UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

The Many Faces of Data Access

Legal and Policy Implications for Research

Ausloos, J.; de Souza, Siddharth Peter

Publication date
2026

Document Version
Final published version

License
CC BY-NC-SA

Link to publication

Citation for published version (APA):
Ausloos, J., & de Souza, S. P. (Eds.) (2026). *The Many Faces of Data Access: Legal and Policy Implications for Research* . Institute of Network Cultures. https://networkcultures.org/blog/publication/themanyfaces/

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

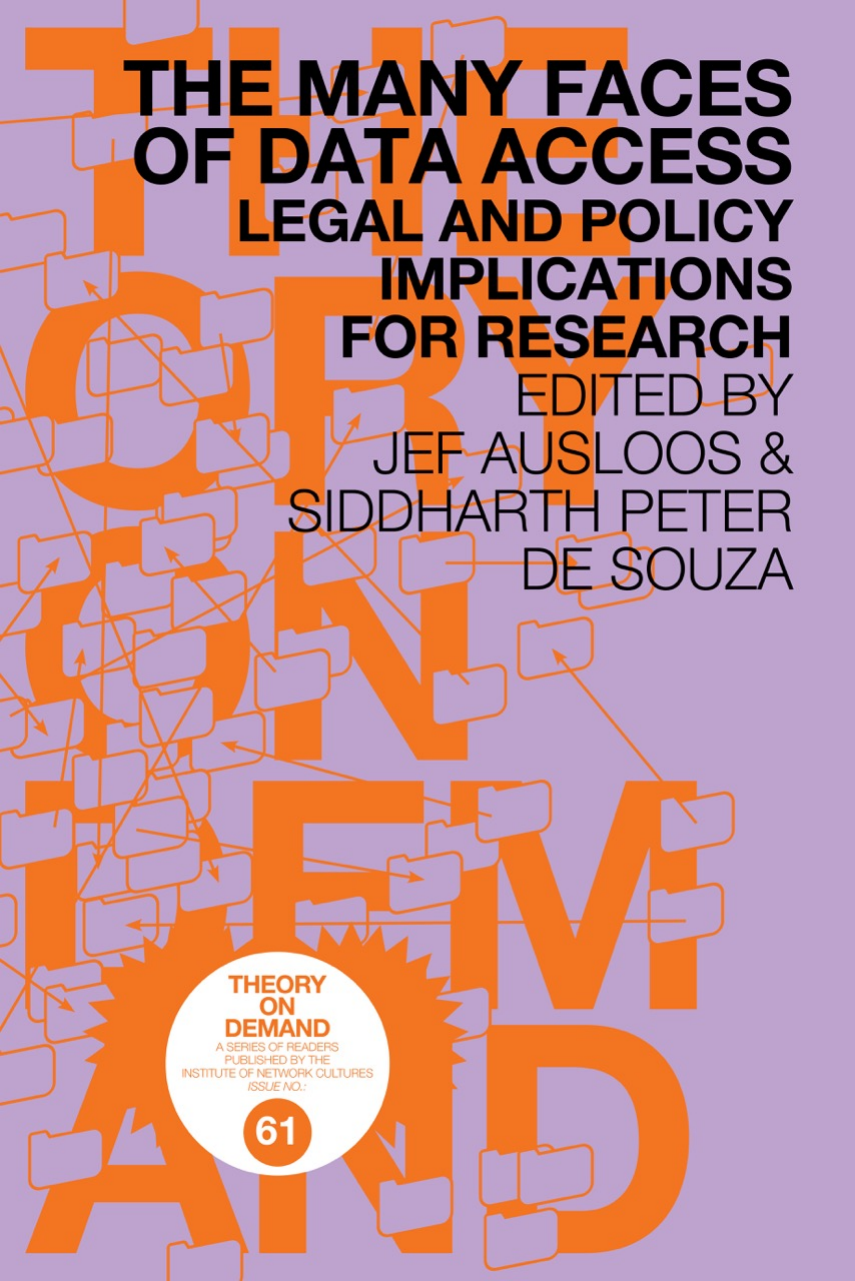
UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:08 Sep 2026

THE MANY FACES OF DATA ACCESS

LEGAL AND POLICY IMPLICATIONS FOR RESEARCH

EDITED BY
JEF AUSLOOS &
SIDDHARTH PETER
DE SOUZA



THE MANY FACES OF DATA ACCESS

LEGAL AND POLICY IMPLICATIONS FOR RESEARCH

Theory on Demand #61
The Many Faces of Data Access: Legal and Policy Implications for Research

Edited by: Jef Ausloos & Siddharth de Souza

Authors: Carolina Aguerre, Frank Kwaku Agyei, Pedro Amaral, Jef Ausloos, Reuben Binns, Lawrence Kwabena Brobbey, Siddharth Peter de Souza, Paul Esselaar, Michalina Kowala, Boateng Kyereh, Matteo Nebbiai, Midas Nouwens, Paul Osei-Tutu, Marcos César M. Pereira, André Ramiro, Jake Stein

Peer review: All chapters in this book have been subject to two rounds of (single-blind) peer reviews, two rounds of editorial reviews, and a writing workshop

Cover Design: Katja van Stiphout

Design and EPUB development: Klaudia Orczykowska

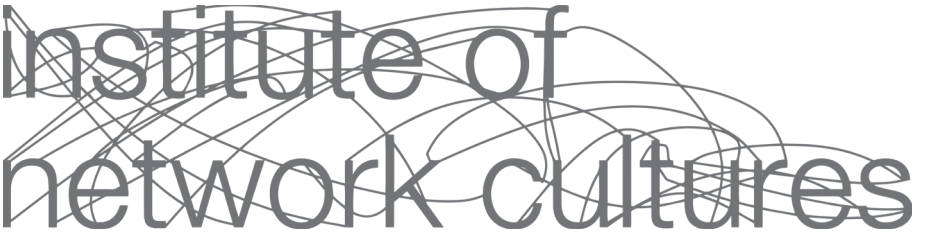
Published by the Institute of Network Cultures, Amsterdam, 2025

ISBN: 9789083672113

Contact Institute of Network Cultures Amsterdam University of Applied Sciences (HvA)
Email: info@networkcultures.org Web: www.networkcultures.org

Order a copy or download this publication for free at: www.networkcultures.org/publications

This publication is licensed under the Creative Commons Attribution NonCommerical ShareAlike 4.0 Unported (CC BY-NC-SA 4.0). To view a copy of this license, visit www.creativecommons.org/licences/by-nc-sa/4.0./



CONTENTS

ACKNOWLEDGMENTS 3

1. DATA ACCESS FOR RESEARCH: IMAGINATIONS, LIMITATIONS AND PROMISES
JEF AUSLOOS AND SIDDHARTH PETER DE SOUZA
7

PART I: IMAGINATIONS 27

2. RE-CONCEPTUALIZING GOVERNANCE POLICIES ON DATA ACCESS FOR RESEARCH
CAROLINA AGUERRE
30

3. VIOLENT PLAINS: CHALLENGES AND STRATEGIES FOR PASTORALISTS' DATA ACCESS IN GHANA
FRANK KWAKU AGYEI, LAWRENCE KWABENA BROBBEY, PAUL OSEI-TUTU, AND BOATENG KYEREH
51

4. FROM RIGHTS TO SKILLS: DATA ACCESS FOR TEACHING DATA LITERACY
MIDAS NOUWENS
67

PART II: LIMITATIONS 91

5. KEYS THROWN AWAY? CHALLENGES IN BRAZIL ON ACCESSING PUBLIC-INTEREST DATA ON STATE SURVEILLANCE TOOLS VIA TRANSPARENCY PORTALS AND REQUESTS FOR INFORMATION
ANDRÉ RAMIRO, PEDRO AMARAL, AND MARCOS CÉSAR M. PEREIRA
92

6. DIGGING INTO EU DATA LAWS AND THEIR IMPACT ON AFRICAN RESEARCHERS
PAUL ESSELAAR
114

PART III: PROMISES 136

7. A SUBJECT ACCESS REQUEST, THEN WHAT?: (UN)STRUCTURING ONLINE ANALYTICS FOR DATA INSTITUTIONS
JAKE STEIN AND REUBEN BINNS
137

8. DATA INTERMEDIARIES FOR GOOD: CAN DATA INTERMEDIATION SERVICES HELP DATA ACCESS IN RESEARCH?
MATTEO NEBBIAI
159

9. ACCESS TO DATA ON DISINFORMATION WITHIN THE CODE OF PRACTICE ON DISINFORMATION
MICHALINA KOWALA
184

ABOUT THE CONTRIBUTORS 202

FIGURES

7.1	Schema, field, and data levels of data structure	156
7.2	The nested logics encoded in DSAR data	161
7.3	Illustration of parallax	165
8.1	Country of establishment of DISs that have been accessible in the territory of the EU and the UK between 2000 and 2022	179
8.2	Frequencies of types and data specialization of DIS	180
8.3	Share of data intermediation services possessing research openness features	182
8.4	Number of research openness features among the DISs	183
8.5	DISs' research openness features disaggregated for DIS type and data sector	184

TABLES

2.1	OS and AI strategies: summary	43
3.1	Category and sample of respondents used for the study	57
4.1	Terms and definitions related to data literacy	75
4.2	A simplified overview of data access rights in the EU	78
8.1	Features used to assess the DISs' research openness	181
8A.1	Dataset of data intermediation services accessible in the territory of the EU and the UK, 2000–2022	188
8A.2	Draft list of features used to assess the DISs' research openness	190

ACKNOWLEDGMENTS

We would like to thank authors and participants from a workshop at the Institute for Information Law, University of Amsterdam, in March 2023, where early drafts of the chapters in this volume were discussed. In particular, we wish to thank Max van Drunen, Paddy Leerssen, Arlette Meiring, and Stefanie Boss for their collaboration in organizing the workshop, and Priyanka Das and Klaudia Orczykowska for their support with the editing of the book. We are also grateful to the University of Amsterdam and the Institute for Information Law for their support of the workshop.

1. DATA ACCESS FOR RESEARCH: IMAGINATIONS, LIMITATIONS AND PROMISES

JEF AUSLOOS AND SIDDHARTH PETER DE SOUZA

Throughout the last decade, concerns have been growing over the (lack of) transparency of digital infrastructures. While these infrastructures have rapidly nested themselves into every part of society – from social and professional contexts to industrial and bureaucratic processes – their inner workings have remained shrouded in secrecy to outside viewers. This fundamentally challenges the ability to observe, and understand, the world around us, whether it be for public scrutiny and accountability purposes or for different kinds of knowledge production more broadly. Indeed, independent researchers – from academia, investigative journalism, and civil society – have decried the many legal, technical, financial, and other obstacles preventing them from observing and scrutinizing the data, algorithms, and general operation of these digital infrastructures.

Over the years, a wide variety of measures have been proposed and tested to overcome the obstacles faced. These range from methodological tools developed by the research community (for example, web scraping) to self-regulatory initiatives (for example, data philanthropy) and policy proposals (for example, Article 40 in the European Union's [EU] Digital Services Act).[1] Regulators across the world have recently stepped up, proposing legal frameworks targeting a range of issues raised by digital infrastructures. The EU regulator in particular has put forward a barrage of legislation purported to boost transparency and facilitate data access.[2] These proposals are primarily conceived in light of internal market goals, often disregarding alternative purposes and interests: from holding different sources of power to account to diverse forms of knowledge production in a digitally intermediated world. Data access for independent research – the central theme in this volume – only features in the margins of these new frameworks, if at all.

The explosion of new and proposed legal frameworks can also lead to regulatory inflation, which is the increase in regulatory requirements, overwhelming under-resourced and under-represented interest groups in trying to make sense of transparency and access provisions. More often than not, overambitious legislative action may effectively result in a denial-of-service attack, hijacking the limited resources, capacity, and imaginaries of change of those they purport to empower. As such, while ostensibly aiming to reign in the power of digital infrastructures, these new legal frameworks when combined may further solidify the power of those with the resources to manage the operationalization of the law.

This tumultuous landscape of multifarious legislative action along with the political economy of digital infrastructures more broadly also calls for a reappraisal of the need and conditions for independent research in the first place. Transparency and data access for independent researchers may seem laudable goals in the abstract, but to what end? What and whom do these debates obscure? What are, or should be, the goals and interests of (academic) researchers demanding access to digital infrastructures? In short, there is a clear need for critical and global reflection on how new data access rules shape research agendas but also on the positionality of different actors in these debates.

This is why, in March 2023, we organized a small-scale workshop at the Institute for Information Law, University of Amsterdam. The explicit goal was to bring together a group of academics from different backgrounds and provide a platform for critically exploring key issues relating to the regulation of access to data for research. Around 30 participants were selected based on a call for participation, with particular attention to geographical, disciplinary, and career-level diversity. To facilitate this, we were lucky enough to secure institutional funding for covering travel expenses of participants from five different continents.

Participants were invited to submit draft papers presenting their ideas and provocations on a range of themes and questions. These were circulated beforehand and discussed in depth during the workshop. During these discussions, we debated issues such as the opportunities and challenges of the law as a tool for observing digital infrastructures; how data access regimes may privilege certain geographical and institutional contexts; and how it may affect different kinds of dependencies. We also discussed work that critically looks at academia's data production, use, and ownership, as well as the political economy of data access for research and the multifarious power dynamics between academia, private/ public sector, and civil society.

Based on these conversations and further written feedback on their work, authors developed their papers into the chapters presented in this volume. We rejoice in the fact that these chapters reflect the exploratory ambitions we had for the workshop, welcoming many different perspectives beyond dominant legalese or Eurocentric narratives. It is precisely by combining these different scopes and angles that we hope to challenge readers in their thinking and expand their imaginations on the topic of (researcher) data access to digital infrastructures. The chapters in this volume cover legal, technical, political, didactic, and social questions, at different jurisdictional levels and deploying different methodologies, from doctrinal and historical work to empirical, auto ethnographic, and participatory action research.

As academics are increasingly demanding for data access to digital infrastructures, and a growing number of data access regimes are seeing the light of day, it is important to keep asking big and critical questions: What is the agenda-setting power afforded by data access regimes? Whom are these regimes for, or whom should they be for in light of the politics of knowledge production? How do they nurture and/or challenge existing structures of power in academia, politics, and industry? While this volume certainly does not purport to answer these vital questions, we hope that it does reveal the richness of the debate, prompt new questions, and inspire to challenge existing paradigms.

Before getting to the different contributions, we want to use this introductory chapter to explore – what we consider – a few pivotal concepts in the debate on researcher data access to digital infrastructures. Specifically, we look at the notion of 'data access' and its different epistemologies. Building on this, we also consolidate different critical perspectives on researcher data access into the notion of the 'academic data gaze'. This, finally, will also allow us to uncover some of the underlying power dynamics at play in these issues and debates.

All in all, we want this chapter to invite the academic community to self-reflect on critical questions brought to the surface in the researcher data access debate. Pluralizing what we understand data to be, data to do, and data access to include; the role of academia as agents to challenge data-driven power; as well as academia's own complicity in allowing structures of power. Put differently, we cannot be content with taking data access provisions and academic claims to data at face value but need to consider their underlying logics and politics.[3] This is particularly necessary because academic data access is a function of different understandings of data, and also because the avenues to access are unequal depending on resources, technical capacity, as well as institutional protections. As academics therefore, it is important to have 'awareness of one's own epistemological commitments', as well as reflect and acknowledge that one's epistemic choices have effects on others.[4] This volume is therefore an opportunity to unpack tensions that emerge in securing data access for researchers, while also being conscious that the contexts within which this work is produced, and our conferences are organized, remain places of immense privilege.

The Epistemic Questions behind Data Access

The intersections between data, access, and knowledge production become important starting points to be able to unpack data access for research. To begin with, in thinking about data, we wish to ascertain the various characteristics, imaginaries, and values that influence how data is understood and regulated.[5] If one is to look at the business models of companies and their digital infrastructures, the regulations emerging from transnational organizations such as the World Trade Organization or at a regional level the EU regulation on data governance, it becomes clear that data is conceived to be an economic commodity – one that can be brokered, traded, and used as a resource to extract value.[6] The notion of 'data as the new oil' has resulted in regulation across the world taking an approach which examines how best to regulate the 'data market'.[7] This approach aims to ensure that regulation of data does not restrict innovation and fosters competition and growth of the tech industry. Such an approach, while being by far the more prevalent, is not the only notion of data that is relevant to pay attention to.

For instance, if one is to look at the work of indigenous data sovereignty experts, then a key characteristic of data is that it is living.[8] This entails that data plays an important role in determining the personhood as well as the embodied experiences that people have. Data is something that is central to a person's autonomy and sense of respect.[9] This emphasis on people's agency confronts motivations of open data, where there is an emphasis on data sharing without adequate reflection on the power structures and historical conditions that underpin how data sharing and exchange has been constructed.[10]

That data is not a resource that can be commodified is powerfully explained through the work of We Are Not Numbers, a Gazan youth group. The world often speaks about Palestinians in terms of numbers – 'specifically, how many killed, injured, homeless and/or dependent on aid. But numbers are impersonal and often numbing. What they don't convey are the daily personal struggles and triumphs, the tears and the laughter, and the aspirations that are so universal that if it weren't for the context, they would immediately resonate with virtually

everyone.'[11] This distinction is important because it situates how treating data as a resource detaches personhood and dehumanizes what is represented in that data into an objective, commensurable fact, devoid of nuance and context.[12]

Another understanding of data is that it is not something that just exists in nature,[13] but it is a consequence of labour that is put in by people, oftentimes in exploitative scenarios. We have seen recently how corporations around the world are exploiting economic conditions to find the cheapest labour to perform 'data work' such as in relation to labelling, curating, and sorting of data.[14] Such workers are paid measly amounts of money and often do not have safe working conditions when they deal with data about violent circumstances, as has been seen in a recent case of OpenAI in Kenya.[15] Amazon Mechanical Turk is an example of a crowdsourcing platform where companies are accessing digital labour at low costs without any obligations towards worker security and social protection.[16]

Data also does not just affect the individual person but also systematizes relations between people and/or objects and digital infrastructures.[17] As Salome Viljoen has argued, '[B]y engaging with datafied systems we place ourselves in relationships to others, which then make it possible to manipulate or monetise both parties, and the relationship itself.'[18] These relations can be between peers at the horizontal level but also at a vertical level between institutions and people. Therefore, if one is to think about data, can one also consider how data does not just affect individual autonomy but also creates social inequality?

This brings us to the question of thinking about the concept of access.[19] If we are interested in thinking about access to data, it becomes important to unpack alternative visions of what access to such data can mean.[20] To make access not just something that is provided but something that is also meaningful, a useful starting point is to consider the pathways to access that people take.[21] Aligned to this is to consider that there are different reasons for which data access is required. Whereas we have discussed different meanings of data, we also argue that it is important to ascertain – data access for what?

The motivation(s) behind access can result in different manifestations, such as 'public access' where it is meant for consumption by the public and in the public interest; 'regulator access' where institutions require it in order to fulfil a bureaucratic function; 'research access' where it is necessary to be able to study the effects of digitally intermediated phenomenon; and 'corporate access' for pursuing of economic objectives. Looking at these different actors demonstrates how varied the purposes of data access can be.[22] At one level, an obvious intention of creating better data access is to ensure that there is greater transparency and accountability, notably of online platforms, for instance, such that one can understand the ways in which they function and operate. But at another level, the purposes are much broader, and entirely subjective, depending on who the person claiming access is. In this way, it is relational, and contextual, and emerges within the location of those making claims.[23]

In the context of this volume, we are interested in exploring data access for research in particular. Research is often defined as a systematic and scientific exercise to study phenomena. It results in the creation of new understandings of existing knowledge or in the creation of new knowledge on its own. The role of a researcher in carrying out such an enquiry is often bound by the constraints of a particular discipline, where scientific rigour becomes a necessary condition to what research is seen as appropriate or valid. However, as decolonial scholars have repeatedly argued, the construction of who a researcher is, and what is valid research, is a function of the political economy of knowledge and knowledge-producing institutions.[24] For several years, the conditions of successful research are embedded in a Eurocentrism, where an othering takes places of people and contexts outside the core of the West.[25] This is an important consideration to keep in mind, even while we see the emergence of new regulations to facilitate data access to research digital infrastructures.

In thinking about the effects of data access on knowledge production, it is necessary to examine the bundle of rights that are affected because of a denial of data access.[26] For instance, constraints to data access not only limit one's capacity to know and understand but it also affect the capacity and ability to participate and act. This is because data becomes a currency to make things knowable and countable. It is used as the predominant medium to describe phenomena, and thereby to make things commensurable.[27] Not engaging through this language therefore limits the agency and autonomy of people, whether individuals or groups, to take actions which are in their interest.

A second aspect in relation to data access is to examine who has the capacity to make sense of the data once access is provided. For access to be meaningful, it also requires the technical know-how to be able to make sense of the data. Consequently, we see an emergence of power hierarchies in terms of not only the form of data that is being made available but also the emergence of those that then act as interpreters, interlocutors, and potentially gatekeepers of the knowledge of the data. In this regard, those that have data access become critical in terms of not just holding knowledge but also determining the governance of data.[28]

A third consequence of such an approach is about narratives around data.[29] This means that how they are interpreted, deployed, and actioned will depend on who controls not only their access but also their interpretation. As a consequence, it becomes imperative to examine who draws the 'abyssal line' between what is considered important, critical, and valid and what is consigned to an abyss and deemed irrelevant.[30]

The Academic Data Gaze

The debate on (academic) researchers' claim to data in, and about, digital infrastructures reveals deeper questions on the politics and economy of knowledge production. Underlying the growing efforts to improve researchers' access to data in the digital society is a strong normative assumption that such access (and data) is generally beneficial. While we certainly sympathize with many such efforts, we believe it is important to critically engage with these underlying assumptions of the impacts of data access by also locating it in the political economy around it. Failing to do so may render academia complicit in a variety of problematic dimensions of digital technology and, in particular, its extractivist practices. Moreover, it may amplify a number of issues already inherent to dominant academic research practices, including the Matthew effect. For the purposes of this introductory chapter, we try to condense some of these issues into the notion of the 'academic data gaze'.

The academic data gaze can be situated along a long trajectory of the hegemony of mainly Western academic research and its underlying logics. The term builds on the concept of the *data gaze* as theorized by David Beer: 'a concept that targets an understanding of the connections, structures, and performances of power within [data] analytics.… This data gaze, and the discourse that facilitates and informs it, is suggestive of how lives are viewed differently through data – in ever more forensic, strategic, predictive, and knowing ways.'[31] Importantly, we believe the data gaze is fundamentally intertwined with modern-day forms of (for example, platform,[32] informational,[33] and surveillance[34]) capitalism and its deeply extractive logics. It acknowledges that data is not a natural element to be observed, but always artificially produced for specific goals in the eyes of the beholder.

Despite the much-acknowledged critique that 'raw data is an oxymoron',[35] (digital) data is increasingly seen as the main source of valid 'evidence' in dominant research methods.[36] This may not come as a surprise considering the broader fetishization of 'big data' in (Euro-American) academia as well as its benefactors in the public and private sectors. Researchers' calls for improved data access therefore cannot be detached from questioning the politics of *knowledge* production more broadly. This means recognizing that knowledge is 'socially constructed, and historically situated', rationalized through research methodologies that are in themselves 'historically produced social formations articulated through particular discourses and systems of signification'.[37]

By prefixing *academic* to the *data gaze*, we wish to emphasize the complicity of academia in establishing the 'data imaginary' as a dominant paradigm structuring contemporary societies. The *data imaginary* determines what is rendered (in)visible in data-led processes, ordering, and governance.[38] And while its risks and constraints have been amply debated in scholarship,[39] datafication processes have developed into constitutive elements of modern-day knowledge production.[40] In other words, we believe academia's growing adoption of data driven research methods has significant political economic reverberations. It lends further legitimacy to the data gaze's salient claims to objectivity, neutrality, rationality, and universality.[41] Yet it often forgoes how ever-expanding processes of metrification, extraction, and abstraction deprioritize or invisibilize certain modes of knowledge (production), reinforce a variety of power dynamics, and may even cause harm.[42]

Through her work on 'missing datasets', academic and artist Mimi Ọnụọha cogently explains the important blind spots of the data imaginary.[43] First, those with the resources to produce or collect certain data often lack the incentives to do so (for example, law enforcement and military operations are some of the most data-driven public sectors, yet there is disproportionately little systematic data collected about police brutality and military abuse). Second, the data imaginary only allows for collecting things that fit its modes of collection, that is, through quantification and datafication processes (for example, some things may resist simple datafication, such as emotions or institutional racism). Third, the act of datafication may involve more effort or resources then the perceived benefits of datafication (for example, the benefit of reporting sexual harassment is often perceived lower than the cost of the process). Fourth, there might also be advantages to non-datafication (for example, protection of situationally disadvantaged groups). Ọnụọha reminds us that data will not solve all problems or scientific inquiries, and this is a good thing, especially as we may risk reductionism based on the choices made in collecting, ordering, and labelling data.

Even so, it is increasingly recognized that surveillance and datafication processes – often (co-)constituted and adopted by academia – have very problematic roots indeed:[44] from oppressive and racist regimes underpinning slavery[45] and Nazi Germany[46] to the persecution of black and brown people in the US,[47] as well as centuries of colonial extractivist practices more broadly.[48] Scientific research has played a crucial role in enabling and validating these regimes, reproducing social hierarchies.[49] In the digital context, pioneering anthropologist and science and technology studies scholar Diane Forsythe already established the sexism and silencing of voices in artificial intelligence (AI) development in the 1990s.[50] This historical trajectory of both science and datafication processes cannot simply be ignored and should actively factor into a constant self-reflective praxis.

More recently, an important strand of critique of the academic data gaze in the social sciences relates to so-called easy-data scholarship. Jean Burgess and Axel Bruns, for example, explain how social media platforms (deliberately) affect research agendas through their technical design.[51] Twitter (now X) in particular has proven to be the 'most (over-) studied social media platform precisely because it offers relatively open data access.… Yet, the non-randomness of data captured via these APIs [application programming interfaces] means that, even in the best of times, many Twitter studies have drawn conclusions based on substantially biased inferences.'[52] This is not to say that all studies using 'easy data' are necessarily reductionist or unrepresentative,[53] but it is a reminder to acknowledge how the academic data gaze is shaped by the socio-economic conditions in which it occurs. New and upcoming legal frameworks will only amplify these external influences on academia's data imaginary, especially in light of how transparency and data access requirements will be interpreted, operationalized, and enforced. Again, this raises questions on what information and research are prioritized/invisibilized and on agenda-setting power more broadly.[54]

Ironically, the ample critique (often coming from academia) on the public and private sectors' embrace of data-driven processes and the numerous issues it presents has not prevented many academic disciplines from going down the same path. All too often 'more data' is automatically seen as leading to 'better research outcomes', without a clear sense of what those

outcomes ought to be in the first place or a reflection on alternative research methodologies.[55] This entails deeper questions on the merits, integrity, and ethics of academic research that we cannot tackle here. At the very least, there is a serious need for individual and collective self-reflection on questions relating to data-driven research methods and disciplines: when they are exploitative; what their blind spots are; how they relate to similar processes in industry or government; and how they affect structural inequalities, injustice, and power.[56] Academia must confront its 'epistemology of ignorance' – that is, academia's ignorance of its own processes of reproduction – and continuously reflect on the emergence and use of dominant research methodologies.[57]

In sum, the academic data gaze is ever-expanding and does not tolerate alternative means of knowledge production.[58] Its internal logics and outcomes are presented as indisputable truths. As such, academia is constitutive of dominant *data assemblages*, described by Rob Kitchin as complex socio-technical systems involving a multitude of actors and processes pursuing the production of data.[59] That is why claims more and better data access should be accompanied by critical inquiries into the underlying logics and rationales of data processes and their agenda-setting power. How, and at what expense, do legal data access frameworks and data-driven research methods reinforce or construct specific frames of truth and knowledge?[60] We see this volume as contributing to this vital broader discussion that we need to have on academic claims to researcher data access and the power dynamics it entails.

Power Dynamics in Data Access for Research

In this subsection, we explore some of the complex power dynamics underlying researcher data access debates. The incremental invasion of digital infrastructures into every part of society has accelerated the accumulation of power and established new forms of data-driven power asymmetries. A quintessential example of this is the emergence of online platforms as dominant actors in today's information society and the focal point of many new legal frameworks.

The rise of platform power manifests primarily through the operational control that vests with Big Tech companies because of their dominance over computational infrastructures.[61] This control over infrastructure emerges in terms of determining which parts of the world can be made visible or legible through providing data access, what values and ideas must be given prominence in governance or policy settings, and how private–public relationships are being reshaped through increased dependency on Big Tech.[62]

Power takes on different forms. There is a technical power by virtue of platforms determining how one can operate within them or in relation to them. This technical power, however, does not operate on its own, because of their infrastructural power to also lobby for self-regulation and, in turn, reshape governance practices. In doing so, technical power gives way to shaping how sectors are being managed. Consider the case of the Oversight Board, which is an entity set up by Meta to decide on matters related to content on its platforms, including Facebook and Instagram. This is an example where an attempt is made to propose a regulatory architecture that informs how to uphold the right to freedom of expression online. However,

as Chinmayi Arun has argued, Meta operates differently in different jurisdictions, engaging flexibly when it deals with states and publics that do not have the regulatory capacity to push back, and therefore ensuring that content moderation has different approaches depending on location.[63] This is relevant because it demonstrates the power that platforms have to mediate between different political and economic entities. More often than not, where it is to their economic benefit to side with states, over people, it is willing to make compromises. WhatsApp, for instance, attempted to change its privacy policy all over the world: whereas in the EU data would be shared only for WhatsApp's purposes, in non-EU regions it could also be used for other Meta companies.[64]

This example demonstrates that along with computational power, regulatory power is mediated by the power of transnational markets, the capacity of institutions to be able to push back, as well as a societal power and awareness about the possibilities of alternative platforms.[65] For instance, Europe, despite priding itself on being at the forefront of tech regulation, has had innumerable challenges with implementation and enforcement. Ireland, which plays a vital role as a data protection regulator in the EU, recently proposed a law that renders confidential all matters before the data protection commissioner. This undertaking, as Amnesty International has argued, is a 'blatant attempt not only to shield Big Tech from scrutiny but also to silence individuals and organizations that stand up for the right to privacy and data protection'.[66]

These decisions by states in turn determine how regulation is perceived and what the capacity of civil society (including academia) is to engage, challenge, and resist. Platform power has increasingly affected the independence of academia which has long been complicit with accepting funding from corporations, including fossil fuel companies and those engaging in surveillance, without the necessary safeguards or institutional mechanisms to ensure transparency and accountability. A recent report of De Jonge Akademie in the Netherlands argues that there is a need for a greater overview of funding flows and how such funding affects researchers, what are the risks at play of accepting such funding while, at the same time, ensuring that institutions have the capacity to be able to report as well as audit funding relationships.[67]

Another telling example of platform power comes from the ongoing genocide in Palestinian territories.[68] On 27 October 2023, Israel cut access to phone and internet services for 34 hours, leaving a majority of over two million Palestinians who live in Gaza as well as aid organizations on the ground with no way to connect to the outside world.[69] It is a move that Israel has since used multiple times as a deliberate strategy, in particular before a military operation. The situation was so dire that, with no alternatives in sight, people appealed to Elon Musk who owns Starlink – a satellite internet venture – to provide internet.[70] This is not the only case. In Ukraine, during the ongoing war with Russia, Musk was asked by the Ukrainian authorities to provide Starlink as a service to keep people online in case the internet infrastructure was destroyed, which he did.[71] However, over the course of the conflict, Musk has variously threatened to withdraw access, at times stating that he cannot fund it, while at other moments making grand claims that, regardless of the financial situation of Starlink, he will keep Ukrainians online.[72] The absurdity that the only way for millions of besieged people to

have internet access is through appealing to one person's largesse demonstrates the dangers that will emerge because of private corporations having control over critical infrastructures.[73]

As this example demonstrates, this is not just technical power but power that manifests in terms of informational power, determining what narratives will be heard and what will be erased.[74] It is an institutional power because it intrinsically influences how emergency teams will be able to respond, ask for help, and document atrocities being committed. It is also societal power, shaping the ways in which people will know about these crises and how they make sense of their role in such a situation.

The criticality of data access is even more apparent in the Palestinian context, when the former president of the United States (US), despite the vast amount of verified images and video material showing shocking amounts of human and material devastation, said, 'I have no notion that Palestinians are telling the truth about how many people are killed.'[75] As Hala Alyan, a Palestinian American author states, this statement was made knowing the power that it would have: 'It would quietly cleverly, delegitimize the dead. Even the dead. He said it and the saying was erasure.' She goes on to say, 'What counteracts erasure? Witnessing', going on to show how a report was then released by the Palestinian Ministry of Health with 7,028 names, of which 2,913 were children.[76] At the time of finalizing this volume, the confirmed death toll stood at 34,367, of which more than 13,800 were children and over 100 were journalists.[77]

Witnessing is a critical component to counteracting the power of platforms. It is a fundamental component of access. Access is not passive; it is an opportunity to be able to challenge structures of power that constitute how platforms are designed and imagined. Having data access allows one to be able to understand what data is being collected, recorded, labelled, and a consequence distributed to offer provides perspective in terms of what information can be considered valid and which must be disregarded – an epistemic choice, which is used to silence those who are already oppressed.[78] As Omar Suleiman writes,

> Open Apple, Google, or any other digital map. Type 'Palestine'. You won't find it. You will only find Israel. If you're lucky, you may be directed to a small patchwork of what is called 'Palestinian Territories' firmly embedded inside Israel lest anyone mistakenly think it is an independent nation-state. And of course, you will find nowhere on any map the keyword that precedes Palestinian Territories to lay bare the ugly, but necessary and harrowing truth: 'Occupied'.[79]

This is but one example of how we are witness to the ways in which platforms construct identities, legitimize versions of history, and dominate the lenses through which we are to view the world. Palestinian erasure is continually supported by platforms with accusations of shadow bans and blocking of several pro-Palestinian accounts.[80] It would be complete without the challenge to platforms by Palestinian people and others across the world who must subvert platforms by interspersing content on dispossession of lands, homes, and people with humour, and the like, completely unrelated to the conflict.

This is where witnessing is also important to challenge the epistemic power of platforms. This power, to draw from Miranda Fricker, can manifest in terms of who has the power to bear testimony (testimonial power), and whether their words are believed and acted upon, and a hermeneutical power where people's capacities to understand themselves are impacted because of being denied the vocabulary to voice their experiences.[81] This power does not emerge in a vacuum but is deeply embedded in the political economy of where platforms operate and what regulatory power they seek to influence, or abide by. Numerous incidents in the past have shown that platforms such as Zoom have actively participated in censoring different voices. Through regulation such as anti-terrorism laws, it has sought to prevent Palestinian activists from speaking on university campuses. In China, it has been accused of violating its own policies to censor discussions about Tiananmen Square at the Chinese government's request.[82] Social platforms not only thereby become gatekeepers to knowledge but also supersede the expertise housed within different domains such as academia to make such decisions.[83]

In these different attempts to constrain access to digital infrastructures, the consequences are significant. This is because these platforms' power not only comes from determining who gets to see what, but also from their infrastructural power and how they prevent anyone from observing their scale or operations, such as the logics for which posts get censored, shadow-banned, or removed. It is these complex and multidimensional power dynamics that ought to be factored into any (policy) debate on researcher access to data.

This volume brings together 8 different chapters within the theme of researchers' access to data in and about digital infrastructures. The collection of chapters also demonstrates the wide variety of angles and issues involved when it comes to the governance and regulation of data access for research. Because of this diversity in angles, the chapters are distributed along three main parts: imaginations, limitations, and promises.

Imaginations

We do not take the contours of what data access is to be a given. Instead, we recognize that what it is and how it is understood are unique in the different chapters in this volume. The chapters in this part offer distinct imaginations of data access by exploring its relationship to research and, as a result, also adjacent issues related to governance, autonomy, rights, and practice. In chapter 2, 'Re-Conceptualizing Governance Policies on Data Access for Research', Carolina Aguerre helps us make sense of the plethora of national and international approaches to govern data access (for research), taking a polycentric governance lens. Her analysis reaches beyond traditional Euro-American perspectives, highlighting majority-of-the-world angles on two specific policy arenas: open science and AI strategies. The following chapter, 'Violent Plains', by Frank Kwaku Agyei, Lawrence Kwabena Brobbey, Paul Osei-Tutu, and Boateng Kyereh, pushes our imagination even further, exploring challenges and strategies for access to pastoralists' data in Ghana. It offers a very concrete and gripping account of the many issues underlying the collection of data about (groups of) individuals in the first place. We see the chapter as an invitation to the academic community to self-reflect on over-reliance on available data, what is invisibilized, as well as various assumptions about data quality and universality.

Through these different imaginations, the chapters in this part challenge us to think of ways of describing, categorizing, and institutionalizing what researchers can do in different contexts. The last chapter in this part also demonstrates this quite well. In chapter 4, 'From Rights to Skills', Midas Nouwens reflects on data access rights in relation to their role as an educator in higher education. Based on their teaching experiences, they propose a variety of ways for students to work with and critically reflect on data access rights, helping them to navigate and co-shape digital societies.

In sum, the imaginations presented in the first part of the volume highlight multifarious aspects of the debate around data access in academic contexts. They invite us to reflect on who owes data access on the one hand and who is owed data access on the other. They help discuss the political economies within which such relations are created, the challenges of enabling and sustaining such access, and who benefits as a result of such data access, and they do so at different scales – from local in focus to transnational in scale – as well as speak to different epistemic communities such as universities, transnational organizations, pastoralists, and regulators.

Limitations

Critical to our understanding of data access is the reality that just providing such access is insufficient. Access requires an enabling environment where people have the capacity, resources, skills, and time to use such access to achieve desirable outcomes. This part of the volume discusses various challenges both to the EU conceptualization of data access rights and more generally in terms of the challenges of researching digital platforms and infrastructures. It explores across contexts, both jurisdictional as well as domain-wide, how using data access for accountability and transparency requires more than just access rights and instead requires institutional capacity. Such capacity oftentimes is not available in academic and civil society institutions, and the chapters caution us whether access rights create dependencies on actors who do not have the capacity to be able to make use of the rights.

Chapter 5 looks at transparency and data access in the context of accountability of public authorities, specifically law enforcement and intelligence agencies. André Ramiro, Pedro Amaral, and Marcos César M. Pereira in chapter 5, 'Keys Thrown Away?', reflect on their own experience of conducting an empirical research project involving the strategic use of freedom of information laws in Brazil. This self-reflective study exposes important limitations to research into government surveillance infrastructures, through convoluted claims to secrecy. As such, the chapter also shows us the dangers of 'pretend-transparency' for democratic oversight.

While chapter 5 focuses on how researchers relate to accountability and democratic oversight dimensions of data access rules, the following chapter zooms in on the market-driven aspects of these rules. Chapter 6, 'Digging into the EU Data Laws and Their Impact on African Researchers', by Paul Esselaar, explains the EU internal market focus of many new provisions, at the expense of researchers, especially those based outside the EU. Exploring the issues and constraints of a whole raft of new EU data access provisions, specifically in relation to African researchers, Esselaar reveals a darker side of the EU's extraterritorial reach and the potentially perverse effects of these rules on data and research originating in the African continent.

Promises

In the last part of the volume, we are interested in unpacking various promises that emerge because of providing data access for research. The dominant presumption with data access is that it will have beneficial consequences for increased transparency, oversight, and accountability to the wider publics, especially when it comes to online platforms. As a result, such access would not only challenge informational asymmetries but also encourage more a meaningful understanding of, and participation in, the digital society. That also appears in much of the debate surrounding the barrage of new policy frameworks tackling the data economy. The chapters in this part take a closer look at different EU policy initiatives involving transparency and data access provisions. While hopeful in their outlook, the authors in these chapters also provide necessary nuance to policymakers' claims. Importantly, they help us think through important issues and pitfalls, tracing concrete methods and practices for overcoming them.

Chapter 7, 'A Subject Access Request, Then What?' by Jake Stein and Reuben Binns, does so by zooming in on data access promises in the context of platform workers. The chapter draws on participatory action research, incorporating legal and technical insights, to build a data architecture for platform worker empowerment. The authors are driven by an aspiration to design public-service data infrastructures, tackling hard questions on how workers, researchers, or other public-interest-driven groups may confront data structures that are imposed by powerful actors such as platform companies. In light of new and upcoming transparency requirements, notably in the EU, Stein and Binns turn the discussion towards lowering the barriers to data analytics for advocates in low-resource environments and propose a lightweight, queryable data institution which relies on existing open-source, unstructured data analytics infrastructures. Such data institutions form the focal point of the next chapter: chapter 8, 'Data Intermediaries for Good', in which Matteo Nebbiai zooms in on the emergence of a particular type of data institution that emerged as a new legal category of actors in the EU's Data Governance Act (DGA). So-called data intermediation services are hailed by the legislator as enabling data sharing between an undetermined number of data holders and users. Through a systematic analysis of 54 already existing data intermediary services, Nebbiai explores to what extent these actors also contribute to better access for research in particular.

The next chapter takes a closer look at questions of researcher data access in the online platform context. Chapter 9, 'From Discretion to Obligation?' by Michalina Kowala, traces the legislative development of the EU's 2018 and 2022 codes of practice on disinformation. Kowala autoethnographically identifies the key issues and shortcomings of researcher data access provisions in the code(s). As such, the chapter demonstrates the gap between theory and praxis when it comes to regulating data access for research, in a very concrete context.

In sum, this part offers various ways in which the promises of data access regulation can be operationalized, for instance, by framing such access as a right to ensure that there are guarantees, but also by creating new institutional frameworks to encourage participation and collaborative communities for research. This part therefore provides pathways with which the goals and values associated with data access can be realized.

***

Finally, we hope this volume may provoke and incite readers to consider the many dimensions of researcher data access. Recent policy discussions have lauded the potential of the data economy, with little attention to deeper questions relating to data access in non-economic contexts or to broader political economy implications and shifting power dynamics. In this introductory chapter, we aimed to lift the veil on some of these broader issues that merit more attention. Notably, what are the underlying assumptions of 'access' and how does it inform how we think about data access in research contexts? How may academia – from the individual to the institutional levels – be complicit in structures of power and oppression? And what do these structures look like in the first place? As such, this chapter invites you, the reader, to reflect on your own positionality in relation to existing data access claims as well. We believe this to be a great starting point to peruse through the chapters in this book. Rather than giving a coherent or comprehensive account of the debate – which we do not deem possible in the first place! – the chapters expose us to very diverse perspectives and approaches: from the specific to the general, the local to the international, technical to legal, doctrinal to empirical. May they inspire you in your own academic practice.

NOTES

1 See Axel Bruns, 'After the "APIcalypse": Social Media Platforms and Their Fight against Critical Scholarly Research', *Information, Communication and Society* 22, no. 11 (19 September 2019): 1544–1566; Daphne Keller and Paddy Leerssen, 'Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation', in *Social Media and Democracy: The State of the Field and Prospects for Reform*, ed. N. Persily and J. Tucker, 220–251 (Cambridge University Press, 2020); Jef Ausloos, Paddy Leerssen, and Pim ten Thije, *Operationalizing Research Access in Platform Governance: What to Learn from Other Industries?* (Algoritm Watch, 2020); Eszter Hargittai, *Research Exposed: How Empirical Social Science Gets Done in the Digital Age* (Columbia University Press, 2021).

2 For a detailed overview of transparency and data access provisions in recent EU 'data law', see Jef Ausloos, Arlette Meiring, Doris Buijs, Mireille van Eechoud, Stefanie Boss, and Joanna Strycharz, *Information Law and the Digital Transformation of the University*, pt 2: *Access to Data for Research*, 15 September 2023 (Institute for Information Law, University of Amsterdam), https://www.uva.nl/en/about-the-uva/policy-and-regulations/ general/preserving-digital-sovereignty-of-universities-and-researchers/ preserving-digital-sovereignty-of-universities-and-researchers.html (27 November 2024).

3 See, in this regard, also Celine-Marie Pascale, 'Epistemology and the Politics of Knowledge', *Sociological Review* 154, no. 58 (2010): 154–165, 163.

4 Diana E. Forsythe and David J. Hess, *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence* (Stanford University Press, 2001), xix.

5 Joan Lopez Solano, Aaron Martin, Siddharth de Souza, and Linnet Taylor, *Governing Data and Artificial Intelligence for All: Models for Sustainable and Just Data Governance* (European Parliament, 2022); Amber Sinha and Arindrajit Basu, 'Why Metaphors for Data Matter', Bot Populi, 13 August 2021, https://botpopuli.net/?post_type=post&p=4069 (accessed 11 September 2023).

6 Nadezhda Purtova and Gijs van Maanen, 'Data as an Economic Good, Data as a Commons, and Data Governance', arXiv, 18 April 2023, http://arxiv. org/abs/2212.10244 (accessed 15 November 2023); Anita Gurumurthy and Nandini Chami, 'Governing the Resource of Data: To What End and for Whom? Conceptual Building Blocks of a Semi-Commons Approach', Data Governance Network Working Paper 23, 2022, https://itforchange.net/sites/ default/files/1741/WP23-Governing-the-Resource-of-Data-AG-NC.pdf (accessed 19 December 2024).

7 Solano, Martin, de Souza, and Taylor, *Governing Data and Artificial*

Intelligence for All.

8 Tahu Kukutai and John Taylor, *Indigenous Data Sovereignty: Toward an Agenda* (ANU Press, 2016).

9 Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson, 'The CARE Principles for Indigenous Data Governance', *Data Science Journal* 43, no. 19 (2020) 1–12, DOI: https://doi.org/10.5334/dsj-2020-043.

10 Tahu Kukutai, 'Indigenous Data Sovereignty: A New Take on an Old Theme', *Science* 382, no. 6674, DOI: 10.1126/science.adl4664.

11 We Are Not Numbers, 'About', https://wearenotnumbers.org/about (accessed 15 November 2023).

12 Yuval Abaraham discusses how artificial intelligence (AI) is being used to identify assassination targets amongst people in Gaza, with little human oversight, creating a mass killing machine. Yuval Abraham, '"Lavender": The AI Machine Directing Israel's Bombing Spree in Gaza', *+972 Magazine*, 3

April 2024, https://www.972mag.com/lavender-ai-israeli-army-gaza (accessed 12 May 2024).

13 Lisa Gitelman, *'Raw Data' Is an Oxymoron* (MIT Press, 2013).

14 Julian Alberto Posada Gutierrez, 'The Coloniality of Data Work: Power and Inequality in Outsourced Data Production for Machine Learning', thesis, University of Toronto, November 2022, https://tspace.library.utoronto.ca/ handle/1807/126388 (accessed 9 October 2023).

15 Billy Perrigo, 'OpenAI Used Kenyan Workers on Less than $2 Per Hour: Exclusive', *Time*, 18 January 2023, https://time.com/6247678/openaichatgpt-kenya-workers (accessed 8 August 2023).

16 Birgitta Bergvall-Kåreborn and Debra Howcroft, 'Amazon Mechanical Turk and the Commodification of Labour' *New Technology, Work and Employment* 213, no. 29 (2014): 213–223.

17 Linnet Taylor, Luciano Floridi and Bart van der Sloot (eds.), *Group Privacy: New Challenges of Data Technologies* (Springer International Publishing, 2017).

18 Salome Viljoen, 'A Relational Theory of Data Governance', SSRN Scholarly Paper ID 3727562, Social Science Research Network, https://papers.ssrn. com/abstract=3727562 (accessed 19 October 2021).

19 Daniel M. Brinks, 'Access to What? Legal Agency and Access to Justice for Indigenous Peoples in Latin America', *Journal of Development Studies* 348, no. 55 (2019): 348–365; Siddharth Peter de Souza, 'A Capability Approach to Access to Justice in Plural Legal Systems', in *Designing Indicators for a Plural Legal World*, by Siddharth Peter de Souza, 164–203 (Cambridge University Press, 2022).

20 Miranda Fricker, 'Introduction', in *Epistemic Injustice: Power and the Ethics of Knowing*, by Miranda Fricker, 1–8 (Oxford University Press, 2007); Sabelo J. Ndlovu-Gatsheni, 'Introduction: Seek Ye Epistemic Freedom First', in *Epistemic Freedom in Africa*, by Sabelo J. Ndlovu-Gatsheni, 1–41 (Routledge 2018).

21 Hazel Genn, 'The Landscape of Justiciable Problems', in *Paths to Justice: What People Do and Think about Going to Law*, by Hazel Genn and Sarah Beinart, 21–66 (Hart Publishing, 1999); Pascoe Pleasence, Nigel Balmer, and Rebecca Sandefur, 'Paths to Justice: A Past, Present and Future Roadmap', Nuffield Foundation, 2013, https://www.nuffieldfoundation.org/sites/ default/files/files/PTJ%20Roadmap%20NUFFIELD%20Published.pdf (accessed 19 December 2024).

22 See, for example, Ausloos, Meiring, Buijs, van Eechoud, Boss, and Strycharz, *Information Law and the Digital Transformation of the University*, pt 2: *Access to Data for Research*.

23 Jef Ausloos and Pierre Dewitte, 'Shattering One-Way Mirrors: Data Subject Access Rights in Practice', *International Data Privacy Law* 4, no. 8 (2018): 4–28; René Mahieu, Jef Ausloos, and Michael Veale, 'Getting Data Subject Rights Right', LawArXiv, 2019, preprint, https://osf.io/e2thg (accessed 15 November 2023).

24 Linda Tuhiwai Smith, 'Introduction', in *Decolonizing Methodologies: Research and Indigenous Peoples*, by Linda Tuhiwai Smith, 1–18 (ZED Books 2012).

25 Edward W. Said, *Orientalism* (Knopf Doubleday, 2014).

26 Amartya Sen, *The Idea of Justice* (Harvard University Press, 2011); Amartya Sen, 'Human Rights and Capabilities', *Journal of Human Development* 151, no. 6 (2005): 151–166.

27 Wendy Nelson Espeland and Mitchell L. Stevens, 'Commensuration as a Social Process', *Annual Review of Sociology* 313, no. 24 (1998): 313–343;

Sally Engle Merry, *The Seductions of Quantification: Measuring Human Rights, Gender Violence, and Sex Trafficking* (University of Chicago Press, 2016); Siddharth Peter de Souza, '"Meanings", "Trust" and "Power": Critical Perspectives on Legal Indicators', in *Designing Indicators for a Plural Legal World*, by Siddharth Peter de Souza, 20–50 (Cambridge University Press, 2022).

28 Sally Engle Merry, 'Measuring the World: Indicators, Human Rights, and Global Gover-

nance', *Current Anthropology* S83, no. 52 (2011): S83–S95.

29 Wendy Espeland, 'Narrating Numbers', in *The World of Indicators: The Making of Governmental Knowledge through Quantification*, ed. Richard Rottenburg, Sally E. Merry, Sung-Joon Park, and Johanna Mugler, 56–75 (Cambridge University Press, 2015).

30 Boaventura de Sousa Santos, 'Introduction: Why the Epistemologies of the South? Artisanal Paths for Artisanal Futures', in *The End of the Cognitive Empire:The Coming of Age of Epistemologies of the South*, by Boaventura de Sousa Santos, 1–16 (Duke University Press, 2018).

31 David Beer, *The Data Gaze: Capitalism, Power and Perception* (Sage Publishing, 2019).

32 Nick Srnicek, *Platform Capitalism* (John Wiley & Sons, 2017).

33 Julie E. Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press, 2019).

34 Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (Profile Books, 2019).

35 Gitelman, *'Raw Data' is an Oxymoron*; Rob Kitchin and Tracey Lauriault, 'Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work', in *Thinking Big Data in Geography*, ed. J. Thatcher, J. Eckert, and A. Shears, 3–20 (University of Nebraska Press, 2018).

36 Pascale, 'Epistemology and the Politics of Knowledge', 154, 157.

37 Pascale, 'Epistemology and the Politics of Knowledge', 163. See also Linda Tuhiwai Smith, *Decolonizing Methodologies: Research and Indigenous Peoples* (Bloomsbury Academic, 2023).

38 S Beer, *The Data Gaze*.

39 See ample references in Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (Sage Publishing, 2014); Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (NYU Press, 2018); Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press, 2018); Andreas Hepp, Juliane Jarke, and Leif Kramp, 'New Perspectives in Critical Data Studies: The Ambivalences of Data Power—An Introduction', in *New Perspectives in Critical Data Studies: The Ambivalences of Data Power*, ed. Andreas Hepp, Juliane Jarke, and Leif Kramp (Springer International Publishing, 2022).

40 See, for example, Bernhard Rieder and Theo Röhle, 'Digital Methods: Five Challenges', in *Understanding Digital Humanities*, ed. David M. Berry, 67–84 (Palgrave Macmillan, 2012).

41 Stefania Milan and Emiliano Treré, 'Big Data from the South(s): Beyond Data Universalism, Television and New Media', *Television and New Media* 20, no. 4, DOI: https://doi.org/10.1177/1527476419837; Rieder and Röhle, 'Digital Methods'.

42 See, in this regard, also David Beer, *Metric Power* (Palgrave Macmillan, 2016); Kitchin and Lauriault, 'Towards Critical Data Studies', 7.

43 Mimi Ọnụọha, 'An Overview and Exploration of the Concept of Missing Datasets', GitHub, 3 February 2016, https://github.com/MimiOnuoha/ missing-datasets (accessed 12 October 2023).

44 Meredith Whittaker, 'Origin Stories: Plantations, Computers, and Industrial Control', *Logic(s) Magazine*, 17 May 2023.

45 Caitlin Rosenthal, *Accounting for Slavery: Masters and Management* (Harvard University Press, 2018).

46 Edwin Black, *IBM and the Holocaust: The Strategic Alliance between Nazi Germany and America's Most Powerful Corporation* (Crown Publishers, 2001).

47 Simone Browne, *Dark Matters: On the Surveillance of Blackness* (Duke University Press,

2015).

48 Donna Cormack and Tahu Kukutai, 'Indigenous Peoples, Data, and the Coloniality of Surveillance', in *New Perspectives in Critical Data Studies: The Ambivalences of Data Power*, ed. Andreas Hepp, Juliane Jarke, and Leif Kramp (Springer International Publishing, 2022), 126–127.

49 For example, academic work on eugenics and early psychoanalysis have been constitutive of systemic racism and sexism across society. See notably Pascale, 'Epistemology and the Politics of Knowledge', 154, 157. See also Smith, *Decolonizing Methodologies*.

50 Forsythe and Hess, *Studying Those Who Study Us*.

51 Jean Burgess and Axel Bruns, 'Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research after the Computational Turn', in *Compromised Data: From Social Media to Big Data*, ed. Ganaele Langlois, Joanna Redden, and Greg Elmer, 93–111 (Bloomsbury Academic, 2015).

52 Rebekah Tromble, 'Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age', *Social Media + Society* 7 (2021): 1–18. Similarly, see Kevin Driscoll and Shawn Walker, 'Working within a Black Box: Transparency in the Collection and Production of Big Twitter Data', *International Journal of Communication* 8 (2014): 1745–1764; David Gunnarsson Lorentzen and Jan Nolin, 'Approaching Completeness', *Social Science Computer Review* 38 (2015): 10–24; Eszter Hargittai, 'Potential Biases in Big Data: Omitted Voices on Social Media', *Social Science Computer Review* (2018), DOI: 10.1177/089443931878832; Cornelius Puschmann, 'An End to the Wild West of Social Media Research: A Response to Axel Bruns', *Information, Communication and Society* 22, no. 11 (2019): 1582–1589.

53 Suay M. Özkula, Paul J. Reilly, and Jenny Hayes, 'Easy Data, Same Old Platforms? A Systematic Review of Digital Activism Methodologies', *Information, Communication and Society* 1, no. 15 (2022): 1–20.

54 Beer, *The Data Gaze*.

55 See, for example, Arielle Hesse, Leland Glenna, Clare Hinrichs, Robert Chiles, and Carolyn Sachs, 'Qualitative Research Ethics in the Big Data Era', *American Behavioral Scientist* (2018), DOI: https://doi. org/10.1177/0002764218805806; Ausloos, Meiring, Buijs, van Eechoud, Boss, and Strycharz*, Information Law and the Digital Transformation of the University*, pt 2: *Access to Data for Research*.

56 Sebastián Lehuedé calls for *radicalizing* reflexivity in critical data studies in 'The Double Helix of Data Extraction: Radicalising Reflexivity in Critical Data Studies', *Technology and Regulation* (22 March 2024): 89–91, DOI: https://doi.org/10.26116/techreg.2024.009.

57 Pascale, 'Epistemology and the Politics of Knowledge', 154, 158. In the context of digital methods, see Burgess and Bruns, 'Easy Data, Hard Data'. For humanities and social sciences research, see also the work of Bernhard Rieder and, in particular, Rieder and Röhle, 'Digital Methods'.

58 See Beer, *The Data Gaze*, 220; Smith, *Decolonizing Methodologies*.

59 Kitchin, *The Data Revolution*; Kitchin and Lauriault, 'Towards Critical Data Studies', 6–7; Hepp, Jarke, and Kramp, 'New Perspectives in Critical Data Studies', 5.

60 Beer, *The Data Gaze*, ch. 1.

61 'Programmable Infrastructures', TU Delft, https://www.tudelft.nl/tbm/ programmable-infrastructures (accessed 29 June 2022).

62 Joan Lopez Solano, A. Martin, F. Ohai, S. P. de Souza, and I. Taylor, 'Digital Disruption or Crisis Capitalism? Technology, Power and the Pandemic', *Global Data Justice* (2022), DOI: https://doi.org/10.26116/gdj-euaifund.

63 Chinmayi Arun, 'Facebook's Faces', 15 March 2021, https://papers.ssrn. com/

abstract=3805210 (accessed 15 November 2023).

64 'We Knew India Was Going to Be a Big Place for Signal: Brian Acton', *Business Standard*, 13 January 2021, https://www.business-standard.com/ article/technology/we-knew-india-was-going-to-be-a-big-place-for-signalbrian-acton-121011300189_1.html (accessed 11 March 2023).

65 See the work of the Non-aligned Technologies Movement, which argues for why we need to pursue alternative platforms. Ulises Ali Mejias, 'To Fight Data Colonialism, We Need a Non-Aligned Tech Movement', Al Jazeera, 8 September 2020, https://www.aljazeera.com/opinions/2020/9/8/to-fight-data-colonialism-we-need-a-non-aligned-tech-movement (accessed 16 April 2024).

66 'Ireland: Draconian Law to Make Data Protection Procedures Confidential', Amnesty International, 28 June 2023, https://www.amnesty.org/en/latest/ news/2023/06/ireland-draconian-law-to-make-data-protection-proceduresconfidential (accessed 15 November 2023).

67 'Space to Think: An Analysis of Structural Threats to Academic Freedom and Integrity', De Jonge Akademie, 20 June 2023, https://www.dejongeakademie.

nl/en/publications/2495737.aspx (accessed 15 November 2023).

68 'Public Statement: Scholars Warn of Potential Genocide in Gaza', *TWAILR: Third World Approaches to International Law Review*, 17 October 2023, https://twailr.com/public-statement-scholars-warn-of-potential-genocide-ingaza (accessed 15 November 2023).

69 Abu Bakr Bashir, Iyad Abuheweila, Vivian Nereim, and Yousur Al-Hlou, 'Gaza Blackout Cut Palestinians' Internet and Phone Service for 34 Hours', *New York Times*, 29 October 2023, https:// www.nytimes.com/2023/10/29/ world/middleeast/gaza-blackout-internet-israel.html (accessed 15 November 2023).

70 Priyanka Shankar, 'Can Elon Musk's Starlink Provide Internet Service to Gaza?' Al Jazeera, 29 October 2023, https://www.aljazeera.com/ news/2023/10/29/can-elon-musks-starlink-provide-internet-service-to-gaza (accessed 15 November 2023).

71 'Elon Musk Activates Starlink Satellites to Give Ukraine Data Backup', *Politico*, 26 February 2022, https://www.politico.eu/article/elon-muskactivates-starlink-satellites-to-give-ukraine-data-backup (accessed 17 April 2024).

72 'Elon Musk U-Turns, Says Will Keep Funding Starlink in Ukraine', Al Jazeera, 15 October 2022, https://www.aljazeera.com/news/2022/10/15/inreversal-musk-says-will-continue-funding-starlink-in-ukraine (accessed 17 April 2024).

