Addressing Big Data and AI Challenges: A Taxonomy and Why the GDPR Cannot Provide a One-size-fits-all Solution

Dr. Maria Tzanou, Senior Lecturer in Law, Keele University, UK

m.tzanou@keele.ac.uk

Abstract

The chapter challenges the assumption that data privacy frameworks in general and the GDPR in particular can provide an appropriate regulatory solution for big data. It argues that in order to be able to properly reflect on regulatory approaches that grasp with big data challenges, closer attention should be paid to these particular challenges. In this respect, this chapter makes three distinct contributions to the debate regarding regulatory approaches to big data: First, it develops a taxonomy of big data challenges that allows a comprehensive overview of the issues at stake. Second, it examines the capabilities and limitations of the GDPR to address the risks identified in the proposed taxonomy. Third, it offers some suggestions on the pathways that regulators should be considering when approaching big data and AI.

1. Introduction

Big data analysed through algorithms are considered the oil of the modern online economy. The value of big data is immense, but they also pose significant challenges to individuals and the society as a whole. It is a common assumption that big data, data mining, algorithmic decision-making and Artificial Intelligence (AI) should be considered as data privacy issues[1]

[1] For instance, Yeung notes: 'The right most clearly implicated by Big Data driven hypernudging is the right to informational privacy, given the continuous monitoring of individuals and the collection and algorithmic processing of personal digital data'. Karen Yeung, '"Hypernudge": Big Data as a mode of regulation by design', (2017) 20 (1) *Information, Communication & Society*, 118, 124.

that should be addressed by data privacy laws. In this respect, the EU's General Data Protection Regulation (hereinafter the GDPR)[2] is considered as the central legislative measure aimed at addressing big data risks[3] and has been hailed as 'the first piece of legislation for AI' that processes personal data.[4]

This chapter challenges the assumption that data privacy frameworks in general and the GDPR in particular can provide an appropriate regulatory solution for big data.[5] It argues that in order to be able to properly reflect on regulatory approaches that grasp with big data challenges, closer attention should be paid to these particular challenges. Searching for appropriate regulatory solutions requires a focus on the problems that need to be addressed. While big data challenges have been extensively discussed in the literature; remarkably, there is a lack of systemisation of big data risks that would permit a reflection about potential regulatory responses to these which goes beyond the commonly proposed data privacy frameworks.

In this respect, this chapter makes three distinct contributions to the debate regarding regulatory approaches to big data: First, it develops a taxonomy of big data challenges that allows a comprehensive overview of the issues at stake. Second, it examines the capabilities and limitations of the GDPR to address the risks identified in the proposed taxonomy. Third, it

[2] 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), Official Journal of the European Union , L 119/1, 4 May 2016.
[3] EDPS, Opinion 7/2015 *Meeting the challenges of big data- A call for transparency, user control, data protection by design and accountability.*
[4] Paul Nemitz, 'Constitutional democracy and technology in the age of artificial intelligence', *Philosophical Transactions. R. Soc.* (2018): 1. The European Commission notes that 'the rules laid down in the [General Data Protection] Regulation provide a general framework and contain specific obligations and rights that are particularly relevant for the processing of personal data in AI'. See Commission, Communication from the Commission to the European Parliament and the Council, Data protection rules as a trust-enabler in the EU and beyond – taking stock, Brussels, 24.7.2019, COM(2019) 374 final.
[5] For instance, the EDPS states: 'The question is not whether to apply data protection law to big data, but rather how to apply it innovatively in new environments.' EDPS, Opinion 7/2015, n 3, 4.

offers some suggestions on the pathways that regulators should be considering when approaching big data and AI.

The chapter is organised as follows: Section 2 proposes a typology of big data and AI challenges. This categorises big data challenges at two levels: *individual* and *societal*. *Individual* level challenges are classified in two further dimensions: those that relate to the *processing* of the data and those that are linked to the *outcomes* of such processing. The taxonomy of big data challenges is necessary if we are to seriously consider regulatory methodologies and analytical prisms to approach these. Rather than relying on the simplistic assumption -which even the GDPR adopts to some extent- that data protection law should catch in its protective net all the issues that big data and AI raise because it applies to the processing of personal data, the proposed taxonomy reveals that big data present challenges -both at the individual and at the societal level- are far from the scope or irrelevant to data protection frameworks due to their complex dynamics, actors, processes and outcomes.

In the light of the taxonomy developed, Section 3 examines the role that the GDPR can play to address big data and AI challenges. It finds that this legal instrument can regulate individual big data risks that relate to the *processing* of personal data, despite the scepticisms raised in this respect. The GDPR can also potentially play an indirect role in addressing individual level challenges that occur as *outcomes* of the processing and *societal* level threats, but this is very limited and clearly not enough taking into account the issues arising in the big data age. Section 4 considers the regulatory pathways and methodologies to address big data issues. Section 5 offers brief Conclusions.

2. Mapping the Challenges of Big Data: A Taxonomy

2.1 Preliminary issues and caveats

Before we proceed with the analysis, a number of caveats and clarifications is required. A first question that arises when attempting to map the challenges of big data is a definitional one. What is meant by big data 'challenges'? Is it the same as speaking of big data 'risks' or 'threats'? Admittedly, the concepts of 'challenges', 'concerns', 'risks' and 'threats' appear semantically similar; nevertheless, this chapter opts to adopt the 'challenges' or 'concerns' terminology to broadly denote the different legal -and to an extent ethical- 'issues' that big data and AI raise.

This is a deliberate choice: I approach big data issues in a neutral way, that steers away from the GDPR's 'risks' terminology. More particularly, it is well-known that the GDPR marks the introduction of a risk-based approach to data protection.[6] This is pursued mainly by data protection impact assessments (DPIAs)[7] that controllers must undertake 'where a type of processing in particular using new technologies, …is likely to result in a high risk to the rights and freedoms of natural persons',[8] and *ex post* notifications of data breaches to the supervisory authorities and the data subject[9] when they are 'likely to result in a high risk to the rights and freedoms of natural persons'.[10] Recital 75 explains in this respect that risks 'of varying likelihood and severity may result from personal data processing' and could lead to 'physical, material or non-material damage' and provides an indicative list of these. The 'risk' of processing, therefore, acquires a specific normative significance under the GDPR that triggers several obligations for controllers. It should be clarified from the outset that the taxonomy

[6] See Maria Eduarda Gonçalves, 'The EU data protection reform and the challenges of big data: remaining uncertainties and ways forward' (2017) 26 (2) *Information & Communications Technology Law*, 90, 99.
[7] Articles 35, 36 GDPR and Recital 76. For an ethical and social impact assessment see Alessandro Mantelero, 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment' (2018) 34 *Computer Law & Security Review*, 754.
[8] Article 35 (1) GDPR.
[9] Articles 33 and 34 GDPR.
[10] Article 34 (1) GDPR.

developed in this chapter understands big data 'issues' and 'challenges' in a way that does not attribute to these any normative value and is not linked to the meaning of 'risks' and 'risk assessments' under the GDPR.

Secondly, big data challenges have been extensively discussed in the academic literature and beyond.[11] The present analysis builds on this literature but aims to move the discussion a step further by providing a classification of these. The classification is proposed as a methodological attempt to systematise these challenges at two *levels* (*individual* and *societal*) and two further *dimensions* at the *individual* level (challenges associated to the *processing* of personal data and challenges associated to the *outcomes* of such processing). The specific challenges falling under each category are depicted in Table 1 and analysed in detail below. The taxonomy provides a useful tool to structure the discussion on big data issues and more importantly to organise the debate on the appropriateness of regulatory approaches to big data both current (the GDPR) and future. In this respect, this chapter fills an important gap in the literature on regulatory responses to big data.

Thirdly, the chapter uses various examples from big health data, but avoids limiting the discussion only to health data for two main reasons. The first reason concerns the problems around defining health data. As seen in Chapter 1[12] and other chapters in this book, it is very difficult to distinguish sensitive health data from apparently innocuous information (like a super-market list) that might reveal important details about a person's diet and therefore health. Furthermore, big data have numerous lifecycles and different datasets are aggregated at different time points, so it is difficult to specify when exactly 'health data' are analysed. Second, the proposed taxonomy applies to big health data analytics -seen here as an application

[11] EDPS, Opinion 7/2015, n 3.
[12] See Maria Tzanou, The GDPR and (Big) Health Data: Assessing the EU Legislator's Choices.

domain of big data analytics- but can be generalised to cover big data issues more broadly. The overall discussion, therefore, adopts a broader scope in this respect.

