We also want to thank the TechgReg reviewers for their incredibly fast and thoughtful feedback and suggestions.

316 e.g., Johnny Ryan, 'Regulatory Complaint Concerning Massive, Web-Wide Data Breach by Google and Other "Ad Tech" Companies under Europe's GDPR' (*Brave Browser*, 9 December 2018) https://www.brave.com/blog/adtech-data-breach-complaint (accessed 1 May 2019); 'Our Complaints against Acxiom, Criteo, Equifax, Experian, Oracle, Quantcast, Tapad' (*Privacy International*, no date) http://privacyinternational.org/advocacy-briefing/2426/our-complaints-against-acxiom-criteo-equifax-experian-oracle-quantcast-tapad (accessed 8 April 2019).

317 One of the authors has undertaken several types of research set-ups, interacting with subjects in different ways. Unsurprisingly, the project with only a limited number (3) of law students, with bi-weekly follow-up calls, appeared the most successful.

318 See the Uber references in n 14.

319 Aim > Data > Legal Approach > Scope > Recruitment Strategy > Interaction Strategy > Data Analysis Strategy.