73 Adam Satariano, Scott Reinhard, Cade Metz, Sheera Frenkel, and Malika Khurana, 'With Starlink, Elon Musk's Satellite Dominance Is Raising Global Alarms', *New York Times*, 28 July 2023, https://www.nytimes.com/ interactive/2023/07/28/business/starlink.html (accessed 15 November 2023).

74 Ndlovu-Gatsheni, 'Introduction'.

75 'Biden Says He Has "No Confidence" in Palestinian Death Count', Reuters, 26 October 2023, https://www.reuters.com/world/middle-east/biden-sayshe-has-no-confidence-palestinian-death-count-2023-10-26 (accessed 15 November 2023).

76 Hala Alyan on Instagram: 'On Witnessing [*sic*] Second Slide Is from @ dianabuttu', Instagram, 26 October 2023, https://www.instagram.com/p/ Cy4KUAOvkjy (accessed 15 November 2023).

77 'Israel-Gaza War in Maps and Charts: Live Tracker', Al Jazeera, https://www. aljazeera.com/ news/longform/2023/10/9/israel-hamas-war-in-maps-andcharts-live-tracker (accessed 17 April 2024).

78 Boaventura de Sousa Santos, *Epistemologies of the South: Justice Against Epistemicide* (Routledge, 2015).

79 Omar Suleiman, 'Erasing Palestine', Al Jazeera, 19 Oct 2023, https:// www.aljazeera.com/

opinions/2023/10/19/erasing-palestine-2 (accessed 15 November 2023).

80 'Platforms Must Stop Unjustified Takedowns of Posts by and about Palestinians', Electronic Frontier Foundation, 8 November 2023, https:// www.eff.org/deeplinks/2023/11/platforms-must-stop-unjustifiedtakedowns-posts-and-about-palestinians (accessed 15 November 2023); 'Digital Blackout: Systematic Censorship of Palestinian Voices', Global Voices, 8 November 2023, https://globalvoices.org/2023/11/08/digital-blackout-systematic-censorship-of-palestinian-voices (accessed 15 November 2023); Kari Paul, 'Instagram Users Accuse Platform of Censoring Posts Supporting Palestine', *The Guardian*, 18 October 2023, https:// www.theguardian.com/technology/2023/oct/18/instagram-palestine-postscensorship-accusations (accessed 15 November 2023); Priyanka Shankar, Pranav Dixit, and Usaid Siddiqui, 'Are Social Media Giants Censoring ProPalestine Voices Amid Israel's War?' Al Jazeera, 24 October 2023, https:// www.aljazeera.com/features/2023/10/24/shadowbanning-are-social-mediagiants-censoring-pro-palestine-voices (accessed 15 November 2023).

81 Fricker, 'Introduction'.

82 'How Zoom Violated Its Own Terms of Service for Access to China's Market', Human Rights Watch, 22 December 2020, https://www.hrw.org/ news/2020/12/22/how-zoom-violated-its-own-terms-service-access-chinasmarket (accessed 17 April 2024).

83 Alice Speri and Sam Biddle, 'Zoom Censorship of Palestine Seminars Sparks Fight over Academic Freedom', *The Intercept*, 14 November 2020, https:// theintercept.com/2020/11/14/zoom-censorship-leila-khaled-palestine (accessed 29 June 2022).

PART I: IMAGINATIONS

2. RE-CONCEPTUALIZING GOVERNANCE POLICIES ON DATA ACCESS FOR RESEARCH

CAROLINA AGUERRE

The last decade has seen an expansion of scholars who are increasingly interested in questioning the current data governance arrangements, its politics, and the very foundations of knowledge and power against the dominance of Big Tech platforms.[1] Much of this literature has claimed that the power harnessed by these organizations should be addressed practically and politically, through resistance strategies as well as through policies, including regulation. Yet often it claims that this is not only old wine in new bottles but rather that the nature of this power deserves a special epistemological status.

The notion of 'data interest' as one that constitutes a need and value of data as a resource is found in different sectors, including the scientific, and involves design and governance practices.[2] Access to data has been a crucial and longstanding concern of academia to fulfill its mission towards advancing science and achieving research goals.[3] The development of the internet in the 1990s and the expansion of digital networked technologies and artificial intelligence (AI) became game changers for the practices of the scientific community in different disciplinary domains.

The possibility of amassing large data troves of big data has been possible due to the internet and the platforms and applications that have connected layers of individuals and devices.[4] The control of these data troves has been well guarded in many of the large internet platforms that have trained their algorithms with these machine learning (ML) models. The emergence of even larger data sets that train the algorithms of most generative AI applications is based on larger volumes of data that are collected and analyzed from the open internet, including platforms.[5] Without the complementarities between different technologies and open infrastructures, the current disruptions of generative AI applications fostered by the private sector and many legacy digital platforms would not have been achieved.[6]

It is well known that the tensions between data access and the interests of global corporations are a pressing issue[7] for governments, civil society, researchers, and small companies, as legitimacy concerns backed by an idea of access to data as a justice issue[8] continue to build up.[9] While the Digital Services Act, 2022 (DSA), and in particular Article 40 have played a relevant role in the debate of access to very large platforms' data for research over the last years, this policy remains a regional instrument. The Digital Markets Act (DMA), 2022, and the DSA seek to address some of the concerns that unfold between the control of Big Tech and large and relevant data sets for society at large. Yet it is an instrument that is notably absent in countries outside the European Union (EU), and even more so in global majority contexts. Despite this, it may have implications for researchers beyond the EU and thence transnational and extraterritorial effects,[10] as evidenced from recent examples with EU regulation on data and digitization.[11]

For stakeholders outside of the potential scope of these measures, there are different initiatives which have treated data access issues for scientific endeavours. These actions embrace different objectives and have been developed by national and supranational organizations with varying degrees of formality, binding nature, and scale over the last decade. National open data strategies, open science (OS) recommendations, and AI strategies are different policies that address data access as a key pillar to achieving different policy goals. Despite the claims made by these policies concerning data access options, these strategies are much more elusive concerning issues of access to data by the research community. For example, OS promotes the open sharing of scientific knowledge, including research data, and open data initiatives aim to make data accessible to the public, including researchers and AI practitioners. By providing open access to relevant datasets, open data may facilitate the development and training of AI models by local groups of researchers. While these claims may seem legitimate, these policy approaches do not always facilitate these objectives. This is largely because they are still disconnected policy instruments, but also because in many countries in global majority contexts, research and innovation is still an aspiration rather than a well-defined and consistent policy.[12] Though the general objectives and the actors for both OS and (national) AI strategies are different, there are some overlaps and a notorious common interest in having access to data.

The dispersed features around digital data policies are characteristics of polycentric governance and governing. Polycentric theories address how scattered sites of governance function, or not.[13] The first theoretical developments concerning polycentric governance occurred with Michael Polanyi in 1951, when he studied science and the evolution of species along the lines of selforganization. Science as such was one of the first locus of polycentric governance conceptualizations. Polycentric theorizing later evolved into the well-known frameworks developed by Vincent and Elinor Ostrom, which addressed the different types of goods, mainly those of a public good nature that tend not to exclude its uses. Many of these are related with natural resource management. This provides a relevant analogy with the recent but contentious analogies of data as a natural resource.[14] Yet it is still relevant to acknowledge since it underscores the factiousness around data governance, including issues of data ownership, access, appropriation, protection, and management, which are all central to the tensions of natural resources.

In the context of the consolidated power of Big Tech,[15] it seems counterintuitive to ascertain digital data governance as polycentric. Yet, as developed by Kate Crawford, 'this large-scale capture (of data) has become so fundamental to the AI field that it is unquestioned. So how did we get here? What ways of conceiving data have facilitated this stripping of context, meaning, and specificity? … What forms of power do these approaches enhance and enable?'[16] How are these data-capture practices enabling certain forms of knowledge? What would expanding access to data mean for researchers in the context of increasing expansion of AI and more centres of power, authority, and legitimacy?

This chapter's fundamental objective is to discuss how access to data for research is being framed in different policy initiatives with an international reach and how should they evolve. There is a normative orientation that is concerned with this evolution, since there is an under-

standing that there are gaps to be attended to promote more equitable, just, and sustainable arrangements for the have-nots for science and research when data is out of reach for so many actors outside of Big Tech. More specifically, this chapter examines two existing policy instruments with a large presence in different countries and regions, with varied resources and capabilities: OS and AI strategies. This work probes polycentric governance concepts and how these different sites of governance and specific policy domains are a case for advancing a multi-pronged agenda on the conditions for access to digital data for scientific purposes.

OS initiatives and AI strategies have been chosen for their focus on access to data for science, development, and innovation. They are also present in many global majority countries. While open data strategies would also potentially touch upon this concern, their focus is on increasing government transparency and citizen and business engagement with mostly open-sector data. As the last decade has shown in terms of the open data movement, it is not enough for the public sector to open data, as the private sector holds the key to the largest data sets.[17] States in different parts of the world and different political regimes are concerned with the asymmetries between local actors' capacity to develop data-driven products on the one hand and develop innovation and economic growth on the other. That is, they have a competitive motivation to help firms in their countries develop the production of data-driven goods and services, and sometimes this is only possible with the integration of science and research synergies with the local business environment[18] in the EU and also beyond it.[19]

The research here is based on different data sources, including analysis of national and international policy documents concerning OS and national and regional AI strategies and plans, nine exploratory interviews, and participant observation in three workshops relevant to the theme (see Appendix 2A for details). The work uses an interpretative approach to identify and clarify meanings and understandings from existing policies. It is complemented with thematic analysis to identify possible integrations among themes.

The first section discusses the state of the art of the policy developments concerning data access for research in the OS movement and AI strategies and plans. The second section develops the conceptual framework that will be used based on polycentric governance theorizing in the digital field.[20] The third elaborates on the landscape of initiatives and themes against polycentric governing. The final section addresses knowledge and power concerns and the way forward for this debate.

Access to Data for Research: OS and AI Strategies

This section addresses two policy approaches concerning access to digital data for scientific purposes: OS and AI strategies. Both these instruments are found in global majority countries, though they are not ubiquitous. Rather than contesting the status quo of digital governance and large platforms, the ventures of OS share the concerns for data access and inequality but are framed with the fuzzier aim of the democratization of science. AI strategies, on the other hand, have interest in including alternatives that promote more equitable conditions for local actors to develop AI, and big data is a crucial need to train deep learning algorithms, even more so with the hype of fundamental models and generative AI applications.

Access to Data in the OS Movement

OS rests on several definitions which embrace different scopes and practices. New demands are emerging from academia and civil society in relation to data access to achieve socio-economic development, advance science beyond a corporate agenda, and address cognitive justice.[21] Ultimately, these different approaches share a common denominator in making knowledge development processes as open and accessible as possible.

> Open science is defined as inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community.[22]

According to the UNESCO (United Nations Educational, Scientific, and Cultural Organization) Recommendation on Open Science, this concept rests on the following pillars: open scientific knowledge, open science infrastructures, science communication, open engagement of societal actors, and open dialogue with other knowledge systems. Of relevance for this work is the notion of open scientific knowledge, which refers to open access to scientific publications, research data, metadata, open educational resources, software, and source code and hardware that are available in the public domain or under copyright and licensed under an open licence that allows access, reuse, repurpose, adaptation, and distribution under specific conditions, provided to all actors.[23]

Other definitions based on literature reviews define OS as 'transparent and accessible knowledge that is shared and developed through collaborative networks'.[24] The term 'knowledge' includes code, data, ideas, information, scientific outputs, scientific publications, and scientific results.[25] Sh. Moradi and S. Abdi conceive 'open science' as a novel approach to science and the process of its generation, monitoring, and dissemination, covering three sectors affected by the notion of 'openness' – that which should be incorporated in research data, in scholarly communication, and in access.[26]

The EU's OS policy is defined as an 'approach to the scientific process that focuses on spreading knowledge as soon as it is available using digital and collaborative technology'.[27] Concerning the data component, the European Commission requires research and innovation funding beneficiaries to make their publications available in open access and 'make their data as open as possible and as closed as necessary', acknowledging a perennial tension of the debate concerning intellectual property and commercialization tensions. The EU's approach stresses timeliness as a key attribute and added value of OS data access policies. In the case of Latin America, the COVID-19 pandemic enhanced the need to embrace a regional scope for OS policies.[28] The region has countries with strong policies concerning scientific data and access to publications (Mexico, Peru, Argentina) and recommendations on different OS dimensions in Chile, Brazil, Colombia, and Uruguay, particularly within their national research agencies but which have not become consolidated policies.[29] Africa has the Africa Open Science Platform

(AOSP), established in 2017, which is supported by several scientific and research authorities in South Africa and the region, such as Bibliotheca Alexandrina. For many African countries, OS remains confined to scientists and research networks rather than being a state policy.[30]

Data governance practices in OS are enshrined in two sets of principles: FAIR and CARE. The FAIR principles, meaning that data should be findable, accessible, interoperable, and re-usable, are a cross-cutting standard for all the data produced within the OS umbrella since they were adopted in 2016.[31] The publication of these principles helped encourage practices that made OS values more visible and applicable for data access for scientific research, and they have also been included in open data strategies. CARE (collective benefit, authority to control, responsibility, and ethics) principles were approved in 2019 and are broader than FAIR but allow to frame the context and purpose of data governance for indigenous data sovereignty,[32] which includes recognizing the different conceptions around science, technology, and culture. Indigenous data sovereignty 'refers to the right of Indigenous Peoples to govern the collection, ownership, and application of data about Indigenous communities, peoples, lands, and resources'.[33] Though the CARE principles are not invoked in the pastoralists' struggle in Ghana to secure their data rights (see chapter 3), they are struggling for their own self-determination concerning their data politics and practices.

OS has had significant effects on open access publications during the last 20 years.[34] Open access and open data publications are becoming increasingly mandatory in research environments with public funding, notably in the EU, Latin America, and Canada. Two broad approaches may be identified in the different policies concerning the availability of open scientific knowledge: publication in open access formats and repositories and open data initiatives for scientific endeavours. There are more nuanced interpretations with regard to the publication of accessible and open data emerging from research. For example, the Organization for Economic Co-Operation and Development (OECD)[35] mapped 15 open data initiatives for open science and 6 for publications, most of them concentrated in Europe and Latin America, the two regions where OS is more heavily discussed and practised.[36] There are fewer incentives to publish open data emerging from scientific research compared with defining the outlets for publishing in open access across the different national and regional initiatives that have been assessed in Latin America, Europe, and Canada.[37]

Several studies suggest that there is a lack of incentives for open science. Even though a vast majority of scientists support open science, few practise it, and there is a call for better enforcement.[38] The Registry of Open Access Repositories Mandatory Archiving Policies (ROARMAP) is an example to redress this issue where policymakers and funders have developed certain policies requiring that the research they have sponsored should be released under open access and open data conditions.

OS is approached differently, not only in diverse geographic settings but also within stakeholders and institutions. The COVID-19 pandemic has underscored the need for much greater collaboration among scientists, and a shared understanding of the policies and practices, including those related to data access, has become more urgent to help address future global crises.[39] The mix of concerns at the heart of the OS movement expose the different interests,

actors' values, and normative understandings. Research on this issue underscores the different policy objectives on how OS is framed, ranging from open data and open access as facilitators of OS to scientific efficiency, infrastructure enablers, strategic advantages, the role of the private sector, beneficiaries of OS, socio-economic progress as a response to global development and as a model to reduce dependencies on current structures.[40] These different issues have multiple centres of decision-making and authority that span national and regional science agencies, data protection authorities, government open data policies, the private sector, academic publishers, cloud providers, universities and research centres, and non-governmental organizations (NGOs). Unless there are clearly defined policies, follow-up mechanisms, and consultations processes nationally and internationally, it can become a buzzword for cosmetic approaches on openness in scientific practices.

Access to Data in AI Strategies and Policies

The integration of AI systems in society has become a concern and a priority for many governments which steer these debates in their national contexts and embed their own vision.[41] The so-called AI gap is characterized by those who have the capacity to design and implement AI applications which are largely based on computational capacity and power, access to relevant data, and skills and knowledge in AI.[42] Access to large amounts of high-quality data provides one of the greatest competitive advantages under the AI paradigm.[43]

During the second half of the 2010s, a growing number of countries and regions developed AI strategies, plans, or agendas as policy instruments, with varying degrees of involvement and participation from stakeholder groups outside government. AI strategies became policy artefacts that have helped frame, prioritize, and steer actions concerning the development, use, and regulation of AI systems in specific contexts. The first of these policy instruments emerged in Canada, but it was closely followed by both consolidated and emerging economies, mostly from Asia and Latin America from 2018.[44] In most cases, a 'first generation' of strategies may be identified between the first period between 2017 and 2020 that aimed to address many needs: (*a*) clarification of the fastpaced changes that were being brought about by AI systems and their potential future scenarios, and definition of roles and responsibilities; (*b*) capacity development to coordinate the stakeholder groups and defining the ecosystem of actors and issues involved; and (*c*) provision of a competitive advantage to the country and their position to compete in the 'AI race'. A 'second generation' of AI strategies was identified after 2021, when the UNESCO Recommendation on the Ethics of AI was approved by the General Assembly of this body, involving its 193 member states, the majority of which did not have any such instrument in place. This 'second generation' of AI strategies differs from the first wave in that they are more comprehensive and involve ethics, regulation, and development issues in the policy repertoire of these strategies in more explicit terms. Many countries with an existing AI strategy are embarking on a revision of these terms.[45]

As data plays such a pivotal role in ML models, its access and sharing are key considerations for sectors involved in the development and uptake of AI. In these AI strategies and related processes, countries continue to focus on providing access to civil society and firms to public sector data, not only open government data but also geodata and transportation data, as

well as data sharing within the public sector. It is becoming increasingly evident that national strategies are focusing on open data access policies and strategies not only for the original purposes of transparency, research, and innovation, but also more forcefully to promote the development of AI systems in those contexts.[46] Examples mapped by the OECD range from provision of access to weather, climate, and marine data following the example of the Danish Meteorological Institute; European co-operation on space data; the production of public data available in an open, reusable, and accessible format for ML in several countries across different regions (consolidated and emerging economies); the creation of several national AI policies plan to develop centralized, accessible repositories of open public data; and the creation of special agencies and institutions devoted to data, as is the case of national data institutes or similar organizations. The OECD AI observatory mapped 70 initiatives contained in different AI strategies and plans concerning data access and sharing as part of the policies defined as 'AI enablers'.

At the same time, countries and regional institutions such as the EU and Economic Commission for Latin America and the Caribbean (ECLAC) seek to incentivize and discuss data sharing in the private sector for AI systems development. For firms, there are some incentives connected with their data sharing for AI systems development, the consolidation of institutional network and ecosystems, open innovation purposes, and public 'good' research purposes. At the same time, firms lead 'data/AI for good initiatives' as part of the repertoire of instruments that are developed in partnership with other local actors (governments, universities, and civil society) mostly attending climate change and environment issues, followed by AI to achieve the Sustainable Development Goals (SDGs) and healthcare as the three main sectors in different national contexts.[47]

Two recent events have spurred the debates concerning data access for science and AI: the COVID-19 pandemic and the widespread emergence of generative AI applications for the larger public, with commercial applications such as ChatGPT, Midjourney, and Gemini. In the first case, AI systems proved to be instrumental in some areas, such as vaccine development.[48] However, despite the many diagnostic and predictive tools, most of them were flawed for various reasons.[49] These weaknesses concerned data that was of poor quality, not standardized, mislabeled, biased, and frequently from unknown sources. Another issue identified in the UNESCO report[50] was 'that frequently neither data nor training models were shared since academic researchers have commonly few career incentives to do so'. The gap in terms of access to vaccines and relevant treatments that could be developed with AI based on large volumes of data exposed the need to include scientists from less favoured environments. A similar perception of asymmetry in power and capacity concerned generative AI systems that have flooded the digital public sphere since the emergence of ChatGPT by the American firm OpenAI in November 2022. These Generative foundation models (GFMs) rely on vast quantities of data. The costs of processing such huge amounts of data to develop these systems and train these models would make it impossible for research ventures in middle- and low-income countries, as well as small and medium enterprises (SMEs), to compete.

Polycentric Governance Theories and Digital Data

The digital network environment is a contentious socio-technical infrastructure due to its increasing power asymmetries, open-ended characteristics, and commons spaces that are exploited by different types of private and state actors, which foster strong disciplinary and policy debates about global governance. Polycentric theories are relevant in governance studies due to their ability to provide insights and frameworks for understanding and analyzing the complexities of governance. This chapter expands on the theorizing of polycentric governance of digital data as one that provides a set of lenses that help tie together the wide range of actors, issues, and processes that are involved in access to digital data for research as a specific and contentious issue concerning digital data governance more broadly.[51] Polycentrism reveals multiple power centres and connections, including formal and informal arrangements, multiple scales (local-to-global), and different sectors (governmental, commercial, civil society, technical, academic).[52] The polycentric condition involves both dispersion and structure.[53] Polycentrism sheds light on both formal and informal attempts to address concerns surrounding data at different levels of activity and across sectors. It enables the identification and assessment of highly fluid arrangements emerging in global digital data governance.

Even in the specific domain of data governance concerning access to data for research, there are tensions, contradictions, and normative expectations which need to come to conversation. Polycentric perspectives help draw together a growing range of insights from different disciplines about digital data governance: how it occurs, how it might occur differently, and how it should occur. It also allows to pinpoint specific sources of tensions, rule formation, and mechanisms. Adopting polycentric perspectives help identify and connect key actors and location of authority across different levels – national, regional, international – in making sense of the complexities of global digital data governance and how to rebalance and question them from the point of view of the needs of researchers more widely.

Most significantly, polycentricity addresses the sources of authority and decision-making centres that are involved in the different data governance issues. Particularly in the case of common goods – that is, those that are more prone to excluding others from their consumption, misuse, and/or depletion – polycentricity allows to trace both micro and macro practices of power and different types of formal and informal authority that shape both data and its governance at a macro level. While data may not be fully considered a pure common good, the tensions surrounding its governance lie in the lowering of barriers to its access that has been part of a process of expansion of digitalization, datafication, and radical productive transformation.[54] The corporate sector, mainly the platformbased technological companies, have managed to capture the largest amounts of data globally.[55]

Up until a decade ago, polycentric governance theory was mainly related with the institutional perspectives from the Bloomington School and Elinor and Vincent Ostrom's work on common-pool resources, which have influenced the studies concerning digital data governance from a data commons approach and related institutional and normative practices, including data trusts and 'data sovereignty'.[56] With an international and complexity lens on polycentric governance, a more explicit focus emerges on the dynamics of power, legitimacy, authority,

cooperation, as well as more day-to-day dispersed forms of technical coordination by firms, citizens, civil society, and other actors.

Both data in its digital form and governance are open and contested notions. There is not one single approach to or understanding of what digital data governance is, either conceptually or from a policy perspective. How data is governed in the digital space is a polycentric issue. Some forms of data have achieved regulatory governance status, as is the case of data protection or data pertaining to intellectual property, though other functionalities of data, particularly those that concern research and innovation, are involved in more fluid notions where the dispersion of interests, actors, and regulatory possibilities at the national and international levels are much more debatable. Regulatory polycentric regimes[57] are defined similarly to how recent theoretical developments have expanded on polycentric theorizing to address transnational issues.[58] A crucial difference is that regulatory polycentric regimes as approached by Julia Black are developed as a set of sustained events to modify the behaviour of others,[59] while polycentric theorizing on digital data governance specifically[60] is at an earlier stage in the construction of a normativity for a regime, probably because of the need for further analysis on existence, overlap, and tensions of multiple regimes that operate on the spectrum of digital data. Polycentric governance involves different actor constellations operating with multiple rationalities, normative orientations, ethical concerns, technologies, and institutional arrangements. Digital data from this perspective is subject to different modes of knowledge, uses, and interpretations, both from a scholarly and policy perspective.

Polycentricity, OS and AI Strategies

In this section, the OS paradigm and the national AI strategies are discussed with a polycentric governance lens to assess their relevance for data access for research. Both policy approaches are examined against the characterization of specific sites of power as those that have the capacity to change other actors' behaviours, and those of authority as those that have influence on others based on some legitimacy claims, rooted in the Weberian distinction that is critical for polycentric governance theories from the Ostrom legacy to newer approaches.[61] This classification is relevant for this work as the policies that are examined – OS and AI strategies – are not binding laws and must navigate different levels, actors, ambiguous hierarchies, and diffused authorities, all of which are typical attributes of polycentric governance[62] which needs some form of legitimacy as a minimum baseline from which to influence behaviours, norms, and expectations.

Against a polycentric governance framework, these disparate arrangements are examined by connecting three different systemic organizing forces that frame the issue of access to data for research purposes from a polycentric governance lens. These systemic ordering forces are *norms*, *practices*, and *underlying orders*. They allow to move beyond an ontological characterization of data governance through a lens of polycentric attributes to one where the relations between these three forces and their interconnections are rendered. Norms may be defined as principles that inform the process of governing, and they guide ideas and behaviours on what is correct.[63] As a specific analytical point of entry, normative approaches take into account both ethical orientations and legal foundations. Practices are what people

do with varying degrees of consciousness in their everyday routines and the regulating effects of these actions on how 'things are done'.[64] Finally, 'underlying orders' are the most invisible and deeply embedded systemic forces that ultimately shape many of the conscious adoption of norms and unconscious practices in a polycentric arrangement. Underlying orders are systemic in that they permeate different connections in a polycentric governance regime.[65] They are powerful driving forces for practices and norms.

Table 2.1 summarizes the findings in the respective approaches with a polycentric framework. Both OS and AI strategies configure specific data interests and are nuanced particularly in what concerns the practices and the underlying orders. The increasing overlap of sites of power and authority in these two domains configures a cross-policy polycentric governance sector of digital data access.

Table 2.1 OS and AI strategies: summary

	Sites of power and authority	Systemic ordering force		
		Norms	**Practices**	**Underlying orders**
Open science	EU	FAIR principles	Support of open access journals	Innovation
	States	Intellectual property	Creation of open indexed repositories	Commercial interests
	UNESCO	Privacy	Generation of data repositories rather than finished publications	Altruistic motives: 'advancing humanity'
	Universities	Open data mandates	Training and education	Data science as new knowledge paradigm
	National research institutions	Safety and security		
	Publishing firms			
	Tech companies			
AI strategies and plans (national, regional)	Regional organizations	National competitiveness	Training and education	Techno-solutionism
	States	Market growth	Use cases	Utilitarianism
	UNESCO	Ethics	International benchmarks	Geopolitics
	OECD	Human rights	Data infrastructures	Innovation
	Tech/AI companies	Responsibility and trustworthiness	Coordination and governance	
	Universities	Data protection		
	Research agencies	Regulation		
	Civil society			

Source: Collated by the author.

Open Science

In terms of sites of governance and authority in OS, since 2021 the UNESCO framework has generated an international and quasi-global instrument that represents a global site of authority for its governance. Yet the OS movement faces strong challenges concerning the application of FAIR principles, 'particularly in the Global South, as countries need to transform the normative incentives within government bodies and funding agencies, as well as the evaluation models for researchers, institutions, and research programs'.[66] OS and data governance approaches in many of these countries need not only to update existing practices but also to change long-standing approaches on how data is conceptualized, including intellectual property issues. On the other hand, the application of FAIR principles to data becomes one of the most consistent practices to guarantee the implementation of OS and is also unequivocally polycentric as claimed by their authors: 'The resulting data ecosystem, therefore, appears to be moving away from centralization, is becoming more diverse, and less integrated, thereby exacerbating the discovery and re-usability problem for both human and computational stakeholders.'[67] While there are inherent tensions that aim to centralize data for historical reasons motivated by intellectual property concerns and even security, greater demands placed on transparency and democratization principles invoke a different perception of the problem.

The multiplicity of policies concerning OS indicates the need for a suitable platform for its implementation in the scientific community; on the other hand, the diversity of these supportive policies reveals the wide range of subjects and the array of the OS ecosystem. Therefore, the distinctive features of the scientific community of the country in question should be considered. Realizing such a systemic change involves an internationally coordinated effort of researchers, universities, research institutes, publishers, research councils, and policymakers.[68]

Ghislaine Chartron spelled out three different and convergent orders of values and beliefs that underly the OS paradigm: 'the *ethical values* of science as defined by Merton (1942 [1973]), *data science* as a new scientific paradigm, and *innovation* as a source of economic growth'.[69] According to this vision, the innovation system is at the heart of the promotion of OS. Knowledge production must cross-fertilize scientific fields, companies, and governments.[70] Possible transfers between science, the economic fabric, and civil society to gain growth and confidence in science become the basic motivation. 'When these three regimes come face to face, a certain confusion gradually sets in. Hardly reconcilable visions are being projected, such as the defense of the common good of science and the quest to create value for the socio-economic world.'[71] These visions underscore the inherent tensions in a polycentric domain with different underlying orders.

Moradi and Abdi's research focused on European OS policies found that these prioritized open access publications, followed by open research data.[72] Their work also noted the conflation of the notion of open access to data included in publications, rather than data for research. Part of the normative underpinnings of access to data by the OS movement concerns different framings around intellectual property provisions and privacy regulations when personal data

is involved, but most relevantly the incentives for researchers to generate open data repositories. While Latin America aims to increase uptake and foster open access indexed journal repositories, the application of the FAIR principles to data has not yet been substantively accomplished.[73]

AI Strategies and Plans

AI strategies tend to approach data governance as a vital resource for local AI development.[74] The creation of large datasets is essential for the development of AI, sector-specific innovation, and capacity creation. Local research communities have been shown to be key actors in many contexts in the formulation of these policy instruments, but their concerns were not met in the first generation of AI strategies in most countries.

AI plans that formulate access to data for research have a strong competitive incentive. Practically all strategies address access to data for research with a utilitarian and geopolitical understanding, concerning the 'AI race' where indicators about number of standards, conference papers, and the like are routinely compared.[75] There are few interdisciplinary approaches or concerns in AI strategies for non-computer approaches to data and AI. Access to data for research is not even considered as part of the repertoire of many AI strategies, and these interests are included in other policy instruments, notably in the science sector.[76] In a recent regional AI index covering 12 countries in Latin America,[77] access to relevant data is confined to the dimensions of the Global Data Barometer which are more relevant for an open government strategy rather than for comprehensive national data strategies that are needed in AI and OS.

The sites for power and authority for AI plans have been developed by different organizational bodies, steered mostly by national and regional governments and supranational organizations such as the OECD and UNESCO. The OECD published in 2019 its document on principles on AI, which was endorsed by other countries outside of this organization. The document maps the development of, and access to, a digital ecosystem for trustworthy AI that includes digital technologies and infrastructure 'and mechanisms for sharing AI knowledge, as appropriate. In this regard, governments should consider promoting mechanisms, such as data trusts, to support the safe, fair, legal, and ethical sharing of data'.[78] Karen Yeung assesses the OECD recommendation as treading between two underlying forces with normative approaches: the ethical imperative to provide safeguards to the rise of AI risks but also as enablers 'to avoid killing the golden goose of technological innovation'.[79]

The UNESCO Recommendation on the Ethics of AI has spurred a more specific normative orientation on the ethics of AI as the centrepiece of this recommendation. It has become the only document with a global reach and though it is not legally binding, it is an authoritative instrument. UNESCO is pursuing this authoritativeness by implementing a readiness assessment methodology (RAM) instrument, which in 2023 spanned over 50 countries, many of which already had national AI plans in place but were looking to further enhance the ethical dimensions and the ensuing policies that should be implemented. The increasing regulatory and legislative measures related with AI development, as is the case of the AI Act of 2024

in the EU as well as the UNESCO initiative on the RAM, are promoting a normative orientation towards AI systems that fully incorporates binding legal measures.

From an underlying orders perspective, utilitarian perspectives predominate in AI strategies as their socio-technical imaginaries[80] are imbued by notions of supremacy in AI capacity, the AI race, and cybernetic futures.[81] Utilitarian approaches look at the maximization of benefits for 'a majority', which have greater difficulty in matching these expectations on the consequences and outcomes of data-driven AI systems at broader social levels. The solutionist approach of national and regional AI strategies is another fundamental underlying order that is aligned with utilitarian values. 'Solutionist beliefs are a particularly important part of the spirit of digital capitalism, which we define as those normative beliefs that play a legitimizing, motivating and orienting role for today's tech companies.'[82] These beliefs are present in AI strategies and inform not only how tech and AI firms see themselves but also how they are perceived by others, notably governments which steer the developments of these national plans. While there is an increasing normative orientation on the inclusion of societal risks and the ethical and human rights principles that should be the foundations of AI plans, their underlying orders are still at odds as the geopolitics of AI is still predominant.

Conclusion

The aim of the chapter has been to develop a case for a more holistic and relational approach to the issue of access to data for research purposes, beyond existing possibilities as those portrayed in the DMA–DSA package, by discussing different data regimes found in many contexts, including global majority regions, with a polycentric theory. This chapter adopted a normative orientation concerning the relevance of access to data for research purposes and developed a case to address this by encompassing two existing policies – OS and AI strategies – as policy alternatives to further advance other scientific interests and uses from a polycentric perspective. This normative stance was adopted as an intellectual response to the increasing centrality of data-driven AI technologies and large platforms power vis-à-vis local research capacity.

A polycentric lens enabled the connection of different sites of power and authority, as well as different normative, practical, and underlying orders of these arrangements. The different policy and regulatory initiatives involving access to digital data are a manifestation of broader theoretical, political, and sociotechnical intersections.[83] Currently, available policy initiatives should address these different centres of power and authority from these specialized fields and build bridges as access claims are shared. This recommendation is not concerned with promoting centralization, that is, hierarchically concentrating power and authority for the different implications concerning digital data and access for research, but rather about developing pathways across these different policy domains, increasing coordination and coherence. This is particularly relevant for many countries in global majority contexts that have consolidated their own trajectories in some domains – for example, in Latin America with national open data strategies and open science initiatives and in Asia with consolidated national AI strategies – but whose regulations will be difficult to implement as those like the DMA–DSA in Europe. The involvement of the global, regional, and national levels in the analysis is the key to understanding how access to data for research is framed and how it should evolve.

Since 2021, there is a more pronounced global configuration around OS and AI strategies led by UNESCO with its two recommendations approved in that same year. These global policy instruments are not only relevant as additional sites of international governance but also more particularly since they provide common ground to conversations that would otherwise be disconnected. This international expansion also spurs the involvement of middle- and lower-income countries. Without a global and mobilized research community, specific instruments such Article 40 of the DSA that is devised for a particular region would not have so much power and authority across other national governance sites.

Policymaking at a national level is essential for the development of data access policies for research. Like other policy innovations, it will require adjustments and redefinitions of governance structures, including reviewing or formulating policies that are critical drivers for implementations in the OS domain, in the AI policies space, and in the complementary strategies that should be considered to address a polycentric issue.[84] Some of these policies will become legally coded, but others will be addressed through practices, as has been the case of the expansion of the FAIR and CARE principles for data and the open repository initiatives that rely on organic processes and different timings.

Polycentric governance theorizing proved to be a relevant lens to frame the different sites of power and authority that are embedded in OS and AI policy initiatives and their connections with other initiatives to reframe the current landscape of data control around large platforms. It helps to develop a position around difference and negotiation, rather than centralization of power as the only alternative solution to either Big Tech predominance or governmental control. Polycentricity also embraces the institutional complexity concerning digital data governance, including the political economy of digital data commons, in an era of mass web scraping to train fundamental AI models, which is one of the major contentions in the contemporary debates about the systemic power of large corporations with implications for data protection and intellectual property concerns.

Two substantive differences between OS and AI national plans stand out in the redefinition of data governance rules. The first is concerned with the conditions for research and how scientists make their own data available, structural access, and the publishing of data. OS and principles such as FAIR and CARE reconfigure the norms and underlying structures of research and data. It has a meta-governance configuration where the norms and practices are subverted to incorporate new conditions 'about' research and access to data. This has implications for the wider ecosystem and governance sites. AI strategies, on the other hand, are more inclined to develop a tactical approach towards data concerned with conditions 'for' research. The provocation is extended to rethink beyond a framing of OS in relation to science policy[85] and to develop a triangular dialogue between OS, science policy, and AI policy to advance access to data for research, among other objectives. This is a vital link where the research community should play an active role in the reconfiguration of norms, practices, and underlying orders about contemporary notions around digital data and knowledge.

Future work could engage more thoroughly on issues concerning 'data access' for research from a justice, sovereignty and sustainability lens. The problems concerning access are rele-

vant as they not only encompass a range of dimensions, including data infrastructures, some of which already have polycentric attributions, and their availability, but also capabilities and skills of local communities and epistemic power. The reification of data and datafication processes should also be examined against more thorough philosophical positions concerning the construction of legitimacy of certain types of knowledge, as the introduction to this volume by Jef Ausloos and Siddharth Peter de Souza has highlighted.[86] Further studies could also address the different disciplinary stakes on the issue of access to data for research, including a global majority perspective that situates practices of science and innovation policies and their implications for sovereign governance capacities.

Appendix 2A

The following public forums and events informed this research through a modality of participant observation: (*a*) 4 May 2023: virtual discussion organized by UNESCO on the readiness assessment methodology (RAM) published by this organization;[87] (*b*) 10–11 May 2023: workshop convened by the National Endowment for Democracy (NED) (Buenos Aires) as part of a consultation on AI and its impact on democracy and human rights; and (*c*) 28–30 June 2023: roundtables organized by the Agency for Electronic Government and Information and Knowledge Society (AGESIC), the national agency for digital government of Uruguay, as part of the revision of the national AI strategy. These spaces were nurtured by the presence of policymakers, experts, and advocates from Argentina, Brazil, Egypt, and Uruguay and organizations such as UNESCO and the Development Bank of Latin America and the Caribbean (CAF), and civil society participants from Canada, Poland, the United States, and various countries in Latin America.

Anonymized Interviews

Senior executive, National Agency for Research and Innovation (ANII), Uruguay (in person), March 2023.
Former advisor to the Ministry of Communications and Information Technology (MCIT), Egypt (virtual), April 2023.
Representative, Latin American digital rights NGO (in person), May 2023.
Senior executive, UNESCO, Paris (virtual), May 2023.
Director, UNESCO regional office (in person), April 2023.
Senior executive, UNESCO, Paris (virtual), May 2023.
Representative, AGESIC, Uruguay (in person), May 2023.
Representative, Ministry of Science and Technology, Argentina (in person), April 2023.
Secretariat for open data, Argentina (in person), March 2023.

Notes

1 O. Nachtwey and T. Seidl, 'The Solutionist Ethic and the Spirit of Digital Capitalism', 13 March 2020, https://osf.io/preprints/socarxiv/sgjzq (accessed 25 June 2023); Robert Gorwa, 'What Is Platform Governance?' *Information, Communication and Society* 22, no. 6, DOI: https://doi.org/10.1080/13691 18X.2019.1573914; S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2019); E. Ruppert, E. Isin, and D. Bigo, 'Data Politics', *Big Data and Society* 4 (2017), DOI: https://doi.org/10.1177/2053951717717749; Danah Boyd and K. Crawford, 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon', *Information, Communication and Society* 15, no. 662 (2012): 662–679.

2 Gry Hasselbalch, *Data Ethics of Power* (Edward Elgar Publishing, 2021), 2.

3 J. Brase and A. Farquhar, 'Access to Research Data', *D-Lib Magazine* 17, nos. 1–2 (2011), http://www.dlib.org/dlib/january11/brase/01brase.html (accessed 3 March 2024); A. Nagaraj, E. Shears, and M. de Vaan, 'Improving Data Access Democratizes and Diversifies Science', *Proceedings of the National Academy of Sciences* 23490 (2020), DOI: https://doi.org/10.1073/ pnas.2001682117.

4 S. A. Aaronson, 'Data Is Dangerous: Comparing the Risks that the United States, Canada and Germany See in Data Troves', Working Papers, 2020, https://ideas.repec.org//p/gwi/wpaper/2020-5.html (accessed 16 June 2021); Kai-Fu Lee, *AI Superpowers* (Houghton Mifflin Harcourt, 2018); A. S. Obendiek, *Data Governance: Value Orders and Jurisdictional Conflicts* (Oxford University Press, 2023).

5 Benj Edwards, 'Sites Scramble to Block ChatGPT Web Crawler after Instructions Emerge', *Ars Technica*, 12 August 2023, https://arstechnica. com/information-technology/2023/08/openai-details-how-to-keep-chatgptfrom-gobbling-up-website-data (accessed 12 October 2023); 'New York Times May Sue OpenAI over Chat GPT Data Scraping', Yahoo Finance, 12 September 2023, https://finance.yahoo.com/news/york-times-may-sueopenai-150048157.html (accessed 12 October 2023).

6 A. Haleem, M. Javaid, and R. P. Singh, 'An Era of ChatGPT as a Significant Futuristic Support Tool: A Study on Features, Abilities, and Challenges', *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 2, no. 4 (2022), DOI: https://doi.org/10.1016/j.tbench.2023.100089.

7 J. Ausloos, P. Leerssen, and Pim ten Thije, 'Operationalizing Research Access in Platform Governance: What to Learn from Other Industries?' Algorithm Watch, https://algorithmwatch.org/de/wp-content/uploads/2020/06/ GoverningPlatforms_IViR_study_June2020-Algorithm-Watch-2020-06-24. pdf (accessed 2 February 2023).

8 J. Black, 'Constructing and Contesting Legitimacy and Accountability in Polycentric Regulatory Regimes', *Regulation and Governance* 137, no. 2 (2008): 137–164, DOI: https://doi.org/10.1111/j.1748-5991.2008.00034.x.

9 L. Taylor, 'What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally', *Big Data and Society* 4, no. 2 (2017), DOI: https:// doi.org/10.1177/2053951717736335; Joan Lopez Solano, Aaron Martin, Siddharth de Souza, and Linnet Taylor, *Governing Data and Artificial Intelligence for All: Models for Sustainable and Just Data Governance* (European Parliament, 2022).

10 M. Husovec, 'How to Facilitate Data Access under the Digital Services Act', 19 May 2023, https://papers.ssrn.com/abstract=4452940 (accessed 8 July 2023).

11 A. J. Carrillo and M. Jackson, 'Follow the Leader? A Comparative Law Study of the EU's General Data Protection Regulation's Impact in Latin America', *Vienna Journal on International Constitutional Law (ICL Journal)* 177, no. 16 (2022): 177–262, DOI: https://doi.org/10.1515/icl-2021-0037; T. M. Bueno and R. G. Canaan, 'The Brussels Effect in Brazil: Analysing the Impact of the EU Digital Services Act on the Discussion Surrounding the Fake News Bill', *Telecommunications Policy* 48, no. 5 (2024), DOI: https:// doi.org/10.1016/j.telpol.2024.102757.

12 E. G. Pacchioni, W. Maloney, and X. Cirera, 'Why Poor Countries Invest Too Little in R&D', Centre for Economic Policy Research (CEPR), 29 November 2017, https://cepr.org/voxeu/columns/why-poor-countries-invest-too-littlerd (accessed 4 June 2023).

13 Frank Gadinger and Jan Aart Scholte, *Polycentrism: How Governing Works Today* (Oxford University Press, 2023).

14 Lisa Gitelman (ed.), *'Raw Data' Is an Oxymoron* (MIT Press, 2013).

15 Joan López, Aaron Martin, Franklyn Ohai, Siddharth Peter De Souza, and Linnet Taylor, 'Digital Disruption or Crisis Capitalism? Technology, Power and the Pandemic', Global Data Justice, 11 May 2022, https:// globaldatajustice.org/gdj/2649 (accessed 8 June 2023).

16 K. Crawford, 'Data: From the Atlas of AI', Missing Links in AI Governance, UNESCO, 2021, https://www.unesco.org/en/articles/missing-links-aigovernance (accessed 6 June 2023).

17 T. Boellstorff, 'Making Big Data, in Theory', *First Monday* 18, no. 10 (2013), DOI: https://doi.org/10.5210/fm.v18i10.4869; J. van Dijck, D. Nieborg, and T. Poell, 'Reframing Platform Power', *Internet Policy Review* 8 (2019), https://policyreview.info/node/1414 (accessed 13 November 2022); Zuboff, *The Age of Surveillance Capitalism*; S. Baack, 'Datafication

and Empowerment: How the Open Data Movement Re-Articulates Notions of Democracy, Participation, and Journalism', *Big Data and Society* 2, no. 2 (2015), DOI: https://doi.org/10.1177/2053951715594634.

18 UNESCO, 'Recommendation on the Ethics of Artificial Intelligence: UNESCO Biblioteca Digital', 2021, https://unesdoc.unesco.org/ark:/48223/ pf0000380455 (accessed 10 April 2023).

19 Índice Latinoamericano de Inteligencia Artificial (ILIA), 'Índice Latinoamericano de Inteligencia Artificial' (ILIA, 2023).

20 Maria Koinova, Maryam Zarnegar Deloffre, Frank Gadinger, Zeynep Sahin Mencutek, Jan Aart Scholte, and Jens Steffek, 'It's Ordered Chaos: What Really Makes Polycentrism Work', *International Studies Review* 1988, no. 23 (2021): 1988–2018, DOI: https://doi.org/10.1093/isr/viab030; Carolina Aguerre, Malcolm Campbell-Verduyn, and Jan Aart Scholte, *Digital Data Governance: Polycentric Perspectives* (Routledge, 2024).

21 Leslie Chan, Angela Okune, Rebecca Hillyer, Denisse Albornoz, Alejandro Posada, 'Contextualizing Openness: Situating Open Science' (University of Ottawa Press, 2019), DOI: https://doi.org/10.20381/ruor-24088.

22 'UNESCO Recommendation on Open Science', UNESCO, 2021, https:// unesdoc.unesco.org/ark:/48223/pf0000379949 (accessed 15 February 2023), 7.

23 'UNESCO Recommendation on Open Science', 8.

24 R. Vicente-Saez and C. Martinez-Fuentes, 'Open Science Now: A Systematic Literature Review for an Integrated Definition', *Journal of Business Research* 428, no. 88 (2018): 428–436.

25 Alejandra Manco, 'A Landscape of Open Science Policies Research', *SAGE Open* 12, no. 4, DOI: https://doi.org/10.1177/21582440221140358.

26 Sh. Moradi and S. Abdi, 'Open Science–Related Policies in Europe', *Science and Public Policy* 50 (2023): 521–530, DOI: https://doi.org/10.1093/scipol/ scac082.

27 'Open Science', European Commission, 10 February 2023, https://researchand-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digitalfuture/open-science_en (accessed 1 March 2023).

28 D. Babini and L. Rovelli, 'Tendencias Recientes En Las Políticas Científicas de Ciencia Abierta y Acceso Abierto En Iberoamérica on JSTOR', Latin American Council of Social Sciences (CLACSO), 2020, https://www-jstororg.eza.udesa.edu.ar/stable/j.ctv1gm02tq (accessed 31 May 2023).

29 Babini and Rovelli, 'Tendencias Recientes'.

30 E. R. T. Chiware and L. Skelly, 'Open Science in Africa: What Policymakers Should Consider', *Frontiers in Research Metrics and Analytics* 7, https://www. frontiersin.org/articles/10.3389/frma.2022.950139 (accessed 21 October 2023).

31 Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, and Barend Mons, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data* 7, no. 3 (2016), DOI: https://doi. org/10.1038/sdata.2016.18.

32 Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson, 'The CARE Principles for Indigenous Data Governance', *Data Science Journal* 19, DOI: https://doi.org/10.5334/dsj-2020-043.

33 S. Rainie, T. Kukutai, M. Walter, O. Figueroa-Rodriguez, J. Walker, and P. Axelsson, 'Issues in Open Data: Indigenous Data Sovereignty', in *State of Open Data: Histories and Horizons*, ed. Tim Davies, Stephen B. Walker, Mor Rubinstein, and Fernando Perini, 300–319 (African Minds and International Development Research Center, 2019), 301.

34 Janne Pölönen, Mikael Laakso, Raf Guns, Emanuel Kulczycki, and Gunnar Sivertsen, 'Open Access at the National Level: A Comprehensive Analysis of Publications by Finnish Researchers', *Quantitative Science Studies* 1 (2020): 1396–1428, DOI: https://doi.org/10.1162/qss_a_00084.

35 A. Paic, 'Open Science: Enabling Discovery in the Digital Age', OECD, 20 July 2021, https://www.oecd-ilibrary.org/science-and-technology/openscience-enabling-discovery-in-the-digital-age_81a9dcf0-en (accessed 11 June 2023).

36 Manco, 'A Landscape of Open Science Policies Research'.

37 Paic, 'Open Science'.

38 Moradi and Abdi, 'Open Science–Related Policies in Europe'; Manco, 'A Landscape of Open Science Policies Research'; M.D. de Rosnay and F. Stalder, 'Digital Commons', *Internet Policy Review* 9, no. 4, DOI: 10.14763/2020.4.1530.

39 'G7 Science and Technology Ministers Commit to Open Science', Ouvrir la Science, 15 May 2023, https://www.ouvrirlascience.fr/g7-science-andtechnology-ministers-commit-to-open-science (ACCESSED 12 June 2023).

40 Denisse Albornoz, Maggie Huang, Issra Marie Martin, Maria Mateus, Aicha Yasmine Touré, and Leslie Chan, 'Framing Power: Tracing Key Discourses in Open Science Policies', in *22nd International Conference on Electronic Publishing*, 1–21 (OpenEdition Press, 2018).

41 J. Bareis and C. Katzenbach, 'Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics', *Science, Technology, and Human Values* 855, no. 47 (2022), DOI: https://doi. org/10.1177/01622439211030007.

42 'IDRC Global Symposium on AI & Inclusion Outputs', International Development Research Centre, January 2018, https://idl-bnc-idrc. dspacedirect.org/handle/10625/56851?locale-attribute=en (accessed 9 February 2023).

43 Lee, *AI Superpowers*.

44 'OECD's Live Repository of AI Strategies & Policies: OECD.AI', OECD, https://oecd.ai/en/dashboards/overview (accessed 23 February 2023).

45 R. Jorge Ricart, V. Van Roy, F. Rossetti, and L. Tangi, 'AI Watch: National Strategies on Artificial Intelligence – A European Perspective: 2022 Edition', JRC Publications Repository, 31 May 2022, https://publications.jrc.ec.europa.eu/repository/handle/JRC129123 (accessed 6 June 2023).

46 Intergovernmental representative, personal communication, May 2023.

47 V. Aula and J. Bowles, 'Stepping Back from Data and AI for Good: Current Trends and Ways Forward', *Big Data and Society* 10, no. 1 (2023), DOI: https://doi.org/10.1177/20539517231173901.

48 Y. Bengio and A. Oh, 'AI for Public Domain Drug Discovery', Global Partnership on Artificial Intelligence, November 2021, https://gpai.ai/ projects/ai-and-pandemic-response/public-domain-drug-discovery/ai-forpublic-domain-drug-discovery.pdf (accessed 4 June 2023).

49 S. Ziesche, 'Open Data for AI: What Now?', UNESCO, 2023, https:// unesdoc.unesco.org/ark:/48223/pf0000385841 (accessed 4 June 2023).

50 Ziesche, 'Open Data for AI.

51 Aguerre, Campbell-Verduyn, and Scholte, *Digital Data Governance*.

52 Koinova, Deloffre, Gadinger, Mencutek, Scholte, and Steffek 'It's Ordered Chaos'.

53 Gadinger and Scholte, *Polycentrism*.

54 S. A. Aaronson, 'Data Is Different: Why the World Needs a New Approach to Governing Cross-Border Data Flows', CIGI Paper No. 197, Centre for International Governance Innovation (CIGI), 2018; M. Flyverbom, R. Deibert, and D. Matten, 'The Governance of Digital Technology, Big Data, and the Internet: New Roles and Responsibilities for Business', *Business and Society* 3, no. 58 (2019): 3–19, DOI: https://doi. org/10.1177/0007650317727540.

55 V. Lehdonvirta, *Cloud Empires: How Digital Platforms Are Overtaking the State and How We Can Regain Control* (MIT Press, 2022).

56 S. Delacroix and N. D. Lawrence, 'Bottom-up Data Trusts: Disturbing the "One Size Fits All" Approach to Data Governance', *International Data Privacy Law* 236, no. 9 (2019): 236–252, DOI: https://doi.org/10.1093/ idpl/ipz014; 'Participatory Data Governance', Ada Lovelace Institute, https://www.adalovelaceinstitute.org/project/participatory-data-governance (accessed 6 June 2023); Anouk Ruhaak, Greg Bloom, Angie Raymond, Willa Tavernier, Divya Siddarth, Gary Motz, and Melanie Dulong de Rosnay, 'A Practical Framework for Applying Ostrom's Principles to Data Commons Governance', Mozilla, 6 December 2021, https://foundation.mozilla.org/ en/blog/a-practical-framework-for-applying-ostroms-principles-to-datacommons-governance (accessed 6 June 2023); A. Powell, R. Johnson, and R, Herbert, 'Achieving an Equitable Transition to Open Access for Researchers in Lower and Middle-Income Countries [ICSR Perspectives]', 11 June 2020, https://papers.ssrn.com/abstract=3624782 (accessed 6 June 2023).

57 Black, 'Constructing and Contesting Legitimacy and Accountability', 138.

58 Gadinger and Scholte, *Polycentrism*; Koinova, Deloffre, Gadinger, Mencutek, Scholte, and Steffek, 'It's Ordered Chaos'.

59 Black, 'Constructing and Contesting Legitimacy and Accountability', 39.

60 Aguerre, Campbell-Verduyn, and Scholte, *Digital Data Governance*.

61 Gadinger and Scholte, *Polycentrism*.

62 J. A. Scholte, 'Polycentrism and Democracy in Internet Governance', in *The Net and the Nation State: Multidisciplinary Perspectives on Internet Governance*, ed. Uta Kohl, 165–184 (Cambridge University Press, 2017).

63 Koinova, Deloffre, Gadinger, Mencutek, Scholte, and Steffek 'It's Ordered Chaos'.

64 Koinova, Deloffre, Gadinger, Mencutek, Scholte, and Steffek 'It's Ordered Chaos'.

65 Gadinger and Scholte, *Polycentrism*.

66 Former government advisor, personal communication, May 2023.

67 Wilkinson, Dumontier, Aalbersberg, Appleton, Axton, Baak, Blomberg, Boiten, Santos, Bourne, Bouwman, Brookes, Clark, Crosas, Dillo, Dumon, Edmunds, Evelo, Finkers, Gonzalez-Beltran, Gray, Groth, Goble, Grethe, and Mons, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'.

68 Moradi and Abdi, 'Open Science–Related Policies in Europe', 529.

69 G. Chartron, 'L'Open science au prisme de la Commission européenne', *Éducation et sociétés* 177, no. 41 (2018): 177–193 (emphasis added).

70 Chartron, 'L'Open science au prisme de la Commission européenne'.

71 Chartron, 'L'Open science au prisme de la Commission européenne'.

72 Moradi and Abdi, 'Open Science–Related Policies in Europe'.

73 Government agency representative, personal communication, March 2023.

74 C. Aguerre, 'National AI Strategies and Data Governance', in *Artificial Intelligence in Latin America: Ethics, Governance and Policies*, vol. 1, ed. C. Aguerre (CETYS, Universidad de San Andres, 2020).

75 Lee, *AI Superpowers*.

76 Intergovernmental representative, personal communication, 4 May 2023.

77 ILIA, 'Índice Latinoamericano de Inteligencia Artificial'.

78 'OECD Recommendation on Competition Assessment: OECD', OECD, 2019, https://www.oecd.org/daf/competition/oecdrecommendation oncompetitionassessment.htm (5 June 2023).

79 K. Yeung, 'Algorithmic Regulation: A Critical Interrogation', *Regulation and Governance* 12, no. 4 (2017): 17.

80 S. Jasanoff and S. H. Kim, 'Sociotechnical Imaginaries and National Energy Policies', *Science as Culture* 189, no. 22 (2013): 189–196.

81 Bareis and Katzenbach, 'Talking AI into Being'.

82 Nachtwey and Seidl, 'The Solutionist Ethic and the Spirit of Digital Capitalism', 3.

83 Aula and Bowles, 'Stepping Back from Data'.

84 Moradi and Abdi, 'Open Science–Related Policies in Europe'; C. Djeffal, M. B. Siewert, and S. Wurster, 'Role of the State and Responsibility in Governing Artificial Intelligence: A Comparative Analysis of AI Strategies', *Journal of European Public Policy* 1799, no. 29 (2022): 1799–1821, DOI: https://doi. org/10.1080/13501763.2022.2094987; R. Radu, 'Steering the Governance of Artificial Intelligence: National Strategies in Perspective', *Policy and Society* 178, no. 40 (2021): 178–193, DOI: https://doi.org/10.1080/14494035.202 1.1929728.

85 Manco, 'A Landscape of Open Science Policies Research'.

86 P. Ricaurte, 'Data Epistemologies, the Coloniality of Power, and Resistance', *Television and New Media* 350, no. 20 (2019), DOI: https://doi. org/10.1177/1527476419831640.

87 UNESCO, *Readiness Assessment Methodology: A Tool of the Recommendation on the Ethics of Artificial Intelligence*, DOI: https://doi.org/10.54678/ YHAA4429.

3. VIOLENT PLAINS: CHALLENGES AND STRATEGIES FOR PASTORALISTS' DATA ACCESS IN GHANA [1]

FRANK KWAKU AGYEI, LAWRENCE KWABENA BROBBEY, PAUL OSEI-TUTU, AND BOATENG KYEREH

The conflict between pastoralists (cattle farmers) and crop farmers constitutes a major source of natural resource conflict in Africa. It was initially thought that the declining availability of grass and water resources and the complex land tenure system in Africa drive the conflict.[2] The depletion of grasslands – to feed cattle – results in competition between farmers and cattle herders for control over grass and farmlands, and that results in intense clashes and violent confrontations.[3] Several instances of the conflict have been reported to show its growing intensity, geographic scope, and effects of displacements, losses, and deaths.[4] Subsequent works in the humanities and social sciences have suggested that the violent clashes between farmers and herders are a result of differences in their ethnic and identity backgrounds – typical in Nigeria where the conflict occurs between Muslim or Fulani herders and Christian farmers.[5] More recent studies suggest political marginalization through land use politics favouring crop farming over pastoralism creates power imbalances and sustains tensions between farmers and herders to increase the prevalence of the conflict.[6]

In Ghana, the conflict has gained considerable media attention and engendered continuing public debate. Despite ongoing efforts by the Ghana government to address the conflict, there are widespread claims of forced population displacements, losses, and destructive reactions. In 2016, youth in southern Ghana killed 80 cattle belonging to Fulani herdsmen when the cattle invaded their farms, and two farmers were shot dead in their farms by Fulani herdsmen in Agogo. Between 2009 and 2012, a total of 12 people were killed in Agogo town.[7] In Berekum (southern Ghana), hundreds of armed residents attacked and evicted herdsmen from their homes, which resulted in the death of 13 herdsmen and 4 farmers.[8] In northern Ghana, a series of fights between Konkomba farmers and nomadic herders resulted in over 562 families becoming homeless as well as loss of livestock, money, and other properties.[9] The trauma and psychological effects of experiences with conflicts hinder farmers and herders from rebuilding their lives and livelihoods as the skirmish weakens the social capital of communities, and these experiences negatively impact the roles that men and women play in societies.

The media is often criticized for being biased towards pastoralists in the framing of the conflict in news reportages, leading to their stigmatization. Journalists covering the majority of Fulani pastoralists' stories rely on secondary accounts and rumoured data. Most journalists are not eyewitnesses to the issues they raise, and there is homogenization of Fulani ethnic groups with herdsmen when individual herdsmen are being referred to.[10] Often, pastoralists are portrayed as 'foreigners' and 'treacherous' by news media outlets, and that has contributed to an unfriendly and unfair discussion against herders and perpetuated an unfavourable opinion of them in Ghanaian society.[11] Through the use of the language of alienation and frames that stigmatize, herders are put at the margins of Ghanaian society. Pastoralist communities

are being marginalized, and this has resulted in their passiveness and poor participation in social interventions, particularly programmes that seek access to their data. This undermines possibilities for generating holistic, secure, and scalable data critical to rebuilding pastoralist heritage and dignity among pastoralists and farmers.