Finally, this contribution follows the definitions of 'big data', 'AI', 'big data analytics' and 'health data' as provided in the first chapter of this book.[13]

<INSERT TABLE 7.1 HERE>

2.2 *Individual* level issues associated to the *processing* of the data

Big data present a number of challenges at the individual level relating to the processing of the data. These are identified and discussed below.

1. Personal data and personally identifiable information

It is normally assumed that the added value of big data is found in its capability to identify novel patterns[14] rather than focus on personal identified information. Indeed, many significant medical trends and conclusions drawn in recent years are based on statistical analysis of huge databases that do not include personal information.[15] Yet, two points should be raised here. Firstly, big data might contain personal information and secondly, it has been shown[16] that modern data analytics can allow for the re-identification of individuals even if seemingly only anonymised data is processed.

Personal data is defined in the GDPR as 'any information relating to an identified or identifiable natural person'; an identifiable natural person 'is one who can be identified,

[13] Ibid.
[14] Viktor Mayer-Schonberger and Yann Padova, 'Regime Change? Enabling Big Data Through Europe's New Data Protection Regulation', (2016) *Colum. Sci. & Tech. L. Rev*, 315, 319.
[15] See Tal Zarsky, 'Incompatible: The GDPR in the Age of Big Data' (2017) 47 *Seton Hall Law Review* 995, 1000.
[16] See Ira Rubinstein and Woodrow Hartzog, 'Anonymization and Risk' (2016) 91 *Washington Law Review* 703, 710-11.

directly or *indirectly*, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.'[17] As processing of personal data triggers the application of data protection law, it is very important to establish when big data is composed of or entails personal information, i.e information that can be linked to a certain individual. In this case, data protection law would apply and govern the processing of the data.

A distinct problem arises when no personal information is contained in the big data, yet there are ways to identify the persons concerned. In such cases, EU data protection legislation is clear. As soon as the data refers to a person who can be identified 'directly' or 'indirectly', this is considered personal data and therefore data protection law applies. In *Breyer,*[18] the CJEU adopted a broad interpretation of 'identifiable person' holding that in order to ascertain this, 'account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person',[19] without requiring that 'all the information enabling the identification of the data subject must be in the hands of one person.'[20] In legal terms, this means that the possibility of re-identification of a person within the context of big data would make data protection safeguards and in particular the GDPR applicable.

2. Legal basis for the processing and the problems and limitations of consent

The literature has identified different ways in which big data can be generated. The LIBE Report on Big Data and Smart Devices[21] distinguishes three categories of generating data

[17] Article 4 (1) GDPR.
[18] Case C-582/14, *Patrick Breyer v. Bundesrepublik Deutschland* [2016] ECLI:EU:C:2016:779. For case-note see Frederik Zuiderveen Borgesius, 'Breyer Case of the Court of Justice of the European Union: IP Addresses and the Personal Data Definition' (2017) 3 (1) *European Data Protection Law Review*, 130.
[19] Ibid, para 42.
[20] Ibid, para 43.
[21] Gloria González Fuster and Amandine Scherrer, 'Big Data and smart devices and their impact on privacy', (2015) Study for the LIBE Committee.

sources: a) volunteered data, that is information volunteered or surrendered by individuals when they 'explicitly share information about themselves or about third parties' (e.g., when someone shares health-related information about herself or other people, such as their children on online social networks or enters information on a health app that predicts ovulation cycles); b) observed data, that arises by observing the activities of users (e.g., Internet browsing habits, GPS and location data when using mobile phones); and, c) inferred data that arise from the result of the automated analysis of other data (e.g., health insurance premiums calculated on the basis of a number of factors and data volunteered or observed).[22] Based on this classification, it can be argued that the processing of different types of personal data might be based on different legal bases.

Consent is a common legal basis allowing for the processing of volunteered, observed (and inferred) personal data in social media, online apps and wearable devices. The GDPR strengthens consent and requires the informed consent of the data subject[23] in order for processing to take place, but there are several problems regarding consent in the context of big data. Drawing on the existing academic literature,[24] these problems can be summarised in two groups: *data subject-related* problems and *systemic* problems. Data subject- related problems refer to concerns about whether individuals sharing their personal information on social media read privacy policies,[25] whether even if they read them they actually understand these in an meaningful way as well as the consequences of sharing personal data and whether even if they

[22] Ibid, 21. Schneier has developed a taxonomy of personal data, on the basis of social media that includes six categories: service data, which is provided to open an account (e.g., name, address, credit card information, etc.); disclosed data, which is entered voluntarily by the user; entrusted data, for example the comments made on other people's entries; incidental data, which refers to a specific user, but is uploaded by someone else; behavioural data, which contains information about the actions users are undertaking when using the site and may be used for targeted advertising; and inferred data, which is information deduced from someone's disclosed data, profile or activities.
[23] GDPR, Art. 7.
[24] See Solove, 'Introduction: Privacy self-management and the consent dilemma' (2013) 126 *Harvard Law Review* 1880.
[25] Aleecia Mcdonald and Lorrie Cranor, 'The Cost of Reading Privacy Policies' (2008) 4(3) *I/S: A Journal of Law and Policy for the Information Society*, 543.

read privacy policies and are indeed interested in protecting their privacy online nevertheless they normally tend to grant their consent almost immediately to proceed receiving the (often free) service/ benefit.[26] While arguably these concerns apply to consent for the processing of personal data in general, the specific attributes of big data[27] present further *systemic* problems regarding consent.[28] In the context of big data, personal information may be processed multiple times in different 'lifecycles' involving different controllers and processors in each 'lifecycle'.[29] The argument, therefore, goes that even if individuals gave their consent for the initial processing of their data, it is impossible to predict –let alone meaningfully consent- to the subsequent virtually limitless possibilities that these can be processed and aggregated. Furthermore, the information imbalances between data subjects and data controllers such as Facebook and Google in the context of big data may undermine the overall validity of the data subject's consent.[30]

Besides consent, the GDPR provides for other grounds for the processing of personal data[31] and the interpretation of certain grounds, such as processing necessary for the performance of a contract (Article 6 (1) (b) GDPR) and processing necessary for the purposes of the legitimate interests of the controller or a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject (Article 6 (1) (f) GDPR).

---

[26] See José van Dijck, 'Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology', (2014) 12(2) *Surveillance & Society*, 199; Nina Gerber, Paul Gerber and Melanie Volkamer, 'Explaining the Privacy Paradox: A Systematic Review of Literature Investigating Privacy Attitude and Behavior', (2018) 77 *Computers and Security* 226.

[27] See Tal Zarsky, 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making' (2016) 41 (1) Science, Technology, & Human Values 118.

[28] Art. 7 (4) GDPR provides 'when assessing whether consent is freely given, utmost account shall be taken of whether, *inter alia*, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.'

[29] See Tzanou, Introduction, n 12.

[30] See Orla Lynskey, *The Foundations of EU Data Protection Law* (Oxford, Oxford University Press 2015), 189 and references therein.

[31] GDPR, Article 6.

3. Implementation and enforceability of fair and transparent processing

Article 5 GDPR contains a number of data protection principles pertaining to the processing of personal data: lawfulness, fairness and transparency of processing, purpose limitation, data minimisation, accuracy and storage limitation. The Regulation requires that the controller who processes personal data is made 'responsible for' and must be 'able to demonstrate compliance' with these principles.[32] It has been argued that these principles are categorically 'incompatible' with big data practices.[33] Similarly to the case of consent discussed above, the alleged challenges or incompatibilities can be grouped in two categories: *systemic* and *actor-related related*.