Much scholarly attention has been devoted to structural explanations of farmer–herder conflicts, pastoralism and rural economies, political economy associated with pastoralism, and institutional interference. The evidence indicates that transhumance and nomadic herders are the main culprits of the conflict. They live in camps in bushy areas and engage in grazing of cattle at night. Further, studies suggest that many herdsmen and their families speak Fulbe and/or other non-Ghanaian languages,[12] and hence poorly integrate with local Ghanaian households. The need to integrate perspectives of nomadic and transhumance herders in conflict management initiatives is widely recognized. Yet pastoralism literature has not systematically accounted for the challenges associated with getting access to pastoralists' data. Data access denotes having access rights in a legal or political sense as well as activities relating to collecting, storing, and acting on data housed in a database.[13] Understanding the challenges as well as the strategies and opportunities associated with pastoralism-related data access is an important starting point for improving the well-being and peace among farmers and herders in Ghana and beyond. In this chapter, we examine (*a*) challenges associated with pastoralists' data access, (*b*) strategies employed by researchers and development actors to get access to pastoralists' data, and (*c*) opportunities for pastoralists' data access. The next section outlines the methods used for data collection and analysis. This is followed by sections on the results and discussion.

Methods

The study targeted six categories of people for data collection: researchers (universities), non-governmental organizations (NGOs), cattle owners, herdsmen, government, and other key informants in the cattle industry (Table 3.1). The researchers look into pastoralism and are based at four public universities in Ghana: Kwame Nkrumah University of Science and Technology, Kumasi; University of Energy and Natural Resources, Sunyani; University for Development

Table 3.1 Category and sample of respondents used for the study

No.	**Category of interviewee**	**Sample**	**Mode of selection**
1	Researchers	11	Purposive
2	NGOs	2	Purposive
3	Cattle owners	7	Purposive
4	*Herdsmen*		
	Settled herdsmen	11	Snowball*
	Nomadic herdsmen	6	Snowball
	Transhumance herdsmen	3	Snowball
5	*Government*		
	Cattle ranch staff	1	Purposive
	District assembly	5	Purposive
	Police service	1	Purposive
6	*Other key informants*		
	Village chief, Zongo chief, farmers	7	Purposive

Source: Collated by the authors.

Note: *We relied on social networks existing among pastoralists to gain access to them. The herdsmen who were first approached connected us to other herders in their network.

Studies, Tamale; and University of Cape Coast. The NGOs – Changing Lives in Innovative Partnerships (CLIP) and Rescue Mission International – engage in community empowerment and climate-resilient strategies of which pastoralist work is a main thematic focus. 'Pastoralist' generally refers to a cattle and other livestock farmer, but in this chapter it specifically denotes a cattle farmer. 'Herdsman' and 'herder' are used interchangeably to mean a person looking after a cattle herd. Pastoralism literature identifies three distinct categories of herders in Ghana: settled, nomadic, and transhumance.[14] The settled herders live among the indigenous agricultural population while the nomadic herders live in mobile camps located in isolated bush areas. The transhumance herders are also nomadic but engage in regular and seasonal movement.[15] Efforts were made to obtain settled, nomadic, and transhumance respondents when sampling

the herdsmen. The sampled herdsmen raised cattle in Agogo in the Asante Akim North district and in Sekyere Afram Plains (Ashanti Region) and Kwahu Afram Plains (Eastern Region). We relied on social networks and pastoralist relations, including ethnic and religious leaders, to gain permission and access to herdsmen and cattle owners. Data collection occurred over five months between 2022 and 2023 in English and the Twi language (a common dialect in Ghana) using interview guides. The interview guides collected data by asking questions such as follows: What is the nature of challenges associated with pastoralism research and data access? What arrangements and practices do you employ to obtain access to pastoralist data and why? How does the cultivation of relations and social networks enable access to pastoralist data? What opportunities exist for pastoralist data access? The empirical work was complemented with desk-based research, including a review of policy and legal documents.

Responses obtained from all categories of respondents were analysed together following content analysis. When analysing responses to 'how' and 'why' questions, it located the reasoning underlying respondents' replies. When asked the question 'How (by what means) do you gain access to transhumant herders for data collection, and why do you employ particular approach(es)?' the underlying reasons given to support the preferred approach(s) were sought from the responses to the aforementioned questions. A list of underlying reason(s) was generated for each respondent after systematically reading the answers of all the respondents. The next stage of the analysis looked for common patterns in the underlying reasons provided by respondents and clustered them according to common reasons such as social relations of friendship. The findings have been presented in themes and relevant quotes have been used to support explanations.

Results

We present findings for the study structured under four themes: (*a*) pastoralists' environment, (*b*) challenges associated with pastoralists' data access, (*c*) strategies enabling access to pastoralists' data, and (*d*) opportunities to access pastoralists' data.

Pastoralists' Environment

The pastoralists engaged in the study narrated their work environment to constitute a dynamic and extreme terrain. Generally, the settled herders live among local farming communities and send cattle to grazeable areas on fallow lands and in village surroundings, but return to villages after feeding the animals. The transhumance and nomadic herders live in camps in distant places in the bush. Nomadic herders are in touch with farming communities, have social exchanges through selling of cow milk and meat with villages, and are more likely to settle in local communities. One transhumance herder explained the rationale for spending the most time in the bush to be that 'the cattle eat a lot, and they need to walk often to avoid contamination and spread of diseases and pests. We will not get adequate feed for our animals if we stay in one place; but in the forest, the animals can feed for a long time' (Herdsman 5, 10 January 2023). A cattle owner added that 'the cattle need much land and forage and water; places with much of those are good for the livestock, but areas with more trees are not good' (Cattle Owner 1, 10 December 2022).

Yet life in the bush is fraught with several challenges, ranging from extreme weather conditions (scorching sun and excessive rainfall) to security and health issues. It was narrated that 'in the bush, the rain falls on me several times when I am out to feed the cattle, and I confront wild animals on regular occasions' (Herdsman 5, 10 January 2023). Most herdsmen in the study share the experience of losing cattle to wild animals, and several others have experienced robbery attacks in the bush. Herders explained that the possibilities of being confronted with dangerous wild animals as they move in the forest encourage them to move along with sophisticated weapons such as guns. Those who keep no weapons move in groups, usually with family members. A herder observed, 'We sometimes get lost in the bush, but when we roam for a long time, we can locate where we are' (Herdsman 1, 5 February 2023). More recently, struggles over access to land have emerged to blur pastoralist–farmer relations. In the past, farmers and herders peacefully coexisted because land was readily available to accommodate small herder populations. Farmers and herders lived apart from each other, so there was minimal interaction between them. Currently, land is largely competed for due to the need for new lands for farming and the rising herd population which demands large pasture lands. A herdsman revealed that 'now we see farms everywhere even in places where there were no farms . . . the farms have now increased in sizes' (Herdsman 8, 4 February 2023). Herdsmen attribute land struggles to its commodification in rural communities.

Challenges Associated with Pastoralists' Data Access

This section outlines challenges constraining access to reliable pastoralist-related data. Pastoralists' data access is challenged by inaccessible locations; health and security concerns; language barrier; non-participatory and biased behaviour; and disorderly, unclean, subjective, and undocumented data.

The distant places where herdsmen keep cattle render their access by researchers a challenge. Being an agrarian country, the majority of rural Ghanaians engage in crop farming as their main livelihood activity. Relatively, few village people add cattle production as their main or alternative livelihood, though city men employ the labour of Fulani herders to look after their cattle. The nature of cattle production where regular feed is needed demands that herdsmen live near savannah areas where there is readily available feed for cattle. Most herdsmen echo that they 'keep cattle at a distance and away from arable farmers due to the nature of the work which demands caring for cattle close to the bush so we can easily feed the cattle and prevent damage to crops' (Herdsman 1, 16 February 2023). Transhumance and nomadic herders are normally located at places several kilometres away from villages. Accessing transhumance herders requires hours of travel, usually by motorcycles. Herdsmen hardly keep cattle at a particular location; rather, they are always on the move due to daily struggles to secure adequate feed and water for their livestock. This is the case for both nomadic herders who roam frequently on a particular terrain as well as transhumance herders who migrate on a seasonal basis. One researcher echoed that 'living in the bushes is their [herdsmen's] cultural trait' (Researcher 2, 10 January 2023).

Lack of footpath and road networks limit possibilities in accessing herder locations in the field. Agricultural lands used for crop farming have footpaths used by farm owners. Footpaths are generally created and maintained by families and individual farmers who use them regularly. Grazing lands do not have footpaths since cattle do not move along cleared paths in the field, but herders can follow animals as they move through bushes. Herdsmen are good at adapting to movement in bushes, but non-herders tend to struggle when walking in bushy areas. Further, the environment where herders keep cattle is dry and hot, and prone to myriad physical and emotional challenges such as wildlife attacks. Such extreme environments are not favourable destinations for most non-herders and deter many researchers from contacting herders in the field.

Further, herdsmen do not want people to know their locations and movements in the field for security reasons. It is common for herdsmen to be silent about the places in the field where they keep their cattle and about their movement routes. They do so to maintain the safety of the cattle and of themselves since police services are not present in the secluded places where nomadic and transhumance herders normally keep their cattle. Transhumant and nomadic herders are exposed to various security risks, including human attacks due to poor identification and invasion by thieves. Some herdsmen observed that farmers put false claims on them for atrocities they have not committed: 'In the bush, we [herdsmen] face several challenges . . . Sometimes we meet armed robbers and we have to fight back; and at times we go to some villages, and when the farmers are not able to identify particular cattle destroying their crops, they blame us' (Herdsman 5, 10 January 2023). Generally, researchers noted that 'sometimes herders are unable to describe their locations in the field because they find it difficult to describe the places since they do not stay in one place, but others also do not want people to know where they keep the cattle' (Researcher 1, 20 January 2023).

Pastoralists' non-participatory behaviour limits the extent to which researchers obtain data from them. Cattle herders are reluctant to speak to researchers, particularly to those who do not seek permission from their leaders, and this makes data collection a challenge. The herdsmen demand researchers and development actors to directly contact the owners of the cattle they are taking care of. Herders posit that the owners make absolute decisions on their cattle and that their leaders are also more capable of delving into conflicts since they often attend skirmish-related meetings and have access to relevant information to enable them to make well-informed contributions. Other pastoralists share the concern that lands they have legally obtained from chiefs and landowners and have occupied for centuries are often contested by individual farmers. One herder narrated, 'My father paid for the land we feed our cattle on; he paid the big chief and there are documents on that, but you can see how farmers have approached the land' (Herdsman 3, 10 January 2023).

The nomadic and transhumance herders have limited social contact with local villages, and that delays their ability to learn Ghanaian languages. The majority of pastoralists in Ghana, particularly herders, are of foreign origin who have migrated from the Sahel region and speak the Fulbe language and/or other Sahel languages. Some herders settle in farming villages, and others maintain nomadic status by moving across different places to feed cattle. Settled herders are better placed to socially integrate with people in the societies they live in and where they

learn local languages. On the other hand, most transhumance and nomadic herders due to their movement do not easily integrate socially with local Ghanaian societies; rather, they live with their families in isolated camps. Occasionally, nomadic and transhumance herders move to cities to sell cow milk and purchase food items and then return to their original locations.

Pastoralists' data are subjective and skewed, and that limits the possibility of accessing it accurately. Different herder types – nomadic, transhumance, and settled – share different stories about the conflict. The field experiences revealed that the accounts of several farmers and pastoralists were mere word of mouth of family members and people living in their neighbourhoods. In instances where farmers and pastoralists were asked to relate to specific and personal stories, most of them could not since the experiences reported on related to other members of their communities. There were instances where figures quoted as the cost of damages to crops, animals, and other kinds of property were largely exaggerated. The standard of accuracy of data presented by cattle actors is dependent on the lens (objective and subjective) that actors involved in conflicts employ to view their situations. Objective accuracy is rooted in prediction and is based on domains of knowledge. There are instances where conflict actors employ objective means – for instance, through the documented conflict cases at the local government and police services offices – to predict the outcomes of conflicts. Subjective accuracy, on the other hand, is rooted in imagination and is based on intuition.[16] The use of a subjective accuracy lens in imagining what is going to happen in a conflict environment dominates in the case we investigate. Perhaps it is practically impossible to separate prediction and imagination in the farmer–herder conflict; yet to achieve peak accuracy and effectiveness requires that both channels are in sync. Intelligent predictions are critical to making creative decisions since accurate observations are a prerequisite for deciding what happens next.

The violent nature of the farmer–herder conflict has caused many victims to be displaced and resulted in the loss of lives and properties. The negative experiences people have cause them to attach deep emotions to conflict narrations. Conflict victims are unable to segregate their emotions from the actual experiences they share. The attachment of emotions to information transfer blurs access to objective data.

The police services and the district assemblies (local government) are the main state institutions receiving complaints about the farmer–pastoralist conflicts. Yet these actors do not have separate data sheets for such conflicts. All cases reported to the institutions are documented in a consolidated sheet, which needs to be segregated to tease out pastoralist-related cases. The difficulty in segregating data makes state institutions reluctant to provide information to researchers. The disorderly institutional database makes the segregation of conflict data from other data reported at public institutions a challenge.

Strategies Enabling Access to Pastoralists' Data

This section outlines the strategies researchers and development actors use to enable access to pastoralists' data. A broad set of social relations, including social identity and membership in groupings by ethnicity and religion, negotiation of social relations to authority and relations of friendship and trust, and access to geographic knowledge of pastoralists, enable access to reliable pastoralists' data.

Access to the geographic knowledge of pastoralists is critical for accessing pastoralists' data. The location of transhumance and nomadic herders are difficult to find, but researchers rely on the snowball approach to get access to new herders. First, researchers need to know the locations of at least one or two herders. Most researchers noted that 'a Fulani herder you know will lead you to another Fulani herder because they know themselves and the environment where they keep their livestock' (Researcher 3, 5 February 2023). The close relations of friendship and trust serve as the first contact to new herder recruits. Usually, the focal persons who are friends with herder groups are those whom herders are comfortable relating to.

Access through social identity and membership in a group, including groupings by ethnicity and religion, shapes access to pastoralists' data. One researcher narrated how being a Muslim he could easily get access to herders for his study. Another researcher observed that 'when not one of their own, they do not want to talk. The herders are unable to differentiate people who visit them so one needs to inform them where they are coming from and which institution they belong to' (Researcher 7, 15 January 2023). Several settled herders live in communities where the dominant religion is Islam and most herders residing in Muslim-dominated communities also observe the religion. Researchers with the same religious background as herders meet and pray together and share similar religious festivities. These shared religious meetings enhance social bonds, causing herders to open up and share information about their livelihood and culture.

The social relations and networks existing among herder groups, including having access to authority and relations of friendship, also shape access to data. When a herdsman settles in a particular community, he makes new friends with people in the community who can either be indigenes or persons belonging to his ethnicity; even the nomadic and transhumance herders have close friends in villages and cities. Pastoralists maintain close relations with their family heads and local leaders, even with those who are physically located in distant places. During family occasions and social and religious gatherings, most herdsmen take advantage of meeting their close relations. Researchers employ the contacts of family heads and local leaders to reach out to cattle headers. One researcher explains, 'You seek approval from cattle owners, and you find a date and someone leads you to the bush to speak with them [herders]' (Researcher 5, 14 February 2023). For herdsmen who are reluctant to speak, one researcher noted that 'some were reluctant to answer because they feel researchers have been there several times to talk to them . . . I relied on key informants to lead me to the Fulani people' (Researcher 6, 10 January 2023). Making clan leaders call herders to inform them of researchers' quest to undertake studies makes herdsmen respond positively to calls for interviews.

Getting access to labour in the form of translators who speak and understand a herder's native language to interpret questions and relay responses is a means people use to get access to data. Some herdsmen, particularly nomadic, speak several languages due to their movement across different societies with different ethnic backgrounds. Most herdsmen speak the Hausa language, which is spoken by a majority of people in the West Africa region. Researchers employ the services of local inhabitants who speak herdsmen's language as interpreters. However, some researchers are lucky to conduct interviews in the English language; one such researcher observed, 'in cases where I settled on people who could speak my language, I talked to them' (Researcher 7, 15 January 2023).

For farmers and pastoralists quoting extreme and ambiguous figures as well as providing untrue narrations, researchers employ the strategies of asking the same questions several times during interviews. They also engage in data triangulation to assess the truthfulness of information given by other actors participating in conflicts. Making farmers and pastoralists show visual evidence of conflict outcomes also validates the data provided. For example, displaced communities are asked to provide practical evidence of pre-displaced communities. Through observations and interactions with community members, researchers can infer the accuracy of the information given. Community members who claim to have sustained injuries through violent attacks are asked to show evidence of injuries to confirm that.

There is always the tendency for victims to overstate the magnitude of the impact of conflicts and be emotional over the issues confronting them. Often, farmer–pastoralist conflicts make displaced people vulnerable as they lose resources, homes, and properties. As they are violently dispossessed of agricultural lands, communities affected by conflicts face the destruction of their cultural heritage, embodied in nature, and the conditions for economic and political dignity. This undermines possibilities for sustainable peace. To stay objective without taking sides, researchers try to be neutral by questioning the narrations of both farmers and pastoralists in equal measure.

Opportunities for Gaining Access to Pastoralist Data

The study identified opportunities to aid access to pastoralists' data, and this includes the existence of strong customary institutions and their traditional role in mediating land access: the local government and police services, NGOs, and pastoralist associations.

In Ghana, chieftaincy is an important cultural heritage and institution, which has strong ethnic support and provides the structures for leadership and exercise of authority. It denotes sacred and sociopolitical power conferred on chiefs and priests in many parts of Ghana.[17] Before colonial rule in Ghana, the customary structure had been the local political leaders acting as chiefs governing the people and resources. In contemporary Ghana, chieftaincy has strong legal and constitutional status. The 1992 constitution of Ghana (article 270[1]) indicates that 'the institution of chieftaincy, together with its traditional councils as established by customary law and usage, is hereby guaranteed'. The Chieftaincy Act, 2008 (Act 759), sets the guidelines for the functioning of chieftaincy institutions and legitimates the National House of Chiefs to 'undertake the progressive study, interpretation and codification of the customary law to

evolve, in appropriate cases, a unified system of rules of customary law, and compiling the customary laws and lines of succession applicable to each stool or skin'. Chieftaincy is an important cultural heritage and institution, which has strong ethnic support and creates the frameworks for leadership and the implementation of authority. As custodians of the land, chiefs emerge as the main institution to mediate land access in Ghana and are very relevant for mediating pastoralist– farmer conflicts. They have customary support to distribute land to indigenous people under their jurisdiction and can allocate land not occupied by families to strangers, including herdsmen who settle and work in their territories. Due to their critical role in granting land rights to both farmers and pastoralists, chiefs wield enormous information on pastoralists and could play an important role in assembling pastoralists.

The local government bodies (metropolitan, municipal, and district assemblies) in Ghana are mandated by article 240 of the 1992 constitution of Ghana to plan and execute policies in respect to all matters affecting the people within their areas, including issues about pastoralist–farmer conflicts. The Police Service Act (Act 350 [1]) of 1970 provides police officials the mandate to 'prevent and detect crime, to apprehend offenders, and to maintain public order and the safety of persons and property'. Both the district assemblies and the police services have played active roles in the processes leading to the mediation of pastoralist–farmer conflicts in Ghana. In Agogo in the Asante Akim North district, the district assemblies and police services have documented massive data on pastoralist– farmer relations which serve as critical resources for pastoralism-related research.

NGOs mediate the pastoralist–farmer conflicts. It has also led to the formulation of various social groups to interfere with the conflict processes. Ghana National Federation of Livestock Inter-Professional (GFLIP) has been doing field-based research on farmer–herder conflicts and has enormous field contacts and experience. Universities and academic institutions undertake research and train students on the conflict. The Department of Silviculture and Forest Management of Kwame Nkrumah University of Science and Technology hosts a Danish International Development Agency (DANIDA)-funded programme dubbed 'Access and Authority Nexus in Farmer–Herder Conflicts' (AAN). AAN trains PhD and MPhil students on pastoralism-related themes. The University for Development Studies hosts the Ghana Development Studies Hub, which provides a space where researchers and development actors join forces to tackle pressing issues of sustainable development that are of specific relevance to Ghana and West Africa. One key focus area is pastoralism and other landscapebased livelihoods.

The Ghana Cattle Ranching Committee is established by the president of Ghana with membership drawn from the ministries of food and agriculture, national security, interior affairs, foreign affairs and regional integration, and the Inner City and Zongo Development; universities; GFLIP; the Ghana National Association of Cattle Farmers (GNACAF); and the Peasant Farmers Association of Ghana (PFAG). The GNACAF has a membership consisting of cattle owners and herders across Ghana.

Discussion

Our results suggest that access to data from pastoralists is entangled with challenges. The transhumance and nomadic herders keep cattle at field locations that are distant from villages and lack marked footpaths and road networks, rendering transhumance herders the least accessible herder category. Data from transhumance herders are subsequently also the least accessible and incorporated into conflict management protocols. The health and security of cattle and herdsmen are challenging issues in bushes. Motivated by security concerns, herdsmen shield information on their locations and movements from the public. Herders are reserved and display non-participatory behaviour to indicate their reluctance to participate in development studies and programmes. The Herdsmen's non-engaging behaviour is a means not to open up about cattle labour, but to exclude and delegitimize others by keeping historically acquired knowledge to themselves. Pastoralist–farmer conflicts are fraught with trauma and intense emotions due to the pain and suffering resulting from the loss of lives, properties, and homes. Hence, experiences shared in such conflicts are entwined with feelings of closeness, affection, and subjectivity. To a large extent, pastoralists' data are undocumented, unclean, and disorderly kept. In what follows, we discuss (*a*) the social, cultural, and political contexts of pastoralists data access; (*b*) the researcher neutrality and trustworthiness; and (*c*) the power relations and politics of identity.

Social, Cultural, and Political Contexts of Pastoralists' Data Access

The findings illustrate constellations of social relationships that enable access to pastoralists' data. Access to geographic knowledge of pastoralists, their network, and the ability to shape discursive terms is a critical condition to accessing pastoralists' data. Researchers and development actors rely mostly on the snowball approach to recruit new herders. Access through social identity and membership in a group, including groupings by ethnicity and religion, shapes access to pastoralists' data. Muslim researchers and development actors gain access to pastoralist groups on rights and claims attached to religious membership. Others use ethnic background and the ability to speak the pastoralists' language(s) as an inclusionary strategy to obtain data. Negotiation of social relations, such as access to authority and relations of friendship and trust, forms a critical strategy in accessing pastoralists' data. This is by far the dominant mechanism enabling access to most herdsmen. We illustrate how access to labour opportunities via entering into a working relationship with herdsmen grants access to pastoralists' data. Researchers and development actors do not labour by themselves; they rather employ the services of people who speak the pastoralists' native language(s).

We document opportunities to aid pastoralism-related research. Chiefs and the customary structure are an important institution, which have strong ethnic support and provide the structures for leadership and exercise of authority. Being custodians of communal land, chiefs know the boundaries of village lands and have informal networks with cattle owners. Through the allocation of land rights, chiefs invoke authority over farmers and pastoralists and play a critical role in mediating conflicts between pastoralists and farmers. The local government structure at district levels (district assemblies) wields formal powers and serves as the central government's representative at the local level mandated to provide formal representation of

constituents within their jurisdiction. District assemblies, particularly in conflict-prone areas, have extensive documented information on pastoralist–farmer conflicts and associated issues. Other actors playing critical roles in pastoralist–farmer conflicts include the academic and research institutions, the Ghana Police Service, the Ghana Cattle Ranching Committee, the GNACAF, the GFLIP, and the PFAG. Due to their long-term engagement in pastoralist–farmer conflicts, there are potential formal and informal networks that could be drawn upon to aid pastoralism-related research and data access.

Researcher Neutrality and Trustworthiness

The findings illuminate the challenge of staying 'neutral' or being 'objective' during data collection and analysis as a researcher. The role and identity researchers take determine to a great degree the nature of information they will get access to, and this is more prominent in our case where researchers have to negotiate along identity and ethnic lines to obtain data. The positionality, trustworthiness, and experiences from the field also play a role in the data analysis. In a broader sense, the researcher always influences how the data is analysed. This starts when making decisions about how to transcribe and what parts to leave out as being not relevant to the research topic.[18] Based on the observation that not all researchers can easily get access to pastoral data, our analysis seems to suggest that processes of data collection will always comprise voluntary or forced choices and will never be able to capture the full extent of reality. Therefore, based on the findings, this chapter argues that the real world cannot be fully captured by data; this could lead to potential biases and problematic outcomes if field data is relied upon excessively. While there is a need to accept this unavoidable reality, our account demonstrates that researchers should always be looking for new and better ways of data collection and analysis.

Power Relations and Politics of Identity

As discussed previously, our findings suggest the non-participatory behaviour of pastoralists in the release of data. They do so by directing researchers to ethnic and membership leaders as well as by indicating an inability to understand languages used by researchers to communicate with them. Such non-participatory behaviour reflects their desire to exercise autonomy over data. Related ideas are shared by the Global Indigenous Data Alliance, which advocates six rights concerning data that indigenous people have.[19] These are the right to self-determination, the right to use, the right to possess, the right to consent, the right to refuse, and the right to reclaim. Indigenous peoples' right to self-determination entails the ability to organize and control data about a collective identity. In our case in Ghana, we observe herdsmen realize that by refraining from engaging with 'outsiders' on an individual basis. Rather, they recognize the critical role of ethnic and pastoralist leaders in maintaining the production of uniform knowledge and collective identity. Our case shows how herdsmen refuse to engage people they have no social ties with as well as the domination of Fulani ethnicities in cattle herding; these depict the pastoralists' exhibition of rights to possess, refuse, and retain data that reflect pastoralists' identities and cultures.

Our findings also illustrate that pastoralists are reluctant to grant data access to people they share no ethnic or membership relations with. This suggests pastoralist groups consider themselves to be wielding higher status over researchers and activists. It is along claims of religious and identity backgrounds that researchers and practitioners avoid facing discrimination in accessing data from pastoralists. We argue that the inaccessibility of pastoralists and discrimination in the release of data could potentially lead to biases and monopolization of data because only a few known people will have access to important data. The possibility of asking limited and skewed questions might lead to a loss of valuable information that could have been gained by asking other questions. Herders' selective behaviour might result in the over-representation of certain elements of a dataset and eventually lead to low accuracy, skewed outcomes, and systemic prejudices. Putting pastoralists' data into domains privilege a few ethnic and religious actors and create barriers to entry into the pastoralism landscape. This monopoly threatens the political economy of pastoral societies and pastoralism. The problems associated with our empirical data access resonate with documented biases and monopolization linked to digital data access.[20] In Ghana, the state institutions dealing with pastoralists' data (district assemblies and police services) are legally mandated to grant public access to documented data. While access to data from public institutions could be entwined with data biases, retrieving herdsmen's empirical data is more likely to be subjected to systemic biases and over-representation of data.

Conclusion

This study has examined the challenges, strategies, and opportunities researchers employ to gain access to pastoralism-related data. To understand this process, the study relied on data from researchers and NGOs engaged in pastoralism work, cattle owners, herdsmen, and other key informants in the cattle industry. The study suggests that pastoralists operate in a harsh social environment of solitude and are exposed to health and psychological dangers. The adverse environment entangles their relations to other societal groups with conflicts and violent extremism, leading to their passiveness and poor participation in social interventions.

First, getting physical access to herdsmen is limited because transhumance and nomadic herders are located in remote places, several kilometres away from villages, and their practice of moving from one place to another to feed their cattle defeats easy identification of and accessibility to their location. The lack of footpaths and road networks to herders' locations in forested areas also limits the possibilities of accessing herdsmen's locations in the field. Herdsmen do not disclose their locations and movement routes for security reasons – that is, to avoid being easily found and attacked by conflict actors. Access to herders' data is compromised by their non-participatory behaviour in conversations and the habit of keeping information restricted to only a few people, often their relations. Conflict victims poorly segregate emotions from the actual conflict experiences they share, and the attachment of emotions to information hampers access to objective data.

A wider range of social relationships, including social identity and membership in groupings by ethnicity and religion, negotiation of social relations to authority and relations of friendship and trust, and access to geographic knowledge of pastoralists, enable access to pastoralists' data. Researchers and practitioners invest in obtaining geographic knowledge of pastoralists, and others cultivate social relations with pastoralist groups via maintaining ties along religious and cultural lines to gain access to them. There is a network of herders and cattle owners in villages and towns. Pastoralists maintain close relations with their family heads and local leaders, even with those who are physically located in distant places. During family occasions and social and religious gatherings, most herdsmen take advantage of meeting their close relations. Researchers employ the contacts of family heads and local leaders to reach out to cattle headers who are difficult to locate.

Finally, the study has identified opportunities to aid access to pastoralists' data, and this includes the existence of strong customary institutions and their traditional role in mediating land access, the local government and police services, NGOs, and pastoralist associations.

Notes

1 This study forms a part of the project 'Access–Authority Nexus in Farmer–Herder Conflicts' (AAN), funded by the Danish Research Council for Development Research (Danida) under grant number 18-14-GHA.

2 T. A. Benjaminsen, F. P. Maganga, and J. M. Abdallah, 'The Kilosa Killings: Political Ecology of a Farmer–Herder Conflict in Tanzania', *Development and Change* 40, no. 3 (2009): 423–445.

3 A. Olaniyan, M. Francis, and U. Okeke-Uzodike, 'The Cattle are "Ghanaians" but the Herders are Strangers: Farmer–Herder Conflicts, Expulsion Policy, and Pastoralist Question in Agogo, Ghana', *African Studies Quarterly* 15, no. 2 (2015): 53–68; S. Soeters, R. Weesie, and A. Zoomers, 'Agricultural Investments and Farmer–Fulani Pastoralist Conflict in West African Drylands: A Northern Ghanaian Case Study', *Sustainability* 9, no. 11 (2017): 1–19.

4 M. D. Turner, 'Political Ecology and the Moral Dimensions of "Resource Conflicts": The Case of Farmer–Herder Conflicts in the Sahel', *Political Geography* 23, no. 7 (2004): 863–889.

5 B. Maiangwa, '"Conflicting Indigeneity" and Farmer–Herder Conflicts in Postcolonial Africa', *Peace Review* 29, no. 3 (2017): 282–288; N. Mikailu, 'Making Sense of Nigeria's Fulani–Farmer Conflict', BBC News, 5 May 2016, https://www.bbc.com/news/world-africa-36139388 (accessed 20

December 2018); M. Moritz, 'Changing Contexts and Dynamics of Farmer– Herder Conflicts across West Africa', *Canadian Journal of African Studies / Revue canadienne des études africaines* 40, no. 1 (2006): 1–40.

6 K. N. Bukari and N. Schareika, 'Stereotypes, Prejudices and Exclusion of Fulani Pastoralists in Ghana', *Pastoralism: Research, Policy and Practice* 5, no. 20 (2015): 1–12; R. Yembilah and M. Grant, 'The Political Ecology of Territoriality: Territorialities in Farmer–Herder Relationships in Northern Ghana', *GeoJournal* 79, no. 3 (2014): 385–400.

7 S. K. Dary, H. S. James, and A. S. Mohammed, 'Triggers of Farmer–Herder Conflicts in Ghana: A Non-Parametric Analysis of Stakeholders' Perspectives', *Sustainable Agriculture Research* 6, no. 526 (2017): 141–151.

8 K. Kyei-Poakwah, 'Understanding Farmers and Herdsmen Conflict: The Case of Crop Farmers and Fulani Herders in the Asante Akim North District', doctoral dissertation, University of Ghana, 2018.

9 'Fulanis Flee Konkomba Farmers Attack at Sene', *Daily Guide*, www. dailyguidenetwork.com/fulanis-flee-konkomba-farmers-attach-sene (accessed 4 January 2022).

10 I. Abubakar, 'Portrayal of Ethnic Minorities in Ghanaian Newspapers: A Case Study of the Fulani', MA thesis, University of Ghana, Legon, 2016.

11 M. Nartey and H. J. Ladegaard, 'Constructing Undesirables: A Critical Discourse Analysis of Othering of Fulani Nomads in the Ghanaian News Media', *Discourse and Communication* 15, no. 2 (2021):184–199.

12 S. Tonah, 'Migration and Farmer–Herder Conflicts in Ghana's Volta Basin', *Canadian Journal of African Studies / Revue canadienne des études africaines* 40, no. 1 (2006): 152–178.

13 T. Desai, F. Ritchie, and R. Welpton, 'Five Safes: Designing Data Access for Research', Economics Working Paper Series, 1601, University of the West of England, Bristol, 2006, 28.

14 J. Kaiser, 'Good Herder, Bad Herder: Understanding the Construction of Fulani Herders' Identities by the Local Community in Agogo, Ghana', master's thesis, University of Zurich, 2019.

15 Olaniyan, Francis, and Okeke-Uzodike, 'The Cattle are "Ghanaians"'; Tonah, 'Migration and Farmer–Herder Conflicts'.

16 G. Pransky, S. Finkelstein, E. Berndt, M. Kyle, J. Mackell, and D. Tortorice, 'Objective and

Self-Report Work Performance Measures: A Comparative Analysis', *International Journal of Productivity and Performance Management 55*, no. 5 (2006): 390–399.

17 B. Bulley, 'Our Chiefs, Their Land Management, and Our Custom: A Case Study of Juaben, Ashanti', MPhil dissertation, University of Ghana, 2014.

18 M. Hammersley, 'Reproducing or Constructing? Some Questions about Transcription in Social Research', *Qualitative Research* 10, no. 5 (2010): 553–569.

19 'Care Principles for Indigenous Data Governance', Global Indigenous Data Alliance, 2023, https://www.gida-global.org/care (accessed 10 December 2023).

20 M. Khan, X. Wu, X. Xu, and W. Dou, 'Big Data Challenges and Opportunities in the Hype of Industry 4.0', paper presented at IEEE International Conference on Communications (ICC), Institute of Electrical and Electronics Engineers, 2017, 1–6.

4. FROM RIGHTS TO SKILLS: DATA ACCESS FOR TEACHING DATA LITERACY

MIDAS NOUWENS

Researchers and policymakers have converged on the idea that operational transparency and data access are necessary to create meaningful change in how digital technologies operate in European societies. The General Data Protection Regulation (GDPR), 2016, set out to 'increase transparency for data subjects' and 'enhance control over one's own data' as a way to empower individuals vis-àvis technology companies.[1] The Digital Services Act (DSA), 2022, attempts to set up a distributed 'data generation machine'[2] that will continuously produce highquality information about the operation and impact of online services in order to curb systemic risks. The Digital Markets Act (DMA), 2022, and the Data Act, 2023, aim to increase consumer power and redistribute data-based value 'in the hands of relatively few large companies'[3] by giving people the right to 'continuous and real-time access' to any data generated by their use of a product or service.[4]

This battery of regulation the European Union (EU) is currently introducing is an attempt to claim power over the way digital technologies affect member states' societies, and the EU's normative commitment to human rights means that those efforts have also translated into the strengthening of existing data rights and the creation of new ones, albeit without a very clear theory of how those rights translate into concrete empowerment and change. How does access to data lead to meaningful transparency, and how does this transparency lead to citizen empowerment and structural change in the power asymmetries of digital societies?

The theory of change underlying access rights differs depending on the context in which it is used. Researchers and activists, for example, have used it to investigate how technology companies operate, reveal the (negative) impact that they have, and then disseminate this knowledge through publications.[5] The implied theory of change of this *access–publish–change* model is that evidence of non-compliance or compliant but harmful behaviour will create enough public pressure for technology companies to alter their practices. Similarly, lawyers have used access rights to force technology companies to be transparent about their data processing practices – mostly in the context of platform labour – as a way to gather evidence for court cases, using an *access–litigate–change* approach to force a change in the way they operate.[6]

One context in which the value of access rights has not been discussed extensively is education: there is no systematic work on this topic and only few accounts of people's pedagogical practice,[7] let alone a theory of how access rights in education contribute to structural change in digital societies (that is, *access– educate–change*). For educators, social change is often one of the explicit goals of their institutions. Formal education has a long (if not uncomplicated)[8] liberal tradition of enfranchising citizens and as a catalyst for structural evolution,[9] and more critical views argue that universities have a responsibility to liberate the oppressed[10] and create a social class fighting for equality and justice.[11] In the context of digital technology, high-

er education has tried to affect what our digital societies look like by, for example, introducing ethics courses for computer scientists[12] or teaching using exclusively open source software.[13] At a higher level of abstraction, educators across disciplines and education levels have started to argue that the structural datafication of societies requires its residents to develop a new kind of literacy: *data literacy*. The pedagogy of data literacy is still unsettled, and it is in the context of this broader effort that we imagine access rights might find a place in education.

This chapter discusses how access rights can be used in higher education as a pedagogical tool to help students develop *data literacy*, a quality of mind and set of competences that can be employed when analysing and reacting to technological phenomena in society, based on our experiences teaching at a (Danish) university. The goal of this chapter is to (*a*) showcase another context in which access rights can be used in a way that might help reconfigure existing power balances under informational capitalism and (*b*) provide educators with a conceptual introduction and hands-on guide that they can use to employ these rights in their own teaching.

First, we will briefly introduce the concept of data literacy and the scholarship around it. Second, we will review existing data access rights in the EU. Third, we will present concrete exercises that use data access and the learning outcomes we believe they resulted in, based on our own experiences over the past three years. Lastly, we will discuss the limitations of using access rights in education based on their legal design and the interpretation of organisations.

Data Literacy

Academics critical of the restructuring of societies as a result of digitalization and datafication have started to advocate that the future citizen needs a certain kind of literacy to co-shape healthy digital societies. Such 'reflexive, active and knowing publics'[14] would have the ability to, for example, understand and critically reflect on data collection,[15] distrust claims of the objective nature of data,[16] or be empowered against harmful algorithmic processing. This interdisciplinary scholarship is loosely organized around the concept of data literacy, a term which has also migrated into policy discussions[17] and (inter)national educational initiatives.[18] Various literature reviews have tried to consolidate the research around data literacy by evaluating definitions and finding common ground, but the diversity of work means that even these are necessarily scoped more narrowly around audiences (for example, educators,[19] researchers and librarians,[20] and citizens[21]).

Variations of the term that have been proposed include 'big data literacy',[22] 'algorithmic literacy',[23] 'data infrastructure literacy',[24] 'data mindset',[25] and so on (see Table 4.2 for a non-exhaustive list of definitions). Inevitably, the definitions and goals of different authors diverge, so these literacies refer to a variety of practices situated across a spectrum of epistemologies and philosophies (from objectivist post-positivism to subjectivist post-modernism). The various approaches can broadly be separated into two camps: instrumental definitions and critical definitions (while keeping the inevitable caveat in mind that there is further diversity within each of those camps and also works that span across this divide).

Instrumental definitions frame literacy as the ability to create or process data, often with the implied goal to upskill the labour force in a datafied society. These prompted more critical and non-technical perspectives, which see literacy as the ability to question the data's neutrality and authority, identify and mitigate various forms of risk,[26] and have the vocabulary to participate in debates about technology design.[27] Infrastructural definitions expand on the critical approaches to include the social relations as well as the technical infrastructures that mediate data creation, extraction, and processing (for example, algorithms, platforms, business models, standards, power relations, and governing bodies).[28] A rough overarching definition of data literacy could be formulated as *competences that allow a person to work with and critically reflect on data and the socio-technical infrastructures surrounding it, with the goal to immunize the individual against informational harm and empower them to create alternative data worlds*.

Table 4.1 Terms and definitions related to data literacy

Source	Term	Definition
D'Ignazio andBhargava (2015)[29]	Big data literacy	'the ability to read, work with, analyze and argue with data', as well as identifying data collection, understanding algorithmic processing of data, and weighing the impacts of data-driven decisions
Bucher (2016)[30]	Algorithmic imaginary	'ways of thinking about what algorithms are, what they should be and how they function'
Crusoe (2016)[31]	Data literacy	'the knowledge of what data are, how they are collected, analyzed, visualized and shared, and … the understanding of how data are applied for benefit or detriment, within the cultural context of security and privacy'
Philip, Olivares-Pasillas, and Rocha (2016)[32]	Racial data literacy	'the set of practices that are necessary for an individual to be racially literate about data and data-literate about race' – for example, 'examining how societal meanings about race are produced, in part, by the possibilities and constraints in the collection, storage, conversion, manipulation, and representation of data sets'
Gray, Gerlitz, and Bounegru (2018)[33]	Data infrastructure literacy	'critical inquiry into datafication, into how datasets are created with certain purposes in mind as well as opening up "infrastructural imagination" … about how they might be created, used and organised differently (or not at all)'
D'Ignazio and Bhargava (2018)[34]	Data mindset	'the ability to think both creatively and critically about what insights and stories might be possible to glean from data'

Source	Term	Definition
Pangrazi and Selwyn (2019)[35]	Personal data literacy	'critical understandings of the reconstitutions and recirculation of data', being able to identify what personal data is, understand how it is processed, reflect on its implications, use data oneself, and tactically resist, obfuscate, and repurpose data
van Es, Coombs, and Boeschoten (2017)[36]	Reflexive digital data analysis	digital data analysis (acquiring, cleaning, and analysing) in which 'researchers consider their own role in the construction of the data' and 'take responsibility to discern how [the tools and platforms they use] shape the data'
Sander (2020)[37]	Critical big data literacy	'awareness, understanding and ability to critically reflect upon big data collection practices, data uses and the possible risks and implications that come with these practices, as well as the ability to implement this knowledge for a more empowered internet usage'

Source: Collated by the author from the sources listed in the 'Sources' column.

The pedagogy of data literacy is still unsettled, in terms of both teaching methods and how to evaluate their effectiveness. Scholars who see data literacy as technical competencies are often more closely connected to educational disciplines and have done more work to operationalize data literacy and study the effects of different teaching approaches.[38] Systematic reviews show that data literacy is being taught at educational institutions across all levels.[39] Courses that integrate data literacy often use practice-based approaches with real-world data. Sometimes data literacy is taught as multiple, successive modules in a stand-alone course, while other times it is integrated in other courses as part of projects.[40] Assessments about the learning outcomes of different pedagogical strategies use either self-report methods (such as surveys and interviews) where students reflect and describe their competences or direct measures (such as competency tests and participant observation).[41] However, these assessment tools are rarely validated and scholars are calling for higher-quality methods.[42]

Critical and infrastructural approaches to teaching data literacy have fewer explicit studies on methods and learning outcomes, and instead mostly include high-level suggestions or autoethnographic descriptions of how data literacy could be taught. The focus is less on programming and data processing, which means teaching methods also include custom-built tools and devices (for example, DataBasic[43] and Stingray[44]), interactive media (for example, Do Not Track[45]), embodied representations (for example, acting out datasets[46] and algorithms[47]), partnerships with more technically capable peers,[48] reflective discussions,[49] and so on. There are few post-hoc evaluations of the learning outcomes of these methods. Rather, suggestions of approaches are often based on the reflection of the teacher(s) after multiple experiences running a workshop or course.

In all the works discussing data literacy pedagogy, there has not been any systematic reflection on the role that access rights can have, either as part of exercise design or as a subject of study in its own right.

Access Rights in EU Tech Law

Access rights over personal data have existed in Europe in some form since the 1970s.[50] They were harmonized for the first time across the EU in 1995 through the Data Protection Directive, 1995,[51] and were subsequently fortified through the GDPR in 2016.[52] The EU's recent DSA, DMA, and Data Act are poised to expand on these rights, although with different scopes and purposes. We briefly summarize the rights that could be used by students and teachers here, which is not intended as a comprehensive legal discussion but instead a gentle and pragmatic introduction for educators who might be unfamiliar with these rights (see Table 4.2 for an overview).

Readers should keep in mind that only the access right in the GDPR has been in effect long enough for us to know something about how it works in practice, whereas the other regulations are still new and untested. Many European countries and the EU have also conferred access rights to citizens over public data through freedom of information legislation, but since their focus is not primarily about increasing transparency or control over digital technologies, these are not included.

Article 15 of the General Data Protection Regulation

The GDPR[53] is a regulation that governs the processing of personal data about people in the EU, and one of its missions is to address the difficulties that people experience to stay in control of their personal data.[54] The right of access in Article 15 of the GDPR is one way it tries to improve this, although it already existed in some form in previous national[55] and EU legislation,[56] making it one of the oldest rights that people have over their personal data in the EU. The right should give so-called data subjects access to three things: (*a*) a confirmation as to whether a data controller has data about them; (*b*) access to a copy of that data; and (*c*) additional meta-information, such as for what purpose the data is processed, what categories the personal data falls into, who else might have access to this data, how long the data will be stored, where the data came from if it was not provided by the data subject themselves, and if and how it is being used for automated decision-making.

Table 4.2 A simplified overview of data access rights in the EU

Law	Type of data	Right holder	Applicable to	Access modality	Response time
GDPR, Article15	Personal data	Any individual	Data controllers	Commonly used electronic form	30 days
GDPR, Article 20	Personal data	Any individual	Data controllers	Structured, commonly used, and machine-readable format	30 days
DSA, Article 40	Data necessary to study systemic risks and mitigation strategies	(Vetted) researchers	Very large online platforms and search engines	Appropriate interfaces specified by researcher, platform, or search engine; real-time (if possible)	15 days
DMA, Article 6(9)	Data provided and generated by an end user	Users/ Authorised third parties	Gatekeepers	Effective, continuous, and real-time	Immediate
Data Act, Articles 3–4	Product and service data, including metadata necessary for interpretation	Users/ Authorised third parties	Data holders	Comprehensive, structured, commonly used, machine-readable, continuous, and real-time	Directly accessible or without undue delay

Source: Collated by the authors.

The European Data Protection Board (EDPB), the institution responsible for making sure the GDPR is consistently applied across the EU, has provided a practical interpretation of this right that describes how organizations should respond when people exercise it.[57] For example, many data controllers will ask data subjects to use a particular form or email address when making their request, but the EDPB explains any request received through a reasonable channel is a valid request, so people cannot be made to use particular communication channels or templates. A data controller should also always interpret an access request as broadly as possible and be as comprehensive as possible with their response. That response should be sent within a month after the request was made, unless it is too complicated to do so in that time. If the identity of the data subject is uncertain (for example, if an access request is made

from a different email address than the one the controller has on file), the data controller can ask for additional information, but this information should follow the principle of data minimization and never be more extensive than is strictly necessary to confirm the identity of the data subject; asking for copies of passports or biometric information, which some data controllers have started to do, is in most cases not necessary or allowed. In terms of the formatting of the data, there are no strict requirements, but it should be assumed that if the request is made by electronic means, then the response should use the same modality in a commonly used electronic format (which in practice often means CSV or JSON files). These instructions from the EDPB are intended to protect the right from constraints that might emerge in practice and to make it as easy as possible for people to exercise.

Article 20 of the General Data Protection Regulation

The right to data portability gives people the right to receive a copy of their personal data that is processed by automated means (that is, not paper files) in a structured, commonly used, and machine-readable format, or have that data sent directly to another data controller. The right was introduced in the GDPR as a response to the issues individuals previously had when exercising their access right under the GDPR's precursor, the Data Protection Directive:[58] people would receive their personal data in whatever format the data controller decided, limiting how they could manage and reuse the data.

Not all personal data can be requested under this right, but only personal data that the controller has based on a person's consent or the contract they have with them. Personal data processed for, for example, complying with legal obligations that the controller might have (such as fraud detection) are not covered. Another limitation is that it only includes data 'provided by' the data subject, although this is interpreted broadly: it includes not only obvious things directly submitted by the user such as account details but also behavioural data generated through user activity, while it excludes data that might be generated by the controller after further processing, such as user profiles.

Data should be provided in a format that is abstracted away from the specific technical implementation of the controller's systems, since the aim is that the data can easily be repurposed by the user; the goal is interoperability. Case law has established that PDF (Portable Document Format) files do not meet this goal.[59] The ability to export data only in small chunks (for example, one email at a time) is also not sufficient.[60] Data should be provided within 30 days, although an extension to three months is possible, as long as the controller explains to the data subject why it needs more time. If there is too much data to transmit digitally within a reasonable time frame, controllers should consider alternative (physical) media.

Article 40 of the Digital Services Act

The DSA[61] regulates online services to try to make them safer and more transparent.[62] Article 40 of the DSA tries to support this by requiring 'very large online platforms' (for example, Facebook, Amazon, and Booking.com) and 'very large search engines' (for example, Google and Bing) to give researchers access to data which can be used to study systemic risks that

might be created because of the service, as well as data that can be used to assess whether the risk mitigation measures that the services themselves have implemented are adequate and efficient.

There is still a lot of uncertainty about who qualifies for access and what data can be requested (see chapter 10 for a more detailed discussion). Much of the Article confers rights to so-called vetted researchers, which are individuals who meet a number of requirements (that is, affiliated with a research organization and independent of commercial interests) and can fulfil certain operational obligations (for example, capable of providing the necessary security and confidentiality for the data and make results publicly available for free). Their application for this status also needs to justify the duration for which they need access to the data and explain how their research contributes to systemic risk assessment and mitigation monitoring. However, paragraph 12 of the Article also opens up access rights to individuals who meet those conditions but are not affiliated with a research organization and do not commit to making their results publicly available for free.

If an individual is officially designated as a 'vetted researcher' by the Digital Services Coordinator (DSC, the main national authority), the individual can ask the DSC to pass on the data access request to the digital service on their behalf. The requested data should be received 'within a reasonable period' and be provided through 'appropriate interfaces' (for example, online databases and application programming interfaces). Those individuals who qualify for access as per paragraph 12 can request data directly without going through the DSC, which they should receive 'without undue delay' and (if possible) in 'real time'.

Article 6(9) of the Digital Markets Act

The DMA[63] tries to address characteristics of digital businesses that result in a lack of contestability (that is, monopolies), such as network effects, multiside markets, vertical integration, and extreme economies of scale. It targets gatekeepers, platforms that, over the last three years, have had a large impact on the EU's internal market and had at least 45 million monthly active end users and 10,000 yearly active business users.

Article 6(9) of the DMA describes data sharing obligations for the 'core platform services' of those gatekeepers that are supposed to stimulate contestability, by making it easier to leave a platform or have it inter-operate with another service. The Article specifies that gatekeepers need to provide end users (or third parties authorized by them) the option to port the data they have provided to the gatekeeper or data that is generated by their use of the service. That data should be provided in a format that can be used 'immediately and effectively'[64] (that is, a commonly used format in a manageable size), and gatekeepers also need to implement 'appropriate and high-quality technical measures, such as application programming interfaces'[65] that make it possible to access this data 'continuously and in real time'.

Articles 3 and 4 of the Data Act

EU policies related to data have longstanding tensions between, on the one hand, wanting to protect fundamental rights while, on the other, furthering economic growth by removing barriers for companies.[66] If the GDPR represents regulation that primarily tries to achieve the first, the Data Act could be seen as the outcome of the second interest, as it sets out to 'maximise the value of data in the economy and society'.[67] It sets down rules to (among other things) make product and related service data available to the users of those products.

Article 3 of the Data Act confers the right to end users to have access to any product data, product-related service data, and metadata necessary to understand and use that data. This data should be made available by default, easily, in a comprehensive, structured, commonly used, and machine-readable format. Users should also, before purchasing a product or a related service, be given information about the type, format, estimated volume, and frequency of data the product is capable of generating, whether it stores this on a device or a remote server and whether the data holders expect to use the data themselves for specific purposes and/or share it with third parties.

Article 4 of the Data Act adds specifications to the data rights described in Article 3 – for example, that data access can be restricted if it undermines the security requirements of the product, that the data holder should notify the competent authorities if it restricts data access, that users can lodge complaints in this case, and that data holders cannot make exercising this right unduly difficult. Arguably, the most important restriction relates to the protection of trade secrets, which is a valid reason to not share data if the user does not take measures to ensure their confidentiality. However, such decisions should be substantiated and the competent authorities should be notified of it, which are burdens placed on data holders to make sure it does not restrict the user's data rights for their own gain.

***

The access rights given to users across these different EU regulations overlap and expand on each other, sometimes explicitly (as in the case of the DMA expanding GDPR rights). They target different spheres of contemporary, digitally mediated life (for example, online services, connected products, and data-hungry governments) and try to rebalance the distribution of power through different approaches (for example, improving fundamental rights, consumer protection, and fair competition). Whether these rights are enough to address remains to be seen. Only the access right to personal data has been in place for long enough to have had an impact, and preliminary signals indicate low levels of compliance[68] and no increase in people's feelings of control over their personal data.[69]

Teaching Data Literacy Using Access Rights

The EU's imaginary of access rights and the goals of teachers advocating for data literacy overlap: both want to empower citizens against the negative externalities of informational capitalism.[70] The EU sees access to data as a precondition for good citizenship in a digital society – for example, by making it possible for people to oversee the fair processing of their personal data or by being able to move their data to other services rather than being locked in. Data literacy advocates think that population-wide competences to understand the generation and analysis of data is necessary to mitigate digital inequities that arise between those who can work with data and those who cannot, which is becoming more pressing in societies where data-based knowledge claims about the world carry more weight.[71] Given this overlap in mission, how can access rights contribute to the development of data literacy in practice?

Over the past three years we have used access rights as part of our teaching in the Department of Information Studies at Aarhus University, Denmark. The department positions itself within the broader humanities faculty as the department for critical reflection on technology and society through theory, empirical research, and the construction of digital artefacts (for example, prototypes, software, and art). Across the different degree programmes, students are broadly taught science and technology studies (STS) theory, human–computer interaction (HCI) design approaches, digital innovation, programming, and qualitative methods. We exercised the right to access under Article 15 of the GDPR (since the other rights were not yet in effect) in three different syllabi: Data Studies, Datafication of Society, and Digital Living. The Data Studies[72] course is a second-year bachelor's degree course that aims to teach the students how to think critically about the production and use of data through theoretical frameworks and hands-on data processing (for example, querying application programming interfaces (APIs) and using machine learning [ML]). The Datafication of Society[73] course is a third-year bachelor's degree course that tries to place the restructuring of societies around data in the context of larger historical and sociological trends. The Digital Living master's programme[74] is an interdisciplinary degree that bridges social theory, business, management, and computer science and includes a one-week cross-course module on surveillance capitalism to show how to combine the perspectives from different disciplines. In all these courses, we have used access rights as part of our teaching to help the students understand what data is collected, how technologies work, or how data is part of organizational practices.

Based on these experiences, we see four ways in which access rights have contributed to the development of students' digital literacy: (*a*) the experience of trying to exercise access rights reveal the power relations between technology companies and nation states; (*b*) the data the students received helps them confront their internalized dataism; (*c*) the meta-information about whom companies receive data from and share it with can be used to trace data flows; and (*d*) comparing outcomes of different data processing methods shows the epistemological impact that mediating artefacts have on the realities and knowledge that are (co-)created by technology companies. We elaborate on the goals, exercises, and learning outcomes of each of these four in the following sections. Although not all exercises were used in all courses, we do present them here as a coherent trajectory in a suggested order where, for example, data received in one exercise can be reused in the next or previous learning outcomes make the next goal easier to understand.

Power Relations of Informational Capitalism

Goal

This exercise seeks to teach students about their digital rights and to demonstrate how current power relations between companies and nation states impact the effectiveness of those rights.

Exercise

Before the session, students are asked to read about their data rights on the website of their national data protection authority.[75] During the session, we first discuss whether the students know about these rights, what they think their purpose is, and if they have ever used them. We then collectively go through the process of exercising their right to access. We brainstorm about possible data controllers they could send a request to, making sure to go beyond just the obvious Big Tech companies so the students stretch their understanding of the extent of datafication. Once we have compiled a list, each student creates and sends a request to a controller of their choice. We use the My Data Done Right (mydatadoneright.eu) tool to generate the request text, which is a user-friendly interactive form created by the Dutch privacy organization Bits of Freedom, but students send the request from their own email accounts. We tell them that they can consult us if they need help in the follow-up process with the company after sending the request. One month later, there is a follow-up poll and discussion where we ask the students how many of them received a response, what that response included, whether they feel empowered, what surprised them, and whether they changed their perception on access rights, technology companies, and the law.

Learning Outcomes

Data access rights are supposed to empower individuals, but a precondition for this is that people know about their rights and actively use them. Incorporating access rights into the curriculum is a concrete and straightforward way to teach larger cohorts of citizens about their rights, which they will hopefully continue to use outside the classroom and throughout their lifetime.

Exercising the access rights, at least right now, also demonstrates to students that laws on paper are different from laws in practice. Research shows that many organizations lack awareness and understanding of the rights,[76] do not respond within legal timeframes,[77] and (when they do) often fail to provide all the required information.[78] These results are confirmed by our classroom experiences, where (in the latest iteration of the Data Studies course in 2022 with approximately 90 students) only around 7 per cent received a complete response (67 per cent received a response, and of those only 11 per cent received all necessary information: confirmation of data processing, meta-information, and a copy of the data). In Denmark, a country with a strong cultural belief in the rule of law[79] and trust in institutions,[80] students are often confused by this experience. Especially those who send data subject access requests to Big Tech or large social medial companies (for example, Meta, Google, Snapchat, and TikTok) are shocked by how they are ignored. These experiences have proven to be a helpful starting point to critically discuss power relations between sovereign nation states and wealthy technology

companies and the effectiveness of law as an instrument to redistribute such power. It raises questions such as following: Why are companies making it difficult to exercise these rights? Why are authorities not able to make them comply with the laws? What is the political philosophy behind individual rights as a remedy to informational power?

Confronting Dataism

Goal

This exercise seeks to uncover and question students' unspoken assumptions about the quality and objectivity of company's data processing and reveal how much data is collected about them.

Exercise

Before the class, students are asked to get copies of their personal data from any digital platforms they use. Ideally, this would be based on copies they receive through their access rights, but that process is currently too slow and unreliable to feasibly build exercises around. Instead, we ask students to use the 'download your data' options that many platforms have started to offer in response to access rights, although we make sure to emphasize that this is not the same as exercising their access right and is likely an incomplete dataset. In the class, the students are asked to review those files, by paying attention to the quantity of the data collected, the quality of that data, and the inferences and classifications made by the company about them. We suggest they can do this in pairs if they feel comfortable revealing personal data, so they can compare with other students how they are perceived differently by the same company. We also provide some scaffold code (in the form of Jupyter Notebooks) that helps them produce aggregates such as counts, averages, distributions, or visualizations. We conclude with a discussion about what they found, what surprised them, and why they were surprised to help surface their unconscious assumptions.

Learning Outcomes

The students we meet are often socialized into the belief that data represents objective, value-neutral measures of the world and that using it, especially in large quantities, will naturally lead to better outcomes, predictions, or products.[81] Confronting this *dataism*,[82] this ideology of the unreasonable effectiveness of data facts, is an important component of the critical approaches to data literacy. However, because this is an ideology, we often encounter some conscious and unconscious resistance from the students to this critique. As Jose van Dijck explains, dataism is not just the belief that data could capture the world 'as is' but also the trust in the institutions and companies that collect, clean, and analyse this data. In our experience, simply mentioning that the full complexity of the world cannot be captured by quantitative data is not that hard for the students to accept. The more difficult barrier to overcome is their assumption that data handlers also know this, have processes in place to overcome these limitations, and surely only draw justifiable conclusions from the data. Irrationality, irresponsibility, pragmatism, and the primacy of profit-seeking are given less weight in their imagination of datafication.

Using access rights has been an effective method to confront the students with the volume of collected data and their assumptions about the quality of its processing. Other scholars have already suggested using 'real data' or data the students have a connection to because it is more engaging,[83] but personal data specifically makes it possible for students to perceive the epistemic distance between their experience and the data double[84] a company has constructed of them. Because the students know themselves, they are often surprised about the amount of data that is being associated with them and incredulous about the (often poor-quality) assumptions, predictions, and categorizations of companies they might hold in high regard. It is hard to drive these points home without personal data: to evaluate whether the volume of data is a lot or not is something that rests on the contrast between the lived experience of using digital services and the data traces they never knew about or reasonably expected. And to evaluate the quality of that data requires that the students know what the underlying reality that data is supposedly capturing.

Justifications for existing technologies and battles over visions of the digital future are often fought through discourse and symbolism.[85] Confronting the students with the actual data that lies behind the imaginaries of technology companies – datasets as a 'higher form of intelligence'[86] – helps dispel some of the myths that are created and hopefully inoculates the students against future hype cycles.

Tracking Data Flows

Goal

This exercise seeks to show how data flows through a network of many different actors and how it gets reshaped at each of those steps.

Exercise

Students are asked to pick a core digital service in their life and trace all the other parties that this organization receives personal data from and shares personal data with. This metadata should be included in responses to an access request under Article 15 of the GDPR, but since transparency obligations in EU regulation (GDPR, DA, DSA) require that such information is also available more publicly, a fall-back option is to look at privacy policies, terms and conditions, consent banners, and any other information that describes the data processing of that organization (keeping in mind that these are not entirely equivalent: access requests should include the exact identity of the recipients of personal data,[87] whereas in privacy policies 'categories of recipients' are enough). Each party can be mapped based on various characteristics (for example, geographical location, type of service, level in the software stack, and annual turnover). Follow-up access requests or investigations into those parties should then provide an insight into what data has been shared with them from the original service, how it is augmented and transformed, what the data is used for, and an additional list of data-sharing partners that could be the seed for the next wave of access requests.