Starting from the principle enshrined in 5 (1) (a), 'lawful' processing means that it should be carried out in accordance with the law.[34] Fairness in processing denotes that data controllers must ensure that the collection and further processing of personal data is undertaken in a fair manner in accordance with the reasonable expectations of the data subjects.[35] This also relates to the requirement of transparency of processing that has been explicitly included in the GDPR. The systemic challenges raised in this respect refer to the usage patterns and methods employed to process big data which are considered to be *de facto* at odds with the lawfulness, fairness and transparency principle as neither the controller collecting the data nor the data subject could have considered or even imagined such methods and patterns at the time of the collection.[36] The actor-related challenges this time refer more to controllers that cannot closely monitor and guarantee transparency in the processing of big data because allegedly they do not know where the information came from especially in processes of merging different databases

[32] GDPR, Art. 5 (2).
[33] Zarsky (n 15), 1020.
[34] See Maria Tzanou, *The Fundamental Right to Data Protection: Normative Value in the Context of Counter-Terrorism Surveillance* (Hart Publishing, 2017) 26.
[35] Tzanou, ibid.
[36] Zarsky (n 15), 1006.

and re-using data[37] that would render any transparency requirements costly, difficult or even impossible to deliver.[38]

The principles of data minimisation and storage limitation which require that data are 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed'[39] and that they are in principle 'kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed'[40] respectively face similar challenges. At a *systemic level*, the 'business model of big data'[41] is perceived to be 'antithetical'[42] to the data minimisation and storage limitation principles.[43] This antithesis is inherent to the nature of big data which provides controllers with a 'clear incentive to collect and retain as much data as they can for as long as possible'.[44] The argument, therefore, goes that 'diligently enforcing' data minimisation principles would limit the success of big data initiatives and undermine their utility.[45] At the *individual/controller level*, the challenges of big data to the principles of data minimisation and storage limitation are drawn from the same utilitarian argument: why should controllers collect and store less data since even more quantities can be easily made available to them? Or to put it differently, even if potential future benefits for collecting more data are uncertain and hypothetical, if the possibility to collect and store as much data as possible[46] exists why should controllers not take advantage of it? In addition, even if data do not have any known utility currently, why erase

[37] Bart van der Sloot and Sascha van Schendel, 'Ten Questions for Future Regulation of Big Data: A Comparative and Empirical Legal Study' *JIPITEC* 1, 10.
[38] Ibid, 1006.
[39] GDPR, Art. 5 (1) (c).
[40] GDPR, Art. 5 (1) (e).
[41] See Colin Bennett & Robin Bayley, Privacy Protection in the Era of 'Big Data': Regulatory Challenges and Social Assessments in Bart van der Sloot et al. (eds.) *Exploring the boundaries of big data* (Amsterdam University Press, Amsterdam 2016)
205, 210.
[42] Ibid.
[43] As put by the Dutch DPA in an empirical study 'Big Data is all about collecting as much information as possible.' See van der Sloot and van Schendel (n 37), 9.
[44] Zarsky (n 15), 1011.
[45] Ibid.
[46] See van der Sloot and van Schendel (n 37), 9.

them permanently if in the future there might be a potential use, for instance, through aggregation with data that are not even currently available?[47]

Finally, concerns arise regarding the accuracy principle. This requires that personal data must be accurate and, where necessary, kept up to date.[48] It is in the interest of controllers to ensure the accuracy of the data they process, but in the context of big data it might prove difficult to take steps to ensure that inaccurate data are erased or rectified without delays,[49] especially when these have been further merged with other data.[50] Indeed, Article 5 (1) (d) GDPR requires that 'reasonable' steps should be taken in this respect which raises questions whether erasing or rectifying the information is a reasonable step that can be taken in the context of big data in order to comply with the principle of accuracy. Indeed, the trend behind big data appears to be 'quantity over quality'.[51]

4. Big data re-purposing

It has been argued that big datasets are likely to contain 'some intrinsic, hidden, not yet unearthed' treasures and, hence, 'the treasure hunt' race is 'on to discover and capture' all of them.[52] The very idea, therefore, that big data often have 'no fixed purpose'[53] but their value and potential use will become clear after their hidden treasures are discovered, is arguably at odds with the purpose specification and purpose limitation principle. According to this, personal data must be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.[54] The common argument in this respect goes that purpose limitation is incompatible and rendered meaningless by the very

[47] As eloquently put by van der Sloot and van Schendel 'Data can always be given a second life.' Ibid.
[48] GDPR, Art. 5 (1) (d).
[49] GDPR, Art. 5 (1) (d).
[50] See relevant comments by Slovenian DPA as reported by van der Sloot and van Schendel (n 37), 9-10.
[51] Ibid, 10.
[52] Victor Mayer- Schoenberger and Kenneth Cukier, *Big data: A Revolution That Will Transform How We Live, Work and Think*, (John Murray, 2013), 15.
[53] Van der Sloot and van Schendel (n 37), 9
[54] Article 5 (1) (b) GDPR.

nature of big data that cannot specify certain purposes for collecting the data; rather 'data collection and analysis are themselves the purposes' for the data collection.[55] This argument links to the systemic challenges identified above regarding fair and lawful processing in general and more particularly, the data minimisation problems.

5. Data Security

In the era of big data and cloud computing, a higher risk of security breaches including both unauthorised access to the data and accidental loss will unavoidably emerge.[56] Numerous, troubling security breaches of health have been recorded over the years.

Safeguarding the security of big data can be a challenging task taking into account the multiplicity of stakeholders (controllers, processors, data brokers, third parties); the diversity of means through which the data is acquired (IoT, mobile phones, wearables, social networks, etc.) and the sheer quantity of data collected, aggregated and stored.[57]

The GDPR includes a general principle that requires data controllers to ensure the appropriate security of personal data (the integrity and confidentiality principle)[58] and introduces new rules regarding the security of processing[59] and data breach notifications.[60]

6. Difficulties in identifying Special categories of data

[55] Lokke Moerel and Corien Prins, Privacy for the homo digitalis, Proposal for a new regulatory framework for data protection in the light of big data and the internet of things < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784123>, 2.
[56] See Ernesto Damiani et al. Big data threat landscape and good practice guide. ENISA, January 2016. https://www.enisa.europa.eu/publications/bigdata-threat-landscape and Rossen Naydenov et al. Big data security. Good practices and recommendations on the security of big data systems. ENISA, December 2015. https://www.enisa.europa.eu/publications/big-data-security.
[57] See Antoinette Rouvroy, '"Of Data and Men" Fundamental Rights and Freedoms in a World of Big Data', Bureau of the Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data [ETS 108], 29.
[58] Article 5 (1) (f) GDPR.
[59] Article 32 GDPR.
[60] Articles 33 and 34 GDPR.

The GDPR provides increased protection to special categories of data, such as those revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation.[61] In fact, the processing of such data is in principle prohibited as a general rule,[62] and only permitted in specific instances.[63]

It is difficult, however, to demarcate what constitutes special and non-special categories of data and these difficulties are exacerbated in the big data context. Take the example of health data that are considered a special category of data under the GDPR. Health data are found in a variety of sources, such as GPs' and electronic health records (EHRs), prescriptions, laboratory tests, patient hospital monitoring, etc. They are generated by the Internet of Things (IoT) through wearables that monitor everything from physical activity to sleep, stress and pulse rate, smart patches, electronic skins and ingestibles.[64] These are further augmented by the huge amount of health and fitness apps where individuals volunteer information among others, about their body, sexual activities, mood, smoking habits, women's intimate health (such as menstrual cycles), etc. Health data are also shared in social media, such as Facebook, Twitter and Instagram and specific health networking platforms and online support communities, such as PatientsLikeMe and are generated by online searches. [65]

These raises fundamental questions as to what type of information will be considered as falling within the sensitive, special category of health data. Even innocuous information, such as supermarket shopping lists, can reveal details about what we eat and -what we do not-

[61] Article 9 GDPR.
[62] Article 9 (1) GDPR.
[63] Article 9 (2) GDPR.
[64] See Nathan Cortez, The Internet of Things (IoT) and Health Big Data- Introduction in Glenn Cohen et al. (eds.) *Big Data, Health Law, and Bioethics* (Cambridge University Press, 2018), 125.
[65] For details on health data sources see Introductory chapter. See also, Urs Grasser, Shifting Paradigms- Big Data's Impact on Health law and Bioethics- Introduction in G. Cohen et al. (eds.) *Big Data, Health Law, and Bioethics* (Cambridge University Press, 2018), 15.

and, therefore, our health condition. Are these also health data protected under Article 9 GDPR? The answer depends ultimately on the particular circumstances, the context and even the purposes for which the data are -or might be- used.[66] This, of course, creates uncertainties as to what types of data should fall under special categories and the relevant grounds of processing applicable in these cases. Such unclear boundaries provide significant discretion to controllers and concomitantly, might result in decreasing the protection of data subjects.