Learning Outcomes

Infrastructural perspectives on data often emphasize its social and material entanglements:[88] the physicality of the internet with its undersea cables and landing sites, the cultural and political incentives that inform categorizations,[89] the public infrastructural responsibilities assumed by private platforms,[90] and the individual and organizational subjectivities that shape processes of data cleaning.[91] This insight is crucial to solidify the understanding that 'raw data is an oxymoron'[92] and that data is always being reprocessed and repurposed as it flows through a complex network of actors. Making those networks explicit – what Geoffrey C. Bowker and Susan L. Star call 'infrastructural inversion'[93] – is the first step in locating power and allocating responsibilities to certain players.

Data access rights include access to metadata, such as where the data comes from (if not provided by the person themself), how long it is stored for, who else the data is shared with, the location of those third parties, and their trading names and contact details. This information can then be used to make consecutive access requests to data held by other parties, allowing students to map out all the different players in the ecosystem, see how the data gets transformed and augmented at each step, what it is used for, and which other infrastructures it touches. What it demonstrates is that data should always also be thought of through the lens of ecosystems or networks, since its shape, its assumed value, and its impact are not inherent to the data itself but instead emerge because of how it flows through a particular chain of players. The same data might be harmless in isolation or when it stays in the hands of a single player (for example, an account on a period-tracking app), but becomes dangerous when combined with other data (for instance, location data) or when shared with other actors (for example, anti-abortion organizations).

Epistemological Impact of Mediation

Goal

This exercise seeks to explain how knowledge is shaped by the digital artefacts that mediate its production and that claims of truth are always based on a particular philosophical position.

Exercise

Following a lecture on the fundamentals of ML, students are provided with interactive code notebooks (Jupyter Notebooks) which explain and demonstrate how ML models for natural language processing and image processing work. The first notebook focuses on sentiment analysis – the prediction of emotional value in text – comparing the VADER and TextBlob models.[94] The second notebook focuses on image classification – assigning a label to an image – and compares the EfficientNetV2 and ResNeXt models.[95] Students are asked to feed the models with their own input data (for example, text messages they sent or images they posted on social media) and reflect on the output they get. As part of their reflection, they are asked to find documentation about the people behind the model, what data it was trained on, what it was created for, where it is being deployed, what claims and decisions are made based on it, and so on.

Learning Outcomes

Mediation theoretical perspectives on technology such as postphenomenology and actor–network theory highlight how our perceptions and actions are coconstituted by the artefacts that sit between ourselves and the world (often blurring ontological separations between subject and object). In the case of digital technologies, mediating artefacts such as algorithms and models help generate a particular view of the world and structure our actions in finite ways. These are quite abstract notions about the nature of being and knowing, but exercises comparing multiple mediating artefacts demonstrate quite concretely how they generate different outputs. Using very specific technologies also makes it possible for the students to trace design decisions that were made (for example, training approaches and parameters) and the other artefacts that are involved (for example, datasets and platforms). Learning how to do this kind of methodological deconstruction is crucial both to evaluate the quality of a knowledge claim made by others and for the students' ability to be intentional and transparent about any knowledge they themselves might create using digital tools.[96] At a higher level, a concrete understanding of the epistemological impact of mediation opens up discussions about whose views are being represented and what kind of values are expressed by design decisions.[97] If real-time and continuous data access is available, a more constructive approach to data literacy could encourage students to build alternative software (for example, apps, websites, and visualization pipelines) that processes their data differently – what Mireille Hildebrandt calls 'agonistic machine learning'[98] or Henrik Korsgaard, Clemens N. Klokmose, and Susanne Bødker call 'computational alternatives'[99] – and, through them, give shape to the digital worlds that they would like to live in.

***

What data literacy means and how it should be taught is not a settled question yet, although broadly speaking it includes competences in working with data and being able to evaluate its value and impact. In our experience, the four exercises presented here help students (*a*) become aware of the politics of informational capitalism, (*b*) confront their internalized dataism, (*c*) track how data flows through global networks, and (*d*) realize how mediating artefacts impact the knowledge that can be derived from data. For educators interested in the instrumental aspect of data literacy, these exercises provide interesting data for students to process and analyse; and for those arguing for a critical perspective on data literacy, these exercises can be used for non-technical students to reflect on and discuss the larger political structures that data is part of.

Limitations of Using Access Rights in Education

Based on our personal experiences and reflections described earlier, access rights can help university students develop data literacy. However, exercising them does not necessarily teach everything that researchers have indicated is important for developing this competence, and it should sit next to other approaches. The legal design of access rights and the way they are interpreted by technology companies place some limitations on how they can be used as part of a pedagogical strategy, which we highlight here. The limitations related to the DSA, the

DMA, and the Data Act are speculative, since these have not gone into effect yet at the time of writing this and we have not had the chance to include them in our teaching.

GDPR access rights are quite easy to exercise in practice: they can be sent to any reasonable contact address, they require no additional technical expertise or financial resources, and there are plenty of text templates available. However, their focus on personal data makes them less suited for creating transparency around the processing methods or technical infrastructure of the organization, since they only reveal the digital double of an individual rather than the various algorithms, models, or databases this data is used in (for example, a person's coordinates within a multidimensional recommender model). Devoid of the context, personal data can become rather arbitrary since decisions and classifications derive meaning from their relative position to other numbers (for example, the Instagram device setting 'face_filter': '14' or Spotify's ad category inference 'dfp_expiration_test_ krishna' does not reveal much). It is possible to get a sense of the larger system by having students compare these values between them, but the personal nature of the data makes such collective aggregation a sensitive exercise. Poor compliance rates and one-month reply durations also make it difficult to build consecutive exercises around them within a three-month semester. At the same time, these rights have also stimulated 'data download' options and public reporting of similar information that provide feasible fallback options.

Access rights under the DSA provide interesting transparency about larger systems because they go beyond personal data and are instead anchored in the goal to detect systemic risks and evaluate risk management strategies, which makes the data that falls inside that scope much more variable. However, these access rights only apply to a handful of digital services (19 at the time of writing) and do not require sharing information about how the data flows beyond the borders of the organization, making it less suited for tracing networks. The access rights are also more restricted because they require an individual to be approved as a 'vetted researcher' by the supervisory authority, because the access is limited to the time it takes to do the research and because the researcher needs to provide 'necessary security and confidentiality' measures (which likely do not include sharing the data with over a hundred students). Rather than sharing their access with students directly, researcher-teachers could instead generate synthetic datasets or replicate processing pipelines that are suitable for teaching (that is, without reidentification and model inversion risks), either for their own courses or to share with other educators. Paragraph 12 of Article 40 of the DSA, which requires platforms to give real-time data access to individuals not vetted by the DSC, might also stimulate platforms to provide student accessible APIs or sandbox environments as a gesture towards compliance, similar to how personal data download options appeared in response to GDPR access rights.

There is considerable overlap in the access rights established by the DMA and the Data Act. Both provide access to data provided or generated by the end user, continuous and in real time (although the Data Act adds more caveats to the access method: 'where relevant and appropriate', 'where applicable'). The main difference between the two is in who it applies to. The DMA rights are, like the DSA, applicable to only a small number of 'gatekeepers' (seven at the time of writing) and only about their 'core platform services'. The Data Act applies to

manufacturers of connected products and providers of related data-generating services on the EU market. Another difference is that the Data Act also requires organizations to provide considerable transparency about what data is generated, how much and in what level of granularity, and who it is shared with. Both the DMA and the Data Act make it possible for users to request that their data is shared directly with third parties, which could include services that analyse and visualize that data in ways that contribute to the development of a student's data literacy. The practical value of these rights for education, however, will depend almost entirely on how they are implemented. A generous interpretation of these access rights would mean a variety of parameterized APIs that output actual data values without any rate limits, while a more restricted interpretation would be a sandboxed environment where computations can be ran 'in situ' but which keeps individual data points obfuscated inside the platform.[100]

Conclusion

New EU regulations are poised to expand access rights to how digital companies collect and process data, doubling down on the EU's governance strategy to combine formal oversight with a human rights approach. How access rights contribute to structural change in the power relations of digitalized societies is not entirely clear. In this chapter, we discussed how access rights can be used in the context of education, connecting to ongoing pedagogical efforts across various disciplines around the idea of data literacy as 'an important part of a strategy in democratic societies to come to terms with living in a digital world'.[101] Based on our experiences teaching at a Danish university over the last three years, we suggest that access rights can contribute to the development of students' data literacy in four ways: (*a*) the experience of trying to exercise access rights and the poor compliance rates reveal the political tension and *power relations between technology companies and nation states*; (*b*) the quality and quantity of the data revealed through data access helps *students confront their internalized dataism*; (*c*) the meta-information about whom companies receive data from and share it with can be used to *trace data flows* and demonstrate the importance of a *network and infrastructure perspective* on data; and (*d*) comparing outcomes of different data processing methods shows the *epistemological impact* that *mediating artefacts* have on the realities and knowledge that is (co-)created by technology companies.

The use of existing and upcoming access rights for education also has limitations, primarily the history of dismal compliance rates (because of either lack of competences or active subversion of the law). Other limitations include constraints placed on who those access rights are for (that is, vetted researchers), who it applies to (for example, very large online services or gatekeepers), what kind of technical competences are required to exercise or analyse them (for instance, understanding of structured data formats or making API calls), or the ambiguous language of the access right obligations that give considerable interpretative freedom to organizations controlling data. While the actual exercise of access rights might not necessarily provide the organizational transparency and individual empowerment that regulators imagine, we expect that they will create more opportunities for opening up the black boxes of technologies broadly. We have seen something similar with current access rights, where even lip service to these obligations (for example, data download buttons and updated privacy policies) have proven to provide accessible opportunities for our in-class activities. In

other words, we do not need every single access right to be respected for them to create new pedagogical opportunities that could contribute to the development of students' data literacy as long as data and insights based on access are shared publicly or between educators.

Future work looking at the intersection of access rights and higher education could include more formal evaluations of specific exercises and learning outcomes for students in order to check whether and to what extent they contribute to data literacy or empowerment more generally. However, this would also require more concretization of the concept of data literacy (what exactly are the most important components and competencies) and a discussion of acceptable evaluation strategies, since the disciplines involved in these efforts span across epistemological traditions and might not be convinced by the same kind of 'evidence'. One elephant in the room that should also be discussed is the fact that centring data literacy as a response to the negative externalities of datafication is a political choice[102] without clear evidence that it can achieve those purposes,[103] and that this choice naturally takes away resources from approaches that do not look at these issues through the lenses of individual resistance, data, or digital technology. Collective approaches or established theoretical traditions of power and political economy (and thus perhaps less glittery than data literacy) should also be considered as legitimate foundations for a pedagogy of the oppressed in digital societies.

Notes

1 EU General Data Protection Regulation (GDPR): 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), OJ 2016 L 119/1.

2 J. Jaursch, 'Transcript for the Background Discussion "New EU-rules for Big Tech: How to Improve the Digital Services Act"', Stiftung Neue Verantwortung, 14 September 2021, https://www.stiftung-nv.de/en/ publication/transcript-background-discussion-new-eu-rules-big-tech-howimprove-digital-services-act (accessed 25 August 2023).

3 EU Digital Markets Act (DMA): Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), OJ 2022 L 265.

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

5 O. Solon, 'How Much Data Did Facebook Have on One Man? 1,200 Pages of Data in 57 Categories', *Wired*, 28 December 2012, https://www.wired.

co.uk/arcle/privacy-versus-facebook (accessed 25 August 2023).

6 *Uber v. Drivers* [2023] ECLI:NL:GHAMS:2023:796 Court of Appeal.

7 As one of the few examples, see S. Yates and E. Carmi*, Developing Citizens Data Literacy: A Short Guide* (London: Nuffield Foundation, 2022).

8 R. Collins, *The Credential Society: An Historical Sociology of Education and Stratification* (Columbia University Press, 2019).

9 In many European countries, for example, higher education institutions were used in the last 50 years as an important 'engine' to bring about the structural evolution of the country into a 'knowledge society', leading to universities dramatically increasing their student populations and changing their management style to be more like a commercial entity. Jussi Välimaa and David Hoffman, 'Knowledge Society Discourse and Higher Education', *Higher Education* 265, no. 56 (2008): 265–285.

10 P. Freire, *Pedagogy of the Oppressed*, 30th anniversary ed. (Continuum, 2000 [1968]).

11 bell hooks, *Teaching Critical Thinking: Practical Wisdom* (Routledge, 2010).

12 C. Fiesler, N. Garrett, and N, Beard, 'What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis', in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (Association for Computing Machinery, 2020), DOI: https://dl.acm.org/doi/10.1145/3328778.3366825.

13 D. Carrington and S. K. Kim, 'Teaching Software Design with Open Source Software', 33rd Annual Frontiers in Education, FIE 2003, 2003.

14 H. Kennedy and G. Moss, 'Known or Knowing Publics? Social Media Data Mining and the Question of Public Agency', *Big Data and Society* 1, no. 2 (2015), DOI: https://doi.org/10.1177/20539517156111.

15 L. Pangrazio and N. Selwyn, '"Personal Data Literacies": A Critical Literacies Approach to Enhancing Understandings of Personal Digital Data', *New Media and Society* 419, no. 21 (2019): 419–437.

16 danah boyd and Kate Crawford, 'Critical Questions for Big Data: Provocations for a Cultural,

Technological, and Scholarly Phenomenon', *Information, Communication and Society* 662, no. 15 (2012): 662–679.

17 Huw C. Davies, 'Rescuing Data Literacy from Dataism', in *Data Justice and the Right to the City*, ed. Morgan Currie, Jeremy Knox, and Callum McGregor, 146–164 (Edinburgh University Press, 2022).

18 Chantel Ridsdale, James Rothwell, Mike Smit, Hossam Ali-Hassan, Michael Bliemel, Dean Irvine, Daniel Kelley, Stan Matwin, and Brad Wuetherick, 'Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report', Dalhouse University, 2015.

19 J. E. Raffaghelli and B. Stewart, 'Centering Complexity in "Educators" Data Literacy' to Support Future Practices in Faculty Development: A Systematic Review of the Literature', *Teaching in Higher Education* 435, no. 25 (2020): 435–455.

20 T. Koltay, 'Data Literacy for Researchers and Data Librarians', *Journal of Librarianship and Information Science* 3, no. 49 (2017): 3–14.

21 Elinor Carmi, Simeon J. Yates, Eleanor Lockley, and Alicja Pawluczuk, 'Data Citizenship: Rethinking Data Literacy in the Age of Disinformation, Misinformation, and Malinformation', *Internet Policy Review* 9, https:// policyreview.info/node/1481 (accessed 25 August 2023); D. Crusoe, 'Data Literacy Defined Pro Populo: To Read This Article, Please Provide a Little Information', *Journal of Community Informatics* 12 (2016), https:// openjournals.uwaterloo.ca/index.php/JoCI/article/view/3276 (accessed 25 August 2023).

22 C. D'Ignazio and R. Bhargava, 'Approaches to Building Big Data Literacy', Proceedings of the Bloomberg Data for Good Exchange Conference, New York, NY, 2015.

23 L. Dogruel, P. Masur, and S. Joeckel, 'Development and Validation of an Algorithm Literacy Scale for Internet Users', *Communication Methods and Measures* 115, no. 16 (2022).

24 J. Gray, C. Gerlitz, and L. Bounegru, 'Data Infrastructure Literacy', *Big Data and Society* 1, no. 5 (2018): 115–133.

25 C. D'Ignazio and R. Bhargava, 'Cultivating a Data Mindset in the Arts and Humanities | Public', *Public* 4, no. 2 (2018), https://public.imaginingamerica. org/blog/article/cultivating-a-data-mindset-in-the-arts-and-humanities (accessed 25 August 2023).

26 Crusoe, 'Data Literacy Defined Pro Populo'.

27 D'Ignazio and Bhargava, 'Approaches to Building Big Data Literacy'.

28 Gray, Gerlitz, and Bounegru, 'Data Infrastructure Literacy'; I. Sander, 'What Is Critical Big Data Literacy and How Can It Be Implemented?' *Internet Policy Review* 9, no. 2 (2020), https://policyreview.info/node/1479 (accessed 25 August 2023).

29 D'Ignazio and Bhargava, 'Approaches to Building Big Data Literacy'.

30 T. Bucher, 'The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms', *Information, Communication and Society* 30, no. 20 (2017): 30–44.

31 Crusoe, 'Data Literacy Defined Pro Populo'.

32 T. Philip, M. Olivares-Pasillas, and J. Rocha, 'Becoming Racially Literate about Data and Data-Literate about Race: Data Visualizations in the Classroom as a Site of Racial-Ideological Micro-Contestations', *Cognition and Instruction* 361, no. 34 (2016): 361–388.

33 Gray, Gerlitz, and Bounegru, 'Data Infrastructure Literacy'.

34 D'Ignazio and Bhargava, 'Cultivating a Data Mindset in the Arts and Humanities'.

35 Pangrazio and Selwyn, '"Personal Data Literacies"'.

36 Karin van Es, Nicholás L. Coombs, and Thomas Boeschoten, 'Towards a Reflexive Digital Data Analysis', in *The Datafied Society: Studying Culture Through Data*, ed. Karin van Es and Mirko T Schäfer, 171–180 (Amsterdam University Press, 2017).

37 Sander, 'What Is Critical Big Data Literacy and How Can It Be Implemented?'

38 Ying Cui, Fu Chen, Alina Lutsyk, Jacqueline P. Leighton, and Maria Cutumisu, 'Data Literacy Assessments: A Systematic Literature Review', *Assessment in Education: Principles, Policy and Practice* 76, no. 30 (2023): 76–96.

39 Bahareh Ghodoosi, Tracey West, Qinyi Li, Geraldine Torrisi-Steele, and Sharmistha Dey, 'A Systematic Literature Review of Data Literacy Education', *Journal of Business and Finance Librarianship* 112, no. 28 (2023): 112–127.

40 Ridsdale, Rothwell, Smit, Ali-Hassan, Bliemel, Irvine, Kelley, Matwin, and Wuetherick, 'Strategies and Best Practices for Data Literacy Education'; Cui, Chen, Lutsyk, Leighton, and Cutumisu, 'Data Literacy Assessments'.

41 Ridsdale, Rothwell, Smit, Ali-Hassan, Bliemel, Irvine, Kelley, Matwin, and Wuetherick, 'Strategies and Best Practices for Data Literacy Education'; Cui, Chen, Lutsyk, Leighton, and Cutumisu, 'Data Literacy Assessments'.

42 Ridsdale, Rothwell, Smit, Ali-Hassan, Bliemel, Irvine, Kelley, Matwin, and Wuetherick, 'Strategies and Best Practices for Data Literacy Education'; Cui, Chen, Lutsyk, Leighton, and Cutumisu, 'Data Literacy Assessments'.

43 D'Ignazio and Bhargava, 'Cultivating a Data Mindset'.

44 S. Vakil, A. Reith, and N. A. Melo, 'Jamming Power: Youth Agency and Community-Driven Science in a Critical Technology Learning Program', *Journal of Research in Science Teaching* 60, no. 8 (2023), DOI: https:// onlinelibrary.wiley.com/doi/abs/10.1002/tea.21843.

45 I. Sander, 'Critical Big Data Literacy Tools: Engaging Citizens and Promoting Empowered Internet Usage', *Data and Policy* 1, no. 2 (2020): e5.

46 Rahul Bhargava, Amanda Brea, Victoria Palacin, Laura Perovich, and Jesse Hinson, 'Data Theatre as an Entry Point to Data Literacy', *Educational Technology and Society* 93, no. 25 (2022): 93–108.

47 D'Ignazio and Bhargava, 'Cultivating a Data Mindset'.

48 D'Ignazio and Bhargava, 'Cultivating a Data Mindset'.

49 L. Poirier, 'Ethnographies of Datasets: Teaching Critical Data Analysis through R Notebooks', *Journal of Interactive Technology and Pedagogy* (2020), https://jitp.commons.gc.cuny.edu/ethnographies-of-datasets-teachingcritical-data-analysis-through-r-notebooks (accessed 25 August 2023).

50 Starting with the Data Protection Act of the German state Hesse (Hessisches Datenschutzgesetz vom 7 oktober 1970 GVBl II 300-10, published at Wiesbaden, 12 October 1970, in *Gesetz-und Verordnungsblatt für das Land Hessen* [Laws and Regulations Journal], part 1, no. 41). For in-depth overviews of the history and emergence of data protection and data rights, see, for example, G. G. Fuster, *The Emergence of Personal Data Protection as a Fundamental Right of the EU* (Springer, 2014) and J. Ausloos, *The Right to Erasure in EU Data Protection Law: From Individual Rights to Effective Protection* (Oxford University Press, 2020).

51 EU Directive 95/46: Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ 1995 L 281/31.

52 GDPR, OJ 2016 L 119/1, Article 15.

53 GDPR, OJ 2016 L 119/1, Article 15.

54 European Commission, 'Commission Staff Working Paper Impact Assessment Accompanying the Document 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) and Directive of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data by Competent Authorities for the Purposes of Prevention, Investigation, Detection or Prosecution of Criminal Offences or the Execution of Criminal Penalties, and the Free Movement of Such Data', 2012, https:// eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52012SC0072 (accessed 1 December 2024).

55 Such as the French *Loi Informatique et Libertés* of 1978 (*Loi n° 78–17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*).

56 Such as the Data Protection Directive, 1995. OJ 1995 L 281/31.

57 European Data Protection Board, 'Guidelines 01/2022 on Data Subject Rights: Right of Access', 2022, https://www.edpb.europa.eu/system/ files/2023-04/edpb_guidelines_202201_data_subject_rights_access_v2_ en.pdf (accessed 1 December 2024).

58 European Data Protection Board, 'Guidelines on the Right to Data Portability under Regulation 2016/679, WP242 rev.01', 2017, https://www.edpb. europa.eu/our-work-tools/our-documents/guidelines/guidelines-right-dataportability-under-regulation-2016679_en (accessed 1 December 2024).

59 District Court of Amsterdam, 11 March 2021, ECLI:NL:RBAMS:2021:1020.

60 Tietosuojavaltuutetun toimisto (Finnish DPA) 22 Marhc 2023, Case 10048/182/20.

61 EU Digital Services Act (DSA): Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), OJ 2022 L 277/1.

62 European Commission, 'Commission Staff Working Document Executive Summary of the Impact Assessment Report Accompanying the Document Proposal for a Regulation of the European Parliament and the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC', 2020, https://eur-lex.europa.eu/legal-content/EN/ ALL/?uri=CELEX-%3A52020SC0349 (accessed 1 December 2024).

63 DMA, OJ 2022 L 265.

64 DMA, OJ 2022 L 265, recital 59.

65 DMA, OJ 2022 L 265, recital 59.

66 Gernot Rieder, 'Tracing Big Data Imaginaries through Public Policy: The Case of the European Commission', in *The Politics and Policies of Big Data: Big Data, Big Brother?* ed. Ann R. Sætnan, Ingrid Schneider, and Nicola Green, 89–109 (Routledge, 2018).

67 European Commission, 'Commission Staff Working Document Impact Assessment Report Accompanying the Document Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Fair Access to and Use of Data (Data Act)', 2022, https://eur-lex.europa.eu/legal-

content/EN/TXT/?uri=CELEX%3A52022SC0034 (accessed 1 December 2024).

68 Alex Bowyer, Jack Holt, Josephine Go Jefferies, Rob Wilson, David Kirk, and Jan David, 'Human–GDPR Interaction: Practical Experiences of Accessing Personal Data', *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), DOI: https://dl.acm.org/doi/10.1145/3491102.3501947.

69 Directorate-General for Justice and Consumers (European Commission) and Kantar, *The General Data Protection Regulation: Report* (Publications Office of the European Union, 2019).

70 M. Castells, *The Rise of the Network Society*, vol. 1, 2nd ed. (Blackwell Publishers, 2009 [1996]), 18.

71 boyd and Crawford, 'Critical Questions for Big Data'.

72 'Data Studies', Aarhus University, https://kursuskatalog.au.dk/en/ course/112914/ Data-studies (accessed 25 August 2023).

73 'Datafication of Society', Aarhus University, https://kursuskatalog.au.dk/en/ course/110859/Datafication-of-Society (accessed 25 August 2023).

74 'Digital Living', Aarhus University, https://kandidat.au.dk/informations videnskab (accessed 25 August 2023).

75 For example, 'Hvad er dine rettigheder', Datatilsynet, https://www. datatilsynet.dk/borger/ hvad-er-dine-rettigheder (accessed 25 August 2023) and 'For the Public', Information Commissioner's Office, https://ico.org.uk/ for-the-public (accessed 25 August 2023).

76 J. Ausloos and P. Dewitte, 'Shattering One-Way Mirrors: Data Subject Access Rights in Practice', 20 January 2018, https://papers.ssrn.com/ abstract=3106632 (accessed 25 August 2023).

77 R. Mahieu, H. Asghari, and M. van Eeten, 'Collectively Exercising the Right of Access: Individual Effort, Societal Effect', 13 July 2018, https://papers. ssrn.com/abstract=3216615 (accessed 25 August 2023).

78 Clive Norris, Paul de Hert, and Xavier L'Hoiry, and Antonella Galetta (eds.), *The Unaccountable State of Surveillance: Exercising Access Rights in Europe*, vol. 34 (Springer International Publishing, 2017).

79 For a historical perspective, see M. F. Jensen, 'Statebuilding, Establishing Rule of Law and Fighting Corruption in Denmark, 1660–1900', in *Anticorruption in History: From Antiquity to the Modern Era*, ed. Ronald Kroeze, André Vitória and Guy Geltner, 197–210 (Oxford University Press 2017). For a quantitative ranking, see World Justice Project, *Rule of Law Index 2022* (World Justice Project, 2022), https://worldjusticeproject.org/sites/default/ files/documents/WJPIndex2022.pdf (accessed 1 December 2024).

80 European Foundation for the Improvement of Living and Working Conditions, *Societal Change and Trust in Institutions* (Eurofound, 2018).

81 boyd and Crawford, 'Critical Questions for Big Data'.

82 Jose van Dijck, 'Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology', *Surveillance and Society* 197, no. 12 (2014): 197–208.

83 For example, Leanne Bowler, Amelia Acker, Wei Jeng, and Yu Chi, '"It Lives All around Us": Aspects of Data Literacy in Teen's Lives', *Proceedings of the Association for Information Science and Technology* 54, no. 1 (2017): 27–35 and R. W. Erwin, 'Data Literacy: Real-World Learning through Problemsolving with Data Sets', *American Secondary Education* 18, no. 43 (2015): 18–26.

84 K. D. Haggerty and R. V. Ericson, 'The Surveillant Assemblage', in *Surveillance, Crime and Social Control*, ed. Clive Norris and Dean Wilson, 61–78 (Routledge, 2006).

85 S. Jasanoff and S. H. Kim, *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power* (University of Chicago Press, 2015).

86 boyd and Crawford, 'Critical Questions for Big Data'.

87 Case C-154/21 *RW v Österreichische Post AG* [2023] ECR I-1.

88 S. Flensburg and S. Lomborg, 'Datafication Research: Mapping the Field for a Future Agenda', *New Media and Society* 25, no. 6 (2021), DOI: https://doi. org/10.1177/14614448211046616.

89 G. C. Bowker and S. L. Star, *Sorting Things Out: Classification and Its Consequences* (MIT Press, 1999).

90 Jean-Christophe Plantin, Carl Lagoze, Paul N. Edwards, and Christian Sandvig, 'Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook', *New Media and Society* 293, no. 20 (2018), DOI: https://doi. org/10.1177/1461444816661553.

91 J. C. Plantin, 'Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science', *Science, Technology, and Human Values* 44, no. 1 (2019): 52–73.

92 G. C. Bowker, *Memory Practices in the Sciences* (paperback edition) (MIT Press, 2008); L. Gitelman, *Raw Data Is an Oxymoron* (MIT Press, 2013).

93 Bowker and Star, *Sorting Things Out.*

94 VADER and TextBlob are two popular natural language processing libraries that include functions which calculate the sentiment of a text string. These libraries are trained on different data and use different ways to calculate sentiment, so the output when applied to the same text is almost always different.

95 EfficientNetV2 and ResNeXt are two popular pre-trained image classification models. When given an image as an input, it provides a list of different words that the model believes describe the content of the image and confidence scores that represent how strongly the model thinks that description matches with the image.

96 K. van Es, M. Wieringa, and M. T. Schäfer, 'Tool Criticism: From Digital Methods to Digital Methodology', *WS.2 2018: Proceedings of the 2nd International Conference on Web Studies* (2018), DOI: http://dl.acm.org/ citation.cfm?doid=3240431.3240436.

97 Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao, 'The Values Encoded in Machine Learning Research', *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), DOI: https://dl.acm.org/doi/10.1145/3531146.3533083.

98 Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao, 'The Values Encoded in Machine Learning Research', *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), DOI: https://dl.acm.org/doi/10.1145/3531146.3533083.

99 H. Korsgaard, C. N. Klokmose, and S. Bødker, 'Computational Alternatives in Participatory Design: Putting the t Back in Socio-Technical Research', *PDC '16: Proceedings of the 14th Participatory Design Conference* (2016), DOI: http://dl.acm.org/citation.cfm?doid=2940299.2940314.

100 Joint Research Centre (European Commission), *The EU Digital Markets Act: A Report from a Panel of Economic Experts* (Publications Office of the European Union, 2021), DOI: https://data.europa.eu/doi/10.2760/139337.

101 L. Pangrazio and J. Sefton-Green, 'The Social Utility of "Data Literacy"', *Learning, Media and Technology* 208, no. 45 (2020): 208–220.

102 Jansen F, 'Critical Is Not Political: The Need to (Re)Politicize Data Literacy', *Seminar.net* 17, no. 2 (2021), https://journals.oslomet.no/index.php/ seminar/article/view/4280 (accessed 25 August 2023).

103 J. Elisa Raffaghelli, 'Is Data Literacy a Catalyst of Social Justice? A Response from Nine Data Literacy Initiatives in Higher Education', *Education Sciences* 10, no. 9 (2020), DOI: https://doi.org/10.3390/educsci10090233.

PART II: LIMITATIONS

5. KEYS THROWN AWAY? CHALLENGES IN BRAZIL ON ACCESSING PUBLIC-INTEREST DATA ON STATE SURVEILLANCE TOOLS VIA TRANSPARENCY PORTALS AND REQUESTS FOR INFORMATION[1]

ANDRÉ RAMIRO, PEDRO AMARAL, AND MARCOS CÉSAR M. PEREIRA

The successive scandals of the Israeli company NSO Group in 2021 and others[2] aroused fears in the international community about the use spyware in a variety of locations. The concern stems from the capabilities of the spying tool, which allows its user to infect the target's device, accessing all stored information, including the ability to remotely turn on microphones and cameras, all without the infected person being aware of it. Usually used against activists and opponents of authoritarian governments, this is not the first time that the public eye has turned to Pegasus. Since 2018, the Citizen Lab organization has published a report that denounces the use of the tool in more than 45 countries, breaking down the locations where the tool was operated. That was the first appearance of the software in national territory, having reported suspicions that signs of the programme in Brazil also existed.[3]

In view of this informational power, Brazilian law enforcement has shown interest in acquiring the NSO Group spy tool. Prior to his election, the defence of the former and current president of the republic, Luiz Inácio Lula da Silva, reported that 'Operation Car Wash' (Operação Lava Jato) even negotiated the acquisition of Pegasus.[4] Former president Jair Bolsonaro's son and Rio de Janeiro city councillor, Carlos Bolsonaro, through an attempt to intervene in the Ministry of Justice, tried to negotiate the spyware for 25 million reais, newspapers reported.[5] However, with the revelations that NSO intended to participate in the public bid, the Israeli company abandoned the negotiations,[6] emphasizing on the reactions of civil society, such as Conectas, Transparency International, and Instituto Igarapé, which questioned the acquisition before the Federal Court of Auditors.[7] Behind the curtains, it has been documented that the objective of the acquisition of Pegasus was to equip the 'ABIN Paralela' (Parallel Abin),[8] an extra official sector whose objective was to snoop into political opponents, including inside the government team, a private secret service to meet the particular agendas of Jair Bolsonaro.[9]

Another tool dealt by Carlos Bolsonaro was the Israeli programme Sherlock, developed by Candiru, a well-known surveillance firm. Confirmed by a former representative of the company, the purpose would be, again, to surveil government officials in order to monitor employees who could be considered 'enemies' of the Bolsonaro administration.[10] For this, the spy programme would only be used on the federal government's internal network, unlike Pegasus, which would be used externally – if purchased. A long and exhaustive timeline of hacking tools acquisitions – including attempts – has emerged in our study, encompassing vendors, resellers, and national and international developers. More recently, it was revealed by the newspaper *O Globo* that ABIN had a contract with the company Cognyte to acquire a tool called FirstMile, which exploited breaches in the Signalling System 7 (SS7) protocol[11] in order to collect location information of citizens.[12]

That landscape suggested that newspapers were sparsely reporting on the existence of disproportionate surveillance cases, especially based on contractual agreements between vendors and the Brazilian government, that needed to be systematized and scrutinized. Our aim, in that case, was to offer a big picture of the 'government hacking' industry and political economy dynamics[13] in Brazil – the involvement in cases of corruption, unjustified surveillance, and, lastly, the lack of regulation that paves the way for these procedures to take place in the country. As shown in this chapter, the pictured environment takes advantage of administrative breaches that go back to a dictatorship heritage that puts intelligence activities away from democratic oversight and challenge the effectiveness of our national transparency regime.

However, considering this global and political scenario, Pegasus has become the main news in the field of surveillance worldwide. That led us to two main hypotheses: (*a*) commercial agreements were in place for many years between a diversity of vendors and Brazilian authorities; (*b*) these surveillance tools, which aimed at extracting data and, as spywares, remotely exploiting vulnerabilities, were being used widely and contemporarily in national territory, in both law enforcement and intelligence spheres, without the requisite judicial and administrative procedure. As the documentation of such agreements was scarce, our objective was to shine a light on the reach of such tools within the organizational investigative structure in Brazil. With that, we aimed to contribute with meaningful data to provide inputs to the making of a proper legal framework that would address such activities with proper assessments considering both government needs and the safeguards of fundamental rights.

Bearing this political and investigative scenario in mind, we published the study *Mercadores da Insegurança: conjuntura e riscos do hacking governamental no Brasil* (Merchants of Insecurity: Conjecture and Risks of Government Hacking in Brazil).[14] We had two approaches: an empirical phase, using the Transparency Portals and Requests for Information (RFI), where we explored the state of affairs related to the debate on government hacking – the exploitation of vulnerabilities in systems and software by government agents to access information that was, until then, inaccessible because of encryption – historically linked to a supposed solution to the clash between increased encryption deployment and the alleged impossibility of investigative agents accessing protected evidence (known as the 'going dark' problem).[15] We then qualitatively analysed the content of the collected documentation, exploring the transparency regime in Brazil and a culture and heritage of secrecy concerning intelligence and law enforcement forces. In a later phase, we analysed the legal framework in Brazil concerning the use of surveillance tools to assess the legality, proportionality, and necessity of this specific practice.[16]

The outcomes and the experience of research itself put in evidence at least two apparently contradictory problems: on the one hand, the transparency regime in Brazil, in theory, organizes and creates technical instruments for citizens to exercise their political rights and demand public entities to comply with their responsibilities, especially based on the oversight of public contracts and spendings; on the other hand, the history of secrecy concerning public security and intelligence sectors makes their activities transparency-proof, contributing for a state of public unawareness and possible arbitrariness against several sectors of society. In this chapter, in light of our findings concerning the use of hacking tools, we aim to critically

explore the development of a national transparency, the limits it faces regarding the historical culture of secrecy of public institutions such as the Brazilian intelligence service, and how it reflects on the ecosystem of exploiting security vulnerabilities as a 'legitimate' procedure that can pose high risks to fundamental rights.

With the political background painted, we have developed the chapter in four main sections: First, we offer an overview of our study, including motivations, methods, findings, and critical conclusions. Then we describe the challenges we have faced when trying to access, through official channels and open government data portals, data related to the subject of our study, namely information regarding contracts with hacking tools' vendors. Furthermore, in order to contextualize the Brazilian political landscape and give a historical perspective to institutions and rights fundamentally dealt with by our experience, we analyse the development of the transparency regime of Brazil and the culture of secrecy inherited from the Brazilian dictatorship period, especially regarding intelligence services. We conclude by stating that, contemporarily, these are two contradictory spheres and that the evolving adoption of hacking tools, without proper regulation, democratic oversight, as well as effective transparency mechanisms, will fertilize a soil that germinates surveillance abuses and violations of fundamental rights in the country.

Our Study: General Motivations, Methodology, Findings, Challenges, and Conclusions

In order to fill the terms of this debate with empirical data, we initially carried out an extensive survey based on quantitative research, encompassing data from the period between 2015 and 2021 of contracts between commercial representatives (either developers or vendors) and the public administration in Brazil. This effort consisted in carrying out approximately 60 *pedidos de acesso à informação* (access to information requests, PAI), the main legal mechanism for requesting information from the public sector under the Law of Access to Information of 2011 (LAI). That included requests to the Ministry of Justice and Public Security, Ministry of Defence, Institutional Security Cabinet (Gabinete de Segurança Institucional, GSI), Federal Police, and Federal Prosecutor's Office, to all 27 secretaries of public security, and to all 27 public prosecutor's offices at the state level. In parallel, an extensive survey of contractual documents was carried out with previously identified companies: 28 transparency portals were analysed, including all 26 portals of the executive branch of the federal states concerning their public security secretaries as well as of the Federal District, and the transparency portal of the federal government. In the second phase of the research, we reviewed critical literature in the field – unfortunately just a few of them produced in Brazil – and analysed, based on qualitative parameters, the information found in the contracts in order to look at three central pillars necessary to understand their use by government authorities: security parameters, legal basis (or conditions) for the use of the tools, and rules concerning the chain of custody.

Among our findings, we identified 209 contractual documents at different levels of the executive and judiciary branches, including all state-level public security entities. The companies found in the work were Cellebrite, Micro Systemation AB (MSAB), OpenText, Magnet Forensics, Exterro/AccessData,[17] and Verint Systems/Cognyte/Suntech,[18] including the two main

commercial proxies in Brazil – Techbiz Forense Digital and Apura Comércio de Softwares e Assessoria em Tecnologia da Informação – and 20 different solutions acquired by public entities.[19] In terms of values, with monetary correction made in November 2023, we saw a growth of 168 per cent in federal spendings regarding such tools and of 514 per cent at the state level. The major part of the analysed database came from information present in the transparency portals, since the vast majority of responses to PAI were totally or mostly negative. The justifications given were based on legal hypotheses of secrecy of public information or even the allegation of non-existence of such contracts and/or technologies.

Institutional Reach of Hacking Tools and Involvement with Corruption Cases

The capillarity of tools illustrated in the contracts we found even points to the existence of data extraction tools in agencies that, typically, do not conduct criminal investigations using cutting-edge forensic techniques. This is the case of the existence of a contract for the acquisition of UFED Touch[20] by the Administrative Council for Economic Defence (CADE),[21] the body typically responsible for, among other attributions, conducting administrative processes and promoting policies regarding fairness in commercial competition.[22] The entity's management report for the year 2021 reaffirms the acquisition of the tool for the 'prosecution of infractions to the economic order', in the amount of more than half a million reais.

One of the contracts, concerning the G12 technology,[23] developed by Verint Systems, was acquired by the Polícia Civil from the state of Pará for 5,000,000,00 reais,[24] and used in the scheme of the governor of the Pará, Helder Barbalho, to spy on investigators of a supposed corruption scheme in which the governor was involved in the purchase of respirators during the COVID-19 pandemic.[25] In another investigation by the Brazilian Federal Police, the 'Operação Chabu',[26] one of Suntech's partners, José Augusto Alves, was arrested on suspicion of being the central organizer, alongside with two other chiefs of police and a police officer of the Federal Highway Police, of an information leak scheme of secret criminal investigations concerning corrupt politicians, the selling of Suntech products, and the exchange of personal benefits.

Another data about the rooted presence of such companies in Brazil relates to how they benefit from military expeditions. In 2016, the state of Rio de Janeiro was targeted with federal military intervention to deal with structural public security violence.[27] According to the transparency portal, Verint Systems is among the 10 companies most financially favoured by the military intervention.[28] A survey conducted by the newspaper *Brasil de Fato* points to Verint as the number one favoured company, with more than 40 million reais paid with public resources for the 'provision of data security and espionage solutions'.[29] In a 2015 army report, contracting with Verint for the acquisition of electronic warfare support measures (MAGE) was already mentioned.

Brief (Lack of) Legal Framework and the Legitimization of a Grey Market

As the objective of this chapter is not to exhaustively analyse each of these law's frameworks, the general conclusion is that none of the regulations establish comprehensive requirements based on legality, proportionality, and necessity to grant to government authorities the legal basis to conduct hacking operations, whether they are for law enforcement or intelligence activities. We analysed these findings through the lens of existing legislation, especially the *Law of Interception of Telecommunication*, 1996;[30] the *Statute for Child and Adolescents*, 1990;[31] the *Law of Organized Crimes*, 1998;[32] the *Brazilian Civil Rights Framework for the Internet*, 2014;[33] the *General Data Protection Law*, 2018;[34] and the *Code of Criminal Procedure*, 2010,[35] which we understand to be the set of laws that currently compose the regulatory framework for hacking procedures. Our conclusion is that both the culture of using such high-risk surveillance tools – highly engaged in an international grey market of vulnerabilities[36] – and the lack of proper protections sedimented in fundamental rights contemporary to our digital realities[37] broaden the space for arbitrariness and authoritarian behaviour. Additionally, it poses risks to democratic participation and networked security in the country, especially portrayed in the persecution and harassment of activists and political figures, as the international panorama shows.[38]

Therefore, the use of these tools is accelerating and being legitimized by the government, gaining the status of proxies[39] or procured authorities, in terms of both diffusion and the technical power of intrusion. They also circumvent security mechanisms of existing devices on targets, such as strong encryption.

If not properly addressed based on risk assessments and transparency protocols, the situation – as well as the scenario of insecurity due to the presence of this unregulated industry[40] and the commercial circuit of hacking tools in Brazil, 'legitimized' by government demands for forensic technologies – poses risks to rights associated with data protection and the secrecy of communications.

Despite the alarming findings, the process of collecting, processing, and analysing the data was illustrative of how access to information on surveillance infrastructures is a huge political and operational challenge. The structure of passive and active transparency, further specified later, was insufficient, and there is a lack of incentives for better implementation of the transparency policy. From offline sites to the almost complete lack of standardization to the disregard for the *Access to Information Act* of 2011 and the *Transparency Act* of 2009, in addition to the large and easily manipulated use of exceptions not to disclose information because of national security reasons that the Brazilian dictatorship has inherited, it becomes clear that there is a long way to go for a more effective oversight mechanism for controlling state resources and practices concerning hacking procedures in Brazil.

The Ethnography of Accessing Surveillance Data: An Empirical Maze Experience within Our Findings

The Case of Transparency Portals

As said before, the research had two approaches based on surveys: one by searches through the public Portais de Transparência (Transparency Portals), enacted by the Decree nº 5.482/2005 and which, theoretically, makes available to every citizen detailed data on public spendings. We sought to investigate in these portals (totalling 28: 26 secretaries of public security at state level, the secretary of public security of the federal district, and the federal government's transparency portal) contractual documents concerning hacking tools that once were, or at the time of the research were, in use in Brazil.[41]

To look at the portals, we first tried to carry out the searches by inserting keywords such as 'extraction', 'access to mobile devices', 'mobile data extraction', 'remote access to digital devices', 'decryption tools', and so on. The outcomes were not considerable enough, making it necessary to rethink the keywords used. We assumed that looking for an exact subject of contractual documents would lead us to an unbearable uncertainty: every company has its own name for its products and for its capabilities. As a solution, we shortlisted a set of 32 companies that were active in the global commercial trading of hacking tools with governments, mentioned in news, mapped by other key stakeholders, or outspoken in the security market. A second reason not to look at functionalities was that the international vendors were 'routing' their tools through resellers in Brazil.

We also found public documents, available on Google, informing acquisitions of Cellebrite tools by various government bodies – whether security departments or state public prosecutions. It is important to remark that the documents highlighted the exclusive resale of *Cellebrite*'s UFED tool by a Brazilian company called *TechBiz Forense Digital*. Due to this exclusivity, the modality commonly used for public acquisitions in which there is no competition – as in this case for exclusivity – are carried out through a *non-enforceability of public bidding*,[42] meaning less public information that government bodies were looking for companies to hire.

From that moment on, we began searching for the names of suppliers to state and federal governments, as well as public prosecutions. We highlight findings of two reseller Brazilian companies: *TechBiz Forense Digital* and *Apura Comércio de Softwares e Assessoria Em Tecnologia da Informação*. For other international companies, such as *Verint Systems*, they have local subsidiaries – in this case, the *Suntech S.A.*, later renamed *Cognyte Brasil S.A.* It showed how surveillance companies often change their names to get off the radar of political scandals and that was also the case in Brazil.[43]

That has posed challenges for us to trace contracts and to find a 'corporate' relationship between *Verint Systems*, *Suntech S.A.*, and *Cognyte Brasil S.A*. We observed that there was a relationship between *Suntech* and *Verint*, in which the former was a company belonging to the Israeli group. However, during the survey of *Cognyte Brasil* tools, we observed that the National Register of Legal Entities (CNPJ), which was used to identify companies as well as

their official emails, was the same as that of *Suntech*. When observing the history of *Verint*, we discover that there was a separation between the intelligence sector – which became *Cognyte* – and the business solutions sector, which kept the name *Verint*. Finally, we noticed that the negotiations made directly with *Verint* were carried out only at the federal level, while acquisitions of its solutions at the state level were carried out through its subsidiary. As a result, relevant international hacking companies find ways to camouflage themselves in the domestic market through resellers or founding a subsidiary with national representatives that has tighter relationships with Brazilian authorities.

Another field of difficulties concerns the process of standardization of the portals. There is no standardization, from the user experience perspective, in the way the searches are carried out – which disrespect common requirements in open data guidelines worldwide.[44] The vendor search could be either on the homepage or hidden within the portal's architecture. It was necessary to learn and relearn with each new portal the engines and paths leading to a given contract. In addition, it was not uncommon for us to find problems on the websites, such as being out of order or being incompatible with browsers.

This experience report on Transparency Portals demonstrates several ways in which, in reality, opacity seems to be the rule, both for political and technical reasons. In a considerable number of cases, the dashboard of contracts had only generic terms, such as metadata which had descriptions that did not inform the exact product being purchased, any documents available, or the amount spent. When the documents were available – when there were any – we had a real notion of the equipment acquired by the government. The lack of standardization of websites, for example, of different states, demonstrates another way of making access to information more difficult. The cultural concealment practices make it more difficult for citizens to oversee state expenditures, their legalities, and consequently confront authorities.

The Case of the Requests for Information (RFI)

The process of requesting information through the LAI was equally difficult and substantially less productive than using the transparency portals. Requests were made to the secretariats responsible for public security policies in the 27 federative entities, as well as to the respective public prosecutor's offices (MPEs) of each state and the public prosecutor's offices at the federal level (MPFs). In addition, requests were sent to federal agencies such as the Federal Police, the Ministry of Defence, the Ministry of Justice and Public Security, and the Institutional Security Office, the institutional body responsible for the Brazilian intelligence agency.

Except for the federal agencies, which use the 'FalaBr' platform, all other entities use their own portals to request public information – which, if not standardized, configures a very dispersed fragmentation of means through which citizens get in contact with the public power. All these other entities use the Citizen Information Systems (SICs), and, in order to access them, registration in each one of them is required. Some of these websites only allowed registration after screening, making the process even slower. Therefore, the first step was to register on all the aforementioned SICs and wait for confirmation, which sometimes took several days.

This first step led to another problem: the number of messages sent from the platforms, ranging from the confirmation email request to the confirmation of completed registration, including the communication of the registered RFIs and, later on, their receipt by the agency. Not rarely, the requests were forwarded to other agencies who did not respond to the requests. The notifications of deadline extension for responding to the information request were also very frequent. Eventually, the requests were answered; but even after that, they continued the communication, whether for archiving our requests or for 'citizens' satisfaction' surveys. In this back and forth, there were 195 communications from the SICs in the inbox of the registered email. However, these were only the remaining ones, as all the initial emails sent for account request and confirmation were deleted. In other words, these 195 emails were only after the completion of the PAI. This entire paragraph serves to highlight a concerning phenomenon: when transparency is buried under an excess of information. Furthermore, the interface is not user-friendly. On the contrary, it often required reading multiple pages of bureaucratic protocol messages to find, within one of them, what was new in that PAI that prompted the email.

We have also identified a constant problem: the low quality of the systems of public service. In one state, for example, an expired security certificate made it impossible to track the information request for months. In another case, one of the systems asked for a set of information that would allow the requester to be identified, which is not legally mandatory, as any citizen should be able to request information while keeping their identity anonymous.

Furthermore, there was usually the option to deny, respond, or extend the deadline for providing the requested information. However, in a strangely convenient manner for the agencies, the denials came in external documents, without the possibility of appealing within the same protocol. This meant that appealing the denial involved, in these cases, an additional effort of collecting documents from the respective systems as well as the original questions in order to file a new request appealing against the denial. In other cases, the very redaction of the denials of different authorities (for example, secretaries of public security of different states, such as Pernambuco, Rio Grande do Norte, and Roraima, as well as the Ministry of Justice and Public Security) were literally the same – which points to a possible coordinated response to our requests.

These problems, however, are parallel or, if viewed as more intentional, complementary to a more or less planned opacity. Neglect, carelessness, and laziness – are they political strategies or operational accidents?

A Necessary Step Back: A Chronology of the National Transparency Regime in Brazil and its Barriers

The institutionalization of a national transparency policy in Brazil occurred gradually over the last 20 years, especially with the federal decree that enacted the *Law of Access to Information* in 2011. The LAI encouraged the implementation of technical guidelines for active transparency – that is, data made available despite any request – and passive transparency – that is, disclosure of information not initially available, based on citizens' request[45] – and has been referenced not only by scholars[46] and by a significant part of civil society[47] as a fundamental tool for the effective instrumentalization of the inspection of public governmental activities, but also considered a reference from an international standing point.[48] That comes from a long-established importance concerning the democratic instruments' access to information to advance the effectiveness of human rights in the country and to shift the long-term censorship rule of law inherited from the military dictatorship period.[49] Nowadays, the set of instruments that configures the national transparency policy is transversal to contemporary struggles towards different variations of social justice,[50] including the memory of disappeared political dissidents by the time of the dictatorship,[51] environment preservation agendas,[52] military activities,[53] corruption in the public sector,[54] and surveillance technologies' programmes, the *locus* of our study.

At first sight, the chronology of the development of a national transparency policy in Brazil suggests a substantial proportion and implementation of this agenda by the past, at least, four president administrations in order to gradually build a set of government initiatives that would provide tools for public oversight. In 2003, it founded the Forum for the Right to Access to Public Information (Fórum do Direito ao Acesso à Informação). In 2004, it was created the Transparency Portal (Portal de Transparência) of the Federal Government – the most substantial policy of active transparency. In 2009, it enacted the Complementary Law (Lei Complementar) nº 131/2009 that altered the text of the Law of Fiscal Responsibility (Lei de Responsabilidade Fiscal) to determine the continuous availability of detailed information about the execution of the public budget. In 2011, a federal decree created the LAI, a landmark regulation that consolidated the transparency regime in Brazil.[55] In 2016, it created the Open Government Partnership, an arrangement between civil society and the government to provide data on human rights violations and corruption, access to quality environment data, and so on, having Brazil as one of its founding members. And, finally, there was the regulation of the Open Data Policy that took place in 2016. According to the Global Information Rating, an initiative that assesses legal frameworks for the right to information globally, Brazil appears in the 28th position, among 139 countries.[56]

Concerning the LAI, the set of guidelines that creates the foundation of the national transparency regime, consolidated in the law, is based on key principles and fundamental rights established in the federal constitution. Among them are Article 5, XIV[57] and XXXIII;[58] Article 37, caput, and §3, II;[59] as well as Article 216[60] that underline the fundamental principles that must guide the public administration, among them the principle of publicity of administrative acts.[61] According to Orion Augusto Platt Neto, Flávio da Cruz, Sandra Rolim Ensslin, and Leonardo Ensslin,[62] the principle of publicity underlines the notion that the people have the right to know about the acts practised by the public administration so that they can exercise social control

and democratic power. To that extent, a strict dialogue is built together with the 'principle of the supremacy of the public interest over the private', typical and foundational of the dogmatic doctrine of administrative law.[63]

The general rule[64] concerning the disclosure of public information defines, as well, exceptions based on which public authorities are not obliged to turn over the requested data. A mosaic of legal exceptions can also be drawn – and, for the purposes of our study, was definitive for the public authorities not to hand over the information required and classify documents about the use of hacking tools. Based on the LAI, for example, Article 7, §1, states that it is not mandatory for public entities to provide information whose secrecy is necessary to the safety of society or the state.[65] Nevertheless, some questions remain: What are the specific parameters for a public entity to conceal information? How do they assess if the information is of public interest or could endanger national security? Would the simple appointment of a *dispositif* of a law be enough in order to justify for the information to be classified, or further and more qualified justification would be necessary?

From Theory to Reality: Three Dimensions of Limits to Transparency Effectiveness

Although we can argue, on the one hand, that this legal framework establishes an ideally proportional and legitimate general rule, on the other hand, the political practice faces challenges that contribute to the continuous effort of the public sector to remain opaque concerning meaningful and public-interest information, which becomes even more noticeable when it comes to surveillance activities.

It becomes clearer that the legal framework deals with a collision of principles in the trajectory of the transparency policy in Brazil,[66] having on the one side a claim for structural changes for the disclosure of information – brought about by civil society organizations and members of the legislative branch – and on the other side an effort to uphold public information pushed by public sectors related to the armed forces[67] and law enforcement activities.[68]

With that background, we highlight three dimensions of limits to the effectiveness of the LAI that have appeared as barriers for the achievement of an effective transparency landscape in the country, especially when the focus is the use of surveillance tools by state actors, and more specifically about hacking tools. The first of these dimensions relies on a misleading understanding of the rule of law with the instrumentalization of the Brazilian General Data Protection Law (LGPD) of 2020, with the objective of systematically denying access to information of public interest.[69] Based on this rationality, for instance, the Military Police recently denied the access to salary information of a policeman suspected of the murder of the city counsellor and human rights activist Marielle Franco and her driver Anderson Gomes[70] in 2018, because the information would supposedly violate the privacy of the suspect. During Bolsonaro's administration, the GSI (directly attached to the presidency) also denied access information,[71] requested via the LAI, about who has visited the Palácio do Planalto (where the official office of the president is located), based on a misleading supposition that it would violate the LGPD,[72] just to name a few cases.

Second, another dimension points to a structural and historical problem regarding an apparent clash between the regulation of the intelligence service due to the nature of intelligence activities – which naturally requires a sufficiently secure degree of secrecy for the effectiveness of the activity – and the reported inefficiency of intelligence accountability bodies, such as the Joint Committee for the Control of Intelligence Activities (Comissão Mista de Controle das Atividades de Inteligência, CCAI).[73] This is fuelled by two other layers of opacity, typical from the intellectual property protections granted to services and technologies traded with the international market and deployed for public– private partnerships. First, companies could not be obliged to provide access to the architecture of their systems because of proprietary rights – although they contract with public actors – and, second, the very content of the public contract instruments – such as the service hired or tool acquired, the budget, their purposes, and their functions – fall under secrecy classifications since they, in theory, could jeopardize 'national' or 'state' security.[74]

Finally, there is another legal dimension, characterized by the 'abuse of right' (*abuso de direito*)[75] that makes excessive and default use of exceptions in order to not provide the requested data. This is in addition to when the requests are simply ignored – a problem we can call a systematic 'bureaucratic disobedience'.[76] That means generalist denials of information without pointing out legal basis or justifications, non-compliance with strict deadlines provided by the LAI, or even a total absence of responses to our requests – which happens to be not circumstantial in the case of information requested to law enforcement agencies.[77] The denials do not have to be applied to the entire content of documents requested (they can, for instance, classify specific sensitive sections) and, if not disclosed at all, they have to be meaningfully seasoned, exposing reasons that base the risk assessment to, for example, national or state security.

In the case of our study, the last two problems – secrecy culture of opaqueness regarding intelligence activities and bureaucratic disobedience – appeared as structural barriers that have been instrumentalized by the inquired public institutions so that they do not have to comply with the RFI. Also, it is very symptomatic that some of the terms of classification were elaborated after our RFI, meaning that the requested information was not classified by the time of the inquiry. While platforms have been developing justified threat assessments that facilitate security claims when deploying privacy-enhancing technologies, such as strong encryption. In the same way, government authorities have to justify, also based on threat assessments, their adoption of secrecy policies that block access to information and mandatory transparency provisions.

For example, when the Secretary of Security of the states of Goiás, São Paulo, and Rio de Janeiro were asked about the existence of such tools and conditions for their use, the denial to grant the documents was based on the Term of Information Classification Information (Termo de Classificação de Informação, TCIs), a legal provision that states the justification for the denial, based on national security reasons, for instance. Nevertheless, several of the TCIs were signed *after* the RFIs. It is possible to conclude that the public conduct points to a systemic disrespect of due administrative process regarding the decision-making procedure about the classification of public-interest critical documents.

Although all the aforementioned challenges compose a networked political mosaic that breaks a further development of the enforcement of a national transparency policy, the last two political arrangements in particular illustratively suggest barriers concerning a democratic institutional that could lead to effective oversight and accountability necessities of intelligence services based on the LAI.

A Second Step Back: The Heritage and Oversight of Intelligence Activities in Brazil as a Structural Problem

Accessing information on government actions is considered a necessary condition to keeping the citizens' will within government activities.[78] In the context of compatibilizing such representativeness with the necessary autonomy governments need to act, mechanisms of control and oversight, especially of intelligence agencies, and activities tend to be fragile and uncertain, as Marco Cepik[79] states. External control is necessary to ensure both the accountability and the responsibility of governments to citizens, but such a requirement enters in a route collision with the secrecy culture that constitutes the intelligence sector.[80] As Cepik reminds us, the problem is that opacity is a pervasive trait of intelligence activities. Thus, such phenomena enter in conflict with the transparency of government actions, which is one of the most defended requisites of contemporaneous political practice and 'democracy principal not fulfilled promise'.[81] Secrecy, as Edward Shils defines, is a compulsory retention of knowledge, secured by sanctions and legal penalties. Shils points out the intentionality and the regulatory elements of secrecy. Such specificity would be useful to define 'public secrecy'. Taking from Kim Lane Scheppele, there would be five information categories regulated by public secrecy: national defence, external policy, judicial process, intellectual property and patents, and citizens' privacy.[82]

Observing the federal intelligence services in Brazil, especially from the time frame that has the military dictatorship period in Brazil as the reference (1964– 1985), a political and institutional heritage from that time can be perceived and suggests the contemporary fragility of the oversight of such activities. Between 1964 and 1985, the intelligence services were mostly operated by the National Service of Information (Serviço Nacional de Informação, SNI),[83] under the Doctrine of National Security (Doutrina de Segurança Nacional).[84] In 1975, the National System of Information (Sistema Nacional de Informação, SISNI) was created and designed to process information. Historically, the intelligence system was focused on political oppositionists, such as anarchists and leftists.[85] The set of human rights violations during the period is widely documented,[86] and the role played by the intelligence system also contributed to the surveillance and persecution of victims of the dictatorship.[87]

During the re-democratization period and, consequently, with the formation of the Constituent Assembly responsible for drafting the federal constitution of 1988, a set of negotiations were made for a gradual transition from the military to a democracy regime.[88] Although the SNI was extinguished and it became the actual Brazilian Intelligence Agency (Agência Brasileira de Inteligência, ABIN), with intelligence services being expected to count on civil control,[89] the regulation of intelligence services was not consistently addressed and remained employing a military modus operandi.[90] The Law nº 9.883/1999 instituted the Brazilian System of

Intelligence (SISBIN) together with the ABIN, and, only 14 years later, Resolution nº 2/2013 has regulated the external control body, the CCAI, created in 2000 and composed by parliamentarians of the Federal Senate and the Chamber of Deputies.[91]

The oversight arrangement of intelligence activities is composed by internal and external bodies.[92] We focus, nevertheless, on the external aspect of it, based on the legislative control through the CCAI as a means to underline the participatory and transparency levels (or the lack of them) of social control. The work of the CCAI has been broadly documented as ineffective.[93] Cepik[94] raises two hypotheses for the historical low effectiveness of the committee: first, the lack of expertise of the member-congressmen, together with the lack of attention to such thematic by the electorate, which leads them not to be sufficiently involved; second, the tendency of the congressmen to be easily co-opted by members of military forces, leading to a cordial relationship between the CCAI and members of the intelligence service, which can jeopardize the necessary distance for a democratic and impersonal oversight to take place. According to a former analyst of the ABIN, there is no criticism to be made about the oversight of the intelligence in Brazil because it simply does not exist.[95] The reports made annually by CCAI are also not accessible to the public, making it impossible to assess their activities from a social perspective.