7. Effective exercise of data subject's rights

The GDPR provides a number of rights to data subjects, such as right to information,[67] access,[68] rectification,[69] erasure ('right to be forgotten'),[70] a right to restriction of processing,[71] and a right to object[72] and not be subject to a decision based solely on automated processing, including profiling.[73]

The meaningful exercise of these rights in the context of big data is considered 'both unrealistic and deeply paradoxical'[74]. How are individuals expected to monitor all this data deluge every day and have control of how their information is imputed in databases, aggregated and used in order to be able to effectively exercise their access, rectification and erasure rights? Even if they are aware of this information, big data and its multiple lifecycles are notoriously antithetical to the very rationale and ways of exercising these rights. Against which controller should individuals turn, on the basis of which jurisdiction and regulatory framework, when and

[66] Moerel and Prins, n 55, 2.
[67] Article 13 GDPR.
[68] Article 15 GDPR.
[69] Article 16 GDPR.
[70] Article 17 GDPR. See also Maria Tzanou, The Unexpected Consequences of the EU Right to Be Forgotten: Internet Search Engines As Fundamental Rights Adjudicators in M. Tzanou (ed.) *Personal Data Protection and Legal Developments in the European Union* (IGI Global, 2020) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3277348.
[71] Article 18 GDPR.
[72] See Article 21 GDPR.
[73] Article 22 GDPR.
[74] Rouvroy, n 57, p. 34.

how? Similar problems arise from the controllers' perspective. In this case, the argument goes that it is not feasible to ask controllers to keep ahead of the big data deluge in order to always be in a position to allow for the effective exercise of data subjects' rights. These *actor- related* challenges for both data subjects and controllers are fed by and explained by the *systemic* issues that arise from the ways big data work. Similar to the case of the various fair information principles, the effective exercise and enforcement of data subjects' rights would allegedly interfere with the successes of big data initiatives.

8. Enforceability and enforcement of (data protection) law

Significant concerns regard the temporal and territorial application of data protection law- for instance, the GDPR- to big data as well as the relevant enforcement mechanisms advanced by this. Given that such data are generated via wearables, IoT, mobile and online apps, and virtually the whole world wide web and stored on different clouds that can be located anywhere, it becomes obvious why enforceability issues are exacerbated in this context even if the GDPR has extended its *ratione materiae* and *loci* jurisdictional bite significantly.

In addition to enforceability problems, there are also significant enforcement questions arising in the big data era. The GDPR has been criticised for placing the individual as a data subject at the core of its implementation and enforcement apparatus. By granting individuals the right to 'informational self-determination' or the right to 'have control over their personal information', through various principles and rights, such as the fair information principles and data subjects' rights, it also passes on individuals the primary responsibility of monitoring the enforcement of such rights.

Enter again the big data systemic issues here. Because of the nature and functions of these, individual data subjects are in most cases virtually unable to maintain effective control of all the pieces of their personal information in this respect. Control over personal information

is, therefore, almost meaningless in the big data context. The problems do not stop here. The GDPR has introduced accountability obligations for controllers, but again their effective enforcement in the big data environment appears questionable. Similarly, Data Protection Authorities (DPAs) are already overwhelmed in their supervisory role and this is going to get worse when they are asked to deal with big data cases.

Overall, enforceability and enforcement problems of legal frameworks in general, and data protection laws, such as the GDPR in particular, that might render the written law 'a paper dragon in the age of the "digital tsunami"'[75] pose serious concerns and cannot be ignored in this context. This chapter classifies the practical difficulties concerning the enforceability and enforcement, as an *individual level* challenge of big data concerning the *processing* of the information. Arguably, matters of enforceability and enforcement of legal regulation touch upon the outcomes of the processing of big data and pose societal challenges as well. This demonstrates that there are no clear boundaries between the different risks as classified in the analytical framework presented in this contribution, but certain risks may fall in more than one categories.

2.3 *Individual* level issues associated to the *outcomes* of the processing

Big data present complexities and challenges that we have not seen before. Data mining and automated processing using algorithms can create inferences and conclusions that we -humans- are unable to grasp. This issue is excellently captured by the 'black box' metaphor put forward by Frank Pasquale.[76] Even if we voluntarily grant access to our data to be processed, we are unable to follow how the outcome of this processing was reached. This opacity of the 'black box' makes it impossible to have a clear understanding of the consequences of the processing

[75] Mireille Hildebrandt and Bert-Jaap Koops, 'The Challenges of Ambient Law and Legal Protection in the Profiling Era', (2010) *Modern Law Review*, 428, 440.
[76] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, (Harvard University Press, 2016).

of our personal data. An infamous example that demonstrates the problem comes from car insurance. People that use pads to protect their furniture are deemed to be more careful drivers and are therefore granted lower premiums than people that do not use furniture pads. How is the one (careful driving) inferred from the other (use of furniture pads)? The workings of the system are 'mysterious'[77] and they are made even more complicated when the algorithm is a machine-learning one.[78]

The analytical framework of the risks developed in this chapter considers that there are certain risks relating to the black box metaphor that arise at the individual level and can therefore affect the data subjects' fundamental rights. These risks are different from the ones discussed above, however, as they refer to the *outcomes* of the processing of big data and not to the *processing* of the data as such. The risks represented on this axis concern the outputs of the data processes and in particular the inferences and conclusions that can be drawn following the automated processing of the data using algorithms and the decisions that can be taken as a result. These challenges in turn are identified and categorised below.

1. Incorrect or unreliable conclusions

Big data analysis is based on correlations. This means that the finding of a mere statistical relationship between two data values is enough to trigger a certain outcome or decision without it being necessary to establish whether there is an actual causal relation that explains this outcome.[79] To take the example mentioned above, more careful driving is deduced from the fact that people use furniture pads. This is deemed enough to incur lower insurance premiums even if it is clear that a causal relationship between the two cannot be established; the

[77] Ibid, 3.
[78] See Tzanou, Introduction, n 12.
[79] Mayer- Schoenberger and Cukier, n 52, 52.

correlation in this regard is spurious (the association might be statistically robust but happens only by chance).[80]

Furthermore, big data come from the aggregation of various datasets and can be often inaccurate or even contain errors. This 'messiness'[81] of the data may lead to unreliable conclusions.[82] Algorithms can, therefore, make incorrect decisions that can have real effects on individuals. In the big data context this is made possible because there is an abundance of data -often inaccurate and erroneous- and correlations are used to gain valuable insights to them that are fast and cheap [83] without the need to investigate causality.

2. Lack of transparency and foreseeability for the individual affected

As seen above, algorithms are 'black boxes', making it impossible for individuals to anticipate their conclusions and to meaningfully comprehend the logic and processes employed to reach these. Algorithms are often protected under intellectual property rights or kept secret for commercial or national security purposes and are thus inaccessible to individuals.[84] In addition, algorithms are highly complex, often containing many different rules and it is difficult or impossible to establish which points of entry determined how the resulting outcome was built. This refers to the algorithms' opacity problem; individuals are unable to meaningfully understand how their conclusions came to be. As Pasquale explains, 'we can observe [their] inputs and outputs, but we cannot tell how one becomes the other'.[85] Machine learning is another reason for this opacity; the algorithm can train itself to learn to predict unknown

[80] González Fuster and Scherrer, n 21.
[81] Mayer- Schoenberger and Cukier, n 52, 32.
[82] 'Garbage in, garbage out' refers to this problem. See Brent Daniel Mittelstadt et al., 'The ethics of algorithms: Mapping the debate', (2016) *Big Data & Society*, 1, 15.
[83] Mayer- Schoenberger and Cukier, n 52, 66.
[84] Mittelstadt et al., n 82, 6.
[85] Pasquale, n 76, 3.

variables and modify its behavioural structure during operation,[86] thus producing outputs untraceable even to programme designers themselves.[87]

3. Unfair and discriminatory conclusions and decisions

The processing of big data can lead to biased and discriminatory conclusions. Indeed, it has been argued that big data create 'unimaginable opportunities'[88] for discrimination. 'Employers, financial institutions, marketers, … and others can now easily obtain a wealth of bid data about individuals' health status and use it to make adverse decisions relating to data subjects'.[89]

Algorithms are created by humans and unavoidably reflect their authors' values, interests and pre-conceptions.[90] More often, discriminatory conclusions, rather than being the result of intentional design of the system, reflect 'widespread biases that persist in society at large',[91] replicating 'pre-existing patterns of exclusion and inequality'.[92] For instance, it was found that algorithms that 'could predict skin cancers better than dermatologists turned out to be less accurate when diagnosing dark-skinned people because the system was predominantly trained with data from white people'.[93] Pre-existing biases are particularly problematic as it is difficult to identify the root of such (unintentional) discriminatory outcomes and effectively challenge them in courts.[94] Biases in the data might also arise for a number of different reasons,

[86] Mittelstadt et al., n 82, 6.
[87] Jenna Burrell, 'How the Machine "Thinks:" Understanding Opacity in Machine Learning Algorithms' (2016) 3 *Big Data and Society*, 1.
[88] Sharona Hoffman, Big Data's New Discrimination Threats: Amending the Americans with Disabilities Act to Cover Discrimination Based on Data-Driven Predictions of Future Disease in Big Data in Glenn Cohen et al. (eds.) *Big Data, Health Law, and Bioethics* (Cambridge University Press, 2018), 85, 85.
[89] Ibid.
[90] Mittelstadt et al., n 82, 6.
[91] Solon Barocas and Andrew D. Selbst, 'Big Data's Disparate Impact', (2016) 104 *California Law Review*, 671.
[92] Ibid.
[93] Allison Gardner, 'Medical AI can now predict survival rates – but it's not ready to unleash on patients', *The Conversation*, 21 November 2019 < https://theconversation.com/medical-ai-can-now-predict-survival-rates-but-its-not-ready-to-unleash-on-patients-127039 > and references therein; Angela Lashbrook, 'AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind', *The Atlantic*, 16 August 2018 https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/>.
[94] Ibid.

including technological errors and failures, inaccuracies in the data that originate from a variety of resources, and other contextual factors present in the context of big data (such as re-purposing of the data, different controllers and stakeholders involved in the data lifecycles, etc).

4. Profiling and the effects of group analysis on individuals

Big data can be used to both uncover (hidden) patterns and correlations and apply these to specific individuals through decisions that affect them directly.[95] Profiling[96] refers to two different phenomena: the creation of a profile and its use in decision-making processes.[97] The creation of profiles requires the identification of general trends, patterns and correlations in the data 'not visible with the naked human eye'[98] through the use of data mining. Such profiles can then be used to categorise and target specific individuals by attributing to them certain characteristics and risks on the basis of which decisions that affect them will be made.[99]

Profiling is problematic for several reasons. Big data profiling is predictive and probabilistic. It is based on the analysis of huge amounts of data from past behaviour or known cases to generate predictions about future behaviour and unknown features.[100] Profiles drawn up using data mining are normally non-distributive meaning that the patterns identified in the data -which make up the specific group profile- do not apply to all members ascribed in the

[95] Zarsky, n 15, 1000.
[96] Art 4 (4) GDPR defines profiling as 'any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements'.
[97] Paul De Hert and Hans Lammerant, Predictive profiling and its legal limits: effectiveness gone forever? in Bart van der Sloot, Dennis Broeders & Erik Schrijvers (eds.) *Exploring the Boundaries of Big Data* (Amsterdam University Press, 2016), 145, 147. The Art 29 WP identifies four stages of the profiling process: 1) collecting data; 2) analysing data; 3) building a profile for an individual; 4) applying a profile to make a decision affecting the individual. See Art 29WP, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, Adopted on 3 October 2017, 12.
[98] Hildebrandt and Koops, n 75, 440.
[99] De Hert and Lammerant, n 97, 147.
[100] Ibid.

group[101] (as opposed to distributive profiles where the characteristics of the profile apply to all members of the group).[102] Non-distributive profiles can stereotype and stigmatise people by categorising and treating them according to how their group profile is assumed to behave.[103] Such profiles can also result in unfair conclusions and unjustified discrimination for certain individuals,[104] based on incorrect or unreliable assumptions[105] regarding the categories the latter are inferred to belong to. Outcomes based on social sorting[106] can subject people to financial, social and cultural segregation and exclusion.[107] Finally, profiling touches upon the autonomy of individuals by '"normalising" them in the kind of behaviour the profile predicts' on the basis of their group characteristics.[108]

2.4 *Societal* level challenges

Big data analytics can affect individuals in several ways, but they have broader consequences on the whole *society* as well. I call these the *societal* (or collective) level risks of big data. A clarification is needed here. As will be seen below, the issues that big data raise at the *societal* level can have serious implications on individuals' fundamental rights as well. However, this chapter argues that this separate classification is necessary for two main reasons: First, it is not always clear what the particular harm of big data might be on specific individuals. I will borrow an example from surveillance here. In 2013, Edward Snowden, a CIA contractor, revealed that the United States of America (USA) had been operating a secret mass electronic surveillance

[101] Ibid.
[102] Ibid.
[103] Article 29WP, Guidelines on Automated individual decision-making n 97, 5.
[104] See above discrimination.
[105] See above incorrect and unreliable conclusions.
[106] See David Lyon, 'Surveillance, Snowden, and Big Data: Capacities, consequences, critique' (2014) *Big Data & Society* 1, 10.
[107] European Data Protection Supervisor (EDPS), EDPS Opinion 4/2015 *Towards a new digital ethics: Data, dignity and technology*, 11 September 2015, 13.
[108] Hildebrandt and Koops, n 75, 434.

programme that granted it access to Internet data, such as email, chat, videos, photos and file transfers held by leading Internet companies, including Facebook, Google, Microsoft, Yahoo, Skype, Apple and Youtube.[109] The revelations caused a public outcry in Europe, but the US side maintained that US intelligence agencies did 'not have the legal authority, the resources, the technical capability or the desire to intercept all of the world's communications. Those agencies are not reading the emails of everyone in the United States, or of everyone in the world'.[110] It is beyond the scope of this paper to delve in the fallacies of this argument,[111] but if we are to take it on its face value, it is indeed extremely difficult for individuals -if not impossible- to know whether and when they are targeted by surveillance measures, by whom surveillance is undertaken and through which means. Yet, even if harm cannot be particularised on specific individuals, surveillance poses significant risks at the *societal* level. There is a second reason explaining the necessity of drawing a separate category of *societal* level challenges of big data. This is required in order to be able to develop appropriate legal and regulatory solutions. These will be touched upon later in the chapter; for the moment, the analysis focuses on a discussion of the *societal* level challenges.

1. Surveillance

Big data analytics have brought forward new concerns of mass surveillance. To be sure, surveillance is not a new phenomenon emerging in the age of big data.[112] Nevertheless, big data have increased exponentially[113] the possibilities for surveillance and have changed to an extent the actors involved in this. Surveillance in the era of big data analytics is ubiquitous,

[109] See Maria Tzanou, 'European Union Regulation of Transatlantic Data Transfers and Online Surveillance' (2017) *Human Rights Law Review*.
[110] Annex VI, Commission Implementing Decision of 12 July 2016 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequacy of the protection provided by the EU–US Privacy Shield, Brussels, 12 July 2016, COM (2016) 4176 final, 93.
[111] See Tzanou, n 109.
[112] Lyon, n 106.
[113] Ibid.

pervasive and goes well beyond the traditional model of state surveillance. Large organisations, like Google and Facebook hold almost every piece of information about our everyday lives. Surveillance permeates every aspect of our existence even our most intimate details. From wearables to smart sex toys and 'spying vibrators',[114] data about our bodies and functions are visible to private companies.[115] Indeed, commercial surveillance by Internet giants 'dwarfs the surveillance conducted by national intelligence agencies.'[116]

The problem of widespread surveillance in the context of big data affects individual persons but it is primarily a societal one. While people's lives are rendered transparent, 'organisations engaged in surveillance are increasingly invisible to those whose data are garnered and used.'[117] As a result of surveillance, profiles are built and predictions are made; behaviours are nudged; surveillance then feeds back the individuals to whom these profiles will apply and so on.

2. Autonomy, equality and human dignity

Big data analytics pose challenges to equality, autonomy and human dignity. The issues that arise from datamining, machine learning and profiling were categorised in the taxonomy presented in this chapter as individual level challenges that relate to the *processing* of the data. I argue here that these same issues raise problems that cannot be only captured by their effects on individuals, but we should consider the broader dynamics they create in the society.