Carl von Clausewitz[96] points to the necessity of an informational asymmetry, which can be transformed for military and war advantages, or, in other words, power imbalances. However, if intelligence activities are occurring in democratic contexts, who can be the enemies? Which groups, organizations, and individuals could be surveilled? Additionally, it seems that citizens' privacy is less and less protected by public secrecy as law enforcement agencies and the intelligence sector gain more and more informational powers over everyday life through intrusive tools of digital forensics, spyware, or something similar. Such an offensive against citizens' privacy seems to place them in the position of enemies, which is coherent with a growing tendency of securitization and militarization of modern society.

The work of law enforcement in Brazil is also very problematic and widely documented in literature as institutionalized racism, militarism, and a systematic violator of human rights.[97] In terms of numbers, in 2021 there were 6,145 deaths as a result of police interventions. These forces also have an authoritarian past and present, as they were important pieces of military dictatorship that took place in 1964 and lasted more than two decades. Nowadays, there are systematic cases of police abuse,[98] including torture, formation of militias,[99] and illegal surveillance.[100] These abusive tendencies in Brazilian police forces include systematic practices of privacy violations[101] and opacity.[102]

Conclusion

In the first two sections, we showed that our study analysed the institutional relations between the Brazilian state and hacking tool vendors, internationally and nationally, offering also a political panorama of the phenomena worldwide. We found several contracts that encompass data extraction technologies and remote hacking. Our two first hypotheses – that commercial agreements were in place for many years between a diversity of vendors and Brazilian authorities, and that these surveillance tools aimed to extract data and remotely exploit vulnerabilities, including spywares, were being used widely in the national territory, in both law enforcement and intelligence spheres – were proved. Qualitatively and quantitatively, we were able to find 219 contractual instruments that showed that there is an ecosystem of hacking tools operating in Brazil, configuring a political economy crucial for but also enabled by Brazilian law enforcement authorities, making it a more established, unregulated international market of vulnerabilities and legitimized by a diversity of public actors. Although not the focus of this chapter, we also concluded that the country does not count on a modern and resilient legal framework that takes into account modern concerns with privacy, data protection, and cybersecurity.

In sections three and four, we spoke about the transparency regime and the historical opacity of intelligence services in Brazil, which became clear in our empirical research. The experience of instrumentalizing the access to information tools, from transparency portals to requests for information, showed that the due administrative process disrespects deadlines regularly, is broadly generic when denying data, and has little standardization. Some of the 'terms of classification' of documents were also done just after the requests for information in many cases, and the very information in the dashboard of portals was very scarce, leaving just metadata that would not grant meaningful information for a democratic oversight by civil society, academics, or even control bodies within the government. It was also clear that the oversight of intelligence and law enforcement activities is far from being lawfully accessible to citizens and control bodies, such as the CCAI. That maintains and even broadens the margin for opaque surveillance activities, from right to privacy and freedom of expression, association, manifestation, and cybersecurity. As we pointed out, transparency has not been fully achieved in Brazil.

On matters of security policies, the scenario is even worse. The legitimate monopoly of violence is often considered the foundation of the modern state as it is essential to exercise control over a population and territory.[103] Thus, managing effective social control over it is probably one of the biggest challenges to securing the right to transparency within modern democracies. It is somewhat ironic or tragic that while access information can even be considered essential to democratic regimes, it is systematically denied to citizens. 'All states have been "information societies", since the generation of state power presumes reflexively monitored system reproduction, involving the regularized gathering, storage, and control of information applied to administrative ends.'[104] In this chapter, we pointed out how difficult it can be to effectively access information about the methods and tools of information collection and production that are easily available to law enforcement agencies.

The conclusion is that, even with a well-recognized transparency regime, Brazil still lacks the means to enable its proper implementation in the real world. Further research, advocacy, and policymaking work must be done to align the country with the world's best transparency standards, improving and respecting not only the legality, proportionality, and necessity principles in the field of surveillance infrastructures but also in every government activity that can negatively impact individual and collective autonomy. Although our case study was based on Brazil and this was not a comparative-oriented work, the tensions between opacity and transparency can also be further explored through the lens of experiences in other countries, including investigative journalists, non-governmental organizations (NGOs) or civil society, and researchers.

Notes

1 The authors would like to express their sincere gratitude to Mariana Canto for her valuable and insightful contribution during the research project, which has greatly enhanced the quality of this work.

2 Brought about by organizations such as Citizen Lab, Forbidden Stories, and Amnesty International. See 'Targeted Subjects', Citizen Lab, 2023, https://citizenlab.ca/category/research/targeted-threats (accessed 27 June 2023); 'About the Pegasus Project', Forbidden Stories, 18 July 2021, https:// forbiddenstories.org/about-the-pegasus-project (accessed 27 June 2023).

3 Bill Marczak, John Scott-Railton, Sarah McKune, Bahr Abdul Razzak, and Ron Deibert, 'Hide and Seek', Citizen Lab, 18 September 2018, https:// citizenlab.ca/2018/09/hide-and-seek-tracking-nso-groups-pegasus-spywareto-operations-in-45-countries (accessed 27 June 2023).

4 Jamil Chade, 'Lava Jato negociou programa espião Pegasus com empresa israelense', UOL Notícias, 2021, https://noticias.uol.com.br/colunas/ jamil-chade/2021/07/26/lava-jato-negociou-programa-espiao-pegasus-comempresa-israelense.htm (accessed 27 June 2023).

5 Lucas Valença, 'Carlos Bolsonaro intervém em compra de aparelho espião e cria crise militar', UOL Notícias, 2021, https://noticias.uol.com.br/politica/ ultimas-noticias/2021/05/19/briga-entre-militares-e-carlos-bolsonaro-rachaorgaos-de-inteligencia.htm (accessed 10 July 2023).

6 Business and Human Rights, 'Brasil: Empresa de software espião Pegasus abandona licitação do governo', Centro de Informações sobre Empresas e Direitos Humanos, 2021, https://www.business-humanrights.org/pt/ latest-news/brasil-empresa-de-software-espi%C3%A3o-pegasus-abandonalicita%C3%A7%C3%A3o-do-governo (accessed 10 July 2023).

7 'Entidades questionam no TCU contratação de software de espionagem', Conectas Direitos Humanos, 2021, https://www.conectas.org/noticias/ entidades-questionam-no-tcu-contratacao-de-software-de-espionagem (accessed 10 July 2023).

8 ABIN is the Brazilian Intelligence Agency.

9 Germano Oliveira, 'O serviço secreto pessoal do presidente', *Isto É*, 29 April 2020, https://istoe.com.br/o-servico-secreto-pessoal-do-presidente (accessed 27 June 2023).

10 Lucas Valença, 'Além do Pegasus, Carlos Bolsonaro queria sistema para monitorar o Planalto', Uol Notícias, 2021, https://noticias.uol.com.br/ politica/ultimas-noticias/2021/08/03/alem-do-pegasus-carlos-bolsonaroprevia-sistema-para-monitorar-planalto.htm (accessed 10 July 2023).

11 On which telecommunication networks are based, such as phone calls and SMS.

12 Rafaela Mansur, 'MPF-MG abre investigação preliminar sobre programa usado pela Abin para monitorar celulares', *G1 Minas Gerais*, 24 March 2023, https://g1.globo.com/mg/minas-gerais/noticia/2023/03/24/mpf-mg-abreinvestigacao-preliminar-sobre-programa-usado-pela-abin-para-monitorarcelulares.ghtml (accessed 30 June 2023).

13 In order to categorize 'government hacking' and considering that these expeditions have received different names in literature, we consider government hacking as a vulnerability, whether intentional or unintentional, known or not, by the manufacturer, that results in unauthorized access to information, whether communication or data at rest or in transit; second, from a behavioural point of view, this exploitation involves intentionality,

14 The study was only published in Portuguese, unfortunately. Nevertheless, the publication has been the focus of many newspaper covertures, among them the biggest in the country, Folha de São Paulo, and also central media vehicles responsible to the most recent revelations of scandals in the country, such as The Intercept Brasil, Agência Pública, Tecnoblog, and Outras Palavras. See, for example, Patrícia de Campos Mello, 'Gastos com sistema espião disparam em estados e no governo Bolsonaro', Folha de S.Paulo, 12 November 2022, https://www1.folha.uol.com.br/poder/2022/11/

gastos-com-sistema-espiaodisparam-em-estados-e-no-governo-bolsonaro.shtml (accessed 8 July 2023); Paulo Motoryn, 'ABIN comprou programa que pode espionar internet', *The Intercept*, 19 April 2023, https://www.intercept.com.br/2023/04/19/abincomprou-programa-que-pode-espionar-internet (accessed 8 July 2023); Caio de Freira Paes, 'ABIN de Ramagem gastou R$ 31 milhões com ferramentas de vigilância secretas e sem licitação', Agência Pública, 25 April 2023) https://apublica.org/2023/04/abin-de-ramagem-gastou-r-31-milhoes-comferramentas-de-vigilancia-secretas-e-sem-licitacao/#:~:text=Desde%20o%20 in%C3%ADcio%20do%20governo,das%20empresas%20que%20as%20fornecem (accessed 8 July 2023); and 'Direito de Resposta', *Outras Palavras*, 8 July 2023, https://outraspalavras.net/outrasmidias/direito-de-resposta (accessed 8 July 2023).

15 Urs Gasser, Jack Goldsmith, Susan Landau, Joseph Nye, David O'Brien, Matt Olsen, Bruce Schneier, and Jonathan Zittrain, 'Don't Panic: Making Progress on Going Dark Debate', Berkman Center for Internet and Society at Harvard University, 1 February 2016, https://cyber.harvard.edu/pubrelease/ dont-panic/Dont_Panic_Making_Progress_on_Going_Dark_Debate.pdf (accessed 10 July 2023).

16 Beforehand, we disclose that the analysis of the legal framework is not the objective of this paper.

17 AccessData was the company acquired by Exterro in 2020 and responsible for the development of the solution Forensic Toolkit (FTK).

18 Suntech was the Brazilian subsidiary of Verint Systems in Brazil. When the parent company split between Verint Systems – focused in business intelligence – and Cognyte – in cyber intelligence – the Brazilian company changed the name to Cognyte Brasil. However, in the transparency portals, the contact email still used '@suntech.com', as well as maintained the same National Register of Legal Entities (CNPJ).

19 The set of hacking tools and services, including the capabilities, are thoroughly detailed in a table in our study.

20 UFED stands for the 'Universal Forensics Extraction Device' series of products. See 'Cellebrite UFED', 2024, https://cellebrite.com/de/cellebriteufed-de (accessed 9 December 2024).

21 The Brazilian equivalent to the Federal Trade Commission in the United States.

22 Governo Federal, 'Conselho Administrativo de Defesa Econômica: CADE', Governo Federal, 12 July 2021, www.gov.br/pt-br/orgaos/conselhoadministrativo-de-defesa-economica (accessed 22 June 2023).

23 Verint Systems, 'Tactical Off-Air Intelligence Solutions', 2013, https://www. documentcloud.org/documents/885760-1278-verint-product-list-engagegi2-engage-pi2 (accessed 24 June 2023).

24 Approximately USD 10,269,900,00.

25 Claudio Dantas, 'EXCLUSIVO: A empresa que vendeu a "maleta hacker" para o esquema de Helder Barbalho', O Antagonista, 2 October 2020, https://oantagonista.uol.com.br/brasil/exclusivo-a-empresa-que-vendeu-amaleta-hacker-para-o-esquema-de-helder-barbalho (accessed 25 June 2023).

26 'Operação Chabu: Prefeito de Florianópolis e mais seis são denunciados por organização criminosa', *G1*, 7 February 2020, https://g1.globo.com/sc/santacatarina/noticia/2020/02/07/prefeito-de-florianopolis-e-mais-seis-pessoassao-denunciadas-por-organizacao-criminosa.ghtml (accessed 25 June 2023).

27 Italo Nogueira, 'Intervenção federal no RJ completa cinco anos com entrega em atraso', *Folha de São Paulo*, 16 February 2023, https://www1.folha.uol. com.br/cotidiano/2023/02/intervencao-federal-no-rj-completa-cinco-anoscom-entrega-em-atraso.shtml (accessed 25 June 2023).

28 Governo Federal, 'Portal de Transparência. Ações Decorrentes da Intervenção Federal no

Estado do Rio de Janeiro na Área de Segurança Pública', Decreto nº 9.288/2018, https://www.portal-transparencia.gov.br/programas-e-acoes/ acao/00QS-acoes-decorrentes-da-intervencao- (accessed 10 July 2023).

29 Jaqueline Deister, 'Intervenção militar: 10 meses depois, medida segue sem solução para a segurança no RJ', *Brasil de Fato RJ*, 6 December 2018, https://www.brasildefatorj.com.br/2018/12/06/intervencao-militar-10- meses-depois-medida-segue-sem-solucao-para-a-seguranca-no-rio#:~: text=Neste%20m%C3%AAs%20de%20dezembro%2C%20a,de%20 seguran%C3%A7a%20p%C3%BAblica%20no%20estado (accessed 25 June 2023).

30 Available at 'LEI Nº 9.296, DE 24 DE JULHO DE 1996', Presidência da República, https://www.planalto.gov.br/ccivil_03/leis/l9296.htm (accessed 9 December 2024).

31 Statute for Child and Adolescents, 1990, Brazil, available at 'LEI Nº 8.069, DE 13 DE JULHO DE 1990, Presidência da República, https://www. planalto.gov.br/ccivil_03/leis/l8069.htm (accessed 9 December 2024).

32 Law of Organized Crimes, 1998, Brazil, available at 'LEI Nº 12.850, DE 2 DE AGOSTO DE 2013', Presidência da República, https://www. planalto.gov.br/ccivil_03/_ato2011-2014/2013/lei/l12850.htm (accessed 9 December 2024).

33 Brazilian Civil Rights Framework for the Internet, 2014, Brazil, available at 'LEI Nº 12.965, DE 23 DE ABRIL DE 2014', Presidência da República, https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965. htm (accessed 9 December 2024).

34 General Data Protection Law, 2018, Brazil, available at 'LEI Nº 13.709, DE 14 DE AGOSTO DE 2018', Presidência da República, https://www. planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm (accessed 9 December 2024).

35 Brazilian Code of Criminal Procedure, 2010, https://www.planalto.gov.br/ ccivil_03/decreto-lei/del3689compilado.htm (accessed 9 December 2024).

36 Lillian Ablon, Martin C. Libicki, and Andrea A. Golay, *Markets for Cybercrime Tools and Stolen Data: Hackers' Bazaar* (RAND Corporation, 2014), DOI: https://www.jstor.org/stable/10.7249/j.ctt6wq7z6; Kelsey Annu-Essuman, 'An Analysis on the Regulation of Grey Market Cyber Materials', *Cornell International Affairs Review* 1, https://journals.library.cornell.edu/tmpfiles/ CIAR_8_1_1.pdf (accessed 26 June 2023).

37 Ana Bárbara Gomes Pereira, André Ramiro, Gustavo Ramos Rodrigues, Pedro Amaral, and Victor Barbieri Rodrigues Vieira, *Decálogo de Recomendações sobre Direitos Digitais e Produção de Provas* (Instituto de Pesquisa em Direito e Tecnologia do Recife e Instituto de Referência em Internet e Sociedade, 2021), https://irisbh.com.br/wp-content/uploads/2021/08/Decalogo-derecomendacoes-sobre-direitos-digitais-e-producao-de-provas-IRIS-IPRECCDR.pdf (accessed 10 July 2023).

38 For an overview, see Ronald Deibert, *RESET: Reclaiming the Internet for Civil Society* (House of Anansi Press, 2020).

39 C. Colleta, L. Heaphy, S. Y. Perng, and L. Waller, 'Data-driven Cities? Digital Urbanism and Its Proxies: Introduction', *Italian Journal of Science and Technology Studies* 8, no. 2 (2017): 5–18.

40 Lilian Ablon, Martin C. Libicki, and Andrea Golay, *Markets for Cybercrime Tools and Stolen Data: Hacker's Bazaar* (RAND Corporation, 2014), https:// www.jstor.org/stable/10.7249/j.ctt6wq7z6 (accessed 10 July 2023).

41 The time frame we were looking for was between 2015 and 2021.

42 For a lack of better translation. But according to the Brazilian Bidding Law, 1993, this exception can happen when, among its legal hypotheses, the product is exclusive to a company so that the government would find it anywhere else.

43 Stephanie Kirchgaessner, 'Management of Five Firms Linked to Pegasus Maker NSO Is

Moved to London', *The Guardian*, 4 January 2023, https:// www.theguardian.com/business/2023/jan/04/management-of-five-firmslinked-to-pegasus-maker-nso-is-moved-to-london (accessed 27 June 2023); Joyce Wells, 'Cognyte Launches as Stand-Alone Company Following SpinOff from Verint', *Database Trends and Applications*, 2 February 2021, https:// www.dbta.com/Editorial/News-Flashes/Cognyte-Launches-as-Stand-AloneCompany-Following-Spin-Off-from-Verint-145031.aspx (accessed 28 June 2023).

44 'Open Data, Software and Code Guidelines', Open Research Europe, https:// open-research-europe.ec.europa.eu/for-authors/data-guidelines (accessed 29 June 2023); 'Open Government Data', OECD, 20 May 2022, https:// www.oecd.org/gov/digital-government/open-government-data.htm (accessed 29 June 2023); 'The Annotated 8 Principles of Open Government Data', Open Government Data, https://opengovdata.org (accessed 29 June 2023).

45 'EBT - Avaliação 360º - 2ª Edição', Controladoria Geral da União, 2023, https:// mbt.cgu.gov.br/publico/portal/metodologia360edicao2/66#:~:text=Os%20 portais%20da%20transpar-%C3%AAncia%20s%C3%A3o,Lei%20de%20 Acesso%20%C3%A0%20Informa%C3%A7%C3%A3o (accessed 1 July 2023).

46 Ethel Capuano, 'Access to Information Law in Brazil: What the Implementation Data Reveal', *American Political Science Association*, 16 August 2021, https://preprints.apsanet.org/engage/apsa/articledetails/6119500e4cb47982992fcfec (accessed 1 July 2023).

47 Juliana Sakai, Manoel Galdino, Jessica Voigt, Hugo Salustiano, and Renata Galf, 'What Do Brazilian Citizens Use the Freedom of Information Law For? a Typology of FOIL Requests', Transparencia Brasil, 2019, https://www. transparencia.org.br/downloads/publicacoes/What%20do%20Brazilian%20 citizens%20use%20Freedom%20of%20Information%20Law%20for.pdf (accessed 3 July 2023).

48 Eneida Bastos Paes, 'A influência internacional na construção do direito de acesso à informação no Brasil', *Revista de Informação Legislativa*, 2012, https://www2.senado.leg.br/bdsf/bitstream/handle/id/496592/000959937. pdf (accessed 3 July 2023).

49 André de Carvalho Ramos, 'Control of Conventionality and the Struggle to Achieve a Definitive Interpretation of Human Rights: The Brazilian Experience', *Revista IIDH*, 2016, https://www.corteidh.or.cr/tablas/r36237. pdf (accessed 6 July 2023); Mariana Paranhos Calderon, 'A evolução do direito de acesso à informação até a culminação da Lei nº 12.527/2011', *Revista Brasileira de Ciências policiais*, July 2013, http://dspace.mj.gov.br/ handle/1/7801 (accessed 9 July 2023); Sergio Adorno and Nancy Cardia, 'The Importance of Access to Information, Past and Present: Human Rights in Contemporary Brazil' *American International Journal of Social Science* 20 (2013), https://nev.prp.usp.br/wp-content/uploads/2015/11/55.pdf (accessed 6 July 2023).

50 Nancy Fraser, 'Abnormal Justice', *Critical Inquiry* 393 (2008), DOI: https:// doi.org/10.1086/589478.

51 Viven Ishaq and André Saboia Martins, 'A importância do acesso às informações funcionais de militares para o esclarecimento da autoria de graves violações de direitos humanos investigadas pela Comissão Nacional da Verdade', *Revista do Aquivo: Arquivo Público do Estado de São Paulo*, 27 April 2016, http://www.arquivoestado.sp.gov.br/revista_do_arquivo/02/index.php (accessed 6 July 2023).

52 Érica Bezerra Queiroz Ribeiro and Bruno Amaral Machado, 'O Acordo de Escazú e o acesso à informação ambiental no Brasil', *Revista de Direito Internacional*, 2018, https://core.ac.uk/download/pdf/211947989.pdf (accessed 6 July 2023).

53 Luiz Fernando Toledo, 'Desclassificação tarjada: O sigilo de documentos das forças armadas brasileira no contexto da Lei de Acesso à Informação', MSc dissertation, Fundação Getúlio Vargas, 2021.

54 Murilo Borsio Bataglia and Ana Claudia Farranha, 'Corrupção, Transparência e CGU: Anal-

isando o Contexto para Implementação do Direito de Acesso à Informação', *Novos Territorios*, 2020, https://periodicos.ufba.br/index.php/ nausocial/article/download/33923/19659 (accessed 6 July 2023).

55 José Maria Jardim, 'A Lei de Acesso à Informação Pública: Dimensões PolíticoInformacionais', Encontro Nacional de Pesquisa em Ciência da Informação, 2012, https://brapci.inf.br/index.php/res/download/181093 (accessed 7 July 2023); Marcio Camargo Cunha Filho, 'Construção da transparência pública no Brasil: Análise da elaboração e implementação da Lei de Acesso à Informação no Executivo Federal (2003–2019)', LL.D thesis, Universidade de Brasília, 2019.

56 'Global Information Rating', Global Right to Information Rating, http:// www.rti-rating.org/country-data (accessed 7 July 2023).

57 Article 5, XIV: It is assured to everyone the access to information, protection of the secrecy of the source when necessary to the professional exercise … (our translation).

58 Article 5, XXXIII: Everyone has the right to receive from public institutions information of their particular, collective, or general interest, which will be granted under the legal due time, under risk of liability, and safeguard those [either public servants or civilians] from which the secrecy is fundamental to the security of the society and the state … (our translation).

59 Article 37, §3º, II: The access to the administrative record and information about the government … (our translation).

60 Article 216, § 2º: The public administration is responsible, in accordance with the law, for managing government documentation and taking steps to make it available for consultation to anyone who needs it (our translation).

61 Article 37: The direct and indirect public administration of any of the Powers of the Union, the States, the Federal District and the Municipalities will obey the principles of legality, impersonality, morality, publicity and efficiency ... (our translation).

62 Orion Augusto Platt Neto, Flávio da Cruz, Sandra Rolim Ensslin, and Leonardo Ensslin, 'Publicidade e Transparência das Contas Públicas: obrigatoriedade e abrangência desses princípios na administração pública brasileira', *Contabilidade Vista and Revista*, 2007, http://www.redalyc.org/articulo.oa?id=197014728005 (accessed 7 July 2023).

63 Humberto Ávila, *Theory of Principles* (Springer, 2017); Alice Gonzalez Borges, 'Supremacia do Interesse Público: desconstrução ou reconstrução', *Revista Eletrônica de Direito Administrativo Econômico*, https://shorturl.at/ gC169 (accessed 10 July 2023).

64 It is worth mentioning literature about 'secrecy as a rule' and how it is incompatible with a democratic regime in Brazil. See Gills Vilar-Lopes, 'Quando o segredo é a regra: Atividade de Inteligência e acesso à informação no Brasil', *Revista Brasileira de Inteligência*, 2017, https://www.academia.edu/35377669/Quando_o_segredo_%C3%A9_a_regra_Atividade_de_ Intelig%C3%AAncia_e_acesso_%C3%A0_informa%C3%A7%C3%A3o_ no_Brasil (accessed 13 July 2023).

65 The definition of such information is considerably broad and illustrated in Article 23 as those that pose risks to national defence and sovereignty, comprise international negotiations, put in risk life and safety of the population, compromise intelligence activities, and so on.

66 Karina Furtado Rodrigues, 'A política nas políticas de acesso à informação brasilieras: trajetória e coalizões', *Revista Administração Pública*, 2020, https://www.scielo.br/j/rap/a/nsqxzWDSh4yVPRLMhNZJkkB/? format=pdf&lang=pt (accessed 7 July 2023).

67 Sakai, Galdino, Voignt, Salustiano, and Galf, 'What Do Brazilian Citizens Use the Freedom of Information Law For?'

68 Gregory Michener, Evelyn Contreras, and Irene Niskier, 'Opacity to Transparency? Evaluating Brazil's Access to Information Law at 5 Years', working paper, FGV Direito Rio, 2016, https://

transparencia.ebape.fgv. br/sites/transparencia.ebape.fgv.br/files/transparencyandopacity_eng.pdf (accessed 7 July 2023).

69 Lara Haje, 'Acesso à informação não pode ser prejudicado por conta de Lei de Proteção de Dados, dizem especialistas', *Câmara dos Deputados*, 18 November 2021, https://www.camara.leg.br/noticias/828370-acesso-a-informacao-nao- pode-ser-prejudicado-por-conta-de-lei-de-protecao-de-dados-dizemespecialistas (accessed 7 July 2023); Fiquei Sabendo, Insper, and Fundação Getúlio Vargas, 'Impactos da LGPD nos pedidos de LAI ao governo federal', 2022, https://fiquemsabendo.com.br/transparencia/relatorio-lgpd (accessed 7 July 2023).

70 Marcelo Bruzzi, 'PM alega sigilo de 100 anos para não informar sobre salários de policial acusado de matar Marielle Franco', *G1*, 27 October 2021, https:// g1.globo.com/rj/rio-de-janeiro/noticia/2021/10/27/pm-alega-sigilo-de100-anos-para-nao-informar-sobre-salarios-de-policial-acusado-de-matarmarielle-franco.ghtml (accessed 7 July 2023).

71 Italo Nogueira, 'Governo Bolsonaro ignora Controladorias e dificulta acesso a dados sobre visitas ao Planalto', *Folha de São Paulo*, 4 September 2021, https://www1.folha.uol.com.br/poder/2021/09/governo-bolsonaro-ignoracontroladoria-e-dificulta-acesso-a-dados-sobre-visitas-ao-planalto.shtml (accessed 7 July 2023).

72 Information about who has access to the Palácio do Planalto is, historically, of public interest according to the Brazilian Office of the Comptroller General (CGU).

73 The set of regulations concerning the intelligence activities in Brazil is documented subsequently in the chapter.

74 See, for example, the case of Wikileaks, Netzpolitik, and other organizations and individuals sued for publicizing public documents. See Emily Cureton Cook, 'Bend Sues Activist in Public Records Fight', Oregon Public Broadcasting, 6 April 2021, https://www.opb.org/article/2021/04/05/bend-sues-activist-in-public-records-fight (accessed 10 July 2023).

75 Milton Flávio de Almeida Camargo Lautenschlager, 'Abuso de direito', *Enciclopedia Jurídica da PUCSP*, December 2021, https://enciclopediajuridica.pucsp.br/verbete/478/edicao-1/abuso-de-direito (accessed 7 July 2023).

76 Borrowing Henry David Thoreau's concept of 'civil disobedience' from the civil rights movement to 'counter-apply' the idea of the government being the actor who disobeys the laws in order to restrict citizens' rights.

77 Neto et al., 'Publicidade e Transparência das Contas Públicas'.

78 Roy Peled and Yoram Rabin, 'The Constitutional Right to Information', *Columbia Human Rights Law Review* 357, no. 42 (2011): 257–401; Deidre Curtin, 'Overseeing Secrets in the EU: A Democratic Perspective', *Journal of Common Market Studies* 52, no. 3 (2014): 1–17.

79 Marco Cepik, 'Segurança Nacional e Controle Público: Mecanismos institucionais existentes', *Contexto Internacional* 23, no. 2 (2001): 295–359.

80 Edward Schils, *The Torment of Secrecy: The Background and Consequences of American Security Policies* (Ivan R. Dee, 1996).

81 Cepik, 'Segurança Nacional e Controle Público', 11.

82 Kim Lane Scheppele, *Legal Secrets: Equality and Efficiency in the Common Law* (Chicago University Press, 1988).

83 Alexandre João Cachoeira and Joel Cezar Bonin, 'A atividade de inteligência no Brasil: uma contextualização histórica', *Ponto de Vista Jurídico*, 23 May 2023, https://periodicos.uniarp.edu.br/index.php/juridico/article/view/3096/1526 (accessed 8 July 2023).

84 The National Security Doctrine was a decree prepared by the Escola Superior de Guerra

(Superior School of War) that allowed the dictatorial regime to pursue and eliminate 'internal enemies' – that is, those who were considered by the dictatorship as threats to the established order for questioning and opposing the authoritarian regime. See 'Doutrina de Segurança Nacional', Memórias da Ditadura, 2023, https://memoriasdaditadura.org.br/saibamais/ doutrina-de-seguranca-nacional (access 10 July 2023).

85 Gibran Ayupe Mota, Henrique Geaquinto Herkenhoff, Pablo Lira, and Erika Ferrao, 'Constitucionalização da Atividade de Inteligência: Perspectivas e Desafios Brasileiros', *Revista Brasileira de Segurança Pública*, 23 December 2018, DOI: https://doi.org/10.31060/rbsp.2018.v12.n1.912.

86 'Comissão Nacional da Verdade', Comissão Nacional da Verdade, December 2014, https://apublica.org/wp-content/uploads/2020/01/relatorio-finalcomissao-nacional-da-verdade.pdf (accessed 8 July 2023); Marcelo Godoy, *A casa de vovó: uma biografia do DOI-CODI (1969–1991), o Centro de Sequestro, Tortura e Morta da Ditadura Militar* (Alameda, 2014).

87 Karin Sant' Anna Kossling, 'As lutas anti-racistas de afro-descendentes sob vigilância do DEOPS/SP (1964–1983)', master's thesis, Universidade de São Paulo, 2007, https://www.teses.usp.br/teses/disponiveis/8/8138/tde01112007-142119/pt-br.php (accessed 9 December 2024); Leonardo Fetter da Silva, 'Sob Suspeita e Vigilância: O monitoramento dos movimentos e grupos de Direitos Humanos pelo Serviço Nacional de Informações (1978– 1985)' *Sillogés*, 2020, https://www.historiasocialecomparada.org/revistas/ index.php/silloges/article/view/97 (accessed 8 July 2023).

88 Maria D'Álva G. Kinzo, 'A democratização brasileira: um balanço do processo político desde a transição', *São Paulo em Perspectiva* 15, no. 4, DOI: https:// doi.org/10.1590/S0102-88392001000400002.

89 According to Jorge Zaverucha, 'Civil control is the ability of constituted authorities (Executive, Legislative and Judiciary) and organized civil society (trade unions, associations, press etc.) to limit the autonomous behavior of the Armed Forces, eliminating, as a result, authoritarian enclaves within the state apparatus' (our translation). Jorge Zaverucha, 'De FHC a Lula: A Militarização da Agência Brasileira de Inteligência', *Revista de Sociologia e Política* 16 (2008): 177–195, 178.

90 After revelations of surveillance abuse by ABIN during Bolsonaro's administration, the agency was removed from the GSI, traditionally commanded by the military, and placed under the Civil Office, a body with the status of ministry that directly assists the presidency in carrying out functions related to the administration of other bodies at the federal level and providing advice on political and institutional relations. We understand this step is, although initial, necessary towards a democratic oversight of intelligence services in Brazil. See Casa Civil, 'ABIN passa a integrar a Casa Civil', 2023, https://www.gov.br/abin/pt-br/centrais-de-conteudo/noticias/ abin-passa-a-integrar-a-casa-civil (accessed 10 July 2023).

91 Resolução nº 2, DE 2013-CN, laying down the Joint Commission for the Control of Intelligence Activities (CCAI). It is a permanent commission of the National Congress acting as an external control and inspection body for intelligence activities, provided for in Article 6 of law nº 9,883 of 7 December 1999.

92 Resolução nº 2, DE 2013-CN, laying down the Joint Commission for the Control of Intelligence Activities (CCAI). It is a permanent commission of the National Congress acting as an external control and inspection body for intelligence activities, provided for in Article 6 of law nº 9,883 of 7 December 1999.

93 Marco Cepik, *Espionagem e democracia* (Editora FGV, 2003); Simone Pereira do Vale, 'A Accountability Horizontal Exercida pela CCAI Sobre a Atividade de Inteligência Realizada pela ABIN no Período 2007-2014', MSc dissertation, Universidade Federal da Paraíba, 2021.

94 Cepik, *Espionagem e democracia* .

95 Marina Marandino, 'O Controle Parlamentar das Atividades de Inteligência no Brasil', LL.B monography, Pontifícia Universidade Católica do Rio de Janeiro, 2014.

96 Carl von Clausewitz, *On War* (Jazzybee Verlag, 1950).

97 Myrian Sepúlveda dos Santos, 'Memória e ditadura militar: Lembrando as violações de direitos humanos', *Tempo Soc.* 33, no. 2 (2021), DOI: https:// doi.org/10.11606/0103-2070.ts.2021.177990; Cecília MacDowell Santos, 'Memória na Justiça: A mobilização dos direitos humanos e a construção da memória da ditadura no Brasil', *Revista Crítica de Ciências Sociais*, 2010, DOI: https://doi.org/10.4000/rccs.1719.

98 Agence France-Presse, 'Shock over Brazil Police "Torture, Executions" in Drug Raid', France 24, 27 May 2022, https://www.france24.com/en/livenews/20220527-shock-over-brazil-police-torture-executions-in-drug-raid (accessed 8 July 2023).

99 Daniel Hirata and Maria Isabel Couto, 'Mapa dos Grupos Armados no Rio de Janeiro', Fórum de Segurança Pública, 19 October 2022, https:// fontesegura.forumseguranca.org.br/mapa-dos-grupos-armados-no-rio-dejaneiro (accessed 8 July 2023).

100 'Doutrina de Segurança Nacional'.

101 Gustavo Rodrigues, 'Acesso policial a celulares no Brasil e a banalização da "criptoanálise de mangueira de borracha"', *Instituto de Referência em Internet e Sociedade*, 26 October 2022, https://irisbh.com.br/acesso-policiala-celulares-no-brasil-e-a-banalizacao-da-criptoanalise-de-mangueira-deborracha (accessed 9 December 2024).

102 'Que arma é essa? Ranking de transparência de dados sobre armas de fogo nos estados', Instituto Igarapé, https://quearmaeessa.igarape.org.br (accessed 8 July 2023).

103 Charles Tilly, *Coercion, Capital and European States, AD 990–1992* (1990).

104 Anthony Giddens, *The Nation-state and Violence* (University of California Press, 1987).

6. DIGGING INTO EU DATA LAWS AND THEIR IMPACT ON AFRICAN RESEARCHERS[1]

PAUL ESSELAAR

There is a well-known saying that when the United States (US) sneezes, the rest of the world catches a cold. While this used to be true for Africa, the European Union (EU) has become the single most important market for African goods, with Africa exporting 33 per cent of its goods to the EU and importing 31 per cent of its goods from the EU.[2] In addition, the EU is the largest source of foreign direct investment in South Africa,[3] and legislative changes to the EU have an inevitable and significant impact on African countries. This has been referred to as the 'Brussels effect' and essentially shows how the EU effectively exports norms and regulations to other countries. Anu Bradford[4] sets out three main preconditions for the Brussels effect, namely (*a*) market size, (*b*) laws that are precise, comprehensive, available in multiple languages, and easy to copy, and (*c*) flexible drafting which allows the laws to work across different legal systems.

While the Brussels effect refers to the influence of Europe on other countries, it has been particularly noticeable in the digital space, where African countries have been in a flurry of activity to enact legislation to protect personal data, most of which has been published in the last 10 years.[5] A good deal of the motivation for this has been to harmonize their laws with the EU's General Data Protection Regulation (GDPR), 2016, to the extent that 'of the 60 countries that have enacted new data protection laws over the last decade, almost all modelled their approach in full or in part on the GDPR'.[6]

The Brussels effect has influenced not only the laws the African countries but also their regulators, and even court decisions have followed the approach of the Court of Justice of the EU.[7] Regional organizations such as the Economic Community of West African States (WAEMU) and the Common Market for Eastern and Southern Africa (COMESA) have also been modelled on the EU, to such an extent that the Court of Justice for WAEMU ruled that the Treaty of Dakar (which established WAEMU) should be interpreted with reference to the Treaty of Rome (which founded the European Community and the jurisprudence of the Court of Justice of the EU).[8]

The Brussels effect refers not only to the *de jure* influence of Europe, but also a *de facto* influence. An example of this effect is African farmers' food safety practices which are largely determined by the EU.[9] Another example was the EU decision to prevent the South African company De Beers from buying rough diamonds from the Russian company Alrosa[10] – effectively resulting in an international prohibition even though neither company was located in Europe. Indeed, if an international merger is prohibited in the EU, then it is effectively banned worldwide, even if it is deemed acceptable by regulators in other jurisdictions.[11] In short, when the EU changes its laws relating to data, there is a good chance that African countries will be forced to follow suit sooner or later.

Bearing in mind the Brussels effect, the remainder of this chapter considers the recent developments in the regulation of data by the EU. The chapter concludes with a call on the EU to

expand the scope of its impact assessment to include countries outside of it and, in so doing, acknowledge the de facto and de jure Brussels effect on African researchers.

The EU Strategy for Data

In 2020, the European Commission released a European strategy for data[12] in which it outlined its vision for the regulation of data. The strategy referenced existing legislation – such as the regulation of the free flow of non-personal data (FFD)[13] which focuses on preventing EU members from implementing data localization – and outlined a strategy to create a single European market for data, which would hold not only data from the EU but also from around the world and include personal data, non-personal data, and sensitive business data, including high-quality product data.[14] In particular, the strategy focused on the flow of data and the incorporation of European values, particularly personal data protection, consumer protection, and competition law.

The strategy further noted that 'sensitive data (e.g. health data) in public databases is often not made available for research purposes, in the absence of capacity of mechanisms that allow specific research actions to be taken in a manner compliant with personal data protection rules'.[15] The European Open Science Cloud is an example of an initiative that provides access to *European* researchers; the creation of the common European Health Data Space[16] aims to assist with preventing, detecting, and curing diseases, as well as allowing for informed, evidence-based decisions to improve the accessibility, effectiveness, and sustainability of healthcare systems.

Although there are many observations to make on the EU strategy for Data, there are two aspects to highlight for our purposes: (*a*) data needs to be managed in a holistic manner, taking into consideration consumer protection, competition law, and so on – not just from a data protection or privacy perspective, and (*b*) the EU is open to receiving data from around the world but is less focused on providing access to the data for non-European researchers.

Putting the EU Strategy for Data into Law

Since the publication of the EU strategy for Data, there have been several EU laws focusing on data regulation that have been published, namely the Digital Services Act (DSA), 2022;[17] the Digital Markets Act (DMA), 2022;[18] and the recent Artificial Intelligence Act (AIA), 2024,[19] as well as the Data Act (DA), 2023[20] (collectively referred to as the EU data laws). Each of these Acts is discussed briefly in order to contextualize the EU's holistic approach to data regulation.

At a high level, the *DSA* seeks to manage illegal content, ensure advertising is transparent, and combat disinformation in large (over 45 million users) platforms,[21] which are designated as very large online platforms (VLOPs) and very large online search engines (VLOSEs). It does this by placing obligations on social media platforms and search engines to address illegal content, disclose how their algorithms work to the regulators, and provide transparency on how decisions are made to remove content and how advertisers target users. It was fully implemented in 2024. In short, the Act is focused on expanding consumer rights with regard to digital services.

The *DMA* is a sister piece of legislation that aims at preventing large companies (gatekeepers[22]) from abusing their market power by preventing self-preferencing (providing prominence to its own products), reuse of personal data, providing business rights to smaller companies, prohibiting contractual requirements to offer the best deals exclusively on the companies' platform, preserving device neutrality (where pre-installed applications can be deleted), and prohibiting the bundling of products. In short, the DMA is focused on encouraging competition in digital markets.

The *AIA*[23] is a relatively groundbreaking piece of legislation and seeks not only to articulate the principles of artificial intelligence (AI) regulation but also to amend various other EU laws to bring them in line with the AIA. The AIA creates and defines 'high-risk' AI systems,[24] provides standards for risk management[25] and data governance,[26] and, interestingly, explicitly mandates the use of human oversight, including the ability to stop the AI using 'a "stop" button'.[27] The AIA also requires that each EU member country have a notifying authority,[28] national supervisory authority,[29] and market surveillance authority[30] to monitor and manage AI use. Generative content created by AI systems – including so-called deepfakes – must be labelled, and summaries of the data used to train the AI must be provided to users.[31]

Finally, the *DA*[32] is designed to regulate the Internet of Things (IoT). It aims to do the following:

Facilitate access to data by consumers and businesses by creating legal certainty as to when data can be requested or must be provided, as well as setting out contractual rules for sharing of data.

Force businesses to provide data to public bodies, primarily in emergency situations, but also when business to government data sharing is justified.

Facilitate switching between cloud and edge services, where cloud (typically with a central data centre) can send data to the 'edge' (typically outside of a data centre) which is closer to the user who is able to collect data.

Control access to data by non-EU countries in order to enhance trust in the EU data economy.

Develop interoperability standards to make it easier to move services between service providers which would include smart contracts that could be based on predetermined conditions set up by the user.

Integrate the DA with other EU legislation, particularly the GDPR, and respect the confidentiality of all data in equipment used by the user (terminal equipment).[33]

These EU data laws are the natural progression of the EU strategy for data and illustrate a holistic approach to managing data, rather than focusing exclusively on personal data. At a fundamental level, these EU data laws rely on data categorization in order to regulate the operation of data, and it is to how data is categorized that we now turn.

One Concept to Rule them All: Data Categorization

A careful look at the legislation governing data reveals that each of the EU Acts governing data relies on data categorization. For example, the GDPR only governs 'personal data', and any data that is not considered to be 'personal' is outside the scope of the GDPR. Similarly, the DA manages 'product data' which is generated from 'connected products' linked to the internet (IoT). Since the correct categorization of data is so fundamental to the operation of the legislation, it is critical that the distinction between the different types of data is clear and unambiguous. In the section that follows, some of the problems with data categorization are highlighted, and in the subsequent section its knock-on impact is discussed.

Beginning with the distinction between 'personal data' and 'non-personal data', this concept has seen some significant sea changes in the EU, which is surprising, considering how entrenched the concept of personal data was thought to be. One particular development occurred on 26 April 2023 in the case of *Single Resolution Board v. European Data Protection Supervisor* in the General Court (Eighth Chamber, Extended Composition)[34] where the court was required to provide guidance on whether pseudonymized data was 'data' or 'personal data'. Up until this case, it was considered to be settled law that 'pseudonymized data' was always 'personal data'.[35]

In this matter, a central resolution authority within the banking union in Europe provided data to Deloitte. The information was pseudonymized by means of a unique and randomized 33-digit alphanumeric code which was assigned to each record. It was later discovered by some of the data subjects who submitted personal data to the Single Resolution Board (SRB) that their data had been provided to Deloitte, and they duly laid complaints with the European Data Protection Supervisor (EDPS) that the SRB had breached its duties in terms of Article 15 of Regulation 2018/1725[36] by failing to indicate that it was providing personal data to third parties in its privacy statement. The SRB refuted this complaint and argued that, in the hands of Deloitte, the data was not personal data as the company had no reasonable prospect of re-identifying the data.

In upholding the SRB's position against the EDPS, the court indicated that in the hands of Deloitte, the data was not personal data, and so there was no obligation of the SRB to disclose the sharing of the data with Deloitte. Specifically, the court referred to the test as set out in Recital 16 of Regulation 2018/1725 which indicates that all objective factors, including (but not limited to) (*a*) cost of reidentifying, (*b*) amount of time required to reidentify, (*c*) available technology at time, and (*d*) expected technological developments, should be considered when determining whether the information is 'data' or 'personal data'. The court also endorsed the approach of *Breyer v. Bundesrepublik Deutschland*, 19 October 2016, Case C-582/14, that re-identification was not reasonably possible if 'the identification of the data subject was prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant.'[37]

The SRB case is a critical watershed in data protection law in that it provides some assistance in how personal data could be differentiated from non-personal data. In particular, the following points are worth emphasizing: (*a*) the test was to determine whether the data was subjective (that is, in the hands of Deloitte) rather than approached from the perspective of a reasonably competent and objective third party, and (*b*) the lack of a 'legal' manner to re-identify the data would result in the data being considered to be de-identified.

The SRB case has since been appealed by the European Data Protection Board (EDPB) on the grounds that the court misdirected itself when it required that the EDPB assess whether the data was personal or not, misinterpreted whether pseudonymized data was personal or not, and placed the onus on the EDPS to prove that the SRB had effectively anonymized the data it was processing.[38]

While the definitive text on personal data is the GDPR, it could be argued that the DA should be confined to non-personal data or 'product data'. While this seems to be mostly the intention, the DA itself does not do this and deals with both personal data and non-personal data which emanates from 'connected products'.[39] The definition of 'connected product' in Article 2(5) is clearly a critical definition on which the DA hinges and is defined thus:

> . . . an item that obtains, generates or collects data concerning its use or environment and that is able to communicate product data via an electronic communications service, physical connection or on-device access, and whose primary function is not the storing, processing or transmission of data on behalf of any party other than the user.[40]

From this definition, it is possible to extrapolate that 'product data' is data which (*a*) emanates from a product, (*b*) is about the product itself or its use in its environment, (*c*) is communicated via the internet, and (*d*) excludes data from, for instance, telecommunications platforms and server farms.

Logically, this definition also means that the DA creates a distinction between 'product data' and 'non-product data', which some commentators have described as both artificial and confusing.[41] This is not the end of the difficulties as there is now a new category of 'exportable data', which refers to input and output data (including data which is co-generated by the customer and the data holder) but excludes data which is protected by intellectual property rights or constitutes a trade secret.[42] Two further new definitions are introduced of 'related service data': (*a*) data generated by the use of the product and (*b*) 'readily available data' which refers to data which can obtained without disproportionate effort.[43]

On a practical level, these definitions demand that data holders – holders of product data – need to undertake a significant amount of work to restructure their underlying framework for data in order to place it into the various categories as required by the DA. An example of these 'categories' which a data holder would be required to implement would be (*a*) product data (data generated by the use of the product)[44] which is personal, (*b*) product data which is personal and exportable, (*c*) product data which is personal and readily available, (*d*) non-product data which is personal, (*e*) non-product data which is personal and exportable, and (*f*)

non-product data which is personal and readily available. (The categorization can continue with product and non-product data which is non-personal.)

While the DA may attempt to differentiate data into these constituent categories, the data itself resists the attempt to fit into these definitions. Moreover, it is unfortunate that businesses will be required to implement significant changes to their underlying information technology (IT) structure without having a clear indication that the manner in which they do so will be consistent with the DA.

In the section that follows, we discuss how difficulties with data categorization affect African researchers.

Barriers to International Data Transfer of Data to African Researchers

African researchers have been highlighted in this chapter as they are less numerous, poorly resourced, and have a greater workload as compared to their peers in the rest of the world. For example, a Malawian qualitative study into the challenges facing African researchers found that lack of funds, mentorship, interest by policymakers, and a heavy workload all contribute to the challenge for African scientists.[45] Although some real progress has been made to develop African researchers, this comes off a very low base. In its report on a decade of development in Sub-Saharan African research, the World Bank noted that international collaboration (and thus access to international data) was a key requirement for African research.[46] Even more telling was the fact that collaboration with extra-regional partners outside of the Sub-Saharan African region amounted to 42 per cent to 79 per cent of research, far exceeding the inter-regional collaboration at 0.9 per cent to 2.9 per cent.[47] This, in turn, results in a tendency for African researchers to have an asymmetric relationship with their international partners and for them to adopt the research agenda of international partners.[48]

In short, African researchers rely on access to the Global North for funding, collaboration, and access to data. It is against this background that the impact of the EU data legislation is considered.

Lack of Clarity on Data Categorization

As illustrated earlier, the DA creates a new category of 'product data', which, in turn, creates a knock-on effect on organizations who will need to firstly define and then implement internal controls to be able to differentiate between different types of data to create controls to manage it. This is a thankless task as not only are there new categories of data, but the boundaries to the categories themselves also are likely to change as understanding of DA matures.

Even what was considered to be the settled concept of pseudonymized data being personal data has been called into question. Additionally, as technology and mathematical approaches are developed, previously de-identified data may become identifiable, making pseudonymization and anonymization techniques obsolete or ineffective.[49]

This lack of certainty means that it becomes difficult for researchers to place the data they need in a category and so be able to determine what rules apply to it.

Lack of Clarity on Adequate Measures

In its explanatory memorandum to the earlier version of the proposed DA, the European Commission notes that 76 per cent of its respondents were concerned about access by foreign authorities to non-personal data based on foreign legislation, with 19 per cent indicating that it was a major risk.[50] This is understandable bearing in mind that there are various countries – such as Russia – which are actively antagonistic towards the EU. While EU data laws have the potential to facilitate access to data by EU member states, it is quite possible – even likely – that they will hamper non-EU states from being able to access data.

For example, Article 32(1) of the DA provides that data processing services must take 'all adequate technical, organisational and legal measures, including contracts, in order to prevent international and third-country governmental access and transfer of non-personal data held in the Union where such transfer or access would create a conflict with Union law or with the national law of the relevant Member State'. It is likely that there will be a great deal of uncertainty as to what 'adequate' measures are that need to be taken and how these measures differ from the requirements of international transfers in terms of the GDPR. In the face of this uncertainty, it will simply be easier (and more legally certain) for researchers to collaborate with their peers in Europe. This has the potential to lump African researchers into the same basket as Russian researchers.

At a high level, Article 32 of the DA is mostly aimed at administrative bodies, such as a regulator or the courts and tribunals of a foreign country, rather than researchers or their organization. But what about local legal and political interference with regard to African researchers?

Political Interference with African Researchers

If an African researcher enters into an agreement with their European counterpart to engage in a joint research project, it would be much easier for the African administrative authority or court to order the local researcher to divulge the product data than to attempt to prosecute the same case in Europe. This then creates another problem, where it is not the integrity of the researcher but rather the political and legislative climate of the country in which the African researcher resides that is of concern. This seems borne out by the Scholars at Risk (SAR) network which found that of the 285 reported attacks on higher education in 2021, 76 of these (or 26 per cent) came from African countries.[51] A specific example can be found in Egypt where four members of the intellectual community – two professors, a human rights activist, and a novelist – were also arrested for demanding that the state take measures to guard against COVID-19 outbreaks in prison.[52] This then raises the – unfortunately realistic – fear that it is not only the integrity of the African researcher that is of concern but also the likelihood of state or political interference that must be considered before deciding to share data with an African researcher.

Administrative Burden

Starting with the DA, the administrative burden on African researchers to get access to data is similar to obtaining a Schengen visa for data. While this may not seem to be a large obstacle, this can amount to a significant barrier for African researchers and could well result in EU–African collaboration becoming undesirable. Certainly, it is less than clear that the claimed benefits of the DA will 'far outweigh the associated administrative costs'[53] for African researchers. On the contrary, it seems likely that the introduction of the EU data laws will amount to something akin to a non-tariff trade barrier for African researchers.

Similarly, the DSA refers to 'vetted' researchers,[54] and it is only a researcher who has achieved this status who has the right to access data from the VLOPs (such as Google). Article 40(8) sets out seven requirements for a researcher to be granted the status of a 'vetted' researcher, which include data security and confidentiality requirements that African researchers may struggle to achieve. However, a more serious obstacle facing an African researcher is the requirement that 'the sole purpose of conducting research [must be] that [it] contributes to the detection, identification and understanding of systemic risks *in the Union*'.[55] This suggests that an African researcher requesting access to data to investigate systemic risks in an African country would never be considered to be 'vetted' as the research does not relate to the EU, which, in turn, suggests that African researchers will only be considered when the subject matter of the research involves the EU. This re-emphasises the point mentioned earlier that African researchers tend to adopt the research agenda of their international partners.

To place this in perspective, a European researcher could request access to Meta's data about election interference in the EU in terms of the DSA, but an African researcher has no similar law to ask Meta for data about election interference in Tanzania.

Agreement Template for Data

One of the key aspects of the GDPR was the guideline on when to share personal data and how to manage the sharing if you, as the data controller, decide to do so. Over the years, this has become relatively mature, and tools, such as those provided by the United Kingdom's Information Commissioner's Office,[56] have become quite sophisticated. In contrast, there is no template for the safe transfer of product data outside of the EU, and this uncertainty as to what should form part of such an agreement will inevitably result in a reluctance to share product data outside of the EU. While personal product data is already protected by the cross-border restrictions in the GDPR,[57] Article 32 of the DA now introduces similar protection for non-personal product data. Indeed, the DA itself has changed substantially from its original proposal to now include multiple references to 'model contractual terms', but whether these will include 'model contractual terms' to be able to send data to an African researcher is unclear.[58] On the positive side, the duty of the commission to provide the 'non-binding' model contractual terms now has a deadline of 12 September 2025.[59] A similar argument can be made for data made available in terms of the other EU Data Laws.

Lack of Hegemony in African Legal Systems

As the initial version of the proposed DA's explanatory memorandum notes,[60] regulating data at an EU member level is simply not effective and would lead to higher transactional costs, lack of transparency, legal uncertainty, and undesirable forum shopping.[61] This is equally true for approaching the regulation of data at an African Union (AU) level, rather than country level, where the differences in legislation are even more pronounced, despite the recent advent of the African Continental Free Trade Area (AfCFTA).[62] Unfortunately, the AU is nowhere close to the type of hegemony that the EU took decades to create. For example, the Convention on Cyber Security and Personal Data Protection[63] seeks to create a common vision of personal data protection and cyber security. While this convention is, with all its flaws,[64] a welcome development, some nine years after it was adopted, it only received the 15 ratifications required for it to come into force.[65] Instead, individual countries have adopted their own data protection laws which have resulted in some odd variations in data protection laws. This disharmony is illustrated by the DS-I Africa (Data Science for Health Discovery and Innovation in Africa) group tool[66] which compares the data protection laws of 12 English-speaking African countries in an attempt to assist a data controller to navigate their way through the disparate data protection laws. Some examples of the differences are as follows: (*a*) 3 of the 12 have a definition for pseudonymization, (*b*) personal data is sometimes referred to as 'personal information' and in some cases includes juristic persons,[67] and (*c*) 'consent' is not a defined term in Ghana's Data Protection Act of 2012.[68]

While this illustrates the problem in the area of data protection, it is worth emphasizing that none of this work has been done from a holistic data regulation perspective as most African countries simply do not have laws that deal with data holistically, let alone tools which facilitate multinational comparisons in data regulation.

Omission of Researcher Rights to Access Data

Unlike the Digital Services Act, the Digital Markets Act simply omits any reference to research or researchers completely, which in turn means that researchers have to attempt to leverage third-party requests for information in order to obtain the data they require. The failure to acknowledge the useful work provided by researchers is a surprising omission in the Digital Markets Act, particularly in the context of its sister legislation.

No Exemption for African Researcher Access to African Data

While the DA does make allowances for the user to receive his or her data from the data holder, it does not facilitate access by African researchers to product data generated from their country. Article 5 allows a user to request that the data holder provide their user information to an African researcher, but obtaining multiple individual consents to access their product data in this way may often not be practical for African researchers. In theory, an African researcher would be subject to Article 6 (obligations of third parties receiving data at the request of the user), but that immediately raises the concern of the enforceability of the DA on an African researcher who is outside EU jurisdiction. The DA does not appear to deal with product data

emanating from a non-EU country. For example, a French multinational company deploying smart fridges in Ghana may get considerable product data from its Ghanian users which is repatriated to France. Once the product data is in France, it is unclear if Ghanian researchers would be able to get access to this product data.

Unfair Access by EU Authorities to Data Produced Outside of the EU

While the DA does require data holders to provide information to EU public sector bodies, it is not clear that the data must emanate from within the EU. Consider a situation where EU health authorities are concerned that there is an outbreak of foot-and-mouth disease in Rwanda, but this is denied by the local Rwandan public authority. The EU health authority may well want to requisition product data from German vaccine producers[69] in order to discover that the number of requests for vaccines in Rwanda has increased substantially, supporting a move by the EU health authorities to ban Rwandan meat products. Even more pernicious is the fact that this product data may have no personal data component, which could result in the EU authorities having better data on situations in the African countries than the African country itself.

The African Union Data Policy Framework

Up to this point, the focus has been on the digital strategy that the EU has adopted and how it has put this into practice. As pointed out previously, there is good reason to believe that African countries will follow a similar path due to the Brussels effect, and this portion of the chapter is dedicated to extrapolating what effect EU data laws will have on African researchers.

As already noted, African countries have finally caught up with the concept of protecting the data of individuals. However, there is a danger that data governance is considered solely from the perspective of the privacy of individuals, rather than from a more complex multidimensional legal approach. Not only does data have privacy considerations, but it also raises questions relating to other areas of law,[70] including competition law, consumer law, intellectual property law, and taxation.

The DA[71] has the potential to have a similar impact on African countries which have a demonstrable difficulty in keeping up with the regulation of technology. Put simply, the DA has the potential to regulate a sector of the (data) economy which African countries have not addressed at all. The purpose of this chapter is, at least in part, to stimulate discussion of the regulation of data holistically in Africa in order to avoid a situation where African countries are, once again, late to the party and caught with their proverbial pants down. That said, this chapter does not attempt to address every issue influencing data regulation, but rather focuses on the impact of insufficient protection of, and access to, data by researchers in Africa.

While African countries have not yet caught up with the developments in data regulation, it is heartening that the recent African Union Data Policy Framework (AUDPF), which was released in February 2022, is fully aware of the multidimensional nature of data.[72] The AUDPF notes that there are no global examples of umbrella laws which regulate every aspect of data, but

rather data is regulated in data protection law, competition law, cyber security law, electronic communications and transactions law, and intellectual property law,[73] and so regulatory bodies in these areas need to coordinate their actions.[74] While acknowledging that African countries have less developed laws on competition, data, and intellectual property, the AUDPF sees this as an opportunity to harmonize legislation between African states.[75]

The AUDPF also confirms the Brussels effect and notes that African countries are largely standard takers, rather than standard makers,[76] and that data-rich and data-intensive developed countries tend to create regulatory precedence.[77] Despite this, the AUDPF does not suggest any method to categorize data, but merely states that this should be done. Indeed, categorization of data is so important that the AUDPF recommends that one of the first actions by the data Information Regulator (IR) is the categorization of data[78] and the establishment of a common data categorization and sharing framework.[79] The AUDPF also recommends that the AU should be actively lobbying the EU regarding standards and laws relating to data.[80] Thus, despite being aware of the importance of data categorization as it is the foundation upon which all controlling regulation is based, Africa still seems set to accept that the categorization of data will be imposed on it by the EU.