Big data analysis leads to predictions that can be consequential, based on past behaviours but also preemptive,[118] aimed to deliberately subvert future choices and equal

[114] Janet Burns, 'We-Vibe Settles For $3.7M In "Spying Vibrator" Data Suit', *Forbes*, 15 March 2017 https://www.forbes.com/sites/janetwburns/2017/03/15/we-vibe-settles-for-3-7m-in-spying-vibrator-data-lawsuit/#3ee371316021.
[115] Janet Burns, 'The "Spying Vibrator" Suit Is Over, But Sex Toys Are Still Talking Data', *Forbes*, 14 December 2016 https://www.forbes.com/sites/janetwburns/2016/12/14/the-spying-vibrator-suit-is-over-but-sex-toys-are-still-talking-data/#692879384417.
[116] Yeung, n 1, 123.
[117] Lyon, n 106, 4.
[118] Ibid.

opportunities.[119] Personal autonomy is stifled if our ordinary, everyday lives are 'determined by algorithms and their continuous variations'.[120] Big data analytics are used to 'seduce, coerce, discipline, regulate and control: to guide and reshape how people, animals and objects interact with and pass through various systems'.[121] Profiling and its group effects can create 'filter bubbles' or 'echo-chambers'[122] undermining other fundamental rights such as freedom of expression and posing threats to democracy. Ultimately, the question that arises for our society is one of human dignity: how will we make sure that individuals are not treated as means to (the big data analytics) end?

3. The power imbalances of big data

Big data processing creates significant asymmetries in our society. Personal data have substantial monetary value[123] and few parties get to reap this. For companies, such as Google and Facebook, personal data are their most valuable asset[124] that is not accounted for on their balance sheet. [125] Indeed, such companies trade in personal data by granting 'free' services to users by harvesting their data and paid services to advertisers and other businesses to use these data. Data brokers mine the web to collect personal information and sell this to interested parties.[126] Significant asymmetries arise as the 'digital dividend'[127] is not shared fairly between users-data subjects and the other stakeholders that trade personal data. As the EDPS observes,

[119] For example, see Samuel Gibbs, 'Women less likely to be shown ads for high-paid jobs on Google, study shows', *The Guardian*, 8 July 2015 < https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study>.
[120] EDPS, Opinion 4/2015, n 107, 13.
[121] Rob Kitchin, 'Thinking critically about and researching algorithms', (2017) 20 (1) *Information, Communication & Society*, 14, 19.
[122] EDPS, Opinion 4/2015, n 107, 13.
[123] See OECD, 'Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value', (2013) OECD Digital Economy Papers, No. 220, OECD Publishing <http://dx.doi.org/10.1787/5k486qtxldmq-en >
[124] Ibid.
[125] EDPS, Opinion 8/2016, Opinion on coherent enforcement of fundamental rights in the age of big data, 23 September 2016, 13.
[126] Federal Trade Commission (FTC), *Data Brokers- A Call for Transparency and Accountability*, May 2014.
[127] EDPS, Opinion 8/2016, n 125, 13.

successful companies operating personal data services platforms are 'able to discriminate by combining knowledge they extract from data with monopoly power and vertical integration in the markets'.[128]

The asymmetries of big data are not merely about monetary winners and losers; more crucially, they are about imbalances of information and -ultimately- of power between those who provide the data and those who exploit them. At the micro level, such imbalances might be imperceptible to individuals, yet have real pernicious effects. For example, the recent revelation that a women's fertility app with which women share the most intimate details about their bodies, including sex and menstruation information is funded by anti-abortion campaigners[129] shows the disturbing extent of the complexities of the information and power asymmetries in the big data society. At the macro/ societal level the accumulation of such degrees of (market) power by private organisations might have serious social, economic, cultural and political consequences. As Kitchin observes, 'far from being neutral in nature, algorithms construct and implement regimes of power and knowledge and their use has normative implications'.[130]

4. The lack of accountability of data stakeholders[131]

The information and power asymmetries in the big data society exacerbate the 'black box' phenomenon discussed above. The aggregation of different sets of data through their constantly evolving lifecycles, machine learning, and AI raise questions as to who bears responsibility for the outcomes of big data. Is it the humans that must be held accountable or the algorithms?

[128] Ibid.
[129] Jessica Glenza, 'Revealed: women's fertility app is funded by anti-abortion campaigners', *The Guardian*, 30 May 2019 https://www.theguardian.com/world/2019/may/30/revealed-womens-fertility-app-is-funded-by-anti-abortion-campaigners and Eva Wiseman, 'Beware the fertility app that wants to share your data with anti-abortion campaigners', *The Guardian*, 9 June 2019 https://www.theguardian.com/lifeandstyle/2019/jun/09/app-creep-and-the-dark-side-of-sharing-private-date-on-our-phones.
[130] Kitchin, n 121, 19.
[131] Accountability is understood differently here from the 'accountability principle under Article 5 (2) GDPR.

And to which degree? This lack of accountability is not only frustrating for the innocent individuals that are negatively affected. It seems to be imposing the 'black box' model on the whole society.[132] An unaccountable society undermines the very principles of democracy and the rule of law.

3. Addressing the Big Data Challenges

3.1 The importance of the taxonomy and why the GDPR cannot provide a one-size-fits-all solution

The previous section has presented a conceptual map of the various issues raised by big data, classifying them at the *individual* and the *societal* level. Individual level concerns were furthermore distinguished depending on whether they relate to the *processing* of personal data or they occur as *outcomes* of the processing. The taxonomy attempts to capture the more important issues raised by big data pulling them together in broader themes, but these might entail further unknown risks. At the same time, several of the issues identified are not clear-cut; they often overlap and recur -albeit in slightly different forms- at the individual and the societal levels. Moreover, similar challenges can be framed in different lights and examined through different lenses, often in such ways that it becomes difficult to properly define the problem and distinguish its effects (is it commercial surveillance or 'surveillance capitalism' understood as a 'privately administered compliance regime of rewards and punishments aimed at modifying and commoditising behaviour for profit'?[133]) Nevertheless, this mapping is not of a merely theoretical, academic value. On the contrary, the taxonomy is crucial in order to be able to reflect on the ways the risks of big data can be properly addressed.

[132] Pasquale, n 76, 217.
[133] Shoshana Zuboff, 'Big other: surveillance capitalism and the prospects of an information civilization', (2015) 30 *Journal of Information Technology*, 75, 83 and 85.

Taking the conceptual map as a starting point, I would like to advance two main arguments: First, regulatory approaches to big data risks cannot be all-encompassing one-size-fits-all measures. Solutions to big data problems should be searched at the particular level -individual or societal- where the issue arises and may vary depending on the specific type of problem. On the one hand, individual level regulatory measures should be aimed at preventing harm to individuals and at safeguarding their basic rights and interests. On the other hand, approaches that aim to deal with the societal risks of big data should address the dynamics and the types of power that big data analytics bring forward.

Against this backdrop, I come now to the main question of this chapter: What is the role of data protection laws in general and the GDPR in particular in regulating big data? The short answer is limited. The GDPR has been at the same time both hailed as the first piece of legislation that deals with big data and AI[134] and accused of failing to cope[135] or being entirely incompatible[136] with big data developments. Despite being diametrically antithetical -the one praises the GDPR's capabilities, the other condemns them as insufficient-, these two views share an interesting common point. They both seem to assume that data protection regulation, and in particular the GDPR, represent the main legal methodological solution attempting to grasp with big data. This assumption is not completely unfounded; indeed, the GDPR seems to make the same self-assertion.[137] Yet, this individualistic methodology is not appropriate to deal with the power configurations and the dynamics that concern us here.[138] Big data certainly entails personal data processing, but not all problems this raises are necessarily data privacy problems that should be dealt with by the GDPR just because they fall *rationae materiae* within

[134] See above.
[135] Gonçalves, n 6, 114.
[136] Zarsky, n 15.
[137] Art 29 WP notes 'The GDPR introduces new provisions to address the risks arising from profiling and automated decision-making, notably, but not limited to, privacy.' Article 29WP, Guidelines on Automated individual decision-making n 97, 6.
[138] Rouvroy, n 57, 22. See also Bart van der Sloot, The Individual in the Big Data Era: Moving towards an Agent-based Privacy Paradigm in Bart van der Sloot et al. (eds.) *Exploring the boundaries of big data* (Amsterdam University Press, 2016), 177.

its scope of application.[139] Such an assumption overlooks mechanisms and outcomes that are often completely irrelevant to data protection frameworks. Big data are less about identifying individuals -although they do have these capabilities- and more about 'algorithmic forms of constantly evolving, impersonal categorisations of risks and opportunities'[140] and generalistic, predictive and even pre-emptive profiles. The data itself has a 'relational' value: 'it is the (cor)relations between data that makes them valuable and useful',[141] sensitive or less so.

This explains why the typology of the big data challenges presented in this chapter is of crucial importance. By ignoring the two levels or conflating big data challenges at the individual and the societal level,[142] we end up trying to understand why data protection legislative measures such as the GDPR are not fit for purpose for big data contexts. The challenge facing us, therefore, is not as has been -erroneously- suggested 'how to take account, in *personal data protection instruments*, of the relational, and therefore also collective, nature of what, *through data*, merits protection?'[143] In my view, the challenge is how to construct a

[139] See also Purtova who argues that 'the material scope of … the GDPR, is growing so broad that the good intentions to provide the most complete protection possible are likely to backfire in a very near future, resulting in system overload.' Nadezhda Purtova, 'The law of everything. Broad concept of personal data and future of EU data protection law', (2018) 10 (1) *Law, Innovation and Technology*, 40, 41.
[140] Rouvroy, n 57, 22.
[141] Ibid.
[142] The indicative list of risks that can arise from the processing of personal data provided in Recital 75 GDPR demonstrates this conflation and confusion of issues. More particularly, Recital 75 states that risks of processing of personal data include 'discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorised reversal of pseudonymisation, or any other significant economic or social disadvantage; where data subjects might be deprived of their rights and freedoms or prevented from exercising control over their personal data; where personal data are processed which reveal racial or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, and the processing of genetic data, data concerning health or data concerning sex life or criminal convictions and offences or related security measures; where personal aspects are evaluated, in particular analysing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, in order to create or use personal profiles; where personal data of vulnerable natural persons, in particular of children, are processed; or where processing involves a large amount of personal data and affects a large number of data subjects.' With the necessary caveats that 'risks' under the GDPR have a slightly different meaning from 'challenges' and 'issues' as discussed in this chapter and they do not concern big data in particular but any type of processing of personal data, Recital 75 contains a mixture of different types of risks (some concern the *processing* of personal data, some the *outcomes* of processing, some are *societal*) put together without any differentiation. This is problematic also for controllers who are required to decide when processing is of 'high risk' or which data breaches entail such 'risks'.
[143] Rouvroy, n 57, 22. Emphasis added.

regulatory framework of what merits protection in the age of big data by going beyond short-sighted visions that everything can be fitted in *personal data protection instruments* just because *personal data* is at stake.

3.2 The limited but important role of the GDPR

Where does all this leave the GDPR? I argue that its role is limited but important. The GDPR is an instrument that regulates personal data processing through its fair information and accountability principles and the rights it grants to data subjects. It cannot be denied, therefore, that by regulating personal data – the raw material of the big data society- it plays an instrumental role in dealing with the risks that big data analytics might incur. This instrumentality of the GDPR, however, in preventing big data risks is only of indirect and probable value. It is possible that some big data risks could be indirectly mitigated if the GDPR principles are properly applied. Considering the seriousness of big data challenges, it is clear that this is not good enough. To put it differently, we should not be expecting the GDPR to address all the big data issues.

That being said, the GDPR has a role to play besides its indirect instrumental value. Going back to the taxonomy of big data challenges proposed above, I argue that the GDPR is able to deal with *individual level* challenges that relate to the *processing* of personal data. These are various and were discussed in detail in Section 2 but can be summarised as follows: some are data-subject- or data controller- related (I call these *actor-related*); some are *systemic*; and, all of them are modelled on and concern specific data protection rules.

Data subject- related issues describe the difficulties of data subjects to take advantage of rights granted to them under the GDPR. For instance, it is alleged that data subjects do not properly exercise their right to inform themselves as they do not read privacy policies and even

when they do it is not certain they understand them. *Data controller- related* problems concern the difficulties controllers face to comply with their GDPR obligations. For example, it is difficult and costly for controllers to closely monitor and guarantee transparency in the processing of big data because allegedly they do not know where the information came from due to the aggregation and re-purposing of the data. *Systemic* challenges refer to the perceived inherent antithesis of big data with data protection principles. In this regard, the argument goes that big data are *de facto* incompatible with GDPR principles, such as data minimisation and storage limitation as their utility is linked to the amount of data collected (the more, the best) and the length of periods retained (the longest, the best).

Despite the particular problems -*actor-related* or *systemic*- all the big data issues analysed above concern specific GDPR provisions (personal data and personally identified information, problems and limitations of consent, implementation of fair and transparent processing, big data re-purposing, data security, etc.) and are addressed by this instrument. In this respect, *actor-related* concerns either from the data subject's or the controller's side assume the non-enforceability of GDPR rights and rules just because it might be difficult to enforce them. These difficulties are not unique to big data, but arguably they augment within this context. Nevertheless, it is wrong to assume that the GDPR does not apply to big data just because it is difficult to enforce its principles. This argument is based on circular reasoning and it ends up denying legal protection for no good reason.

The *systemic* challenges raised against the GDPR look more serious, but they are not insurmountable. The argument this time seems to be deontological (rather than ontological) in nature: the GDPR principles should not apply because they undermine the utility of big data. This implies a somewhat fake dilemma.[144] The GDPR does not contain abstract rules aimed to

[144] The ICO notes 'It's not a case of big data *or* data protection, it's big data *and* data protection; the benefits of both can be delivered alongside each other.' Information Commissioner's Office, *Big data, artificial intelligence, machine learning and data protection*, 20170904 Version: 2.2, para 28.

be antithetical to the big data business model. On the contrary, it has two equally important objectives: to ensure that personal data move freely while the rights and freedoms of natural persons with regard to the processing of personal data are protected.[145] The GDPR aims to act as an 'enabler of big data services in Europe'[146] by providing for context- specific rules[147] that could open up 'some concrete doors for big data to flourish'[148] while safeguarding at the same time the rights of data subjects. The *systemic* challenges argument is also based on the assumption that big data should remain an unregulated space because any regulation limiting these risks might undermine their utility. As Zuboff has astutely put it, this somehow implies 'the attribution of agency to "technology." "Big data" is cast as the inevitable consequence of a technological juggernaut with a life of its own entirely outside the social. We are but bystanders.'[149] This cannot be accepted. Big data should be able to produce their potential while respecting relevant regulatory frameworks applicable to them.

To sum up, the GDPR has a role to play when individual -level issues relating to the *processing* of personal data in the age of big data are at stake. This is because the GDPR sets out a comprehensive legal framework and contains a variety of mechanisms to address these. These are flexible and allow for context-specific solutions. They include legal rules (and their exemptions and exceptions thereof) stipulating the conditions under which processing is lawful and the obligations of data controllers. They also entail technological solutions; data protection by design and by default[150] aim at developing systems, programmes and applications that incorporate data protection principles such as data anonymisation, pseudonymisation,

[145] Art. 1 GDPR.
[146] Commission, 'The EU Data Protection Reform and Big Data, Factsheet' March 2016. See also Gonçalves n 6, 114 who argues that the GDPR demonstrates 'the EU's deliberate, actually explicit intent to simplify rules for companies in the digital age' and 'caught between its twofold objective of strengthening the rights of the data subjects, and facilitating business, the EU legislator ended up favouring the latter to the detriment of the former.'
[147] See Art. 5 (1).
[148] Mayer-Schonberger and Padova, n 14, 318.
[149] Zuboff, n 133, 75.
[150] Art 25 GDPR and Recital 78.

transparency, enhanced security into the design of the system.[151] Finally, alongside mandatory requirements, the GDPR allows for 'soft law' tools such as codes of conduct and certification mechanisms that can take account 'of the specific characteristics of the processing carried out in certain sectors and the specific needs of micro, small and medium enterprises'.[152]

Whether specific solutions enshrined in the GDPR are fit for purpose and effective in practice in the age of big data remains an open issue, allowing for the possible re-working of certain models and approaches in the future. For example, the concept of consent and its seemingly enhanced emphasis under the GDPR is an issue that should be revisited in the future.[153] This, however, does not negate the fact that the GDPR is a legislative instrument well-placed to deal with the individual- level challenges relating to the *processing* of personal data identified above.