The AUDPF also recognizes that there has been little restraint from competition or data regulators on the rise of monopolistic global platforms which are producing and extracting massive amounts of private data which has been commodified with seemingly little regard for the negative impacts on data subjects,[81] effectively making competition impossible for smaller players.[82]

In order to combat this, the AUDPF provides some recommendations on steps that can be taken by African countries, including:

1. The AU should be actively lobbying the EU regarding standards and laws relating to data.[83]
2. Fair contractual standards for public organizations should be created.[84]
3. Codes of practice for using data need to be developed.[85]
4. Data protection rights should be considered more important than intellectual property rights,[86] and contracts that give up digital rights, ignore personal data protection, and inhibit competition should be unenforceable.[87]
5. Novel regulatory ideas could be tested in 'regulatory sandboxes'.[88]
6. Data trusts should be created to manage control of data rather than cede complete control to the collecting entity.[89]
7. Universities should be included as relevant policy stakeholders to help establish a knowledge base from which the local data economy can draw scientific and technological knowledge.[90]
8. The competition chapter of AfCFTA negotiations should set minimum standards to ensure that putative proprietary non-personal data is available to innovators, entrepreneurs, and others in the value chain to encourage competition across the continent.[91]

In short, what the AUDPF has done is begin the conversation about regulating other aspects of data aside from the perspective of privacy, but African countries are years away from implementing legislation that deals with data holistically. Even if African countries were able to implement legislation similar to that put into place by the EU, it is unclear whether this would be desirable as African countries typically have far fewer funds available for regulatory bodies, and the EU itself is not sure what the impact of the data legislation in the EU will be; following too quickly in EU footsteps may ironically result in African countries taking the wrong path.

Consider, for example, the South African IR, which is the implementation mechanism of the South African Protection of Personal Information Act (POPIA) of 2020. In 2021–2022, the IR had a budget of approximately EUR 4.3 million[92] while the French data protection agency, Commission nationale de l'informatique et des libertés (National Commission on Informatics and Liberty, CNIL), had a budget of EUR 24 million.[93] What is particularly startling about this disparity is that France and South Africa have a very similar population size (approximately 63 million people), and yet the IR is expected to do the same work as the CNIL but with a fifth of the budget. Even a cursory look at each of the EU data laws already dealt with makes it clear that a crucial role is played by the regulatory bodies tasked with enforcing them. How precisely should African countries implement similar data laws if they will never have the funds to enforce them?

Cutting through the Red Tape: Enabling African Researcher Access

Up to this point, this chapter has focused on the – possibly unintended – barriers to data access for African researchers. The following section provides some suggestions on how to reduce the barriers for African researchers and further argues that, due to the Brussels effect which has introduced a kind of legislative neo-colonialism, Europe has a moral duty to expand its impact analysis of the EU data laws on the effect of these laws outside EU borders.

Standard Contractual Clauses for Transfers of Data

Despite there being considerable notice of the transition from Directive 95/46/ EC[94] to the GDPR in 2018, the standard contractual clauses for data transfers were only updated three years later.[95] Thereafter, in 2023, the first model template for the sharing of personal data by researchers, which combines the GDPR and the POPIA, was published in February 2023 by the DS-I Africa Law project.[96] While this effort is to be commended, this template does not, understandably, even attempt to either define or deal with the concept of 'product data' or 'exportable data' which researchers may need to access. In order to facilitate African researcher access, it would be helpful if this template were to be updated to consider the requirements of different types of data, so it is considered holistically rather than only from a privacy perspective. It would also be helpful if the standard contractual clauses for the transfer of product data would be made available by the EU in a much shorter time frame, in order to provide some assistance to African researchers wishing to access data.

Code of Conduct for Researchers

One way to ease the burden on African researchers would be for the EU to provide a guideline that researchers, governed by an approved code of conduct, would be able to receive data; or, put differently, researchers governed by that code of conduct would be considered to pass the test of Article 32(1) of the DA that says 'all adequate technical, organisational and legal measures, including contracts, in order to prevent international and third-country governmental access and transfer of non-personal data held in the Union where such transfer or access would create a conflict with Union law or with the national law of the relevant Member State'.

At present there are not many codes of conduct, and those that do exist are intended to address data protection, rather than data holistically. One such example is the draft Code of Conduct for Research promulgated in terms of the POPIA.[97] While this code does introduce concepts that are not present in its enabling legislation (such as pseudonymization),[98] it is clearly focused on personal data[99] and does not consider other data and how this would be managed for the purposes of research. As with the standard contractual clauses mentioned earlier, this may well be the time to start considering a holistic code of conduct that deals with all data types.

AU Conventions Dealing with Data

The AUDPF is a very useful and necessary step in the development of African policy on data, but it does come several years after the same step was taken by the European strategy for data. In a manner similar to the Malabo Convention, it may be useful for the AU to propose a convention which would incorporate the concepts put forth in the DSA,[100] the DMA,[101] the AIA,[102] and the DA.[103] That said, the pace of the ratification of the Malabo Convention would suggest that it would take decades before conventions of this nature would be adopted and ratified.

Interestingly, the AUDPF suggests that the use of a 'regulatory sandbox' could be appropriate for situations such as these.[104] Regulatory sandboxes are a regulatory approach, typically summarized in writing and published, that allow live, timebound testing of innovations under a regulator's oversight. Novel financial products, technologies, and business models can be tested under a set of rules, supervision requirements, and appropriate safeguards. This has the benefit of encouraging experimentation, reducing barriers to entry, and allowing regulators to get valuable insight into how to regulate the sector. In 2018, approximately 20 countries were actively exploring the concept of regulatory sandboxes.[105]

Evaluation of Impact of Data Laws on African Countries

Bearing in mind the novelty of the data laws, the EU wisely commissioned impact assessments and conducted participant studies and questionnaires over several years – for example, on the impact of the DA – but all of these were from an EU member country's perspective.[106] Unsurprisingly, this impact assessment did not meaningfully address the possible impact that the EU data laws could have on Africa and on African researchers.[107] This does not mean that EU data laws will not have an impact on Africa – just that there is no plan to measure it.

The EU also created a mechanism of ex-post evaluations in order to assess whether the objectives of EU data laws were, in fact, being realized. Once again, this is an entirely logical approach to developing novel regulations and appears to be standard practice for the EU. In contrast, while it is undoubtably a wise plan to commission both impact assessments and ex-post evaluations of legislation, in practice this tends not to be done for African countries, even for legislation that they themselves are implementing. As a result, there is a vanishingly small chance that the impact of EU data laws has been assessed by any African country, and it is also unlikely that any plans exist to evaluate the impact of EU data laws once they come into force. Instead, African countries and researchers are likely to start experiencing difficulties in an anecdotal way, similar to what they probably experienced when first encountering personal data protection laws.

While it may be too late to commission an impact assessment, an ex-post evaluation on the impact of EU data laws on African countries, companies, and individuals would be most useful to understand the local conditions and implications of regulating data and also the results of failing to do so. This has the potential added benefit of increasing regulatory harmony between African countries and the EU, which, in turn, would facilitate data transfer. Indeed, the implications of failing to consider and implement similar laws within African countries could well lead to a missed opportunity, particularly when, for example, the EU estimates that unused product data has the potential to unlock EUR 1.5 trillion (USD 15.842 trillion)[108] of value by 2027.[109] To put this in some perspective, the country with the highest gross domestic product (GDP) in Africa is Nigeria at USD 441 billion,[110] and the combined GDP of all African countries in 2021 amounted to USD 2.7 trillion. This means that the value of unused product data in Europe is equal to approximately half the GDP of the African continent.

Conclusion

By pioneering the need to unlock the value of data, the EU has a first-mover advantage over African countries. African countries, on the other hand, are desperately trying to keep pace with the speed of changes in the regulation of data. The EU should have, at least, some empathy for Africa because it was the EU that suffered from not being the first mover when it came to technical innovations, which may account for so many US tech companies and so few EU tech companies in the top 20 global tech companies.[111]

The introduction of EU data laws seems to have the potential to have the (unintended) consequence of further alienating African researchers and creating new barriers to cooperation with their EU counterparts. Moreover, their EU counterparts will have greater (if not perfect)[112] access to data. This means that EU researchers are likely to feel the benefits of EU data laws while the African researchers are likely to suffer the detrimental effects. This is ironic, particularly in light of the moral imperative for the EU to deal fairly with countries that it colonized and bearing in mind the amount of aid the EU provides to Africa.

If the EU, as a highly literate and technically advanced society, can be said to be only unlocking a fraction of the value of the data available to it, then that statement is surely even more applicable to African countries. There must be tremendous potential for researchers to

make meaningful differences in Africa, provided they can get access to data which could be unlocked by EU data laws. For example, it could be hugely beneficial if the EU would support African countries being able to access their own product data from global tech companies, such as Google and Amazon.

African countries should be closely watching the success (or failure) of EU data laws as facilitating access to data has the potential of bringing about an even more dramatic change in Africa than it would in the EU. This is more apposite when considering the relative lack of sophistication in existing African laws. To amend a well-known adage, give a researcher the answer and you solve the problem of the day, but give her access to data and you help her solve problems for a lifetime.

Notes

1 This work was supported by the United States's National Institute of Mental Health and the National Institutes of Health (award number U01MH127690) under the Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa) programme. The content of this chapter is solely the author's responsibility and does not necessarily represent the official views of the aforementioned institutes.

2 Eurostat Statistics Explained, 'Archive: Africa–EU – International Trade in Goods Statistics', February 2022, https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Archive:Africa-EU_-_international_trade_in_ goods_statistics (accessed 25 February 2023).

3 European Commission, 'EU Trade Relations with South Africa: Facts, Figures and Latest Developments', EU Directorate-General for Trade, https://policy.trade.ec.europa.eu/eu-trade-relationships-country-andregion/countries-and-regions/south-africa_en#:~:text=The%20EU%20represents%20the%20most,country's%20industrialisation%20and%20 transformation%20agenda (accessed 25 February 2023).

4 Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (Oxford University Press, 2020).

5 'Data Protection and Privacy Legislation Worldwide', UN Trade and Development, 14 December 2021, https://unctad.org/page/data-protectionand-privacy-legislation-worldwide (accessed on 25 February 2023).

6 Michael Pisa, Pam Dixon, and Ugonma Nwankwo, 'Why Data UNCTAD Protection Matters for Development: The Case for Strengthening Inclusion and Regulatory Capacity', Centre for Global Development, December 2021, https://www.cgdev.org/sites/default/files/why-data-protection-matters-fordevelopment.pdf (accessed 25 February 2023).

7 Bradford, *The Brussels Effect*, 80.

8 Bradford, *The Brussels Effect*, 80.

9 Bradford, *The Brussels Effect*, 94.

10 See, for example, Commission Decision in Case No. COMP/B-2/38.381 (De Beers), C(2006) 521 final (22 February 2006), cited in 2006 O.J. (L 205) 24.

11 Bradford, *The Brussels Effect*, 80.

12 'A European Strategy for Data', European Commission, 19 February 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A 52020DC0066 (accessed 7 June 2023).

13 Regulation (EU) 2018/1807.

14 Bradford, *The Brussels Effect*, 4–5.

15 'A European Strategy for Data', 7.

16 'A European Strategy for Data', 22.

17 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act)', https://eur-lex.europa.eu/ legal-content/en/TXT/?uri=CELEX%3A32022R2065 (accessed 8 June 2023).

18 'Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)', https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1925 (accessed 8 June 2023).

19 'European Parliament Legislative Resolution of 13 March 2024 on the Proposal for a Regulation of the European Parliament and of the Council to Lay Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and to Amend Certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106 (COD)) Regulation (EU) 2021/206 of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts', https://www.europarl.europa.eu/doceo/document/ TA-9-2024-0138_EN.pdf (accessed 8 April 2024).

20 'Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on Harmonized Rules for Fair Access to and Use of Data and Amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)', https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32023R2854 (accessed 28 March 2024).

21 A list of the very large online platforms (VLOPs) and very large online search engines (VLOSEs) was published on 23 April 2023. European Commission, *Digital Services Act: Commission Designates First Set of Very Large Online Platforms and Search Engines*, https://ec.europa.eu/commission/presscorner/ detail/en/IP_23_2413 (accessed 24 October 2023).

22 On 6 September 2023, the European Commission determined that six companies – Alphabet, Amazon, Apple, ByteDance, Meta, and Microsoft – were gatekeepers in terms of the DMA. 'Digital Markets Act: Commission Designates Six Gatekeepers', European Commission, 6 September 2023, https://perma.cc/KR8U-JQ8D (accessed 22 October 2023).

23 'European Parliament Legislative Resolution of 13 March 2024 on the Proposal for a Regulation of the European Parliament and of the Council on Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM (2021) 0206 – C9-0146/2021 – 2021/0106 (COD))', https://www.europarl.europa.eu/ doceo/document/ TA-9-2024-0138-FNL-COR01_EN.pdf (accessed 8 April 2024).

24 AIA, Article 6.

25 AIA, Article 9.

26 AIA, Articles 10-12.

27 AIA, Article 14(4)(e).

28 AIA, Article 28.

29 AIA, Article 70.

30 AIA, Article 74. The national authority carrying out the activities and taking the measures pursuant to Regulation (EU) 2019/1020.

31 AIA, Article 13(3)(b)(v).

32 'Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023'.

33 'Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Fair Access to and Use of Data (Data Act) COM/2022/68', https://eur-lex.europa.eu/legal-content/EN/TXT/? uri=COM%3A2022%3A68%3AFIN (accessed 8 June 2023); Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023'. Although this refers to the proposal of the DA rather than its final version, this vision of the DA remains apposite.

34 Case T-557/20 *Single Resolution Board v. European Data Protection Supervisor* [2023] ECLI:EU:2023:219, https://eur-lex.europa.eu/legal-content/EN/ TXT/?uri=CELEX%3A62020TJ0557 (accessed 15 December 2024).

35 See, for example, the statement by Thomas Zerdick, head of technology and privacy at the EDPS: 'Unlike anonymised data, pseudonymised data qualifies as personal data under the

General Data Protection Regulation (GDPR)'. Thomas Zerdick, 'Pseudonymous Data: Processing Personal Data while Mitigating Risks', European Data Protection Supervisor, 21 December 2021, https://edps.europa.eu/press-publications/press-news/blog/pseudonymousdata-processing-personal-data-while-mitigating_en (accessed 28 July 2023). For a South African perspective on whether pseudonymized data is always personal data, see D. W. Thaldar, 'Does Data Protection Law in South Africa Apply to Pseudonymised Data?' *Frontiers in Pharmacology*, 23 November 2023, DOI: https://doi.org/10.3389/fphar.2023.1238749.

36 'Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the Protection of Natural Persons with Regard to the Processing of Personal Data by the Union Institutions, Bodies, Offices and

Agencies and on the Free Movement of Such Data, and Repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC', [2018] OJ L 295, https://eur-lex.europa.eu/eli/reg/2018/1725/oj (accessed 15 December 2024).

37 'Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018'.

38 'Appeal Brought on 5 July 2023 by the European Data Protection Supervisor against the Judgment of the General Court (Eighth Chamber, Extended Composition) Delivered on 26 April 2023 in Case T-557/20, *Single Resolution Board v European Data Protection Supervisor* (Case C-413/23 P)', https://eurlex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62023CN0413 (accessed 15 December 2024).

39 DA, Article 2(5,15).

40 DA, Article 2(5).

41 Bertin Martens, 'How to Fix the European Union's Proposed Data Act', Bruegel, 4 December 2022, https://www.bruegel.org/blog-post/how-fixeuropean-unions-proposed-data-act (accessed 28 February 2023).

42 DA, Article 2(38).

43 DA, Article 2(16–17).

44 Note that this is owned by the data subject.

45 Save Kumwenda, El Hadji A. Niang, Pauline W. Orondo, Pote William, Lateefah Oyinlola, Gedeon N. Bongo, and Bernadette Chiwona, 'Challenges Facing Young African Scientists in Their Research Careers: A Qualitative Exploratory Study', *Malawi Medical Journal* 29, no. 1 (2017): 1–4.

46 World Bank Group and Elsevier, *A Decade of Development in Sub-Saharan African Science, Technology, Engineering and Mathematics Research* (2014), 48, https://documents1.worldbank.org/curated/en/237371468204551128/ pdf/910160WP0P126900disclose09026020140.pdf (accessed 15 December 2024.).

47 World Bank Group and Elsevier, *A Decade of Development*, 34.

48 J. Cerdeira, J. Mesquita, and, E. S. Vieira, 'International Research Collaboration: Is Africa Different? A Cross-Country Panel Data Analysis', *Scientometrics* 128 (2023): 2145–2174.

49 Emily M. Weitzenboeck, Pierre Lison, Malgorzata Cyndecka, and Malcolm Langford, 'The GDPR and Unstructured Data: Is Anonymization Possible?' *International Data Privacy Law* 12, no. 3 (2022): 184–206.

50 'Proposal for a Regulation of the European Parliament and of the Council'.

51 'Attacks on Scholars a Threat to Democracy in Africa: The Link between Decreased Academic Freedom and the Stagnation of Democracy', Reliefweb, 25 May 2022, https://reliefweb.int/report/world/attacks-scholars-threatdemocracy-africa-link-between-decreased-academic-free-

dom-andstagnation-democracy (accessed 23 October 2023).

52 'Attacks on Scholars a Threat to Democracy in Africa'.

53 'Proposal for a Regulation of the European Parliament and of the Council'.

54 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act)', https://eur-lex.europa.eu/ legal-content/en/TXT/?uri=CELEX%3A32022R2065 (accessed 8 June 2023), Recitals 92, 124, and 137, and Articles 39(3) and 40(8).

55 ''Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022', Article 40(4) as read with Article 40(8)(e) (emphasis added).

56 Elizabeth Denham, 'Data Sharing: A Code of Practice', Information Commissioner's Office, May 2021, https://ico.org.uk/for-organisations/ukgdpr-guidance-and-resources/data-sharing/data-sharing-a-code-of-practice (accessed 15 December 2024).

57 European Union, '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, Articles 44–50.

58 DA, Article 41.

59 DA, Article 41.

60 'Proposal for a Regulation of the European Parliament and of the Council'; DA, 7–8.

61 'Proposal for a Regulation of the European Parliament and of the Council'.

62 'Creating One African Market', AfTACF Symposium, August 2023, https:// au-afcfta.org (accessed 15 December 2024).

63 'African Union Convention on Cyber Security and Data Protection', https:// au.int/sites/default/files/treaties/29560-treaty-0048_-_african_union_ convention_on_cyber_security_and_personal_data_protection_e.pdf (accessed 15 December 2024).

64 Paul Esselaar, Alison Gillwald, Ashly Hope, Gavin van der Nest, and John Stuart, 'Aligning Data Protection Laws in Africa to Facilitate E-Commerce', Tralac, 1 June 2020, https://www.tralac.org/publications/article/14641trade-in-the-digital-economy-a-tralac-collection.html (accessed 15 December 2024).

65 Also known as the Malabo Convention. Only 13 countries have ratified the convention as of 28 February 2023. See https://au.int/sites/default/ files/treaties/29560-treaty-0048__african_union_ convention_on_cyber_ security_and_personal_data_protection_e.pdf (accessed 15 December 2024).

66 'Search and Compare Data Protection Legislation', DS-I Africa, https:// www.datalaw.africa/law/search_compare (accessed 15 December 2024).

67 South Africa's Protection of Personal Information Act (Act no. 4 of 2013).

68 Ghana, Data Protection Act, 2012 (Act 843), published in the Ghana Gazette No. 39, 16 October 2012.

69 In terms of DA, Articles 14 and 15.

70 D. W. Thaldar, B. A. Townsend, D-L. Donnelly, M. Botes, A. Gooden, J. Van Harmelen, and B. Shozi, 'The Multidimensional Legal Nature of Personal Genomic Sequence Data: A South African Perspective', *Frontiers in Genetics* 13 (2022), DOI: https://doi.org/10.3389/fgene.2022.997595.

71 'Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 Decem-

ber 2023'.

72 'AU Data Policy Framework', African Union, February 2022, https:// au.int/sites/default/files/documents/42078-doc-AU-DATA-POLICYFRAMEWORK-ENG1.pdf (accessed 29 July 2023).

73 'AU Data Policy Framework', 11.

74 'AU Data Policy Framework', 31.

75 'AU Data Policy Framework', 12.

76 'AU Data Policy Framework', 1, 17.

77 'AU Data Policy Framework', 1.

78 'AU Data Policy Framework', vii.

79 'AU Data Policy Framework', x, 20, 59.

80 'AU Data Policy Framework', 12.

81 'AU Data Policy Framework', 8.

82 'AU Data Policy Framework', 24.

83 'AU Data Policy Framework', 12.

84 'AU Data Policy Framework', 25.

85 'AU Data Policy Framework', 28.

86 'AU Data Policy Framework', 34.

87 'AU Data Policy Framework', 35. In *Discovery Limited and Others v. Liberty Group Limited*, a solution to the multitude of interests in data was defined, upholding both data protection and competition. In essence, the court held that in such disputes, if the data is personal in nature, it is 'owned' by the data subject, and competitors may not exclude others from accessing this information. ZAGPJHC 67, [2020], https://www.saflii.org/za/cases/ ZAGPJHC/2020/67.html (accessed 30 July 2023).

88 'AU Data Policy Framework', 39.

89 'AU Data Policy Framework', 47.

90 'AU Data Policy Framework', 56.

91 'AU Data Policy Framework', 57.

92 *Annual Report of the Information Regulator 2021/22*, https://www. inforegulator.org.za/wp-content/uploads/2022/10/Info%20Regulator%20 Annual%20Report%202021-22-compressed.pdf (accessed 24 October 2023).

93 'Overview on Resources Made Available by Member States to the Data Protection Supervisory Authorities', European Data Protection Board, 5 September 2022 https://edpb.europa.eu/system/files/2022-09/edpb_ overviewresourcesmade_availablebymemberstatestosas2022_en.pdf (accessed 24 October 2023).

94 This is the predecessor of the EU GDPR. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (the Data Protection Directive).

95 'Standard Contractual Clauses (SCC) Standard Contractual Clauses for Data Transfers between EU and Non-EU Countries', European Commission, 4 June 2021, https://commission.europa.eu/law/law-topic/data-protection/ international-dimension-data-protection/standard-contractual-clauses-scc_ en (accessed 15 December 2024).

96 Lee Swales, Paul Ogendi, Marietjie Botes, Beverley Townsend, Dusty-Lee Donnelly, Lukman Abdulrauf, and Donrich Thaldar, 'A Data Transfer Agreement (DTA) Template for South Africa', Zenodo, 6 February, 2023, https://zenodo.org/record/7537396#.Y_iYDexBxB (accessed 15 December 2024). See also L. Swales, M. Botes, D-L. Donnelly, and D. W. Thaldar, 'Towards a Data Transfer Agreement for the South African Research Community: The Empowerment Approach', *South African Journal of Bioethics and Law* 16, no. 1 (2023): 13–18.

97 'Government Gazette Vol. 695', Republic of South Africa, 12 May 2023, https://inforegulator.org.za/wp-content/uploads/2020/07/GovernmentGazzette-dated-12-May-.pdf (accessed 15 December 2024). For a discussion of the draft Code of Conduct for Research, see D. W. Thaldar and B. Townsend, 'Protecting Personal Information in Research: Is a Code of Conduct the Solution?' *South African Journal of Science* 117, nos. 3–4 (2021), DOI: https://doi.org/10.17159/sajs.2021/9490; A. Gooden and D. W. Thaldar, 'Despite Good Progress with Regard to the Proposed Code of Conduct for Research in South Africa, Unresolved Issues Remain', *Humanities and Social Sciences Communications* (2024), DOI: https://doi.org/10.1057/s41599-024-02715-0.

98 See, for example, clause 4.2.2.3 of the Code of Conduct, which states that researchers must 'ensure that the Personal Information is Pseudonymised unless there is a compelling reason why it is not feasible or appropriate'.

99 Referred to as 'personal information'.

100 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act)', https://eur-lex.europa.eu/ legal-content/en/TXT/?uri=CELEX%3A32022R2065 (accessed 8 June 2023).

101 'Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)', https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1925 (accessed 8 June 2023).

102 European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council to lay down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and to amend certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), https://www.europarl.europa.eu/doceo/document/TA9-2024-0138_EN.pdf (accessed 8 April 2024).

103 Ibid.; 'Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023'.

104 'AU Data Policy Framework', 39.

105 'Briefing on Regulatory Sandboxes', United Nations Secretary-General's Special Advocate for Inclusive Finance for Development (UNSGSA), https://www.unsgsa.org/sites/default/files/resources-files/2020-09/Fintech_ Briefing_Paper_Regulatory_Sandboxes.pdf (accessed 15 December 2024).

106 'Impact Assessment Report and Support Studies Accompanying the Proposal for a Data Act', https://digital-strategy.ec.europa.eu/en/library/impactassessment-report-and-support-studies-accompanying-proposal-data-act (accessed 8 April 2024).

107 'Public Consultation on the Data Act: Summary Report', https://ec.europa.

eu/newsroom/dae/redirection/document/81599 (accessed 15 December 2024).

108 This is as of as of 25 February 2023. https://www.xe.com/currencyconverter/

convert/?Amount=15000000000000&From=EUR&To=USD (accessed 25 February 2023).

109 '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 February 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593073685620&uri=CELEX%3A52020DC0066 (accessed 25 February 2023).

110 'African Countries with the Highest Gross Domestic Product (GDP) in 2021', Statista, 2021, https://www.statista.com/statistics/1120999/gdp-ofafrican-countries-by-country (accessed 25 February 2023).

111 'Market capitalization of Apple (AAPL)', Apple, 2023, https:// companiesmarketcap.com/tech/largest-tech-companies-by-market-cap (accessed 25 February 2023).

112 It is unclear why the right to access product data is limited to the user and the person nominated by the user in the DA. The first reference to 'research' is located in Article 21 which focuses on access by researchers to public bodies, rather than 'data holders'. Interestingly, the final version of the DA introduced Article 44(3) which indicates that the DA is without prejudice to union and national law, providing for 'access to and authorizing the use of data for scientific research purposes' which may turn out to be a backdoor into access by African researchers to data held in the EU.

PART III: PROMISES

7. A SUBJECT ACCESS REQUEST, THEN WHAT?: (UN)STRUCTURING ONLINE ANALYTICS FOR DATA INSTITUTIONS

JAKE STEIN AND REUBEN BINNS

Collecting, storing, analysing, and distributing information derived from disparate data sources have become radically easier in the past decade. The proliferation of big data analytics and algorithmic technologies is thanks to advances at every stage in the data analytics pipeline from ingestion to visualization, and at each layer of the data analytics stack from hardware to user interface.[1] Organizations with access to sufficient capital have access to tools which quickly and reliably convert upfront investment into valuable insights and system optimizations. Indeed, a vast and growing set of systems have themselves been progressively optimized to the needs of large organizations — resulting in generalizable workflows and plugand-play architectures for turning big data into material gains for enterprises with access to valuable troves of data.[2]

At the same time, it has become increasingly apparent that resisting the (mis) use of big data and algorithmic technologies requires the use of sophisticated data analytics too. Advocates seeking transparency, privacy, or agency with respect to these systems often analyse the very same data that platform corporations rely on for their own operations.[3] For efforts that resist platform information asymmetries through collecting and analysing data in pursuit of algorithmic transparency or regulatory action, achieving sufficient scale is not straightforward. The cost of technologies and contracted expertise necessary far exceed grassroots budgets. Big data analytics software has been developed with the needs of centralized, hierarchical structures of corporate practitioners in mind and is poorly suited to the distributed governance and open access to which activists and academics aspire. Despite new technologies' potential to be used in challenging information asymmetries, novel analytics technologies demand enormous upfront investment and sustained overhead cost. Big data is only valuable when it scales, and big is expensive.

One approach to mitigating the inherent gulf of resources between platforms and advocates who seek transparency and agency within corporate systems is to pool resources or form coalitions across civil society and academia under novel data governance models — what the Open Data Institute (ODI) has called 'Data Institutions'.[4] The ODI defines these broadly as 'organizations whose purpose involves stewarding data on behalf of others, often towards public, educational or charitable aims'.[5] Similarly, participatory action research (PAR) has allowed academia to become an important host for infrastructures supporting adversarial data aggregation, working with advocates to audit algorithmic systems or create tools capable of computationally mediating solidarity against overreaching algorithmic control.[6] These functions align closely with the ODI's specification of 'Bottom-up Data Institutions', referring to organizations containing 'processes that enable people — usually those that have generated the data or that the data is about — to actively take part in those data governance processes'.[7]

However, the academy faces a platformization crisis of its own, with universities increasingly supplanted as centres of computational power by cloud-computing providers that offer cheap flexible compute, while academic research budgets are meagre compared to those of corporations and university administrations closely guard valuable intellectual property.[8] While of course the flexibility of cloud computing can also accelerate computation-heavy research, the institutional power granted by a standing reserve of computational power is a card now held closely by industry.[9] Furthermore, access to enormous datasets generated on platforms leaves academic research at a disadvantage in data access. Data subjects, activists acting in their behalf, and academics all find themselves in similarly undesirable positions vis-à-vis platform hegemons, but also possess complementary resources for developing alternatives to serve the public interest.

Regulators have begun to recognize and address information asymmetries, but provide only meagre tools for data subjects or data institutions to combat asymmetries in practice. In many jurisdictions, data access rights face headwinds as they wrongfully pit data subject rights as restraints against governments' efforts to energize the development of artificial intellience (AI) and boost their constituent economies.[10] Individual data access regimes contained in data protection regulations like the General Data Protection Regulation (GDPR), 2018, and the California Consumer Privacy Act (CCPA), 2018, and aggregate data access rights for researchers contained in successive regulation like the Digital Services Act (DSA), 2024, as well as voluntary access to research application programming interfaces (APIs) provide a convenient side door, allowing activists and researchers to access data without facing the costs of constructing net-new data collection infrastructure themselves. Indeed, scholarship and litigation both have made significant progress in asserting the collective use of subject access rights for this purpose.[11] Data collection (or access) is, however, only the first step. Information asymmetries can only be counteracted when subjects have access to *information* contained in their data too, requiring suitable data analytics systems to complement their access.[12] Financing, creating, and governing these systems require navigating the incentives, constraints, and priorities of all of the parties involved, ranging from data subjects themselves to civil society funders, advocacy organizations, and academic researchers.

We examine the politics of this coordination as we navigate the challenges of co-constructing data pipelines alongside activists from Worker Info Exchange (WIE) which seek to use data subject access requests (DSARs) to afford workers transparency and agency within algorithmic management by platform work apps.[13] In particular, we address the obstacles that arise while contending with the structure that DSAR data contains in order to draw attention to the conflicting epistemologies layered in the schemas, ontologies, and metric definitions involved in its analysis.

Our collaboration with WIE exemplifies the broad variety of stakeholders necessary for building data institutions capable of counteracting platform control. Our engagement with WIE is partly supported by the 'Ethical Web and Data Infrastructure in the Age of AI' project – a research programme funded by the Oxford Martin School to design new web protocols and data architectures which can better promote data subject autonomy in the platform-dominated web.[14] We benefit from the collaboration as a rich case study in which to prototype and test our designs.

WIE is a non-profit organization closely aligned with the App Drivers and Couriers Union (ADCU), who represent platform workers in the United Kingdom (UK) and the European Union (EU). Accordingly, WIE is interested generally in providing drivers with better information about how to manage their own activities and income, often by helping drivers seek evidence to contest mistreatment mediated via algorithmic systems. WIE's association with the ADCU means it is also active in collective advocacy. In a landmark finding, one of WIE's suits with the ADCU established the precedent that DSARs can be used for the purpose of sourcing aggregate data in favour of collective advocacy in addition to privacy.[15] At the time of our collaboration, the law firm AWO was engaged with WIE in investigating algorithmic pricing and work assignment for platform workers. AWO is a UK 'law firm and consultancy that empowers individuals and organizations to uphold data rights, comply with the law and effect change in data protection and digital policy'.[16] Finally, the most critical stakeholders are the data subjects themselves – workers who use gig economy apps. Workers participate in this institution by requesting access to their data via WIE. WIE then processes the request with Uber on the driver's behalf, in exchange pooling that data in aggregate form. Driver data is held by WIE and has been shared with us for the purpose of conducting research and providing insights.[17] Workers stand to gain both in terms of greater transparency into their own data and from the results of collective advocacy that aggregate analyses of their data supports.

Taking the design decisions that we observed and made ourselves when building data structures as the site of our research, we illustrate the impacts of data structure on the meaningful access to information contained in it. With the hope of setting out a blueprint for future architectures for data institutions, we evaluate how various data structures serve or privilege individual perspectives within the tangle of incentives existing among the collaborating parties and stakeholders mentioned.

In the design of multi-stakeholder public-service data architectures, we emphasize the need to consider that the same underlying data might simultaneously serve numerous purposes or communicate divergent interpretations depending on its structure (how data is categorized in fields, organized into tables, and aggregated into metrics). By 'infrastructure', we mean the technical systems, databases, and organizational processes through which this data flows.[18]

To capture the polysemous quality of data, we borrow Slavoj Žižek's term 'parallax view' to analogize how data staged in a unified underlying architecture can support many (even conflicting) pursuits and epistemologies.[19] Put simply, a parallax effect occurs when two observations of the same event differ due to perspective. By borrowing the concept of the parallax, we explain that despite having access to the same underlying data (the *observed event*), the structure in which that data is provided (the *perspective*) can radically alter the information accessible within it. Žižek uses this phenomenon as an allegory for his system of reason, while here we use the term to describe the relationship between data and its structure, which is relevant to our pursuit in several ways. First, at the lowest level, it helps illustrate the contingency of data's meaning upon its context and which other data or knowledge it is enriched with. Second, the parallax could be regarded as an essential quality for data structures meant to serve different, even conflicting end goals – in our case, the empowerment of the individual to use their data to act more effectively *within* a system, but also for the collective critique of

that system with the goal of changing its rules and configuration to benefit the data subject. We see this as a divergence from the typical design goal of analytics systems geared toward establishing a unified interpretation of data. The economic and socio-technical demands of large-scale data analytics systems are in turn co-constructed and co-used by academics, advocates, and ultimately data subjects. Accordingly, supporting several contradicting uses or interpretations of data cannot be avoided, even if the overall end goal is the same. We conclude by suggesting which technical implementations can simultaneously support conflicting interests and epistemologies that different parties bring to the data, without engineering a single perspective into the data's structure.

The Implications of Data Structure on Meaningful Data Access

Though any data pipeline may contain a similar mix of technologies, their configuration depends on the volume, velocity, sensitivity, and target uses of data.[20] This section aims to illustrate how decisions about *where* structure is added along the data analytics pipeline and *what* structure is added (or left out) can obfuscate meaning in the data we have processed with WIE. Further, by evaluating the ways data is obscured in the structure returned by Uber, we demonstrate the importance of avoiding imparting restrictive structures of our own when designing technical architectures for data institutions.

Unpacking Obfuscation

The most straightforward ways data structures can impede meaningful access for subjects is through *obfuscation*. Finn Brunton and Helen Fay Nissenbaum famously define obfuscation as a tactic for data subjects to resist surveillance by making data 'more difficult to act on, and therefore less valuable . . . adding to the cost, trouble, and difficulty of doing the looking'.[21] In the context of analysing DSAR responses, obfuscation acts in the reverse, with advocates and academics 'doing the looking' into the behaviours of algorithmic systems on behalf of data subjects and platforms making analysis more difficult for them.

Well-documented examples of obfuscation may be as simple as returning data in non-machine-readable formats like PDFs (Portable Document Format) or providing too little contextualizing information for data to be interpreted.[22] Here we briefly inventory some of the obstacles to analysing data we encountered in data structures. The motivation of this analysis is two-fold: First, it demonstrates that access, human or machine readability, interoperability, and portability are all concepts that should be defined contingently on the intended use of the data in question rather than being held to an objective standard or format. This point hopefully also emphasizes the role data structure plays in influencing the meaning one can ultimately derive from the very same data. Second, this analysis warns that even enhanced access requirements which stipulate API or database availability under progressive transparency regulations like the DSA still leave data controllers many tools to obfuscate important meaning in data that they are compelled to share.

Presumably, unintentional obfuscation abounds in the data that WIE received via DSAR responses. For example, data exporting mistakes led to file errors when processing data – such as extra commas included in CSV (comma-separated values) files, where commas act as cell delimiters, thus preventing files from being immediately readable, or the inconsistent use of data fields with the same names.

Identifying Problem Structures

The structures that impeded our analysis of data occurred at three levels: the schema (or table), field (or column), and value (or cell) (Figure 7 9.1). Our analysis was obstructed by features of the data which were present in all of the responses and separately by inconsistencies in the structure within responses themselves or between responses when aggregated or compared. These structures and inconsistencies among them can be attributed to any number of origins. We expect some might arise from Uber's own internal systems changing over time to support different features which may require more, less, or different forms of data. They might, likewise, emerge from differences between formats convenient to the functional needs of Uber's systems and the transformations performed to prepare data request responses. Whether these layers of structure were intended to obscure elements of data or make the response compliant with the requirements of DSARs under the GDPR is not discernable. After all, a DSAR is far from the natural state of data otherwise meant to function as part of a living, moving operational system.

Further, Uber arguably cannot be expected to foresee all the calculations that the requester aims to conduct and to process data accordingly (unless of course they are instructed); nor are data controllers like Uber expressly obligated to deliver data in its most disaggregated form. However, we might argue that supplying data in a disaggregated format is a straightforward and uncontroversial requirement of interoperability and portability that should require little marginal effort.[23] One could even go as far as attributing some obfuscating data structures provided by Uber as good-faith yet misguided attempts to provide drivers with the metrics they may more easily interpret. Intended or not, inconsistencies in data structure had the effect of obfuscating some of the most contested and controversial aspects of Uber's algorithmic management practices and ultimately demonstrate the need for greater specificity in data standards or access provisions, while emphasizing the valuable lesson that counter-architectures should put off structure as long as possible to facilitate the maximum possible number of further aggregations. With that in mind, we briefly unpack some of the schema-, field-, and cell-level obfuscation which held back our analysis.

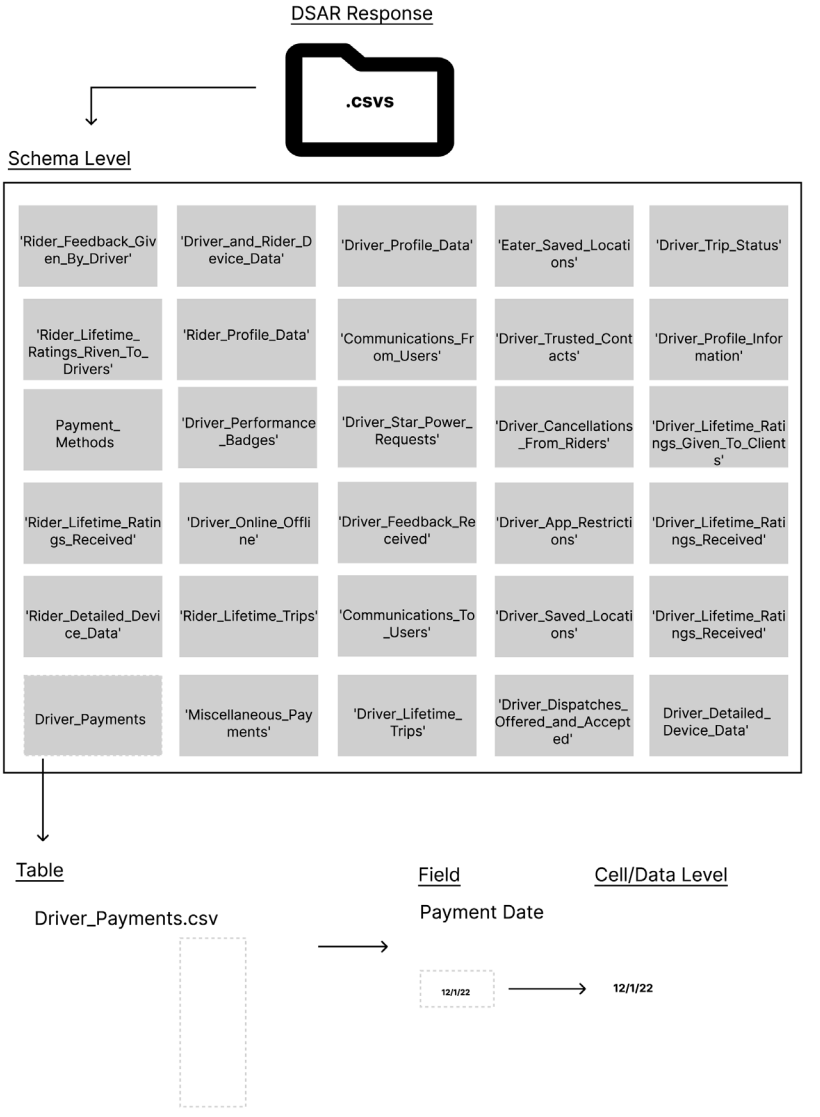


Figure 7.1 *Schema, field, and data levels of data structure.*

Source: Prepared by the authors

The Schema Level

Schema-level structure refers to the decisions made about how many independent tables a database is composed of, which fields are held in each of those tables, and the semantic connections (or lack thereof) that can be made between tables. Data architects meticulously design schemas in order to reduce the latency of queries when writing or retrieving data, and to serve the critical purpose of separating data with varying levels of sensitivity to support table-level access restrictions. The schema of the tables returned in DSAR responses provided some immediate clues and artefacts of its designers' intentions.

On the schema level, two kinds of data structure presented difficulties in analysing data. First were the differences in the schemas of separate DSAR responses preventing the software we produced from processing responses in an interoperable fashion. For instance, some requests received a file titled 'Payments', while other responses updated the name of that file to read 'Driver Payments'. Second, and of much greater consequence, was a lack of semantic structures to connect records of different types within individual DSAR responses. Most notably, there was no unique identifier allowing for linkages between payments (represented independently in 'Driver Payment' and 'Miscellaneous Payments' tables) and trips, which were assigned their own table. As drivers are not paid on a trip-by-trip basis, it is impossible to sufficiently disentangle the data structure to determine exactly what a worker was ultimately paid for a specific trip, including all costs, reimbursements, incentives, and other auxiliary costs. This was further complicated by the existence of the separate 'Miscellaneous Payments' table which provides records of reimbursements or payments for sick leave, holiday pay, pension contributions, and incentives like rewards for accepting consecutive trips. This table likewise had no ability to be linked to individual trip records, even when payments were associated with the completion of particular incentive or goal.

The Table Level

A similar link is lacking between the dispatch and trip tables. This missing link also illustrated how table-level data structures interact with structures within specific tables. In the dispatch table, rather than providing individual records of each dispatch offered, accepted, or rejected, Uber provides hour-by-hour aggregations of dispatches. Naturally, this aggregation makes links to the trip records (whether or not they were completed) impossible. Furthermore, it stands directly in the way of efforts to detect changes in work offered to drivers depending on their selectivity towards work offered – a behaviour that drivers anecdotally reported to WIE.

WIE suspects that drivers are implicitly penalized by the algorithmic systems which assign trips to drivers when they refuse or cancel unfavourable jobs. Left uncontested, such algorithmically mediated punishment would afford Uber the ability to maintain their position that drivers are independent providers free to work as they wish, while still allowing the platform to influence driver behaviour. In any event, the ability to independently audit the behaviour of algorithms which are essential to data subjects should fall within the realm of transparency expected from data access provisions. This is precisely the grounds argued for and won by WIE in the Netherlands and continues to be pursued.[24] The goal of this article, however, is not

to draw conclusions about which structure should be lawfully provided by data controllers in DSAR responses. Rather, we aim to highlight the role of data structure in limiting and enabling different analyses with the goal of producing designs that facilitate the simultaneous use of the same data by several parties, thus amortizing the technical infrastructure on which it relies. The barriers found in data structures provided by Uber are an excellent example where this is not the case, providing invaluable lessons for data architecture design.

If individual records of jobs offered, accepted, and rejected by workers ('dispatches' in Uber's terminology) were accessible, auditing the existence of algorithmic punishment would be a straightforward calculation. However, with individual records only provided to drivers aggregated into hourly totals and with no link (no unique identifier) to trip records, evaluation of any causal link between cancellations and work assignment becomes impossible.

To the critical researcher, the decision to aggregate data appears a frustrating instance of obfuscation, perhaps made to prevent inference of the logics contained within Uber's work assignment algorithms. Giving Uber the benefit of doubt, this aggregation might also be explained as an intended choice made to facilitate better compliance with GDPR by providing human readable metrics. The catch-22 of providing more disaggregated data versus human readable metrics here highlights the consequences of vague definitions for data portability. However, this is in many ways a false dichotomy; by simultaneously providing the logic necessary to aggregate metrics like acceptance rate (or aggregate metrics themselves) and the most disaggregated data, Uber could satisfy both needs. The structure of the dispatch records is exemplary of how data structures become palimpsests of the nested perspectives inscribed into the final data presented in the response. They overlay the intentions of the data architect who originally orchestrated the data collection system to be most effective for Uber's own purposes, with the data analyst subsequently tasked with compiling data for a DSAR response. This layers what Uber is willing to disclose on top of their interpretation of how to fulfill obligations for access to data subjects, and of course the artefacts of the source system. Elsewhere in the DSAR response, records similar to those in the dispatches table are left disaggregated as is the case of individual trip records.

The Field and Cell Levels

Finally, data on a cell-by-cell level significantly hampered analysis. As with schema- and table-level obfuscation, this was the result of both errors or inconsistencies in data quality and choices made in the design of the documents. With respect to data quality, we encountered fields in which the format of data was different even within the same column. For example, some fields contained timestamps of different machine and human readable formats, suggesting those resulted from the combination of other tables. There were also missing values (such as in the vehicle unique identifier field) in several responses or inconsistent values between responses, most likely associated with changes to Uber's own data formats over time – for example, changes in spelling from 'en route' to 'enroute'.

Other cell-level structures that obfuscated our analyses were more indicative of the systems out of which data were sourced. One of WIE's primary objectives was to better understand

the calculation of fares and driver pay for individual rides, both to empower drivers with greater awareness in planning their own activities and to better understand Uber's algorithmic 'dynamic pricing' scheme. In the 'Driver Lifetime Trips' table, Uber supplied a record for each trip taken by drivers. The table contains a variety of fields detailing the fare, from 'Original Fare', which Uber's guidance document clarifies is 'based on time + distance', to 'Upfront Fare', which applies to trips with a quoted fare before acceptance, to 'Base Fare', which is defined only as 'base fare of the trip'. The table also contains the individual components one would presumably need to reproduce the end fare paid by passengers, such as the trip distance, time, and their associated rates, as well as the surge fare applied to trips. Despite providing these aspects of the trip, we were unable to consistently reproduce the fares provided; for instance, in each of the responses we analysed, the formulas for determining original fare (base fare + distance at distance rate + time at time rate + surge + service fees) only occasionally lined up with the fare figures provided elsewhere in the table. Again, these inconsistencies could equally be attributed to fields being meant for use in Uber's system, rather than retrospective analysis, as well as the actual formula used to generate the fare changing over time. This is supported by announcements by Uber that they switched from fares calculated according to rates to a 'dynamic pricing' system and that drivers can access their take-home payments through their app directly.[25]

Coping with Obfuscating Structures

The structure of data in Uber's request responses illustrates how one can gain access to data without necessarily gaining access to all of the information contained in that data. We cannot, however, know whether the intent of the data analysts' chosen structures was goodwilled clarity of communicating the driver's behaviour or obfuscation of the algorithm's logic. The effect nonetheless reveals a fundamental tension for the application of data access policy that goes beyond arguing the definition of personal data. The importance of data structure results in situations where a data controller can return personal data from their systems without returning all of the *information* that data might contain. In the case of WIE, a Dutch appeals court ruled that more granular data and specific explanations of algorithmic systems should be given to drivers, brushing aside claims that this would require platforms to divulge trade secrets. Such a finding, however, only emerged from very specific requirements provided upfront by WIE, not on any independent judgement about data initially returned.

The more granular data and explanations WIE have been able to win are a step in the direction of a more complete picture of the system. From the perspective of a system designer, the outcome is, however, bittersweet. The situation demonstrates that transparency into the decisions of algorithms or the logic of the systems through which data flows is not straightforwardly deducible from granular data alone. Further, it is difficult to expect data consumers to specify the ends they expect to use data for at the outset, or to know enough about the system's underlying structures to specifically request explanations. Indeed, this suggests that perhaps there is value in preserving the structure controllers return data in, rather than opting for the most granular form.

The choice of granular or structured can also be a false dichotomy. With respect to dispatch data, obfuscation could have been avoided if aggregation was kept to later stages in the data pipeline and semantic structures left intact. In any event, in the first instance Uber neither provides a polished report meant to be understandable to workers nor a 'raw' export of the way data exists in their system. Their DSAR responses are something in-between, demonstrating how the political and economic interests of parties providing data are inscribed in the data structure, intended or not.

The obstacles Uber's DSAR data structures present provide a valuable example for data institutions that attempt to counteract them. Providing data available at a lower level of aggregation could help such infrastructures to avoid repeating this obfuscation while allowing for greater flexibility in data's use. A constellation of parties will be needed to actualize data institutions which aim to provide governance and transparency to data subjects while also benefitting other stakeholders like advocates, organizers, and researchers. When data institutions assume the role of imposing new structures on data for their own purposes, they will need to be wary about how balancing the interests of many more parties than merely the platform and data subject may create similar effects to those we observed here.

DSAR responses are artefacts of the systems out of which they emerge. This exposes a somewhat expected mismatch between the positivist imagination of data enshrined in data access regulation and latent (yet overly optimistic) hope that access to it might be sufficient for transparency into or means to act against the systems out of which it flows. It also is a stark reminder of data's innate subjectivity and thus the misguided assumption that we can impose one-size-fitsall 'interoperability' to data.

Rather than rehash that there is no such thing as raw data or that 'all data are local', we take this analysis as the design prompt for our own data architectures.[2266] In data institutions functioning as public-service data architectures or counterarchitectures, should we strive for data to occupy the role of reflecting the ground truth — an honest, unadulterated witness to the events it records — or is it more valuable (or even existentially necessary) that the data still carries the fingerprints of the developers, analysts, and *systems* through which it has already passed? The answer, as we see it, depends on purpose. In the next section, we reflect on the architectures at our disposal for creating architectures which may support a diverse set of stakeholder needs without imparting structure which itself obfuscates or narrows the information available in the data it holds.

Mapping Data Structures and Their Politics

Now that we have grappled with some examples of how data structure restricts and enables the potential information one can extract from underlying data, we can begin to explore how the interests of the parties involved in co-constructing data institutions may themselves manifest in the infrastructures they create. This section proceeds by exploring how data structure factors into goals of our present work alongside WIE: to deliver information from DSARs directly to drivers in a convenient format, to investigate the behaviour of algorithmic pay and work assignment, and to minimize the workload for WIE to self-manage their data infrastructure.

We first sought to make workers' data more accessible and understandable to them. However, even defining basic metrics for drivers such as pay-per-hour is inherently political, requiring that we select which perspective to encode into data structure and present to workers (Figure 7.2). WIE sees hours worked as including time signed on to the app awaiting new jobs, while Uber discounts this time. As a result, pay-per-hour-worked looks radically different depending on the logic used to make the calculations. By way of example, one worker's data we examined would have their hourly wage between GBP 23 and GBP 30 per hour, according to Uber's formula for the metric, which only includes time on the way to passengers or with passengers on rides. Accounting for time waiting, the same driver saw hourly rates hover between GBP 13 and GBP 20 per hour for most of their time working for Uber, with their gross wage per hour dropping into single digits on occasion. This is, of course, before costs covered by the driver, including but not limited to insurance; vehicle maintenance; licence; congestion charges; and fuel. Indeed, according to their DSAR data, this driver typically spent around 30 per cent of their time online waiting for rides. One important complication to this metric is the practice of 'multi-apping', when drivers simultaneously use several gig-work apps in hopes of capturing the best offers and minimizing dead time. This practice is often against the terms of service for gig-work apps like Uber depending on the location, but is frequently brought up as a counterargument for paying for between-ride time.

Returning to the goal of delivering valuable transparency to workers, generating metrics like the aforementioned wage calculation is best suited to highly structured, relational data stores akin to those often used in basic business analytics systems. However, highly structured data infrastructures only accommodate rigid definitions for metrics to be repeatedly applied, and as a result they are highly sensitive to even small changes in underlying data.

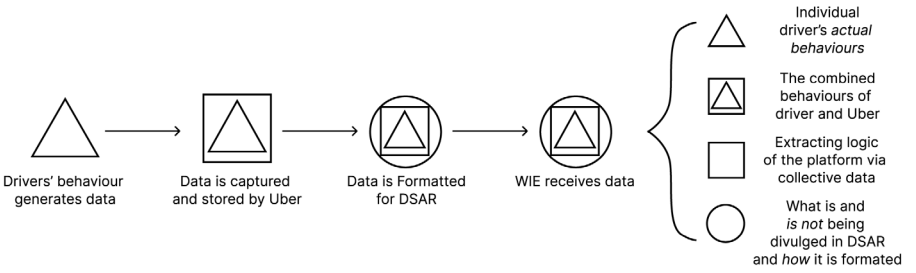


Figure 7.2 *The nested logics encoded in DSAR data.*

Source: Prepared by the authors.

Furthermore, encoding rigid definitions (like Uber's and WIE's competing definitions of pay-per-hour) into the structures inherently narrows the interpretation of the data. DigiPower academy provides an example of this approach, producing polished and digestible figures for individuals based on DSAR data.[27] The advantages of a highly structured environment lie in the convenience and accessibility of outputs to data subjects. However, highly structured databases' sensitivity to changes in data sources makes them labour- and resource-intensive for organizers, placing the parties responsible for their maintenance primarily as intermediaries between platforms and subjects, while leaving fewer resources to inductively investigate systemic problems. Furthermore, the risk of entrenching specific narratives or interpretations into data structures and thus limiting potential interpretations is a deal-breaker to researchers committed to open science or advocates like WIE who are interested in uncovering more systemic patterns within aggregate data without frequently refactoring their systems.

Our overall design decisions quickly became defined by a core trade-off. Either we privilege repeatable processes, dedicating time to building end-user-friendly results backed by a highly structured infrastructure, or we leave data in a lessstructured form, thus preserving the ability to conduct deeper iterative analyses. Understanding which route provides more benefit to the data subject is a fraught exercise of acting within versus against the greater system. WIE's investigation of pay and work assignment algorithms requires an open infrastructure for inductive analysis and, if successful, could achieve considerable gains in rights for workers. However, prioritizing the collective interests of data subjects over individual access to metrics poses a challenge for data stewardship and their individual agency. Britt S. Paris, Corinne Cath, and Sarah Myers West have highlighted the penchant for technological solutions to tack towards value ethics prioritizing efficiency and perceived good from the top down in place of an approach based in care ethics which might honour the individual autonomy of data subjects.[2828] Dubal's ethnographic work, likewise, warns about placing perceived collective gains ahead of individual agency in gig work.[29] In our case, privileging aggregate analysis ahead of individual data access risks making qualifying all workers as a homogenous set, rather than empowering them with information directly. On the other hand, though an algorithmic audit requires diversion of resources from immediate data analytics outputs directed at data subjects, it could lead to much more impactful legal victories or proof of algorithmic coercion, exploitation, or discrimination when completed. Ultimately, any sustained effort requires both forms of advocacy, at which point the question becomes: how might we support both?

Dilemmas like these force introspection about the role of researchers in PAR and as infrastructure developers. Should researchers concentrate efforts towards providing the most information (albeit limited in its scope) in its most understandable form to the most workers as fast as possible, or should infrastructure seek to facilitate larger-scale inductive aggregate analyses which possess the potential for paradigm shifts via strategic litigation, regulation, and ultimately better policy outcomes – the goals of WIE? In this sense, the design of the infrastructure can be reframed as a prioritization of stakeholders' needs. Even so, each stakeholder's contribution is just as necessary to fulfilling any of the other's goals as the next: the subject by providing their data, WIE in aggregating the data for collective advocacy, and researchers for development of the analytics systems, defining what will result in the most agency for data subjects remains muddled.

Undoubtedly, data subjects' interests should come first. Selecting which interests and how to serve them best is a more difficult question. Furthermore, implementing this prioritization via data infrastructure without cementing a single politics of change is not nearly so clear. Our ultimate goal as researchers is to generate data architectures which allow advocates with large-scale goals to shape policy and win rights via collective advocacy, while also helping them seek immediate relief and agency within poor working conditions. Building infrastructure that does not overtly set priorities in pursuit of allowing both stakeholders to access and analyze data autonomously also runs the risk of failing to execute fully on either goal, potentially leading the effort astray from the immediate needs of data subjects.[30]

The process of building infrastructures necessitates difficult trade-offs: should these infrastructures seek to help data subjects engage within the game set by the platform or should they seek to reveal and change the rules through collective advocacy?

Achieving both of these goals requires simultaneously accommodating each of the many, often conflicting requirements we have just enumerated. Provided sufficient access to data, these goals do not need to be mutually exclusive; no technical constraint prevents all of these architectures to be implemented independently on top of a common source of data. The resources and coordination needed to construct these infrastructures are, however, desperately scarce. To make a case for specific architectures that aspire to these goals, we inherently need to make them and their governance as resource-efficient before also critically considering what underlying politics we ourselves are at risk of encoding into our own data infrastructures.

Parallax

Nick Srnicek points to a peculiarity that sets platform firms apart from their predecessors: cross-subsidization – the unique ability of platform companies to internally hedge individually unprofitable services on the knowledge gained through the data they generate or other services they facilitate.[31] Gmail, Facebook, and of course Uber all fit this design pattern. Gmail offers free email hosting in exchange for data used for marketing or AI training. Facebook's social network is likewise the façade of a web-wise AdTech empire, while Uber consistently posts losses in pursuit of squashing competition, logistics optimization, and even autonomous vehicles.

Despite gains made in accessing the collective data which underpin these systems, no institution, whether government, academic, or public-interest, can realistically support data infrastructures that are able to match the scale and sophistication of venture-backed platform companies, promising boundless optimization and a future of AI – that is, if they do so alone.

In the preceding sections, we considered the impacts of data structures on the availability of information in data. First, we drew attention to how data structures within DSAR responses obfuscated the information contained in the data it communicated. Considering how data structure inherently tailors the perception of underlying data based on the politics and aims of its source systems, we critiqued the role played by academics or others building underly-

ing data infrastructures. We concluded that any data institution meant to serve such diverse stakeholders as WIE's, and supported by resources pooled from data subjects, activists, and academics, must avoid encoding similar constraints on data. More specifically, this means being able to regard data as a record of events in order to give clarity to data subjects, so they can act with greater individual autonomy *within* a system where information asymmetries left them without tools to resist. At the same time, our infrastructure must also enable a critical lens on the very same data, instead viewing it as trace of Uber's system, the actions of models operating on that system, and their diversion from workers' and advocates' version of the truth. We recognize the complex and contradictory nature of proposing a data analytics infrastructure detached from any one analytical aim, while also supporting two epistemic readings of the same information. In order to unpack this, we borrow Žižek's concept of the 'parallax' to account for both the inherent conflicts that arise in designing this system.[32] In optics parlance, parallax refers to the phenomenon of observing equally accurate movements of an object in the opposite directions caused by the displacement between two points of view. In this sense, two measurements can be both entirely accurate yet conflicting with one another only because they are taken from two distinct perspectives.[33]

In Žižek's *The Parallax View*, he reconsiders a basic Hegelian tenet – the notion that analyses approaching the truth through a see-saw action of contradicting thesis and antithesis, ultimately resolving in an inevitable synthesis. Instead, Žižek abandons the need for any resolution, and instead draws attention to the process of 'putting two incompatible phenomena on the same level' where 'the illusion of being able to use the same language for phenomena which are mutually untranslatable and can be grasped only in a kind of parallax view, constantly shifting perspective between two points between which no synthesis or mediation is possible'.[34] Drawing examples from history like those between economic conditions and political movements, or between individual psychoanalysis and social dynamics, Žižek points his readers' attention to the 'parallax gap, the confrontation of two closely linked perspectives between which no neutral common ground is possible'.[35] For some truths or relationships taken as fact – such as the relationship between individual psychoanalytic conditions and social movements or between economic conditions and political conflicts – our understanding exists in a certain tensile stasis. Each perspective is untranslatable and fundamentally different yet somehow resolve as one event (Figure 7.3).

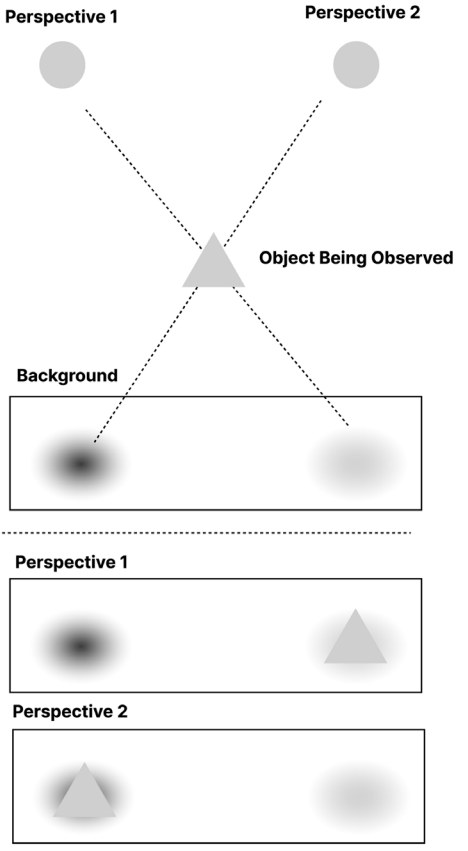


Figure 7.3 *Illustration of parallax.*

Source: Prepared by the authors based on Justin Wick's example of parallax, Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Parallax_Example.png (accessed 12 December 2024).

Our choice of design for public-service data architectures resonates with this framing. First, it accurately embodies our dilemma of selecting which outcome (or form of solidary) our system should aspire to – should it champion individual autonomy for workers *within* the system or should it prioritize aggregate data for their collective cause by working *against* that very same system? The parallax also resonates with a confounding question around the epistemic lens with which we might analyze data in the pursuit of empowering workers. Having pointed to the unavoidable warping of meaning in data through its structuring, can an architecture simultaneously use data from DSARs as evidence of exploitation within a system *and* also show that the data that same system produces is flawed, obfuscating or betraying the experience of those operating within it? Put differently, can it champion solidarity *within* Uber's system

through analysis of data, while using that same data to critique that system's own validity? Žižek's operationalization of the parallax offers that both are not only possible but also often necessary within historical, philosophical, and scientific narratives.

Data subjects, activists, academics, and their respective funders see the same data through different lenses. Some of these lenses do not necessarily gaze towards a ground truth captured in data, but instead critique the behaviours and intentions couched in the structural artefacts left by prior manipulations and ontologies. In this sense, suitable analytics infrastructures must be *parallax* in nature. That is, though relying on the same material data as their basis, they must support inquiry which pursues divergent epistemologies of the data and interpretations of autonomy. Accommodating various perspectives is a significant challenge for any analytical system, especially given such systems inherently adopt a positivist viewpoint. Though relying on a common infrastructure may offset some technical and organizational overhead, the admitted weakness remains as to who maintains such systems and for what reward.

We posit that the best approach is not to conduct individual, teleological efforts affirming a single perspective (and thus structure), but rather architectures that can enable the diversity of goals stakeholders seek and the epistemologies these investigations require. In the next section, we describe the features of the architecture we find best approximates this function given the technical tools we possess today and the regulatory climate we operate within.

Ways Forward: Architectures and Policy

You cannot buy a parallax data architecture from Amazon Web Services (AWS) or fork one from GitHub (at least in name). We put forward, however, that existing data analytics architectures used in open-source and industry practice might be configured to embody many of the values a parallax architecture might support. In this final section, we point to examples of architectures and how we imagine data institutions might use them to this end. Further, we place these technical systems in the context of their regulatory climate – specifically pointing out the role such architectures could play in the techno-legal ecosystems that new data protection regulation has created. On the flip side, we also describe the needs left unaddressed by today's growing data protection regimes but potentially aided by data institutions.

We started with an assumption: should data infrastructures be able to lend transparency into their vastly more complex corporate counterparts, they will need to pool resources and share knowledge as seamlessly as the platforms integrate their ecosystems. We have argued that a primary obstacle to this goal will be accommodating data structures which are each independently necessary to analyse data for separate stakeholders needs, but may also lead to conflicting demands on data infrastructures.

In our designs, we theorize that unstructured data analytics infrastructures could provide a suitable technical basis to simultaneously support the diverse analytical needs.[36] When maintained exclusively by a single academic or publicinterest institution, these infrastructures may closely resemble current corporate practice, and their cost might likewise become

unjustifiable. In a shared setting, however, the flexibility of unstructured architectures could allow parties with common interests in underlying data but conflicting data structures to still pool their resources and make them available as a collective resource. This approach constitutes a departure from much of the existing literature, which calls for strict data interoperability standards from the outset.[37] Instead, having considered interoperability as something deeply contextual in its definition, rather than a single objective format, we advocate for data institutions to take up an unstructured approach to underlying data storage, facilitating the maximum possibilities to query the same data resources, without putting in place structures which obstruct other analyses.

Unstructured or NoSQL data architectures like Elasticsearch have a variety of features which make them suitable for the demands of data institutions. Most relevant to our discussion is their natural orientation towards 'schema on read' (SoR) analytics strategies.[38] SoR systems stipulate that the greatest part of data structure is added at time of analysis, rather than upon ingestion, while persistent data remains in its unstructured form. While SoR is by no means a new idea, it is rarely considered with a view respect to its potential impacts on data stewardship and collective governance.

Pushshift is an excellent example of how unstructured data analytics can be used at scale in the public service.[39] The site maintains massive repositories of comments and posts from Reddit and other social media outlets as a research resource. Using an Elasticsearch architecture, Pushshift makes data available via a search API. Elasticsearch's pre-indexing of data allows for data consumers to construct complex searches, adding their own desired structure to data at the time of their query with little latency, and the option to work iteratively, without Pushshift needing to pre-compute statistics or add structures which might obfuscate or prevent further analysis. The downside is that more work may have to be done by the data consumer to extract their desired information. Once this work is done the first time, however, it is easy to replicate and build upon when queries are shared. In our case, it is easy to imagine sharing a definitive query among organizations for common metrics like pay-per-hour. Though Pushshift demonstrates how a vast volume of data can be made available for many purposes with relatively low overhead and maintenance cost (it is maintained almost entirely by one person),[40] its use case does not contain sensitive personal data, and thus has not required the implementation of collective or personal data governance controls as platform worker data does.

Some of the natural resonance that unstructured data analytics systems have with shared or collective governance originate in their semi-decentralized architecture. Unstructured data analytics deployments were engineered to rely on distributed data storage to facilitate queries from many data analytics access points with different levels of access, increasing data availability while limiting latency and actively balancing workloads. Drastically simplifying their technical design, technologies like Elasticsearch replicate data across isolated indexes to reduce the time it may take to access any single record and to allow multiple parties to simultaneously access the same data. The 'horizontal scalability' lent by the distributed nature of these data stores means different organizations, researchers, or even platforms could independently contribute their data without fully relinquishing governance over it.[41] For example, WIE might maintain an index of worker data controlled as a data trust through

agreements with drivers as they do today. WIE could make their index searchable by data protection regulators in the UK, while an EU-based equivalent of WIE might allow queries from WIE under a mutual agreement but block queries from UK regulators. Similarly, the SoR stance of these architectures would mean that stewardship decisions about anonymization or aggregation could be made at the time of analysis, rather than being persistently hard-coded into the functional structure of data stores. Queries from particular parties could be limited to indexes or subsets of data; and depending on their use or alternatively, data stewards could review the results of queries submitted before allowing results to be returned.

Indexing data in unstructured data stores allows it to remain in its original structure while also making it query-able for those consuming it. At the same time, the effort placed into making information available from it creates a tangible record of what is necessary to take data from its original state into the necessary form for end users. We have already seen an instance of how this might be useful to regulators in our work with WIE. The code we created inherently chronicles a running record of each deviation between DSAR responses we need to resolve and the logs produced by our code record each time it is applied to DSAR data, counting each row and field normalized and the frequency of discrepancies between tables. For example, on each occasion a timestamp was corrected from a human-readable time format to a machine-readable format or vice versa, or the title of a column was made to conform to other responses, a log was recorded. Logs like these could be a crucial resource not only for authorities enforcing interoperability requirements but also for pinning down what common data formats should look like, or for tuning or training machine learning models to automate the normalization of data formats we have thus far performed manually. Other approaches to aggregating data from data subjects for use in research and compliance also produce similar artifacts. The Open-Source Data Donation Framework or (OSD2F) elicits scripts from researchers to be tested in the private environment of a data subject's browser before allowing data donors to elect to contribute their data to research.[42] The scripts contributed by researchers both play a similar role to logs we produced, showing the necessary transformations to make data interoperable, and the queries we mentioned earlier, by providing open-source or shared resources for extracting information from the underlying data.

There are also undoubtedly weaknesses of relying on unstructured data analytics technologies. One aspect is that they are only semi-decentralized when compared to other, fully decentralized alternative data architectures. Tim BernersLee's Solid project (to which our team actively contributes) facilitates one-off permissions to data held by subjects, making it possible for data to be accessed by services without ever leaving the data subject's personal data store.[43] Solid envisions a future in which data stored in a decentralized setting is organized according to a common semantic format (Resource Description Framework, or RDF), allowing for perfect interoperability. The price of Solid's decentralization is strict adherence to this common data format. We see unstructured data infrastructures not as competing with such decentralized ecosystems but as a layer which might enhance them. If (or until) such adoption is achieved, we see unstructured data analytics deployments as a necessary bridge to ease the pains of data normalization and deliver needed analytical insight now. Indeed, subjects might consent to share data directly from their Pod with a data institution which has a governance regime complementary to their values. Sylvie Delacroix and Neil D. Lawrence imagine a similar tech-

no-legal configuration when they describe 'a plurality of bottom-up data trusts', meant to suit the highly heterogeneous sets of data subject preferences – something they viewed as an obstacle to the initial data trust model proposed by Lilian Edwards and championed by Dame Wendy Hall and Jérôme Pesenti.[44]

One of the inherent disadvantages to the model we propose is its reliance on DSARs for its data. Much of our earlier sections covered specifically how access to data via DSARs does not necessarily lead to the transparency or agency sought from aggregating it. This point becomes even more critical in the context of new regulations that grant more direct research access to platform data for researchers. The added access granted by Article 40 of the EU's DSA constitutes a significant advance in data access for evaluating systemic harms and risks of the type we examine in our work with WIE.[45] This provision of the DSA is unfortunately not applicable to Uber as it does not break the 45-million-user thresholds set by the policy.[46] With laws like the DSA opening new avenues to aggregate data access and recent cases won by WIE establishing a precedent in which Uber cannot deny data access for the purpose of aggregation, we might expect obfuscation through the use of data structure to become an even more common tactic of data controllers. Remaining optimistic, however, the added access researchers may be able to achieve for other platforms via the DSA could also become a critical resource to enhance data institutions drawing from aggregated DSAR responses. The vastly different nature of this data – likely aggregate, higher velocity (more regularly refreshed), and hopefully more readily machine readable – only strengthens the need of architectures that can quickly accommodate the combination or comparison of differing data structures with existing data collected directly by data subjects through their own self-tracking or through DSARs.

In the absence of greater legal tools to access data, grassroots efforts have also been successful in sourcing data to give workers immediate insights and audit the behaviour of algorithms. Dan Calacci and Alex Pentland creatively combined a chatbot with computer vision techniques to collect and process screenshots of workers for the US platform Shipt.[47] Through this technique, the authors worked with data subjects and activists to reveal how alterations to the design of dynamic pay systems on the part of the platform resulted in changes to the wages of workers. Commercial implementations that source data directly from individual workers have also emerged such as Argyle, a service that uses workers' credentials to fetch their data and build verifiable labour records across platforms. Rodeo, an app that allows workers to compare gig-work offers across platforms, uses Argyle to source data, but is ultimately at risk of losing access to data should platforms cut off data flows as Deliveroo, a UK-based food delivery platform, has recently done.[48]

Conclusion

The technological innovation that has facilitated platform dominance could also be used to supplant the information asymmetries it has created. Unfortunately, platforms' aggregation of data and the powerful algorithmic technologies that emerge from it are cumulative in their momentum, and they have been given a multi-billion-dollar head start. Further, hegemonic patterns of control have also made institutions in academia, civil society, and government slow to adequately retool and cooperate to reverse escalating concentrations of power.

Regulation, however, is beginning to catch up, creating cracks in the veil between data subjects' and controllers' views of the systems on which they both rely. Further, new generative AI methods and analytics architectures, though conceived to power companies' data-grabs, could be put to use to erase much of the overhead that separates platforms' technical capabilities from those of organizers who seek to hold them to account. While some regulations like the DSA might help organizers catch up in the analytics race through better data access and mandated algorithmic transparency, others also slow down or limit the analytics practices that set platforms apart. The Digital Markets Act (DMA), 2022, prevents large platforms from pooling data sourced from disparate services undercutting or disincentivizing 'cross-subsidization'.

Together, newfound access to data, combined with prohibitions of platforms' data pooling practices, presents a distinct opportunity for advocacy organizations and research bodies to produce sustainable alternatives that are also accountable to data subjects. Despite this promising shift in the regulatory landscape, the most treacherous obstacles for organizers may soon come not from barriers to accessing data or technology but from understanding how to redraw deeply intrenched institutional boundaries, while also including data subjects to participate in shifting power.

Notes

1 Rob Kitchin and Gavin McArdle, 'What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets', *Big Data and Society* 3, no. 1 (2016), DOI: https://doi.org/10.1177/2053951716631130.

2 A whole sub-industry has sprung up not only offering data analytics as a service but also in the optimization of organizational structures surrounding analytics development in 'AI operations' (AIOps). So-called process-mining tools have also become widely available to streamline data structuring, pipeline building, and exploratory analysis by non-technical users. See Yingnong Dang, Qingwei Lin, and Peng Huang, *AIOps: Real-World Challenges and Research Innovations*, 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion) (Institute of Electrical and Electronics Engineers, 2019), https://ieeexplore. ieee.org/document/8802836 (accessed 28 February 2023).

3 Dan Calacci, 'Organizing in the End of Employment: Information Sharing, Data Stewardship, and Digital Workerism', 2022 Symposium on HumanComputer Interaction for Work, Durham, NH, 2022.

4 Jack Hardinges and Jared Robert Keller, 'What Are Data Institutions and Why Are They Important?' Open Data Institute, 29 January 2021, https:// theodi.org/article/what-are-data-institutions-and-why-are-they-important (accessed 7 November 2022).

5 Hardinges and Keller, 'What Are Data Institutions'.

6 Lilly C. Irani and M. Six Silberman, 'Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk', Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2013, https://dl.acm.org/ doi/10.1145/2470654.2470742 (accessed 13 July 2022).

7 Jack Hardinges and Jared Robert Keller, 'What Are "Bottom-Up" Data Institutions and How Do They Empower People? – the ODI', Open Data Institute, 25 June 2021, https://theodi.org/article/what-are-bottom-up-datainstitutions-and-how-do-they-empower-people (accessed 7 July 2021).

8 Tobias Fiebig, Seda Gürses, Carlos H. Gañán, Erna Kotkamp, Fernando Kuipers, Martina Lindorfer, Menghua Prisse, and Taritha Sari, 'Heads in the Clouds: Measuring the Implications of Universities Migrating to Public Clouds', 27 July 2021, http://arxiv.org/abs/2104.09462 (accessed 20 June 2022).

9 Fiebig, Gürses, Gañán, Kotkamp, Kuipers, Lindorfer, Prisse, and Sari, 'Heads in the Clouds'.

10 Wendy Hall and Jérôme Pesenti, 'Growing the Artificial Intelligence Industry in the UK', Department for Digital, Culture, Media and Sport and Department for Business, Energy and Industrial Strategy, GOV.UK, 2017, https://www.gov.uk/government/publications/growing-the-artificialintelligence-industry-in-the-uk (accessed 11 November 2024).

11 René Mahieu and Jef Ausloos, 'Recognising and Enabling the Collective Dimension of the GDPR and the Right of Access', preprint, 29 April 2020, https://osf.io/b5dwm (accessed 29 October 2021).

12 Dan Calacci and Jake Stein, 'From Access to Understanding: Collective Data Governance for Workers', *European Labour Law Journal* 14, no. 2 (2023), DOI: https://doi.org/10.1177/20319525231167981.

13 Worker Info Exchange (WIE) is a non-profit data rights and worker advocacy organization in the United Kingdom (UK). The organization assists workers with individual claims, including algorithmic discrimination, data access, unfair working conditions, and unjust dismissals. Closely aligned with the App Drivers and Couriers Union (ADCU), the organization is active in collective advocacy, winning several landmark cases establishing standards for data access in the UK and the European Union (EU). For information about WIE, visit workerinfoexchange.org.

14 Our project regards data autonomy first and foremost as data subjects' ability to control, manage, and maintain their personal data encompassing both personal data privacy and transparency into how data is used and processed. We also consider data autonomy to go beyond these core pillars of data governance to include data subjects' access to 'mutual legibility', or the ability for data subjects to understand data in the same contexts (aggregate and individual) as data collectors.

15 *Applicants 1-4 v. UBER BV* [2023] Rechtbank Amsterdam 200.295.742/01; *Applicants 1-6 v. UBER BV* [2023] Rechtbank Amsterdam 200.295.747/01.

16 'AWO', https://awo.agency (accessed 31 May 2023).

17 This research project has been approved by the Departmental Research Ethics Committee for Computer Science, University of Oxford, under reference CS_C1A_23_016_001.

18 Britt S. Paris, Corinne Cath, and Sarah Myers West, 'Radical Infrastructure: Building beyond the Failures of Past Imaginaries for Networked Communication', *New Media and Society* 26, no. 11 (2023), DOI: https:// doi.org/10.1177/14614448231152546.

19 Slavoj Žižek, *The Parallax View* (paperback) (MIT Press, 2009 [2006]).

20 See Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J. Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, and Sam Whittle, 'The Dataflow Model: A Practical Approach to Balancing Correctness, Latency, and Cost in MassiveScale, Unbounded, Out-of-Order Data Processing', *Proceedings of the VLDB Endowment* 8, no. 12 (2015): 1792–1803. For example, an analytics system to evaluate the performance of a jetliner will introduce structure early on in the pipeline in order to monitor complex machine-generated data in real time whereas social media analytics systems might introduce structure later in the pipeline to allow for inductive analysis.

21 Finn Brunton and Helen Fay Nissenbaum, *Obfuscation: A User's Guide for Privacy and Protest* (MIT Press, 2015), http://public.eblib.com/choice/ publicfullrecord.aspx?p=4093096 (accessed 3 December 2020).

22 Jef Ausloos and Pierre Dewitte, 'Shattering One-Way Mirrors: Data Subject Access Rights in Practice', *International Data Privacy Law* 4, no. 8 (2018): 4–28.

23 EU General Data Protection Regulation (GDPR), 2016, Article 4.

24 *Applicants 1-4 v. UBER B.V.; Applicants 1-6 v. UBER B.V*; Paris, Cath, and West, 'Radical Infrastructure'. Please note the research described in this article occurred prior to the conclusion of these cases.

25 Jessica Phillips, 'How Uber's Dynamic Pricing Model Works', *Uber Blog*, 21 January 2019, https://www.uber.com/en-GB/blog/uber-dynamic-pricing (accessed 31 May 2023).

26 Geoffrey C. Bowker, *Memory Practices in the Sciences* (MIT Press, 2005); Yanni A Loukissas, *All Data Are Local: Thinking Critically in a Data-Driven Society* (MIT Press, 2019).

27 Alex Bowyer, Jessica Pidoux, Jacob Gursky, and Paul-Olivier Dehaye, 'Digipower Technical Reports: Auditing the Data Economy through Personal Data Access', Zenodo, https://zenodo.org/record/6554178 (accessed 7 October 2022).

28 Paris, Cath, and West, 'Radical Infrastructure'.

29 Veena Dubal, 'Wage Slave or Entrepreneur? Contesting the Dualism of Legal Worker Identities', *California Law Review* no. 105 (2017): 65–123.

30 Paris, Cath, and West, 'Radical Infrastructure'.

31 Nick Srnicek and Laurent De Sutter, *Platform Capitalism* (Polity Press, 2017).

32 Žižek, *The Parallax View*.

33 Kaj Strand, 'Parallax', *Encyclopedia Britannica*, 2023, https://www.britannica.com/science/parallax (accessed 12 December 2024).

34 Žižek, *The Parallax View*.

35 Žižek, *The Parallax View*.

36 Rick Cattell, 'Scalable SQL and NoSQL Data Stores', *ACM SIGMOD Record* 12, no. 39 (2011): 12–27.

37 Gianclaudio Malgieri and Frank Pasquale, 'From Transparency to Justification: Toward Ex Ante Accountability for AI', Brooklyn Law School, Legal Studies Paper 712, 2022, 27; Milagros Miceli, Tianling Yang, Adriana Alvarado Garcia, Julian Posada, Sonja Mei Wang, Marc Pohl, and Alex Hanna, 'Documenting Data Production Processes: A Participatory Approach for Data Work', 9 August 2022, http://arxiv.org/abs/2207.04958 (accessed 12 August 2022).

38 Sladjana Jankovic, Snezana Mladenovic, Dušan Miodrag, and Mladenović Slavko Vesković, 'Schema on Read Modeling Approach as a Basis of Big Data Analytics Integration in EIS', *Enterprise Information Systems* 1180, no. 12 (2018): 1–22.

39 Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn, 'The Pushshift Reddit Dataset', *Proceedings of the International AAAI Conference on Web and Social Media* 14, no. 1 (2020): 830–839.

40 'Contributors to Pushshift/Api', https://github.com/pushshift/api/graphs/ contributors (accessed 31 May 2023).

41 Venkat N. Gudivada, Dhana Rao, and Vijay V. Raghavan, 'NoSQL Systems for Big Data Management', 2014 IEEE World Congress on Services, Institute of Electrical and Electronics Engineers, 2014, https://ieeexplore.ieee.org/ document/6903264 (accessed 28 February 2023).

42 Slađana Janković, Snežana Mladenović, Dušan Mladenović, Slavko Vesković, and Draženko Glavić, 'Schema on read Modeling Approach as a Basis of Big Data Analytics Integration in EIS', *Enterprise Information Systems* 12, no. 8–9 (2018): 1180–1201.

43 Essam Mansour, Andrei Vlad Sambra, Sandro Hawke, Maged Zereba, Sarven Capadisli, Abdurrahman Ghanem, Ashraf Aboulnaga, and Tim BernersLee, 'A Demonstration of the Solid Platform for Social Web Applications', Proceedings of the 25th International Conference Companion on World Wide Web, 2016, 223–226.

44 Sylvie Delacroix and Neil D. Lawrence, 'Bottom-Up Data Trusts: Disturbing the "One Size Fits All" Approach to Data Governance', *International Data Privacy Law* 9, no. 4 (2019): 236–252; Lilian Edwards, 'The Problem with Privacy', *International Review of Law Computers and Technology* 18, no. 3 (2004): 263–294; Hall and Pesenti, 'Growing the Artificial Intelligence Industry in the UK'.

45 Martin Husovec and Irene Roche Laguna, 'Digital Services Act: A Short Primer', 2022, DOI: https://dx.doi.org/10.2139/ssrn.4153796.

46 'Uber', https://www.uber.com/legal/de/document (accessed 31 May 2023).

47 Dan Calacci and Alex Pentland, 'Bargaining with the Black-Box: Designing and Deploying Worker-Centric Tools to Audit Algorithmic Management', *Proceedings of the ACM on Human-Computer Interaction* 1, no. 6 (2022): 1–24.

48 Oscar Hornstein, 'Deliveroo Accused of Blocking Courier Access to Gig Economy Finance App', *UKTN*, 21 March 2023, https://www.uktech.news/ mobility/deliveroo-blocks-rodeo-20230321 (accessed 31 May 2023).

8. DATA INTERMEDIARIES FOR GOOD: CAN DATA INTERMEDIATION SERVICES HELP DATA ACCESS IN RESEARCH?

MATTEO NEBBIAI

Data is a crucial competitiveness factor for an increasing number of organizations in the global economy. At the same time, asymmetries in data availability are generating debates on the effect of digital technologies on the functioning of markets,[1] competition policy,[2] and capitalism.[3] This asymmetry is also reflected by the limited benefits that the academic community obtained from the 'big data flood'. Various scholars have highlighted how online platforms such as Facebook and X (formerly Twitter) act as gatekeepers of data that are crucial for the advancement of various disciplines.[4] Despite the emergence of research strategies vis-à-vis big data holders, such as data scraping and the use of legal tools,[5] researchers' access options seem still limited in comparison with the possibilities available to the companies.

In the last decade, the European Union (EU) has adopted a variety of measures to improve data sharing among businesses, consumers, and researchers.[6] The General Data Protection Regulation (GDPR), 2016, introduced the right to data portability, which allows data subjects to obtain their personal data from controllers and transmit them to a different one.[7] The Digital Markets Act, 2022, obliges large online platforms to share with other businesses certain categories of data generated by consumers,[8] and the Digital Services Act, 2022, sets a procedure for researchers to access the data of large online platforms or search engines.[9]

In addition to this top-down approach, EU legislators increasingly see 'data intermediaries' as potential facilitators for consumer-to-business and business-to-business data transactions.[10] While a shared definition remains debated, 'data intermediaries' can be described as actors positioned between data holders and data users that facilitate data sharing.[11] It has been suggested that the focus on intermediaries is attractive because it 'lies between the two extremes of self-regulation and detailed binding statutory obligations'.[12]

The Data Governance Act (DGA), adopted on 30 May 2022, introduces a series of rules for intermediary actors, by creating the categories of 'data intermediation services' and 'data altruism organisations'.[13] Despite the ambitiousness of the DGA, many doubts and criticisms about its clarity and impact have been raised.[14] Particularly, it is not clear whether the regulation will help to make intermediaries competitive with well-entrenched 'Big Tech' firms[15] and which actors will ultimately benefit from their functions.

The DGA proposes a separation between data intermediation services (DISs) as purely commercial activities and data altruism organizations as non-profit projects supporting research purposes;[16] however, empirical observation of extant DISs show that the picture might be more nuanced. As illustrated later, some commercial intermediaries offer data access opportunities for researchers as well. For this reason, the chapter focuses on the following question: do DISs make data more accessible to researchers?

The first section analyses the text of the DGA to set the foundations for the empirical investigation. The second section reveals the methodology followed to build an original dataset including 54 DISs. The third section investigates the data access options offered by DISs through a qualitative and quantitative analysis of the dataset. The fourth section discusses how the DGA provisions may impact researchers' data access options. Finally, the conclusions sum up the findings and propose policy interventions and further research paths.

The Data Governance Act

The DGA aims to promote data sharing among governments, businesses, and individuals.[17] In this chapter, I focus on the DGA provisions regulating 'data intermediation services' and 'data altruism organisations'.

'Data intermediation services' are defined as services aiming to 'establish commercial relationships for the purposes of data sharing between an undetermined number of data subjects and data holders on the one hand and data users on the other, through technical, legal or other means'.[18]

Article 10 specifies that the following typologies of DISs are subject to regulation: (*a*) services intermediating data exchange or joint use between data holders and potential data users, also by offering technical, infrastructural, or other means; (*b*) services enabling data subjects to make their personal and non-personal data available to potential data users, in particular enabling the exercise of the GDPR rights (that is, PIMSs[19]); and (*c*) data cooperatives, which are organizations composed of multiple data subjects having one of their main objectives the support of their members in the exercise of their data rights (that is, making informed choices on data processing consent and negotiating terms and conditions for data processing on behalf of their members).[20]

The DGA establishes mandatory rules for DISs, aiming for data standardization[21] for fair and transparent data access[22] and the neutral use of data[23] (see the fourth section for more details).

'Data altruism' is defined as voluntary data sharing by data holders, without monetary rewards, for purposes of general interest, such as healthcare and scientific research.[24] The DGA establishes an *optional* registration procedure to become a 'recognised data altruism organisation'.[25]

The DGA considers data altruism organizations and DISs as distinct types of initiatives, whose line of demarcation is the operation on a for-profit or not-for-profit basis.[26] Moreover, it recognizes in data altruism organizations the purpose of supporting academic research,[27] whereas the DISs are only defined by their commercial purposes.

However, the empirical observation reveals that some DISs also offer data access opportunities for researchers. For instance, Quli is a Dutch service to store health data and transmit them to healthcare providers, and it allows the sharing of such data with researchers.[28] CitizenMe offers an app allowing citizens to share data for monetary rewards; such data is then

made available in a marketplace which is used for research purposes.[29] These observations back this chapter's research question: can DISs make data more accessible to researchers, thus redistributing part of the data sharing benefits to the academic community?

DISs Dataset

To answer the research question, I have created a dataset that aims to include the highest possible number of DISs available in the EU. To build such a dataset, it is firstly necessary to understand what a DIS is. This is no easy task, because the DGA provisions are far from straightforward,[30] and the definition of 'data intermediation services' is open to various interpretations.

First, the service must 'aim to establish commercial relationships'.[31] It is unclear if mere providers of technical data sharing infrastructure fall under this definition because it might be difficult to assess when they are aware of the existence of commercial relationships occurring on the infrastructure they built. In the dataset, I exclude the projects which do not host data sharing procedures that involve the use of monetary transactions by any of the parts (data holders, data users, intermediary).

Second, data sharing must happen 'between an undetermined number of data subjects and data holders on the one hand and data users on the other'.[32] Read in conjunction with Article 2(11)(c) and Recital 28, this specification seems to exclude services that are used (*a*) by a closed group of entities and (*b*) by a single data holder to enable the use of its data.[33] However, we currently lack specifications on the features that make a group of entities sharing data 'closed'. In the dataset, I exclude all projects whose business model is to require data sharers to pay consultancy-like services to implement data sharing infrastructure within their organization (thus creating a 'closed system').

Third, data sharing intermediation can occur 'through technical, legal, or other means'.[34] In the absence of a *de minimis* threshold, this provision possibly enlarges the scope of the DGA to entities that intermediate data sharing only sporadically and without any technical sophistication. In the dataset, I take under consideration only entities that have, among their declared goals, the systematic enabling and/or facilitation of data sharing.

This chapter must be read taking into account the dynamic nature of DISs regulation: First, the interpretation of the DGA will evolve following the European Commission's delegated acts,[35] judiciary interpretations, and further regulation or soft law. Second, the governance and business models of DISs will evolve as well, also in the attempt of service providers to comply with (or circumvent[36]) the DGA. For these reasons, the chapter must be considered as an exploration of trends that shall be verified by further empirical research, especially after the DGA enters into force.

To create the dataset, I extracted the projects that can be considered DISs according to the criteria mentioned here from five initiatives aggregating data intermediaries. These are, to my knowledge, the existent projects that systematically created datasets of data intermediation

initiatives: (*a*) the MyData organisation,[37] (*b*) Mozilla's 'Data Stewardship Landscape Scan';[38] (*c*) the 'Data Stewardship Explorer' by Aapti Institute;[39] (*d*) the Ada Lovelace Institute's list of case studies;[40] (*e*) the 'Data Collaboratives Explorer' from New York University's GovLab;[41] and (*f*) the Data Institutions Register from the Open Data Institute.[42] Finally, through a snowballing technique, I have added to the dataset all the projects that I found mentioned in the empirical analysis of DISs and the academic literature consulted to write the paper.

The included projects must have been accessible by both data providers and users in the territory of at least one of the EU members or the United Kingdom (UK) between 1 January 2000 and 31 December 2022. Projects that never passed the stage of prototype (that could not be used by a data holder or user simply by accessing its webpage) were not included. Projects whose material was not available online[43] on the date of 31 May 2023 have been also excluded. The final dataset is available in Table 8A.1 (Appendix 8A).

The following features have been collected for each DIS:

- *country of establishment* – that is, where the organization running the DIS is legally established;

- *type of DIS* – that is, the category identified by the DGA[44] into which the organization falls:

  a. infrastructure: services offering infrastructure intermediating data exchange between data holders and users (that is, data marketplaces);

  b. subject-centred: services enabling data subjects to make their data available to data users and exercise their GDPR rights (that is, PIMSs); and

  c. cooperative: organizations supporting their members in the exercise of their data rights. For some DISs, more than one category could be applied; the 'dominant' one has been decided by analysing in which order the provider states its main goals; and

- *data specialization* – that is, some DISs focus on data related to a specific disciplinary sector (health data, agricultural data, and so on), whereas others are non-specialized (that is, data marketplaces and PIMSs that collect data from other services or platforms).

Empirical Analysis

The dataset contains 54 DISs. As shown in Figure 8.1, most of them are based in the Netherlands, the United States (US), the UK, and France. Particularly, the Netherlands has such a high number because of its ecosystem of 'personal health environments' used by citizens to share data with the healthcare system.[45]

Figure 8.2 shows the frequency of the features collected in the dataset. The figure on the left shows the 'types' of DISs, according to the taxonomy proposed by the DGA (see the previous section). The majority of DISs in the dataset are subject-centred (that is, PIMSs and

services helping individuals to manage their data and exercise their data rights), and there is a significant number of DISs enabling the infrastructure for data exchanges. No operating data cooperatives have been found in the data collection. The figure on the right shows how many specialized (focusing on data related to a certain disciplinary sector) and unspecialized DISs are contained in the dataset. The majority of DISs gather data from a specific sector – particularly, 2 collect agricultural data and 34 collect health data.

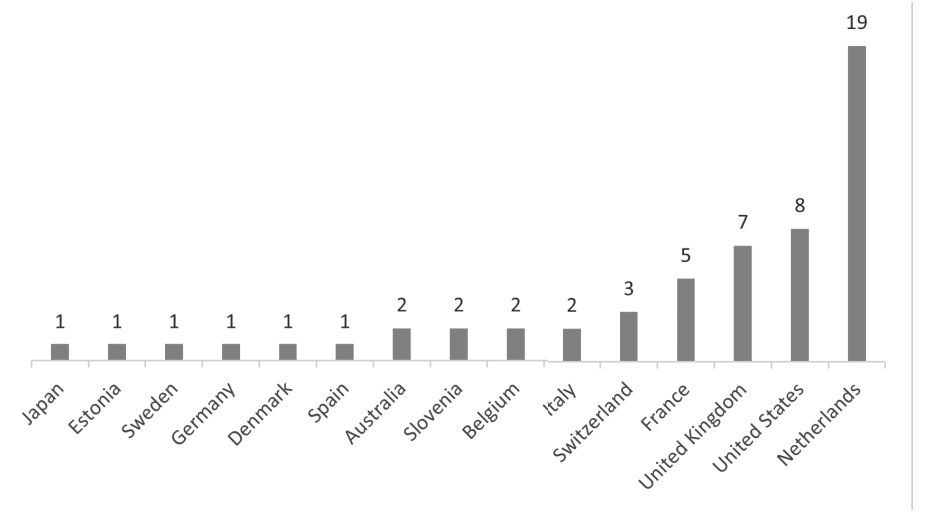


Figure 8.1 *Country of establishment of DISs that have been accessible in the territory of the EU and the UK between 2000 and 2022*

Source: Prepared by the author.

To understand whether these intermediaries can improve data access for researchers, I first need to conceptualize which features indicate how 'open' a DIS is to researchers' data access. The concept of openness includes whether academic researchers (affiliated with a university or a research institution) can access data from a DIS, how easy is it for them to access data, and whether some particular procedures are dedicated to facilitating or improving their work.

To operationalize these concepts, I took inspiration from the literature that measures the governments' data openness. DISs, like governments, are gatekeepers that regulate how large amounts of data are distributed.[46] Therefore, similar features can be taken into account to measure how much these organizations are open to sharing data with researchers. Particularly, I scrutinized the literature on the *indexes* measuring how open government data are to third parties, allowing them to create services, products, and research.[47]

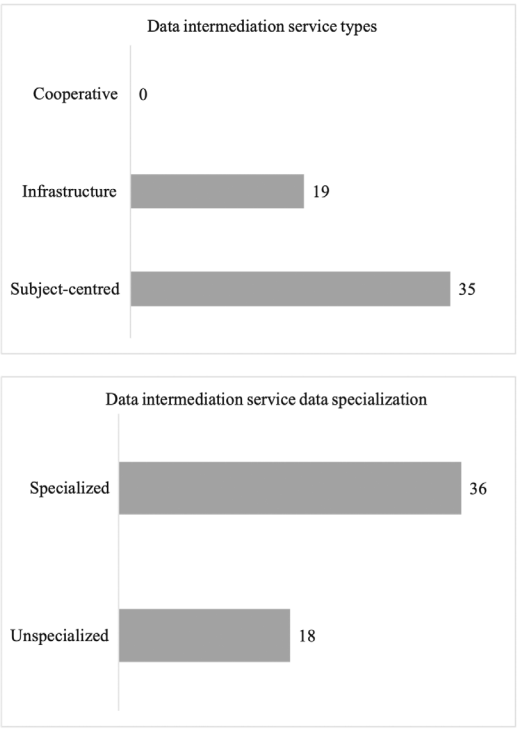


Figure 8.2 *Frequencies of types and data specialization of DIS.*

Source: Prepared by the author.

Widely cited examples of these indexes are the Five-Star Model of data availability,[48] composed of five levels depending on whether data is available on the web with an open licence, machine-readable, using non-proprietary file formats, using open standards, or linked to other people's data for context; the Four-Stage Model of data availability,[49] composed of four levels depending on whether a government offers a description of the procedure to access information, data available in a non-reusable and non-machine-readable format, data available in a reusable and machine-readable format, or data visualizable with predefined tools; and the Open Knowledge Foundation model, which scores data openness according to whether data is produced in digital format, publicly available, free of charge, available online, openly licensed, machine-readable, available in bulk, or updated.[50] Finally, in the last years, the FAIR principles have diffused as guidance for scientific data management and stewardship; they promote findable, accessible, interoperable, and reusable data.[51]

Drawing inspiration from these models, I first created a draft list of features that could be observed in DISs to measure their 'openness' vis-à-vis academic researchers (the draft list is

available in Table 8A.2 [Appendix 8A]). Then I analysed the website and documentation (that is, white papers, slideshows, explanatory material, and so on) of every DIS in the dataset. To make the research more accurate, I used the Google search engine to scrape the websites' content by using the following keywords: 'research', 'academic', and 'university'. Where no English version of the material was available, I relied on online translation services. Through an inductive process, I removed from the list the features that did not appear in any DISs and added features that were not included in the initial list. The final list includes three features that have been observed in existing DISs and are outlined in Table 8.1.

Table 8.1 Features used to assess the DISs' research openness

Feature	Description	Rationale
Data access mention	The possibility for researchers and academic institutions to access data from the project is explicitly mentioned.	Mentioning the possibility of data access for research purposes suggests that academic needs have been considered during the design of the DIS and support for researchers is available.
Dedicated data access	Research or academic institutions are provided with dedicated procedures to access data.	Dedicated procedures facilitate accessing and tailoring the data for research purposes.
Documented implementation in a research project	The data accumulated by the initiative has been employed in documented research activity by researchers or academic institutions.	The existence of projects using the DIS's data demonstrates the accessibility of such data for research purposes and possibly suggests dedicated tracks to access data.

Source: Compiled by the author.

I report here some examples, drawn from the dataset to better explain each feature, together with snippets taken from the DISs' websites:

Data access mention: Patients Know Best and Gezondheidsmeter are services through which patients can create a record of their health data and share them with healthcare providers. The websites of these services explicitly mention that patients can give consent to share their data with medical research groups and institutions.

> With Patients Know Best patients can give direct consent to share their information for research purposes. Organisations can select the institutions to work with and help share this data seamlessly in real-time with those selected research groups and associations.[52]

Dedicated data access: Seedlinked is a platform that allows growers to run seed variety trials. The platform offers a dedicated option to allow university researchers to create and study the trials. The service Selfcare allows users to integrate the measurement data from various devices in a 'personal health environment'. With the consent of the users, researchers can obtain data on a targeted population across a certain period.

> Research using wearables with which participants can carry out measurements themselves in their own (home) environment is becoming increasingly important. With the use of SelfcareResearch, this data comes to the researcher in an automated manner. As a researcher you have access to a dashboard where you can export the research data 24/7.[53]

Documented access for research purposes: BitsaboutMe and CitizenMe are services that allow consumers to provide data about their purchasing behaviour and lifestyle, and to share anonymized information to earn money. These services have been used by the University of Zurich and University of Leeds to conduct research on the changes in consumer habits and preferences.

> To participate in the research project . . . you agree to share your purchase history of Coop and/or Migros (or REWE in Germany) with the research group. This will be done completely anonymously, so no personal data such as name, email, etc. will be transmitted.[54]

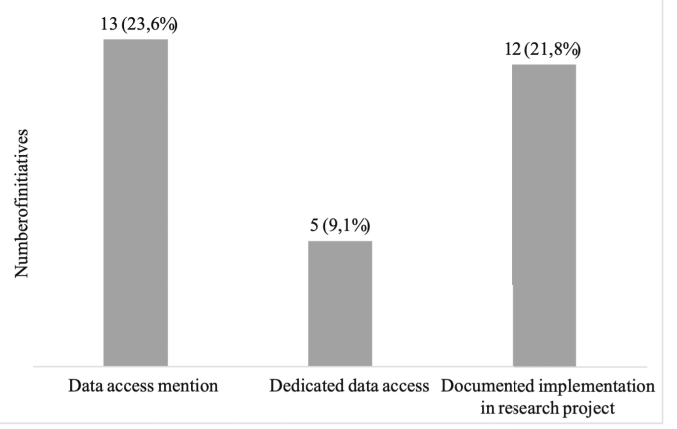


Figure 8.3 *Share of data intermediation services possessing research openness features.*

Source: Prepared by the author.

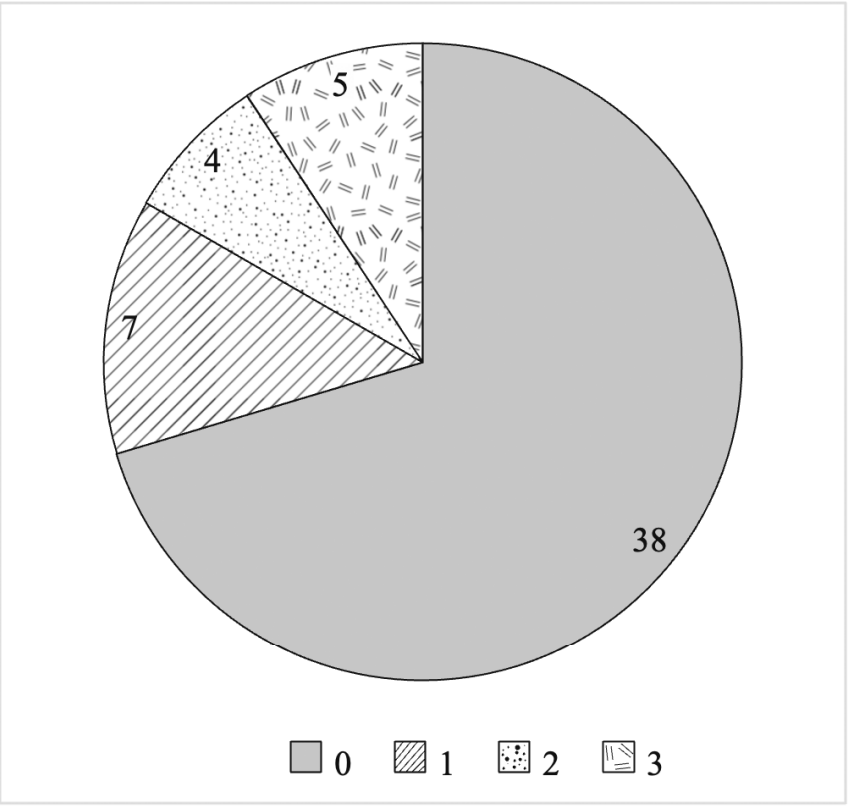


Figure 8.4 *Number of research openness features among the DISs.*

Source: Prepared by the author.

Figure 8.3 shows the number of DISs that possess research openness features. One out of four DISs mentions that their data can be accessed for academic purposes. One out of five DISs offer proof that research projects have been pursued thanks to the access to their data. On the other end, only a small minority of projects offer data access options dedicated to researchers. Figure 8.4 shows how many DISs possess one or more data openness features. Overall, about 30 per cent of DISs in the dataset possess at least one data openness feature.

This data shows that the majority of DISs do not offer any data openness features. Unsurprisingly, the majority of DISs are businesses that have more interest in establishing commercial relationships than in pursuing research partnerships. In the analysed cases, most of the partnerships with researchers attempt to promote the quality of the service, rather than suggesting that the interaction with the scholar community is a core feature of their organization. Research-friendly features are rarely at the forefront of the promoting material (that is, the website's homepage). Nevertheless, the share of DISs having at least one data openness

feature remains significant. This is surprising given the for-profit nature of DISs and the fact that these services have no official connection with academic institutions.

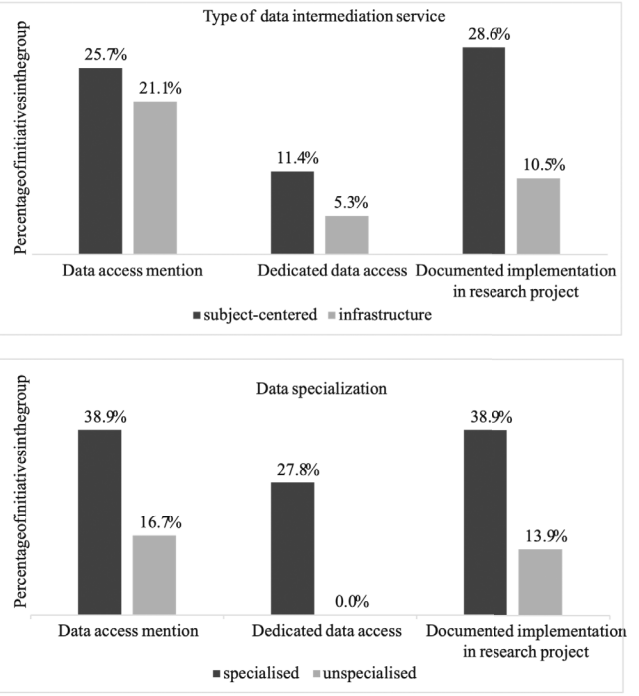


Figure 8.5 *DISs' research openness features disaggregated for DIS type and data sector.*

Source: Prepared by the author.

Note: Each bar represents the share of DISs belonging to a certain category that have a data openness feature.

Figure 8.5 shows the frequency of DISs' research openness features disaggregated for DIS type and data specialization. When looking at the difference between subject-centred and infrastructure DISs, we find that both types offer data access options; however, subject-centred DISs offer data access options more often. As regards the difference between specialized and non-specialized DISs, the specialized ones seem more likely to offer data openness features. Also, only 'specialized' DISs offer data access options dedicated to researchers. Overall, the low number of DISs does not allow for any causal inferences. However, neither the DIS type nor the data specialization seems to completely foreclose the possibility of offering data access options for researchers.

This descriptive data allows the elaboration of two hypotheses to be tested with further research: (*a*) infrastructure DISs are less likely to offer data openness features because they host 'neutral' and 'technical' infrastructures for the exchange of data and do not have any interest in favouring the access to specific categories of data users; (*b*) DISs specialized in a certain type of data are more likely to offer data openness features because they have a particular interest in gaining legitimacy from the scholars of the discipline and/or it is easier or less costly for them to provide useful data to the academic community.

The DGA Impact on Research Openness

This research investigates the choices of DISs providers *before* the implementation of the DGA, whose provisions on DISs entered into force in September 2025 (Article 37). In this section, I report which provisions are more likely to interact with the research openness features of the DISs and comment on what such impact might be.

Article 12(1)(a) introduces a 'neutrality obligation' by requiring every DIS provider to 'not use the data for which it provides data intermediation services for purposes other than to put them at the disposal of data users'.[55] The impact of such a provision on data access for research purposes is difficult to foresee. On one hand, neutrality prevents DIS providers from developing vertically integrated business models[56] that could facilitate offering services useful for researchers (that is, data analytics and manipulation services). On the other hand, limiting the competition between DISs to the modalities they put data at the disposal of data holders might push providers to improve data access openness.

Article 12 (1)(d) requires every DIS to 'facilitate the exchange of the data in the format in which it receives it from a data subject or a data holder':[57] through this measure, the European Commission aims to prevent service providers from imposing their own data standards, thus causing lock-in effects.[58] Standardization measures can improve the efficiency of data usage by researchers because they reduce barriers and prevent the loss of valuable information when merging multiple datasets.[59]

On the other hand, forced standardization might hinder the diversification strategies of new companies, thus entrenching the regulatory power of incumbents.[60] For instance, this could happen if the standard chosen by a firm becomes dominant in a certain sector and is mechanically replicated by the mediation of DISs. The DIS providers 'shall convert the data into specific formats only to enhance interoperability … or if requested by the data user or where mandated by Union law or to ensure harmonisation with international or European data standards'.[61] Therefore, the providers maintain various options to convert data formats, but the 'exceptional' framing of such options might have chilling effects.

Finally, Article 12 (1)(f) requires DIS providers to allow access in a 'fair, transparent and non-discriminatory' (FRAND) way for both data subjects and data holders.[62] This provision aims to limit the discriminatory treatment of data holders and to support organizations in weak bargaining positions.[63] This could facilitate data access for researchers, even though DIS providers could still use forms of 'pure secondary line differentiation' (that is, a DIS may

be willing to transfer data only to research projects that it deems of sufficient 'quality').[64] A possible solution is to specify a list of contractual terms that are considered 'unfair' when imposed by data holders on certain categories of data users (that is, researchers), like the initial proposal of the Data Act, 2023, protected small and medium enterprises from unilaterally imposed data sharing clauses.[65] For this purpose, a broader safe harbour for researchers could be created by enshrining a 'right to research' in future legislation.[66]

Conclusions

This chapter has explored the following question: do DISs make data more accessible to researchers? To answer this question, I have identified a set of 'research openness features' that facilitate data access for researchers when they interact with a DIS. Then I built an original dataset of 54 DISs and investigated how many of them present such features. The findings show that the majority of DISs do not offer any research openness features; nevertheless, there is a significant share of services (30 per cent) that at least mention data access for research as one of their goals, offer dedicated data access option to researchers, or have collaborated with academic projects. It is possible to draw three reflections from this analysis.

First, these results show that, in a significant minority of cases, DISs are a valuable tool to access data for academic researchers. However, in the future, this situation might significantly change due to the DGA enforcement and the evolution of the market. Further research should investigate the causes driving the supply of data access options to researchers, by testing the hypotheses on DIS types and data specialization, but also the role of the actors governing the intermediaries[67] and the corporate interests in data sharing.[68]

Second, the empirical analysis shows that to understand how the DGA will impact the distribution of data among societal actors,[69] it is necessary to take into account how data intermediaries interpret, exploit, or circumvent the new rules. A growing literature has investigated how the preferences of targeted actors affect the outcome of personal data regulation[70] and data access regulation.[71] The DGA stresses this aspect even more because it explicitly assigns regulatory functions to some types of intermediaries.[72] For these reasons, the DGA should be interpreted as a case of 'decentred' regulation, which is defined by Julia Black as the recognition of 'a shift … in the locus of the activity of "regulating" from the state to other, multiple, locations'.[73]

Third, the chapter shows that the approach of the DGA, which considers 'data altruism' operators as the only type of data intermediaries that can favour academic research, is too simplistic. The DGA allows only non-profit organizations to be identified as 'recognised data altruism organisations'. In the long term, this might discourage for-profit initiatives from sharing data for research purposes, because they cannot obtain such a certification[74] and might have unintended consequences in transnational data sharing partnerships.[75] Also, it might jeopardize their economic growth because some of their services cannot flourish at their fullest potential.[76] In such a scenario, the DGA would ultimately foreclose – instead of empowering – the researchers' access options to a greater amount of data.

Hence, this study shows how data access regulation requires a more careful assessment of the political economy of targeted actors. To tackle some of the identified issues, I advance three policy proposals that could be introduced in the European Commission's DGA rulebook,[77] delegated acts, new regulations, or soft law.

First, the interpretation of DGA provisions illustrated in the fourth section should favour academic research; for instance, the European Commission should ensure that the FRAND requirements are applied consistently in the case of academic research and could mandate compliance to particular sectoral standards to facilitate a standardization process that is not driven by private actors and where scientists are heard.

Second, the European Commission could develop tools (that is, sandboxes, research grants, and so on) to incentivize DISs to experiment with forms of data sharing with researchers, with the aim of making these relationships self-sustaining in the long term.

Finally, the European Commission should amend the DGA to make for-profit projects recognizable as 'data altruism organisations' – this would prevent chilling effects on data sharing with researchers and allow more effective monitoring of (real) data altruism initiatives.

Appendix 8A

Table 8A.1 Dataset of data intermediation services accessible in the territory of the EU and the UK, 2000-2022

Name	Data access Mention	Dedicated access researchers	Documented access for research purposes	Country	Type	Specialised
Bitsaboutme	1	0	1	Switzerland	subject	no
CGM LIFE	0	0	0	Netherlands	subject	yes
CitizenMe	1	0	1	United Kingdom	subject	no
City Data Exchange	0	0	0	Denmark	infrastructure	no
clture	0	0	0	United States	subject	no
Cozy	0	0	0	France	subject	no
Datafund	0	0	0	Slovenia	subject	no
Dawex	1	0	1	France	infrastructure	no
Digi.me	0	0	0	Australia	subject	no
DJustConnect	0	0	0	Belgium	infrastructure	yes
Drimpy	0	0	1	Netherlands	subject	yes
Fair Data Society	0	0	0	Slovenia	infrastructure	no
Gezondheids-meter	1	1	1	Netherlands	subject	yes
HAT	0	0	0	United Kingdom	infrastructure	no
iDataSwift	1	0	0	United Kingdom	subject	no
iGrant	0	0	0	Sweden	subject	no
ishare	0	0	0	Netherlands	infrastructure	no
Ivido	0	0	1	Netherlands	subject	yes
JoinData	1	0	0	Netherlands	infrastructure	yes
Lotame	0	0	0	United States	infrastructure	no
Luna	1	1	1	United States	subject	yes
Marketplace. city	0	0	0	United States	infrastructure	no
Medsafe	0	0	0	Netherlands	subject	yes
medxpert	0	0	0	Netherlands	subject	yes

Meeco	0	0	0	Australia	infrastructure	no
MIJNPGO	0	0	0	Netherlands	subject	yes
MyDataMood	0	0	0	Spain	subject	no
Mydex	0	0	0	United Kingdom	infrastructure	no
Myfairdata	0	0	0	France	subject	no
Ockto	0	0	0	Netherlands	subject	no
OneCub	1	0	0	France	infrastructure	no
Ozone.ai	0	0	0	United States	subject	no
Patients Know Best	1	1	1	United Kingdom	subject	yes
Personium	0	0	1	Japan	subject	no
PolyPoly	0	0	0	Germany	subject	no
Prifina	0	0	0	United States	subject	no
Qii	0	0	0	Netherlands	subject	no
QIY	0	0	0	Netherlands	infrastructure	no
Quli	1	0	0	Netherlands	subject	yes
Seedlinked	1	1	1	United States	infrastructure	yes
Selfcare	1	1	1	Netherlands	subject	yes
Snowflake	0	0	0	United States	infrastructure	no
Solid	1	0	1	United Kingdom	subject	no
Streamr	0	0	0	Switzerland	infrastructure	no
Swarm	0	0	0	Switzerland	infrastructure	no
swashapp	0	0	0	United Kingdom	subject	no
Uw Zorg online	0	0	0	Netherlands	subject	yes
Visions	0	0	0	France	infrastructure	no
VitaalBank	0	0	0	Netherlands	subject	yes
Weople	0	0	0	Italy	subject	no
X-Road	0	0	0	Estonia	infrastructure	no
Yivi (ex IRMA)	0	0	0	Netherlands	subject	no
Zodos	0	0	0	Netherlands	subject	yes
Zorgdoc	0	0	0	Netherlands	subject	yes

Souce: Prepared by the author.

Table 8A.2 Draft list of features used to assess the DISs' research openness

Feature	Description	Rationale
Data access mention	The possibility for researchers and academic institutions to access data from the project is explicitly mentioned.	Mentioning the possibility of data access for research purposes suggests that academic needs have been considered during the design of the DIS and support for researchers is available.
Dedicated data access	Research or academic institutions are provided with dedicated procedures to access data.	Dedicated procedures facilitate accessing and tailoring the data for research purposes.
Documented implementation in a research project	The data accumulated by the initiative has been employed in documented research activity by researchers or academic institutions.	The existence of projects using the DIS's data demonstrates the accessibility of such data for research purposes and possibly suggests dedicated tracks to access data.
Free of access	Research or academic institutions can obtain data for free or with a monetary discount, contrarily to other actors.	Economic incentives facilitate the use of DIS data for research purposes.
Governance representatives	Research or academic institutions are formally allowed to influence the governance of the organization.	Academic representatives can represent research interests within the DIS.

Source: Prepared by the author.

Notes

1 Katharina Pistor, 'Rule by Data: The End of Markets?' *Law and Contemporary Problems* 101, no. 83 (2020): 101–124.

2 Maurice Stucke and Allen Grunes, *Big Data and Competition Policy* (Oxford University Press, 2016).

3 Viktor Mayer-Schönberger and Thomas Ramge, *Reinventing Capitalism in the Age of Big Data* (Basic Books, 2018).

4 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', *Digital Journalism* 532, no. 7 (2019): 532–540; Axel Bruns, 'After the "APIcalypse": Social Media Platforms and Their Fight against Critical Scholarly Research', *Information, Communication and Society* 1544, no. 22 (2019): 14–36.

5 Jef Ausloos and Michael Veale, 'Researching with Data Rights', *Technology and Regulation* 136, no. 2 (2020): 136–157.

6 Marco Botta, 'Shall We Share? The Principle of FRAND in B2B Data Sharing', Robert Schuman Centre for Advanced Studies Research Paper, 2023, 25–32.

7 '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 L 119, Article 20.

8 'Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) [2022]', OJ L 265, Article 6.

9 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act)', OJ L 277, Article 40.

10 European Commission, 'Commission Staff Working Document Impact Assessment Report Accompanying the Document Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act)', European Commission, 2020, SWD/2020/295 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020SC0295; Bertin Martens, Alexandre de Streel, Inge Graef, Thomas Tombal, and Néstor Duch-Brown, 'Business-to-Business Data Sharing: An Economic and Legal Analysis', JRC Digital Economy Working Paper Series (2020), https://ec.europa.eu/jrc/sites/jrcsh/files/jrc121336.pdf (accessed 4 April 2023).

11 Alina Wernick, 'Defining Data Intermediaries: A Clearer View through the Lens of Intellectual Property Governance', *Technology and Regulation* 2 (2020): 65–77.

12 Heiko Richter and Peter R. Slowinski, 'The Data Sharing Economy: On the Emergence of New Intermediaries', *IIC: International Review of Intellectual Property and Competition Law* 4, no. 25 (2019): 4–29.

13 'Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance and Amending Regulation (EU) 2018/1724 (Data Governance Act) [2022]', OJ L 152 (DGA).

14 Julie Baloup, Emre Bayamlıoğlu, Aliki Benmayor, Charlotte Ducuing, Lidia Dutkiewicz, Teodora Lalova-Spinks, Yuliya Miadzvetskaya, Bert Peeters, 'White Paper on the Data Governance Act', white paper on the Data Governance Act, CiTiP Working Paper, 2021, https://lirias.kuleuven.be/3473884 (accessed 20 August 2023); Gabriele Carovano and Michèle Finck, 'Regulating Data Intermediaries: The Impact of the Data Governance Act on the EU's Data Economy', *Computer Law*

and Security Review 105830, no. 50 (2023), DOI: https://doi.org/10.1016/j.clsr.2023.105830.

15 Raphaël Gellert and Inge Graef, 'The European Commission's Proposed Data Governance Act: Some Initial Reflections on the Increasingly Complex EU Regulatory Puzzle of Stimulating Data Sharing', in *Digitalisering en conflictoplossing*, ed. P. T. J. Wolters, André Janssen, Pietro Ortolani, Pieter Theo Jozef Wolters, and Rudolf Martinus Hermans, 11–13 (Wolters Kluwer, 2021); Jathan Sadowski, 'The Political Economy of Data Intermediaries', Ada Lovelace Institute, 1 June 2022, https://www.adalovelaceinstitute.org/ blog/political-economy-data-intermediaries (accessed 7 February 2023).

16 DGA, Article 18(c).

17 DGA, Recitals 3, 6, 27, 45.

18 DGA, Article 2(11).

19 The Personal Information Management Systems (PIMSs) are services that aim to put users in control of their personal information. To do this, they usually facilitate the exercise of data rights such as the GDPR right to access, withdraw consent, and data portability. European Data Protection Supervisor, 'Opinion 9/2016-EDPS Opinion on Personal Information Management Systems', https://www.edps.europa.eu/sites/default/files/ publication/16-10-20_pims_opinion_en.pdf (accessed 8 December 2024).