4. Addressing issues that relate to the *outcomes* of the processing and *societal* level challenges of big data

As mentioned above, the GDPR has only an indirect instrumental role to play when issues relating to the *outcomes* of the processing and *societal* challenges of big data are at stake: by

[151] As the EDPS correctly recommends 'Designers and manufacturers should apply the same level of creativity and dynamicity they usually display in introducing attractive devices and apps to also provide individuals with effective and user-friendly privacy notices and setting options. As a result, individuals should be able to set options relevant to their privacy and data protection with the awareness that this is an important element of the devices and apps' use, in their own personal interest, and not a boring formality or a useless burden.' EDPS, Opinion 1/2015 Mobile Health- Reconciling technological innovation with data protection, 13. See also EDPS, Opinion 7/2015 (n 3), 14.

[152] Recital 98 GDPR. See also Art 40 GDPR.

[153] The Centre for Information Policy Leadership (CIPL) observes that 'Despite the fact that there are six legal basis contained in the GDPR, none of which are privileged over the other, there is a general feeling among data protection practitioners, lawyers and DPOs that DPAs, lawmakers and policymakers in the EU place strong emphasis on consent as a more important legal basis… For the GDPR to serve as a modern privacy law, its consent requirements cannot be emphasised as the principal legal ground for processing, nor should the other legal bases be continuously construed narrowly.' Centre for Information Policy Leadership, 'GDPR One Year In: Practitioners Take Stock of the Benefits and Challenges', 31 May 2019, < https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_report_on_gdpr_one_year_in_-_practitioners_take_stock_of_the_benefits_and_challenges.pdf>, 8.

regulating personal data it could potentially minimise or prevent some of the risks analysed above. Yet, the GDPR itself aims at a much higher objective: it purports 'to address the risks arising from profiling and automated decision-making, notably, but not limited to, privacy'.[154] For instance, Article 22 GDPR 'establishes a general prohibition for decision-making based solely on automated processing'[155] that produces legal effects concerning the data subject or significantly affects her.[156] The GDPR also grants data subjects rights to access[157] and be informed[158] about the existence of automated decision-making, including profiling and to receive meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing and the right to contest the decision.[159]

Admittedly, automated decision-making involving personal data constitutes processing within the meaning of the GDPR[160] but it is not the application of the *general* data protection principles and data subject rights to this that is contested.[161] What is problematic is the *specific* rules on automated decision- making and profiling seen above. Leaving aside questions regarding the effectiveness of such provisions in the age of big data, there is another fundamental problem that arises here. This is about regulatory overreach and perhaps more crucially, what Brownsword termed 'normative disconnection'[162]: the GDPR essentially attempts to control the outcomes of algorithmic processing and grant the persons affected by this due-process rights. But, even if AI risks arise at the individual level, they are hardly issues that concern the data subject. They concern decision-making processes, dynamics and outcomes that should not be confounded with data-processing issues. Therefore, the problem

[154] Art 29 WP Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 6.
[155] Ibid, 19.
[156] See also Recital 71.
[157] Article 15 (1) (h) GDPR.
[158] Articles 13(2) (f) and 14(2) (g) GDPR.
[159] Article 22 (3) GDPR.
[160] Article 4 (2) GDPR.
[161] See Article 29WP, Guidelines on Automated individual decision-making n 97, 9-19.
[162] Roger Brownsword, *Rights, Regulation, and the Technological Revolution* (OUP, 2008), 166.

is not that these GDPR provisions are not fit for purpose; the problem is that they are not appropriate at all to deal with outputs irrelevant to the GDPR's scope.[163]

The individual-level challenges that occur as *outcomes* of the processing and the *societal* challenges of big data require regulatory methodologies that go beyond the atomistic, data subject- centred GDPR approach. How would these look like? The answer to this could be the subject of a whole monograph, but four preliminary points could be advanced here.

First, regulatory responses should correspond directly to the level and the types of risks we are facing. It is for this reason that the taxonomy of challenges presented in the first part of this chapter becomes relevant. For instance, the problems raised by profiling at the individual level should be addressed by '*sui generis*'[164] (and often sector-specific) measures that mandate the audit of algorithms and grant interested groups -and wherever possible individuals- specific profiling due-process rights[165] that are independent to whether the personal data of a certain data subject were processed at a certain time point by unknown controllers.

Second, AI and big data challenges require a combination of different approaches rather than a single legislative measure. In this respect, the methodological approach of the EDPS of viewing big data as posing data protection, competition and consumer law problems and therefore requiring a holistic inter-sectorial methodology is certainly a step towards the right direction.

---

[163] As Koops observes 'This… requires stretching the concept of personal data (sometimes to the point of breaking, or perhaps rather of becoming void of meaning), or stretching the regulatory problem so that it becomes a problem of processing personal data.' Bert-Jaap Koops, 'The trouble with European data protection law', (2014) 4 (4) *International Data Privacy Law*, 250, 258.

[164] Ibid, 260.

[165] See Danielle Keats Citron and Frank Pasquale, 'The Scored Society: Due Process for Automated Predictions', University of Maryland Francis King Carey School of Law Legal Studies Research Paper No. 2014 – 8; Crawford and Schultz propose 'procedural data due process' for big data that 'rather than attempt regulation of personal data collection, use, or disclosure ex ante, procedural data due process would regulate the fairness of Big Data's analytical processes with regard to how they use personal data (or metadata derived from or associated with personal data) in any adjudicative process, including processes whereby Big Data is being used to determine attributes or categories for an individual.' Kate Crawford and Jason Schultz, 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms', (2014) 55 *B.C.L. Rev.*, 93, 109.

Third, solutions require innovative thinking that includes both a transformational evolution of current ways of regulating ever-existing problems such as discrimination as well as the adoption of completely new rules addressing specific AI and big data challenges. For example, in the US context Sharona Hoffmann proposes that in the era of big health data the Americans with Disabilities Act (ADA) (the federal law that prohibits public and private bodies from discriminating against individuals because of their disabilities) be amended to expand its anti-discrimination scope to include also a prohibition of 'future' physical or mental impairments.[166] Given that predictive medicine profiling has been a reality for several years now,[167] legislators should proactively reflect on how to protect people from predictive, inferential discriminatory decisions as well.

Fourth, separate, specific rules might be needed for different stakeholders in the big data society, such as data brokers,[168] platforms,[169] the insurance sector, etc. These should ensure the accountability of these actors and account for the societal responsibilities of big data players.

5. Conclusion

[166] Hoffman, n 88, 85.
[167] EDPS, Opinion 1/2015 *Mobile Health: Reconciling technological innovation with data protection*, 21 May 2015, 10.
[168] Data brokers should comply with data protection legislation, but specific legislative measures could also be applicable to them. See Privacy International, Why we've filed complaints against companies that most people have never heard of – and what needs to happen next, 8 November 2018, < https://privacyinternational.org/advocacy/2434/why-weve-filed-complaints-against-companies-most-people-have-never-heard-and-what>; Amit Katwala, 'Forget Facebook, mysterious data brokers are facing GDPR trouble', *Wired*, 8 November 2018 < https://www.wired.co.uk/article/gdpr-acxiom-experian-privacy-international-data-brokers> .
[169] See Christian Sandvig, et al, 'An Algorithm Audit.' In: Seeta Peña Gangadharan, ed., *Data and Discrimination: Collected Essays*. (New America Foundation, Washington, DC, 2014). <http://www-personal.umich.edu/~csandvig/research/An%20Algorithm%20Audit.pdf >. The authors propose audits for online platforms that 'will ascertain whether algorithms result in harmful discrimination by class, race, gender, geography, or other important attributes.'

Big data and AI raise a number of different challenges that legislators are still trying to grasp. In this respect, the GDPR is often presented as an example of a legislative solution aimed at addressing big data risks such as automated decision-making and profiling. This chapter has made three distinct contributions to the debate regarding regulatory approaches to big data: 1) it has provided a taxonomy of big data challenges; 2) it has examined the capabilities and limitations of the GDPR to address the risks identified; and, 3) it has offered some suggestions on the pathways that the regulators should be considering.

Big data challenges are mapped in three categories: individual issues that relate to the *processing* of personal data; individual level issues relating to the *outcomes* of the processing; and *societal* level challenges. It was argued that the GDPR is well-placed to deal with individual issues relating to the *processing* of personal data, but its atomistic, 'privacy self-management',[170] data-subject- focused approach is completely inappropriate to address risks that occur as the result of processing and *societal* level challenges. These require innovative ways of thinking that approach the players and complex dynamics of the big data and AI society in a holistic way, while presenting targeted solutions to the specific risks identified at each level.

[170] Solove, n 24, 1880.