20 DGA, Article 10.

21 DGA, Article 12 (1)(d).

22 DGA, Article 12 (1)(f).

23 DGA, Article 12 (1)(a).

24 DGA, Article 2(16).

25 DGA, Article 19.

26 DGA, Article 18(c).

27 DGA, Article 2(16).

28 Quli, 'Quli Voor Patiëntgroepen', Health Bank, 2023, https://www. healthbank.coop (accessed 25 February 2023).

29 'Case Studies', CitizenMe, 2023, https://www.citizenme.com/case-studies (accessed 4 March 2023).

30 Baloup, Bayamlıoğlu, Benmayor, Ducuing, Dutkiewicz, Lalova-Spinks, Miadzvetskaya, and Peeters, 'White Paper on the Data Governance Act'; European Data Protection Board and European Data Protection Supervisor, 'EDPB–EDPS Joint Opinion 03/2021 on the Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act) | European Data Protection Board', 2021, https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-jointopinion/edpb-edps-joint-opinion-032021-proposal_en (accessed 27 May 2023).

31 DGA, Article 2(11).

32 DGA, Article 2(11).

33 Carovano and Finck, 'Regulating Data Intermediaries', 7.

34 DGA, Article 2(11).

35 See, for instance, the rulebook mentioned in Article 22 of the DGA.

36 Carovano and Finck, 'Regulating Data Intermediaries', 6–9.

37 'Organisation Members', MyData, 2023, https://www.mydata.org/ participate/members

(accessed 23 February 2023).

38 Stefan Baack and Madeleine Maxwell, 'Alternative Data Governance Approaches: Global Landscape Scan and Analysis', Mozilla, 16 September 2020, https://foundation.mozilla.org/en/data-futures-lab/data-for-empowerment/ whos-trying-global-landscape-scan-and-analysis (accessed 23 May 2023).

39 Siddharth Manohar, Astha Kapoor, and Aditi Ramesh, 'Understanding Data Stewardship: Taxonomy and Use Cases', Aapti Institute, 2020, https:// thedataeconomylab.com/wp-content/uploads/2020/06/UnderstandingData-Stewardship-Aapti-Institute.pdf (accessed 5 December 2024).

40 Valentina Pavel, 'Rethinking Data and Rebalancing Digital Power', Ada Lovelace Institute, 17 November 2022, https://www.adalovelaceinstitute. org/report/rethinking-data (accessed 23 January 2023).

41 Stefaan Verhulst and David Sangokoya, 'Data Collaboratives: Exchanging Data to Improve People's Lives', Medium, 23 April 2015, https://sverhulst. medium.com/data-collaboratives-exchanging-data-to-improve-people-slives-d0fcfc1bdd9a (accessed 12 April 2023).

42 'The Data Institutions Register', Open Data Institute, 30 June 2021, https:// theodi.org/insights/tools/the-data-institutions-register (accessed 10 April 2024).

43 Unavailable pages have been searched on the Internet Archive's Wayback Machine, a non-profit project to preserve archived copies of defunct webpages. Only one of the entries in the database has been added through the information found on the archive (Ozone.ai).

44 DGA, Article 10.

45 Particularly, the Netherlands has such a high number because of its ecosystem of intermediaries used by citizens to share data with healthcare providers. For more information on how such data intermediaries are integrated in the Dutch healthcare system, see the website of its standardization body: 'Homepage', MedMij, 2023, https://medmij.nl (accessed 20 February 2023).

46 Gellert and Graef, 'The European Commission's Proposed Data Governance Act', 11–13; Karine Barzilai-Nahon, 'Toward a Theory of Network Gatekeeping: A Framework for Exploring Information Control', *Journal of the American Society for Information Science and Technology* 1493, no. 59 (2008): 1493–1512.

47 Marwah W. Alrushaid and Abdul Khader Jilani Saudagar, 'Measuring the Data Openness for the Open Data in Saudi Arabia E-Government: A Case Study', *International Journal of Advanced Computer Science and Applications* 7, no. 12 (2016): 113–122, 115.

48 Tim Berners-Lee, 'Linked Data: Design Issues', W3C, 2006, https://www. w3.org/DesignIssues/LinkedData.html (accessed 28 February 2023).

49 David Osimo, 'Benchmarking eGovernment in the Web 2.0 Era: What to Measure, and How', *European Journal of ePractice* 37, no. 4 (2008): 1–11.

50 'Methodology: Global Open Data Index', Open Knowledge Foundation, 2016, https://index.okfn.org/methodology/index.html (accessed 16 February 2023).

51 Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and

Barend Mons, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data* 160018, no. 3 (2016): 1–9.

52 'Commissioners and Payers', Patients Know Best, 2023, https:// patientsknowbest.com/commissioners (accessed 20 March 2023).

53 'SelfcareResearch: Real World Data', SelfCare, 2023, https://selfcare4me. com/selfcareresearch-real-world-data (accessed 25 March 2023).

54 'Contribute to the Research Project of UZH on the Economic Impact of SARS-CoV-2', BitsAboutMe, 2023, https://bitsabout.me/en/contribute-tothe-research-on-the-impact-of-sars-cov-2 (accessed 19 March 2023).

55 DGA, Article 12 (1)(a).

56 Aline Blankertz and Louisa Specht, 'What Regulation for Data Trusts Should Look Like', Stiftung Neue Verantwortung, 2021, https://www.stiftung-nv. de/sites/default/files/regulation_for_data_trusts_0.pdf (accessed 1 June 2023).

57 DGA, Article 12 (1)(d).

58 Lukas von Ditfurth and Gregor Lienemann, 'The Data Governance Act: Promoting or Restricting Data Intermediaries?' *Competition and Regulation in Network Industries* 23, no. 4 (2022) 270–296, 284.

59 Michael Mattioli, 'The Data-Pooling Problem', *Berkeley Technology Law Journal* 179, no. 32 (2017), https://www.repository.law.indiana.edu/ facpub/2663 (accessed 5 December 2024); European Commission, Catarina Arnaut, Marta Pont, Elizabeth Scaria, Arnaud Berghmans, and Sophie Leconte, *Study on Data Sharing between Companies in Europe: Final Report* (Publications Office of the European Union 2018), 80.

60 Fabrizio Cafaggi and Katharina Pistor, 'Regulatory Capabilities: A Normative Framework for Assessing the Distributional Effects of Regulation: Regulatory Capabilities', *Regulation and Governance* 95, no. 9 (2015): 95–107.

61 DGA, Article 12 (1)(d).

62 DGA, Article 12 (1)(f).

63 von Ditfurth and Lienemann, 'The Data Governance Act', 286.

64 Inge Graef, 'Differentiated Treatment in Platform-to-Business Relations: EU Competition Law and Economic Dependence', *Yearbook of European Law* 38 (2019): 448–499, 456–458.

65 European Commission, Proposal for a Regulation of the European Parliament and of the Council on Harmonized Rules on Fair Access to and Use of Data (Data Act) 2022 [COM/2022/68 final], Article 13.

66 See chapter 11 in this volume.

67 Kenneth W. Abbott and Snidal Duncan, 'Strengthening International Regulation through Transnational New Governance: Overcoming the Orchestration Deficit', in *The Spectrum of International Institutions*, ed. Kenneth W Abbott and Duncan J. Snidal, 95–139 (Routledge 2021).

68 Abraham L Newman, 'What You Want Depends on What You Know: Firm Preferences in an Information Age', *Comparative Political Studies* 1286, no. 43 (2010), DOI: https://doi.org/10.1177/0010414010369068; Christine Trampusch, 'Regulating the Digital Economy: Explaining Heterogenous Business Preferences in Data Governance', *Journal of European Public Policy* 31, no. 7 (2023): 1902–1926.

69 Angelina Fisher and Thomas Streinz, 'Confronting Data Inequality', *Columbia Journal of Transnational Law* 829, no. 60 (2021): 829–956.

70 See Ausloos and Veale, 'Researching with Data Rights', 138–139, on the politics of APIs, Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, andIgnacio Sanchez, 'The Right to Data Portability in the GDPR: Towards User-Centric Interoperability of Digital Services', *Computer Law and Security Review* 193, no. 34 (2018): 193–203, and Matteo Nebbiai, 'Intermediaries Do Matter: Voluntary Standards and the Right to Data Portability', *Internet Policy Review* 11, no. 2 (2022), on the interpretation of the right to data portability.

71 See chapters 5 and 9 in this volume.

72 Gellert and Graef, 'The European Commission's Proposed Data Governance Act', 11–13.

73 Julia Black, 'Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a "Post-Regulatory" World', *Current Legal Problems* 54, no. 1 (2001): 103–146, 112.

74 Winfried Veil, 'Data Altruism: How the EU Is Screwing Up a Good Idea', Algorithm Watch, https://algorithmwatch.org/de/wp-content/ uploads/2022/01/2022_AW_Data_Altruism_final_publish.pdf (accessed 4 April 2023).

75 See chapter 7 in this volume.

76 Carovano and Finck, 'Regulating Data Intermediaries'.

77 DGA, Article 22.

9. ACCESS TO DATA ON DISINFORMATION WITHIN THE CODE OF PRACTICE ON DISINFORMATION

MICHALINA KOWALA

From Discretion to Obligation? The Researchers' Access to Data on Disinformation in the Framework of the Code of Practice on Disinformation

The phenomenon of disinformation is not new. As far back as antiquity, Julius Caesar used the tools of propaganda to demonstrate his power in order to convert people to the Roman way of life.[1] Today, disinformation, defined as 'verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm',[2] is propagated above all in the digital environment. It contributes to polarization and spread of extreme ideas,[3] leading to rise in populism and major social tensions. It 'undermines trust in institutions and in traditional and digital media, and hinders citizens' ability to make informed decisions'.[4] It impacts the policymaking process by giving a distorted image of the actions of public authorities.

Disinformation has intensified in the run-up to the presidential elections in the United States in 2016 and the Brexit referendum. According to 83 per cent of Europeans, this phenomenon constitutes a threat to democracy.[5] Spread mainly by online platforms,[6] it required an immediate legal response at the European Union (EU) level. The Code of Practice on Disinformation, established in 2018, has been the first such framework worldwide, setting out the commitments by platforms and industry to fight disinformation. One of its objectives was to empower the research community. However, the assessment of the implementation of the commitments enshrined therein conducted by the European Commission, the European Regulators Group for Audiovisual Media Services (ERGA), and other independent consultancy bodies revealed the multitude of shortcomings. To address them, the strengthened Code of Practice on Disinformation was adopted in 2022. The objective of this chapter is to discuss the changes in data access for researchers studying disinformation since the adoption of the new code.

First, the framework of the researchers' access to data provided in the 2018 code will be presented. Second, I will refer to my research experience in accessing data from Facebook through the mechanisms implemented by the platform following the adoption of the 2018 code. Third, I will discuss the commitments included in the 2022 code regarding researchers' access to data. I will look at the types of data that researchers are entitled to access, the process of applying for and obtaining data, and the adopted enforcement mechanisms. Further, the focus will be on the tools that Facebook has put in place to implement the commitments included in the new code. I will once again refer to my research experience in order to assess whether my goals would be achievable under the 2022 code. It will involve the comparative analysis between the code (voluntary initiatives) and binding law (notably the Digital Services

Act [DSA], 2022). The objective of the chapter is also to determine how the relationship between researchers and the platform as regards the access to data on disinformation has changed and what characterizes it today. I make use of the autoethnography method of research. It allows for the assessment of the tools made available by Facebook as a result of the implementation of the codes and leads to the determination of the scope of researchers' access to data on disinformation. I decided to choose Facebook[7] to study the mechanisms for granting researchers' access to data since it was one of the first signatories of the 2018 code. Moreover, it is considered as the most used social network for news[8] and 'the worst perpetrator' as regards the spread of disinformation.[9]

The 2018 Code of Practice on Disinformation

The European Commission in 2018 declared that

> there are growing expectations that online platforms should not only comply with legal obligations under the EU and national law but also act with appropriate responsibility in views of their central role so as to ensure a safe online environment, to protect users from disinformation, and to offer users exposure to different political views.[10]

It called upon platforms to step up their efforts to tackle online disinformation. The European Commission considered that self-regulation could contribute to these efforts, provided it is effectively implemented and monitored. To this end, the commission supported the development of the Code of Practice on Disinformation,[11] which was signed in October 2018. It constitutes a set of standards agreed by online platforms, tech companies, and representatives of the advertising industry to address the spread of disinformation.

The 2018 Code: An Instrument of Soft Law

The use of the soft law is often explained by the greater ease with which stakeholders can formulate and reach agreement, as well as by lower administrative costs. The adoption of such an instrument is seen in some cases as a temporary alternative to binding legislation, which may already be in the pipeline.[12] Furthermore, soft law constitutes a mean of stimulating progress and is preferred when member states have considerable interests that they do not wish to jeopardize. It is often proposed to regulate the online environment, which is characterized by rapid technological change.[13] Contrary to hard law, it contains non-binding measures that cannot be legally enforced and is of voluntary nature.

Commitments under the 2018 Code

The code consists of commitments divided into five groups and related to following five areas: (*a*) scrutiny of ad placements, (*b*) political advertising and issue-based advertising, (*c*) integrity of services, (*d*) empowering consumers, and (*e*) empowering the research community. It bound only its signatories who were free to select the commitments they wanted to sign up for. Due to the nature of the code, no mechanism of enforcement or for action in case of non-compliance had been provided.[14]

As regards the empowerment of researchers with tools enabling the scrutiny of the phenomenon of disinformation and measures implemented by platforms to address it, the signatories of the 2018 code committed to taking the reasonable measures to enable privacy-compliant access to data for research activities. They recognized the purpose of scientific cooperation and committed to provide relevant data on the functioning of their services, including data for independent investigation by academic researchers and general information on algorithms.[15] In order to fulfil these commitments, signatories declared their support for 'good faith independent' efforts to track disinformation and to understand its impact. This support included sharing privacy-protected datasets, undertaking joint research, or partnering with academics and civil society organizations.[16] To measure and monitor the code's effectiveness, the signatories committed to write an annual account of their work to counter disinformation. However, no specific reporting scheme has been established for this purpose, and neither the metrics that should be reported by the signatories. The latter agreed to cooperate with the European Commission, including making available appropriate information upon request or responding to the commission's questions and consultations.

Researchers' Access to Data under the 2018 Code of Practice on Disinformation

In 2020, I conducted research on measures to combat disinformation introduced by Facebook. My objective was to assess whether the implemented mechanisms corresponded to, and could contribute to, the achieving of specific goals enshrined in the 2018 code.[17] I decided to analyse the code through the lens of my personal experience as a researcher examining the issues related to disinformation. I examined the data provided in the reports submitted monthly and yearly by the platform to the European Commission as well as their assessment conducted by the latter.[18] I had limited myself to only this source of information.

I faced several difficulties in establishing a complete landscape of implemented measures based on data supplied by the platform. In many cases, the measures that were presented in the reports did not even indirectly address the problem of disinformation or were not specifically tailored to tackle this issue.[19] The platform reported on harmful practices and adopted solutions selectively, by referring to the global level, and sometimes, but rarely, to the European level or to certain member states without any justification for such a mode of communication. This made it difficult, if not impossible, to assess the risks and the adequacy of the measures implemented to address them. Although the reports followed the structure of the code and referred to the measures taken in relation to the five pillars included therein,[20] the lack of an established structure and the specific metrics, indicators, or elements that should be communicated left platforms and, in the discussed case, Facebook absolutely free to decide on the presented data and the way it will be done. Therefore, the researcher could not follow the development of the implemented measures or assess its effectiveness. This was due to several reasons, including inconsistent presentation of data,[21] the repetition of the same data over several months, the presentation of data for a different country each month for a single programme or policy, or the restriction to general statements and the failure to present any figures or percentages.

My impressions as regards the access to data on disinformation in the framework of the 2018 code were reflected in other research projects and reports. The EU DisinfoLab pointed to the complexity of the reports published by platforms and the difficulty in making the comparison between them. According to the organization, it hindered the meaningful analysis of the measures adopted in the context of the disinformation on COVID-19.[22] It was due to the own reporting style of each platform filling in the metrics according to its preferences, due to the lack of specific data such as country-specific metrics, especially regarding the audience of disinformation (clickthrough rate, and so on), or due to the lack of uniform presentation as regards all the countries where the measures have been implemented.[23] The EU Disinfo Lab called for more detailed guidelines on common metrics and streamlined reporting, which were needed to allow for meaningful comparison of platforms' responses to disinformation.[24] In the assessment of the implementation of the 2018 code conducted by the European Commission, the need for 'more consistent reporting adhering to certain minimum information standards that could allow for an even better assessment of the effectiveness of the implementation of the Code'[25] was expressed. Moreover, it has been pointed out that 'the independent auditing of the data delivered by the platforms in their reports could eliminate the debate on whether this data is correct and representative'.[26]

Indeed, in the 2018 code, the signatories did not commit to set up an independent body to ensure the transparency of the presented information. The European Commission was the only interlocutor of platforms. Since no enforcement measures have been foreseen, it could only respond to signatories' reports by encouraging the provision of more granular data, by expressing its concerns, by urging to take further action, or by regretting that the signatories did not supply sufficient information.[27] However, the platforms' response varied depending on whether the provision of specific data corresponded to its transparency policies.

Although the essential role of researchers in providing the proper understanding of the phenomenon of disinformation and in contributing to the development of risk-mitigation mechanisms has been recognized in the 2018 code, the assessment of the implementation of the commitments contained therein reveals multiple shortcomings.[28] The incomplete, inconsistent, out-of-context, and selective provision of data by platforms and the lack of an enforcement mechanism made researchers entirely dependent on platforms. The very act of data provision, and the way in which it was done, was mostly based on voluntariness and discretion of platforms. Researchers were unable to access information that would allow them to reconstruct the context, to establish an overall picture, to determine the cause–effect sequence, or to assess the effectiveness of the implemented measures. It also influenced the quality and the effectiveness of my research. The dynamic of the relationship between researchers and platforms was marked by one-sidedness and the powerlessness of the latter.

The 2022 Code of Practice on Disinformation

The new code,[29] signed in June 2022, intends to address these shortcomings. It constitutes an answer to the calls to reinforce the 2018 code in areas such as larger participation of platforms with more tailored commitments, demonetization of disinformation, provision of a comprehensive coverage of forms of manipulative behaviour, empowering users to flag

disinformation, as well as increase in the coverage of fact-checking and access to data to researchers and robust monitoring framework.[30] The purpose of the 2022 code is to become a more effective tool for countering disinformation. It was issued based on the guidance provided by European Commission[31] and took into account the proposal (at that time) for the DSA in the regulatory framework of which the code would be implemented. However, as underlined by the commission, 'The 2022 Code of Practice is the result of the work carried out by the signatories. It is for the signatories to decide which commitments they sign up to and it is their responsibility to ensure the effectiveness of their commitments' implementation.'[32]

The new code is a part of a broader regulatory framework. It is aligned with the DSA. The latter, adopted in October 2022, constitutes a common set of rules on intermediaries' obligations and accountability across the single market.[33] The connection of the code with the DSA entails specific obligations arising especially for very large online platforms (VLOPs) for which the code becomes a mitigation measure for systematic risks, one of which is disinformation.

Provision of Access to Different Types of Data

The 2022 code includes commitments to set up a framework for robust access to platforms' data by the research community and adequate support for the researchers' activities.[34] To achieve it, different categories of data were identified and different ways to access them were laid out. First, platforms committed to provide access to non-personal and anonymized, aggregated, or manifestly made public, continuous, real-time, or near-real-time data pertinent to undertake research on disinformation. Its provision takes place without any application procedure. Second, they committed to provide data on the signatories' services such as accounts belonging to public figures (for example, elected officials), news outlets, and government accounts. In this case, the provision of data occurs through a specific procedure. The access should be provided through automated means such as application programming interface (APIs) or other open and accessible technical solutions. Third, the 2022 code provides a governance structure for access to data for research purposes requiring additional scrutiny. It will be interesting to see whether the researcher will be able to expect a specific action from platforms in terms of the requested data, such as collection, comparison, and summary of data or whether the role of platforms will be limited to providing data in raw form only.

Involvement of Third-Party Body

The relevant signatories declared to set up, fund, and cooperate with a third-party body. Its role will be to vet researchers willing to scrutinize disinformation.[35] Once vetted, they should be able to access the personal data shared by signatories in accordance with protocols to be defined by the independent third-party body. Signatories committed to also 'support good faith research into disinformation' that involves their services by maintaining an open dialogue with researchers to keep track of the types of data that are likely to be in demand and by ensuring transparency on data types that are currently made available to researchers across Europe.[36] Finally, signatories committed to conduct research based on transparent methodology and ethical standards, as well as to share datasets, research findings, and methodologies with relevant audiences.[37]

The establishment of the third-party body to help oversee and even implement the processes envisioned by the code was already recommended in relation to the 2018 code.[38] Its role, according to the 2022 code, will be to vet researchers and to determine who should have access to data requiring additional scrutiny. This should eliminate platform discretion when deciding who should be granted access to data. Although granting access still ultimately depends on signatories, they have committed to cooperate with the said body. The question arises as to the scope of this cooperation. It is not clear whether the role of the third-party body would consist also of checking whether the data provided by platforms corresponds with the one requested by the researcher. It does not stem from the commitments included in the 2022 code as to whether the researcher will have a possibility of appeal when the data would not correspond to what they have requested and to whom they could turn in such a case. It is not clear whether the consideration of appeals would also be the role of the third-party body.[39]

Alignment of the Code with the DSA

The 2022 code, like its predecessor, is voluntary. It means that the adherence to the code does not imply legal consequences in case of the lack of implementation of the commitments included therein. However, it is considered as a possible risk-mitigation measure under Article 35 of the DSA. For VLOPs,[40] this means that the complete abandonment of the adoption of voluntary measures is not possible in practice.

The Code of Practice on Disinformation is considered as a code of conduct under Article 45 of the DSA and disinformation is considered as a systematic risk.[41] The latter, according to Recital 79 of the DSA, can stem from 'the design, functioning and use of the services of very large online platforms, as well as from potential misuses by the recipients of the service'. The EU legislator has distinguished four categories of systematic risks and classified the dissemination of disinformation in one of them. Platforms should diligently mitigate the systemic risks identified in the risk assessments, in observance of fundamental rights, for example, by initiating and joining the codes of conduct.

According to Article 45 of the DSA where significant systemic risk emerges, the European Commission may invite the providers of VLOPs concerned or the providers of very large online search engines (VLOSEs) concerned and other actors to participate in the drawing up of codes of conduct, including setting out of commitments to take specific risk-mitigation measures, as well as a regular reporting framework on any measures taken and their outcomes. The commission should also aim to ensure that participants report on any measures taken and their outcomes. According to Recital 104 of the DSA, the refusal without proper explanations by an online platform of the commission's invitation to participate in the application of such a code of conduct could be taken into account, where relevant, when determining whether the online platform has infringed the obligations laid down by the DSA. It has been specified in the same recital that the mere fact of participating in and implementing a given code of conduct should not in itself presume compliance with the DSA. Therefore, the signatories must be proactive in addressing the systemic risks, including the circulation of disinformation.[42]

The 2018 code was criticized for the lack of enforcement measures.[43] As for the new code, the enforcement mechanisms can be identified as a result of its alignment with the DSA. According to Article 37 of the DSA, in order to monitor and assess the compliance of VLOPs with certain DSA obligations and, where relevant, the commitments undertaken pursuant to *codes of conduct*, the independent audits should be conducted at least once a year.[44]

The signatories of the 2018 code could choose the commitments they wanted to sign up for. As for its updated version, the signatories agreed to sign up for commitments that are relevant to the products, activities, and services they offer. In case when they do not sign up to a commitment because it is not relevant or pertinent to their services, they will explain the reasons for this. It should be noted that this mechanism was not foreseen in the previous version of the code. In my opinion, it obliges the signatories to be transparent when it comes to subscription to the commitments and their further implementation and leaves less room for discretion in this respect. The conversion of the voluntary 2022 Code into the Code of Conduct within the framework of DSA took effect starting from 1 July 2025, making commitments included therein auditable from that date onwards.

Reporting Scheme

To address the problem of fragmentation and lack of uniformity of reported data, the 2022 code includes an intensified reporting scheme. Signatories committed to provide the baseline reports to the European Commission within one month after the end of the implementation period.[45] After that, VLOPs committed to provide regular reporting on service level indicators and qualitative reporting elements every six months and other signatories yearly, at service and MEMBER_ STATE level.[46] This should allow for a thorough assessment of the extent of the code's implementation. VLOPs are confronted with more demanding requirements as to the frequency of reporting since they are considered as posing particular risk in dissemination of illegal content and societal harms.[47]

Signatories also commit to participate in a permanent task force[48] chaired by the European Commission and including representatives from the European Digital Media Observatory (EDMO), the ERGA, and the European External Action Service (EEASS). The task force's role is to establish the harmonized reporting templates for the code's implementation[49] which the signatories undertake to apply.[50] In February 2023,[51] the signatories published their first baseline reports on how the 2022 code's commitments are implemented.[52]

The new reporting mechanism clearly determines the elements that should be presented and the way in which this should be done. In theory, and from the perspective of researchers studying the problem of disinformation, it should enable access to more specific data that can be analysed and compared. The signatories should also be less likely to present the results of activities not aimed at combating disinformation by plugging them in as such or to provide data selectively, for example, only for a few member states, since they should follow the agreed reporting scheme and include the said indicators. A specific infrastructure regarding reporting has been put in place, and the establishment of the Task Force gives hope that that discretionary reporting will be reduced or even eliminated.

Researchers' Access to Data under the 2022 Code of Practice on Disinformation

After the brief analysis of the commitments and measures enshrined in the 2022 code, it is worth discussing the instruments made available for researchers by Facebook as a result of its implementation. I will refer also to my original research project aimed at assessing whether implemented mechanism by the platform contributes to achieving the goals enshrined in the code. I will analyse whether my research purpose would be successful under the 2022 code. My objective is to verify what, if anything, has changed in the 2022 code that would allow achieving the objectives of the code more effectively.

Access to Data Not Requiring the Application Process

With regard to the commitment to provide access to non-personal data and anonymized, aggregated, or manifestly made public data for research purposes on disinformation, Facebook, in its recently published baseline report, refers to the data provided in its 'Community Standards Enforcement Report',[53] 'Widely Viewed Content Report',[54] and 'Quarterly Adversarial Threat Report'.[55] It declares to support independent research that will enhance understanding of the impact that platforms like Facebook have on society. Moreover, it claims that its policies are based on years of experience and expertise in trust and safety, combined with external input from experts around the world.[56] However, although the access to provided data should serve research on disinformation, there is little data on this particular problem in the indicated reports. The platform repeats and rephrases what has already been said publicly and adds nothing new.

Access to Data Requiring the Application Process

Regarding access to data that requires an application process,[57] Facebook refers to a tool called CrowdTangle. Launched in 2019, it is described as a content discovery and social monitoring platform that provides access to a small subset of public data on Facebook. Researchers used CrowdTangle to study a variety of key topics of social interest, including misinformation, elections, and societal impact of COVID-19.[58] To gain access to this data, researchers must complete an application process.

CrowdTangle has been considered as an unparalleled tool for 'analysing trends, tracking article sources and understanding virality on Facebook'.[59] In general, it is regarded as a vital instrument for researching Facebook's transparency.[60] Despite some rumours that Facebook plans to phase it out,[61] on the CrowdTangle website updated at the beginning of February 2023, the platform declared continued support for the research community and its plans to update this page and expand its support into new areas.[62] It is also indicated that the university researchers (faculty students, PhD students, postdoctoral research fellows) focused on topics such as misinformation, elections, or COVID-19 are prioritized regarding the provision of access to the CrowdTangle interface and APIs, as well as to trainings and resources.

I attempted to apply for this access twice, in December 2022 and in February 2023. In both

cases, the result was the same. I was informed that specific research topics were prioritized and that if my research falls outside that scope, I may not be onboarded. I was advised to wait for further information.[63] Unfortunately, my request has not been approved despite the fact that I am a PhD student conducting research on disinformation and presented my research problem in detail as well as the purpose of having access to CrowdTangle since I received neither an answer nor access to CrowdTangle. I was not informed about the reasons Facebook refused my access request. However, it appears that I am not the only researcher who has been denied access to CrowdTangle's resources. Facebook in 2022 stopped accepting any new-user application due to 'staffing constraints'.[64] As of August 14, 2024, CrowdTangle is no longer available.

As to the commitment to provide vetted researchers with access to data necessary to undertake research on disinformation by developing, funding, and cooperating with an independent third-party body, Facebook declared to have actively engaged in the EDMO Working Group on Platform to Researcher Data Sharing to develop standardized processes for sharing data with researchers. Further implementation measures are planned.[65] The baseline report specifies that the technical standards and safeguards as well as standards for researchers' eligibility still need to be established.[66] It is expected that another code of conduct will be adopted for both platforms and researchers to balance the need for more transparency and research with the protection of personal data.

The third-party body has not been set up yet and it is difficult to predict when it will be done. In theory, it should have important resources, independence, and a strong mandate. However, practical questions regarding how the cooperation will evolve and whether the involvement of an independent body will reduce the arbitrariness of platforms' decisions in the context of sharing personal data remain unanswered. Moreover, in the face of such a threat of disinformation and given that the code was adopted in June 2022, the waiting time for the implementation of some of its measures is long. The code, unfortunately, still remains ineffective as regards some commitments and measures included therein.

Achievability of the Research Goals under the 2022 Code of Practice on Disinformation

To analyse whether my original research purpose would be successful under the 2022 code, the evolution of two factors, namely *discretion* of platforms and *lack of enforcement measures*, which contributed to the non-achievability of my research goal under the 2018 code, should be examined.

Discretion in Reporting of Data

Platforms' discretion is likely to be limited under the new Code in many aspects. The signatories provided assurance of its ability to be transparent as to the subscription to the commitments and their further implementation, which means that they have to explain and justify why they decided not to sign up for certain commitments if it was the case. The regular reporting was foreseen in both versions of the code – with the difference that in the new

code, the frequency of reporting has been increased in relation to the VLOPs, which allows for a more thorough scrutiny of the adopted measures. Signatories committed to participate in a permanent task force whose role is to establish a harmonized reporting template and to develop the structural indicators to measure the code's overall impact on disinformation. This is the most significant change when it comes to limiting the discretion of platforms under the 2022 code, since this is the incomplete, inconsistent, out-of-context, and selective provision of data by platforms which hampered many researches on disinformation.

The publication of the first baseline reports in 2023 showed that unfortunately not all signatories complied with the newly agreed reporting scheme. X (formerly Twitter) has been criticized for providing little specific information and no targeted data in relation to its commitments.[67] As to Facebook, there are legitimate concerns that the platform will continue to report on the adopted measures in a discretionary and incomplete manner. To give some examples from the report published by the platform, Facebook did not provide any information as regards the methodology of data measurement or data concerning the number of academic accounts granted with access to the CrowdTangle as of January 2023.[68] Yet it should do so within the framework of the implementation of measure 26.1 of the code.[69]

Provision of Access to Different Types of Data

With regard to the commitment to provide access to non-personal data and anonymized, aggregated, or manifestly made public data for research purposes on disinformation, which has been already discussed in the chapter, Facebook referred to its general policies[70] but provided little data on how they translate into access to information on disinformation, on the scope of this access, or on the kind of data researchers have access to. These two examples (although an analysis of the entire report would give rise to more) are the source of doubts as to whether my research, directed on whether the implementation of the 2022 code contributes to the achievement of its objective, would be successful, even if only because of the gaps in the report identified here.

It should, however, be pointed out that the baseline report provided by Facebook is considerably coherent as regards, for example, the provision of data per countries,[71] the lack of which constituted a significant weakness under the 2018 code. Moreover, platforms, in the new code, committed to cooperate with members of the task force as regards the consultation with researchers, the development of the third-party body, or sharing of datasets, research findings, or methodologies. On the one hand, giving independent actors a role of intermediaries, watchdogs, and cooperators with platforms in the discussed field may be a guarantee of greater transparency. On the other hand, the question as to whether each platform will engage in this cooperation with the same intensity and what consequences they will face if they do not do so at all should be asked. This is not the only question that arises as regards the compliance with the declared commitments. It is relevant to ask what consequences and which mechanism will be applicable to platforms if they do not cooperate with researchers, for example, for not providing data to those who will be vetted by the third-party body or if they do not properly fulfil their reporting obligations. It is worth to mention that in 2025 the European Commission has issued a fine of €120 million to X (Twitter) for breaching its trans-

parency obligations under the DSA. The breaches include the deceptive design of its 'blue checkmark', the lack of transparency of its advertising repository, and the failure to provide access to public data for researchers. The Commission explains that "X fails to meet its DSA obligations to provide researchers with access to the platform's public data. For instance, X's terms of service prohibit eligible researchers from independently accessing its public data, including through scraping. Moreover, X's processes for researchers' access to public data impose unnecessary barriers, effectively undermining research into several systemic risks in the European Union"[72]. It remains to be seen how this decision will impact platforms practices as regards granting researchers access to data and whether fining X will incentivise other VLOPS to ensure more transparent environment for researchers.

Enforcement Measures

Much hope lies in the enforcement measures. Nevertheless, since the code is of voluntary nature, there were none in the first or in the current version of the code. The status of the 2022 code is, however, different. As it has been already explained, the code is understood as a part of co-regulatory framework foreseen in the DSA which defines certain objectives and criteria that should be respected and applied in particular to the VLOPs. According to Article 45 of the DSA, in the case of systematic failure to comply with the codes of conduct, the European Commission and the European Board for Digital Services may invite the signatories to take the necessary action. The EU legislator has not further specified the term 'necessary action', and the rule that the commission *invite* the signatories to take necessary action suggests that the intervention of soft nature of the European Commission is foreseen. In case necessary action is not undertaken by the signatory, it should be presumed that the provisions on the non-compliance, fines, and penalties from Articles 73–79 of the DSA will apply.

Moreover, according to Article 37(b) of the DSA, the VLOPs should be subject to independent audits to assess compliance with any commitments undertaken while complying with the codes of conduct. The EU legislator specified that the audits should be effective, efficient, and timely. The term 'independent' has been defined to ensure the transparency of the auditing process and the credibility of its results. In case the outcomes of the audit are not positive, the VLOPs should take due account of the operational recommendations with a view to taking the necessary measures to implement them.[73] According to Article 37(6) of the DSA, where the signatories do not implement the operational recommendations, they shall justify in the audit implementation report the reasons for not doing so and set out any alternative measures. If the audit will not be carried out or if recommendations following the audit or the said alternative measures will not be implemented, it could be assumed that the provisions on the non-compliance, fines, and penalties from Articles 73–79 of the DSA will apply. The question arises as to what would be the consequences of partial or inaccurate implementation of recommendations in case it would, for example, hamper the researchers' access to information. On 5 May 2023, the European Commission published the draft of the delegated regulation which was aimed at setting out the necessary rules for the procedures, methodology, and templates used for the audits. The first audit reports under the DSA were published at the end of 2024. The next round of audits, including the initial evaluation of the Code's implementation, is expected to be released by the end of 2025.

The alignment of the 2022 code with the DSA provides the mechanism of the enforcement of the 2022 code. The use of binding measures from the DSA in combination with the implementation of the commitments from the code of voluntary nature focused on the specific problem of disinformation allows me to believe that my original research conducted today would have been more successful. However, it should be noted that the procedure of the enforcement of the code is delimited in a general way. It lacks the researcher-centred mechanism that would allow him, for example, to effectively contest platforms' decisions.

Conclusion

The 2018 code was the first, albeit unsuccessful, attempt to provide a legal framework to address the problem of disinformation. The lack of success of this voluntary tool lies in the very problem of disinformation and the business models of online platforms such as Facebook. Disinformation is a phenomenon difficult to grasp due to the ever-new ways and techniques of its dissemination and the rapid development of new technologies. Platforms provide the environment for its diffusion. The circulation of disinformation brings them increased traffic, which translates into higher profits. Therefore, they are reluctant to engage actively in countering this phenomenon.

In the face of this emerging complexity, the adoption of legal measures to combat disinformation is challenging. The self-regulatory solution chosen in 2018 offered the possibility of adapting the flexible measures to different operating systems of the platforms and to target the specific threats posed by the spread of disinformation. Such instrument allowed a rapid response tailored to evolving new technologies. However, this chapter revealed quite a long list of its shortcomings, which, first, did not allow to effectively combat disinformation[74] and, second, did not facilitate research on the phenomenon.

The 2022 code addresses the shortcomings and fills in the gaps of the 2018 code. The commitment to cooperate with researchers and with the third-party body, as well as other organizations, regarding the provision of data and the transparency of this process, means that the discretion of platforms in this regard could be reduced. While researchers remain dependent on platforms to obtain data, as the latter are the source, the 2022 code balances this dependency with elements such as the involvement of intermediaries, a defined reporting scheme, and enforcement measures. The alignment of the code with the DSA makes the voluntariness of the implementation of the commitments no longer unlimited, especially for the VLOPs. The binding law allows for enforceability of agreed measures and provides a defined liability scheme in case of non-compliance. Therefore, the combination of the commitments enshrined in the code focused on the specific problem of disinformation with the binding law, the DSA, is an added value. It allows for greater flexibility and provides better mechanisms of control and ensuring transparency. However, the question of whether such a combination would translate into a better quality of ongoing research on disinformation and greater research opportunities in this area should be asked. I identified some gaps as regards the tools entrusted to researchers to signal the platforms' inaction or incomplete fulfilment of commitments or to challenge their decisions on data provision. Moreover, the analysis of the first baseline report submitted by Facebook shows some incompleteness and vagueness

in the provision of data. It may be the first sign that the solutions adopted in the 2022 code aligned with the DSA are not sufficient.

It is necessary to wait for the implementation of all mechanisms to which the signatories have committed in the 2022 code. Under the 2018 code, researchers were not provided with effective instruments to access data on disinformation circulating on platforms and to analyse the measures taken by the latter to counter it. The 2022 code aligned with the DSA is an important, although for the moment still not sufficient, step towards ensuring greater data access for researchers studying disinformation.

Notes

1 Garth S. Jowett and Victoria O'Donnell, *Propaganda and Persuasion* (Sage Publications, 2012).

2 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Online Disinformation: A European Approach', COM (2018) 236 final, European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0236 (accessed 28 July 2023).

3 J. Bayer, Bernd Holznagel, Katarzyna Lubianiec, Adela Pintea, Josephine B. Schmitt, Judit Szakács, and Erik Uszkiewicz, 'Disinformation and Propaganda: Impact on the Functioning of the Rule of Law and Democratic Processes in the EU and Its Member States', European Parliament, 2021, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653633/EXPO_STU(2021)653633_EN.pdf (accessed 28 July 2023).

4 'Communication from the Commission'.

5 'Fake News and Disinformation Online', European Commission, March 2018, https://europa.eu/eurobarometer/surveys/detail/2183 (accessed 28 July 2023).

6 'Communication from the Commission'.

7 Despite the name change of the platform in 2021 to Meta, the name Facebook is used throughout this chapter.

8 Nic Newman, Richard Fletcher, Craig T. Robertson, Kirsten Eddy, and Rasmus Kleis Nielsen, *Reuters Institute Digital News Report 2022* (Reuters Institute, University of Oxford, 2022).

9 Mark Travers, 'Facebook Spreads Fake News Faster than Any Other Social Website, According to New Research', *Forbes*, 21 May 2020, https:// www.forbes.com/sites/traversmark/2020/03/21/facebook-spreadsfake-news-faster-than-any-other-social-website-according-to-newresearch/?sh=-3deca4b56e1a (accessed 24 February 2023).

10 'Communication from the Commission'.

11 'Communication from the Commission'.

12 Linda Senden, 'Soft Law, Self-Regulation and Co-Regulation in European Law: Where Do They Meet?' *Electronic Journal of Comparative Law* no. 9 (2005): 1–27, 24.

13 Senden, 'Soft Law, Self-Regulation and Co-Regulation', 1, 23.

14 Iva Plasilova, Jordan Hill, Malin Carlberg, Marion Goubet, and Richard Procee*, Study for Assessment of the Implementation of the Code of Practice on Disinformation Final Report*, SMART 2019/0041, 2019, https:// www.imap-migration.org/sites/default/files/Publications/2020-07/ Study-fortheassessmentofthecodeofpracticeagainstdisinformation.pdf (accessed 8 February 2023).

15 Code of Practice on Disinformation, 2018, https://ec.europa.eu/newsroom/ dae/redirection/document/87534 (accessed 18 December 2022).

16 The partnerships concluded with academics reported by Google, Meta, X (now Twitter), Mozilla, and Microsoft consisted of offering the training of fact-checkers, making available of datasets, launching campaigns on transparency, or building the infrastructures to provider researchers with access to non-personally identifiable data. See 'Analysis Code of Practice Annual Report', 11–12, https://ec.europa.eu/newsroom/dae/document. cfm?doc_id=62698 (accessed 28 July 2023).

17 The principal purpose of the code was to identify the actions that signatories could put in place in order to address the challenges related to 'disinformation'.

18 See 'Staff Working Document', SWD (2020)180 final, 2020, European Commission, https://ec.europa.eu/newsroom/dae/document.cfm?doc_ id=69212 (accessed 16 February 2023).

19 'Staff Working Document'. For example, the platform provided data linked to the restriction of misleading advertising, unsupported commercial claims, or deceptive business practices which were not related to the policies against disinformation.

20 The signatories recognized specific goals which were expressed under the form of commitments and divided into five groups and related to five following areas: (*a*) scrutiny of ad placements, (*b*) political advertising and issue-based advertising, (*c*) integrity of services, (*d*) empowering consumers, and (*e*) empowering the research community.

21 In the report on September and October 2021 actions, 'Fighting COVID-19 Disinformation, as Regards Supporting Media Literacy in Europe', Facebook reported the results of the 'Together Against Covid-19 Misinformation' campaign. See 'Reports on September and October Actions: Fighting COVID-19 Disinformation Monitoring Programme', European Commission, 2 December 2021, https://digital-strategy. ec.europa.eu/en/library/reports-september-and-october-actions-fighting-covid-19-disinformation-monitoring-programme (accessed 20 February 2023). In the report on November and December 2021 actions, 'Fighting COVID-19 Disinformation', neither this action nor its results are mentioned so the assessment of the effectiveness of this action in the long term is not possible. See 'Fighting COVID-19 Disinformation: Reports on November and December Actions', European Commission, 27 January 2022, https:// digital-strategy.ec.europa.eu/en/library/fighting-covid-19-disinformationreports-november-and-december-actions (accessed 20 February 2023).

22 Trisha Meyer, Alexandre Alaphilippe, and Claire Pershan, 'The Good, the Bad and the Ugly: How Platforms are Prioritising Some EU Member States in Their COVID-19 Disinformation Responses', EU DisinfoLab, 28 April 2021, https://www.disinfo.eu/publications/the-good-the-bad-and-the-ugly-how-platforms-are-prioritising-some-eu-member-states-in-their-covid-19disinformation-responses (accessed 15 May 2023).

23 Meyer, Alaphilippe, and Pershan, 'The Good, the Bad and the Ugly'.

24 Meyer, Alaphilippe, and Pershan, 'The Good, the Bad and the Ugly'.

25 Plasilova, Hill, Carlberg, Goubet, and Procee, *Study for Assessment*.

26 Plasilova, Hill, Carlberg, Goubet, and Procee, *Study for Assessment*.

27 See, for example, 'Code of Practice against Disinformation: Commission Takes Note of the Progress Made by Online Platforms and Urges Them to Step Up Their Efforts', European Commission, 20 March 2019, https:// ec.europa.eu/commission/presscorner/detail/en/STATEMENT_19_1757 (accessed 17 February 2023).

28 'Staff Working Document'.

29 The new code relates to the following areas: (*a*) scrutiny of ad placements, (*b*) political advertising, (*c*) integrity of services, (*d*) empowering users, (*e*) empowering the research community, (*f*) empowering the fact-checking community, (*g*) transparency centre, (*h*) permanent task force, and (*i*) monitoring of the code. See *The Strengthened Code of Practice on Disinformation 2022*, https:// ec.europa.u/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

30 'Commission Presents Guidance to Strengthen the Code of Practice on Disinformation', 26 May 2021, European Commission, https://ec.europa. eu/commission/presscorner/detail/en/ip_21_2585 (accessed 15 December 2024).

31 'Guidance on Strengthening the Code of Practice on Disinformation', European Commission, 26 May 2021, https://digital-strategy.ec.europa.eu/ en/library/guidance-strengthening-code-practice-disinformation (accessed 28 July 2023).

32 'Signatories of the 2022 Strengthened Code of Practice on Disinformation', European Commission, 16 June 2022, https://digital-strategy.ec.europa. eu/en/library/signatories-2022-strengthened-code-practice-disinformation (accessed 28 July 2023).

33 'The Digital Services Act: Ensuring a Safe and Accountable Online Environment', European Commission, https://commission.europa.eu/ strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digitalservices-act-ensuring-safe-and-accountable-online-environment_en (accessed 28 July 2023).

34 'VI. Empowering the Research Community', in *The Strengthened Code of Practice on Disinformation 2022*, 26–30, https://ec.europa.eu/newsroom/ dae/redirection/document/87585 (accessed 14 February 2023).

35 The independent procedure of vetting researchers by a Digital Services Coordinator to allow them access to data by VLOPs is provided in Article 40 of the DSA. The access to data awarded within the framework of Article 40 of the DSA is of general nature and not limited to data on disinformation.

36 'Commitment 28', in *The Strengthened Code of Practice on Disinformation 2022*, https://ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

37 'Commitment 29', in *The Strengthened Code of Practice on Disinformation 2022*, https://ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

38 'Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access', European Digital Media Observatory, 31

May 2022, https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-onPlatform-to-Researcher-Data-Access-2022.pdf (accessed 13 February 2023).

39 To compare, an EU legislator in Article 40 of the DSA foresaw the establishment of the Digital Services Coordinator to assess compliance of VLOPs with the DSA, to vet researchers, and to intermediate as regards the provision of data by VLOPs to researchers. Many similarities can be noticed between its role and the role of third-party body within the 2022 Code of Practice on Disinformation. However, the scope of the third-party body is shaped narrowly and relates only to the disinformation's issues.

40 Online platforms and online search engines whose number of average monthly active recipients of the service in the EU is equal to or higher than 45 million. See Article 33 of DSA. In this chapter, I refer only to VLOPs and omit the very large search engines (VLSEs). See '#FindYourVLOP', https://docs.google.com/spreadsheets/d/1H89uABJZCg0BQlUdpDPE0XBpdtX WPGQbwLW4Ug_hmNo/edit#gid=1177757099 (accessed 25 February 2023).

41 Point 'h' of the preamble to *The Strengthened Code of Practice on Disinformation 2022*, https://ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

42 Point 'h' of the preamble to *The Strengthened Code of Practice on Disinformation 2022*.

43 See, for example, 'ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice', European Regulators Group for Audiovisual Media Services, 2020, https://erga-online.eu/wp-content/ uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf (accessed 13 February 2023).

44 Article 37(1) of the DSA.

45 Six months after the code's signature which took place on the 16 June 2022.

46 'Commitment 40', in *The Strengthened Code of Practice on Disinformation 2022*, https://ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

47 See 'The Digital Services Act'.

48 'Commitment 41', in *The Strengthened Code of Practice on Disinformation 2022*, https://ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

49 'Measures 37.2', in *The Strengthened Code of Practice on Disinformation 2022*, https://

ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

50 'Commitment 43', in *The Strengthened Code of Practice on Disinformation 2022*, https://ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

51 Transparency Centre, https://disinfocode.eu (accessed 14 February 2023).

52 'Signatories of the Code of Practice on Disinformation Deliver Their First Baseline Reports in the Transparency Centre', European Commission, 9 February 2023, https://digital-strategy.ec.europa.eu/en/news/signatoriescode-practice-disinformation-deliver-their-first-baseline-reportstransparency-centre (accessed 14 February 2023).

53 'Community Standards Enforcement Report', Meta, https://transparency. fb.com/data/community-standards-enforcement (accessed 16 February 2023).

54 'Widely Viewed Content Report', Meta, https://transparency.fb.com/pl-pl/ data/widely-viewed-content-report (accessed 16 February 2023).

55 'Meta's Adversarial Threat Report', Meta, November 2022, https://about.fb.com/news/2022/11/metas-adversarial-threat-report-q3-2022 (accessed 16 February 2023).

56 'Code of Practice on Disinformation: Meta Baseline Report', Meta, January 2023, https://disinfocode.eu/reports-archive/?years=2023 (accessed 16 February 2023).

57 'Measure 26.2', in *The Strengthened Code of Practice on Disinformation 2022*, https://ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

58 'Code of Practice on Disinformation: Meta Baseline Report'.

59 Seth Smaley, 'Meta Won't Comment on Its Plans to Abandon CrowdTangle', Poynter, 18 August 2022, https://www.poynter.org/reporting-editing/2022/ meta-wont-comment-on-its-plans-to-abandon-crowdtangle (accessed 14 February 2023); Maxime Mohr, 'Meta va éteindre CrowdTangle, son outil de mesure des interactions, SiecleDigital', SiecleDigital, 29 June 2022, https:// siecledigital.fr/2022/06/29/meta-va-eteindre-crowdtangle-son-outil-demesure-des-interactions (accessed 14 February 2023).

60 Gemma B. Mendoza, 'Why Possible Loss of CrowdTangle Worries FactCheckers and Disinformation Researchers', Rappler, 11 July 2020, https:// www.rappler.com/technology/social-media/disinformation-crowdtangledata-access (accessed 16 February 2023).

61 John Albert, 'Facebook's Gutting of CrowdTangle: A Step Backward for Platform Transparency', AlgorithmWatch, 3 August 2022, https:// algorithmwatch.org/en/crowdtangle-platform-transparency (accessed 15 February 2023).

62 Christina Fan, 'CrowdTangle for Academics and Researchers', CrowdTangle, 2023, https://help.crowdtangle.com/en/articles/4302208-crowdtangle-foracademics-and-researchers (accessed 15 February 2023).

63 The exact phrasing of Facebook's response was as follows: 'Thank you for submitting an application for CrowdTangle access. Please note that we are prioritizing specific research topics as noted in the application form, and may not be able to onboard you if your research falls outside that scope. If we are able to onboard you, we will be in touch soon. Thanks, the CrowdTangleTeam.'

64 'Meta Pauses New Users from Joining Analytics Tool CrowdTangle', Reuters, 29 January 2022, https://www.reuters.com/technology/meta-pauses-newusers-joining-analytics-tool-crowdtangle-2022-01-29 (accessed 17 May 2023).

65 'Code of Practice on Disinformation: Meta Baseline Report'.

66 'Code of Practice on Disinformation: Meta Baseline Report'.

67 'Code of Practice on Disinformation: New Transparency Centre Provides Insights and Data

on Online Disinformation for the First Time', European Commission, 9 February 2023, https://digital-strategy.ec.europa.eu/en/ news/code-practice-disinformation-new-transparency-centre-provides-insights-and-data-online (accessed 14 February 2023).

68 Facebook did not provide data on the number of monthly users, number of received applications, number of applications rejected, and so on. 'Code of Practice on Disinformation: Meta Baseline Report'.

69 'Measure 26.1', in *The Strengthened Code of Practice on Disinformation 2022*, https://ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed 14 February 2023).

70 'Code of Practice on Disinformation: Meta Baseline Report'.

71 As regards measure 18.2 of the 2022 code, which concerns the development and enforcement of publicly documented, proportionate policies to limit the spread of harmful, false, or misleading information. Facebook provided information on contents removed for violating the 'harmful health misinformation' or voter or census interference policies as per each member state and then as per EU as a whole. 'Code of Practice on Disinformation: Meta Baseline Report'.

72 European Commisson, https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2934, accessed: 10.12.2025.

73 Article 37(6) of the DSA.

74 'Left Behind: How Facebook is Neglecting Europe's Infodemic', Avaaz, 20 April 2021, https://secure.avaaz.org/campaign/en/facebook_neglect_ europe_infodemic (accessed 19 May 2023).

ABOUT THE CONTRIBUTORS

Carolina Aguerre is Associate Professor in the Department of Humanities and Communication, Universidad Católica del Uruguay (UCU), Montevideo (Uruguay), and honorary co-director at the Centre for Technology and Society (CETYS), Universidad de San Andrés, Victoria, Buenos Aires. Her research interests include theories and practices around the governance of communications technologies and infrastructures, including the internet and artificial intelligence (AI), the intersection with political economy, and north–south perspectives. She edited the volume *Digital Data Governance: Polycentric Perspectives* (with Malcolm Campbell-Verduyn and J. A. Scholte) (2024). In 2020, she was part of the United Nations Educational, Scientific and Cultural Organization's (UNESCO) ad hoc expert working group on the Recommendations on the Ethics of AI. She was a resident fellow at the CGR21 (2020–2021), University of Duisburg-Essen. She has a PhD in Social Sciences from the University of Buenos Aires.

Frank Kwaku Agyei is a scholar of environmental justice and rural wellbeing. He studies the social and political–economic causes of the precarity and suffering of natural-resource-dependent communities. His work is crucial for developing policies and practices that promote justice and sustain ability for such communities, ensuring their well-being and resilience in the face of socioeconomic and environmental challenges.

Pedro Amaral is a PhD candidate in Sociology and a researcher at the Crime and Security Policies Center (NEPS), Federal University of Pernambuco, Recife. Member of the Surveillance in the Majority World Research Network and the Surveillance Studies Network. Formerly, worked at the Digital Rights Secretariat, Ministry of Justice and Public Security, Brazil. He was a former researcher and project leader at Law and Technology Research Institute of Recife (IP.rec).

Jef Ausloos is Assistant Professor at the Institute for Information Law (IViR), University of Amsterdam. His work centres around various information law issues – specifically data rights, transparency, and governance of digital infrastructures – and the broader political economy in which they operate. His research can be situated at the intersection of law, critical data studies, and the politics of knowledge production.

Reuben Binns is Associate Professor of Human Centred Computing, working between computer science, law, and philosophy, focusing on data protection, machine learning (ML), and the regulation of and by technology. Between 2018 and 2020, he was a postdoctoral research fellow in artificial intelligence (AI) in the Information Commissioner's Office, addressing AI/ ML and data protection. He joined the Department of Computer Science, University of Oxford, as a postdoctoral researcher in 2015. He received his PhD in Web Science from the University of Southampton in 2015.

Lawrence Kwabena Brobbey is Lecturer in the Department of Silviculture and Forest Management, Kwame Nkrumah University of Science and Technology (KNUST), Kumasi (Ghana). He combines academic work with practice and has worked with government agencies and international non-governmental organizations (NGOs). His research spans natural resources tenure and access, environmental justice, and commodity chain analysis.

Siddharth Peter de Souza is Assistant Professor in AI and Society at the Centre for Interdisciplinary Methodologies, University of Warwick, UK. His research looks at developments in law and technology from a legal pluralist, data justice, and decolonial perspective. He is the founder of Justice Adda, a law and design social venture which seeks to build legal literacy and awareness in India.

Paul Esselaar is a practising attorney and notary in Cape Town, South Africa, with over 20 years' experience in advising commercial clients. He is an appeal adjudicator for two telecommunications industry bodies within South Africa which focus on implementing voluntary codes of conduct. He has been appointed as Honorary Research Fellow at the University of KwaZulu-Natal, where he is pursuing his PhD. On the publications side, he is a co-author of several articles on data protection as well as two books, including *Overthinking the Protection of Personal Information Act* (2021). More recently, he is a co-author of the Model Law on Health Data Governance (2024), which seeks to introduce health data governance into local legislation worldwide.

Michalina Kowala is an assistant professor at the Faculty of Law and Administration of Adam Mickiewicz University of Poznań in Poland and a lawyer in the Freedom of Expression Program at the Helsinki Foundation for Human Rights.

In 2024 she defended her doctoral thesis 'Publishers' Rights and Copyright Law. Safeguarding Access to Information and Media Pluralism' published in 2025 by Routledge.

Awarded the French government's scholarship, she conducted her research on protection of press sector in cooperation with L'Institut de Recherche en Droit Privé, Nantes.

She was also a visiting researcher at the Centre for IT and IP Law in Leuven, the Max Planck Institute for Innovation and Competition in Munich and the Center for International Intellectual Property Studies, Strasbourg.

She was associated, in a professional capacity, with the European Parliament in Brussels and the Commissioner for Human Rights Office in Warsaw. In 2025 she worked for the Polish Presidency in the Council of the EU.

Boateng Kyereh is Professor in the Department of Silviculture and Forest Management, Kwame Nkrumah University of Science and Technology (KNUST), Kumasi (Ghana), and a natural resource management expert. He provides scientific inputs into policy debates on critical issues in natural resource management and tries to bridge the gap between academia and civil society. He also serves society by contributing to capacity building of stakeholders at both formal and non-formal levels for improved decision-making in natural resource governance and management in Ghana.

Matteo Nebbiai is a PhD student in the Department of Political Economy, King's College London. He focuses on European Union (EU) politics, politics of digital regulation, and lobbying. He has been a Global Political Economy Project (GPEP) affiliate fellow at Georgetown

University, Washington, DC, and holds both a Master of Arts (MA) degree and a Bachelor of Arts (BA) degree in Political Science from Scuola Superiore Sant'Anna and the University of Florence. Alongside his academic career, he has contributed to research projects for the European Commission and the European Parliament. Before pursuing his PhD, he worked as a digital policy analyst for the St. Gallen Endowment for Prosperity through Trade.

Midas Nouwens is Associate Professor in the Department of Digital Design and Information Studies, Aarhus University. His research explores the power dynamics inherent in informational capitalism and the potential of countervailing forces to challenge these structures. This includes examining academia's role in providing evidence for journalists and regulators, fostering digitally empowered citizens through higher education, and leveraging software artefacts as tools for grassroots, adversarial action. He also investigates the strategic and practical challenges faced by regulators as normative institutions, as well as the use of legal mechanisms like rights enforcement and strategic litigation. His primary thematic focus lies on online tracking and algorithmic systems.

Paul Osei-Tutu is Lecturer in the Department of Forest Resources Technology, Kwame Nkrumah University of Science and Technology (KNUST), Kumasi (Ghana). His areas of expertise are local/decentralized natural resource management, formal and informal institutions of local forest management, small forest enterprises, bamboo resource management and utilization, and natural resource conflicts.

Marcos César M. Pereira is Research Coordinator at the Law and Technology Research Institute of Recife (IP.rec), specializing in cryptography, government hacking, and children's rights. He has a master's degree in Anthropology and a bachelor's degree in Social Sciences from the Federal University of Pernambuco, Recife. In 2022, he was a fellow of the Leaders 2.0 programme of the Latin America and Caribbean Network Information Centre (LACNIC), Montevideo (Uruguay).

André Ramiro is Fellow at the Digital Civil Society Lab, Stanford Center on Philanthropy and Civil Society (PACS), California, and a PhD candidate in Informational Law at University of Hamburg. He was a former fellow at the Alexander von Humboldt Institute for Internet and Society, Berlin, and at the Weizenbaum Institute, Berlin; the former director and the co-founder of the Law and Technology Research Institute of Recife (IP.rec); a member of the Latin American Network of Surveillance, Technology and Society Studies (Lavits).

Jake Stein is a researcher, designer, and developer. He is currently pursuing PhD in Human Centered AI in the Department of Computer Science, University of Oxford, while working on the Oxford Martin School's 'Ethical Web and Data Infrastructure in the Age of AI' project under Nigel Shadboldt and Max Van Kleek. He received his Bachelor of Arts (BA) degree from Yale University, focusing on digital media critical theory, as well as a Master of Science (MSc) degree from the Oxford Internet Institute, University of Oxford, and worked as a product researcher in San Francisco.

Theory on Demand #61

The Many Faces of Data Access: Legal and Policy Implications for Research
Edited by: Jef Ausloos and Siddharth Peter de Souza

The Many Faces of Data Access: Legal and Policy Implications for Research provides a rich and interdisciplinary critique of regulation and, in the process, opens the 'black box' of technology companies to researchers. It brings together scholars from across the globe, working in varied fields including critical legal studies, science and technology studies, critical data studies, and digital humanities. The book explores questions of data access – to acquire and use data meaningfully as well as resist power. It covers a variety of themes, including the opportunities and challenges of the law as a tool for observing digital infrastructures, the political economy of data access for research, and the power dynamics between academia, private/public sector, and civil society. In doing so, the book also examines these questions in terms of the politics of knowledge production and investigates whether there is a privileging of geographical and institutional contexts in data access regimes.

Jef Ausloos is Assistant Professor at the Institute for Information Law, University of Amsterdam.

Siddharth Peter de Souza is Assistant Professor of AI and Society at the Centre for Interdisciplinary Methodologies, University of Warwick.

Printed on demand
ISBN: 9789083672113

Institute of network cultures