Data Usability as a Parameter of Rights and Obligations under the EU Data Act

by **Daria Kim and Man Wai Kwok** *

Abstract: As an instrument for advancing the data economy, the EU Data Act aims to enhance the accessibility of data generated through the use of connected products and related services, thereby unlocking the potential of data for the benefit of society. This article focuses on data usability as an equally crucial factor in harnessing value from data, an aspect that gained recognition only in the later stages of the legislative process. In particular, we examine the technical state of data, which is both a technical factor for realising the value of data and a legal parameter delineating the scope of data access and usage rights, along with the respective obligations introduced by the Data Act.

Our analysis finds that data usability is not thoroughly considered in the Data Act and is only minimally addressed within the framework of its data-sharing regime. We identify several concepts bearing on the technical state of data – including the notions of 'pre-processed data', 'readily available data', 'simple operation', 'insignificant investment', and 'disproportionate effort' – that remain unclear, leading to uncertainties regarding the scope of data-sharing obligations. Attaining the policy goals will to a significant extent hinge on the interpretation and application of these criteria. While acknowledging that the final version of the Data Act represents an improvement over the initial proposal in terms of addressing data usability, we contend that the imposition of restrictive criteria on the scope of 'readily available data' and 'pre-processed' data is not justified, whether viewed from the perspective of technical necessity, legal certainty, or a balance of interests.

Keywords: data access and usage rights; data-driven economy; EU Data Act; data usability; readily available data

Recommended citation: Daria Kim and Man Wai Kwok, Data Usability as a Parameter of Rights and Obligations under the EU Data Act, 15 (2024) JIPITEC 139 para 1.

A. Introduction

1 The vision of a thriving data economy and the question of which measures can fulfil it have been debated extensively in the European Union (EU) in recent years. Several legislative initiatives at the EU level have been underway, pursuing the overarching objective of unlocking the value of digital data for society, particularly by facilitating access to data as a multi-purpose input for innovation and a determinant of competition.[1] The regulatory thinking has undergone a notable shift, transitioning from the idea of conferring a data producer's right in relation to sensor-generated data[2] towards an

* Daria Kim (M.A., LL.M., Dr. iur.) is Senior Research Fellow at the Max Planck Institute for Innovation and Competition (Munich).

Man Wai Kwok is a holder of MSc in Engineer (Data Science) and BSc.

1 European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 'A European strategy for data' COM(2020) 66 final (19.2.2020).

2 European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, 'Building a European Data

appreciation of the need to establish a legal basis for claiming access to data and its further utilisation.

2 The Data Act of 13 December 2023[3] presents an unparalleled statute worldwide that has introduced cross-sectoral access and usage rights as regards data generated by connected products[4] or related services.[5] Thereby, the EU legislature aspires to promote the data economy by enabling the broad utilisation of such data,[6] recognised as 'a core component of the digital economy, and an essential resource to secure the green and digital transitions'.[7] Data subject to new data-sharing obligations should serve as input for aftermarket services and downstream use cases that may extend beyond the products or services through which that data was initially collected.[8]

3 By introducing data access and usage rights, the legislature intends to mitigate contractual imbalances and legal uncertainty identified as 'problem drivers' leading to the suboptimal realisation of the value of data.[9] However, equally important is the technical state of the data in which it has to be made available for subsequent use. Such a state should allow for subsequent meaningful processing and analysis of the shared data. This aspect seems to have been overlooked in the initial proposal by the European Commission (hereinafter, the Commission).[10] Only once does the Commission mention usability in its ex-ante impact assessment accompanying the proposal for a data act when stating that it 'aims to make more data in the EU usable to support sustainable growth and innovation by [...] removing barriers for access to data'.[11] In other words, the Commission associated data usability with opening up access to data and focused on overcoming the restrictive effects of the de facto exclusive control by device manufacturers and service providers over product and service data.[12] Unsurprisingly, the initial proposal did not say much about the technical state of data subject to the obligations to make data available, except for limiting such state to 'the form and format in which [data] are generated by the product'[13] and excluding 'derivative data'[14] and 'information derived or inferred' from data.[15] Though not explicitly stated, one would understand it as referring to 'raw' data,[16] which, as keenly pointed out by critics, would fall short of fulfilling the policy objectives.[17]

Economy' COM(2017) 9 final (10.1.2017) 13; European Commission, Commission Staff Working Document on the free flow of data and emerging issues of the European data economy, SWD(2017) 2 final (10.1.2017) 33-34.

3 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 L, 2023/2854 (22.12.2023).

4 Defined as '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' (art 2(5) Data Act).

5 Defined as 'a digital service, other than an electronic communications service, including software, which is connected with the product at the time of the purchase, rent or lease in such a way that its absence would prevent the connected product from performing one or more of its functions, or which is subsequently connected to the product by the manufacturer or a third party to add to, update or adapt the functions of the connected product' (art 2(6) Data Act).

6 recs 2, 4, 5, 6, 15, 16 and 21 Data Act.

7 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 final (23.2.2022) 1.

8 rec 6 Data Act: 'the data recorded by connected products or related services are an important input for aftermarket, ancillary and other services'; rec 15 Data Act: 'data [covered by the Data Act] includes data collected from a single sensor or a connected group of sensors for the purpose of making the collected data comprehensible for wider use-cases'; 'such data [...] support innovation and the development of digital and other services to protect the environment, health and the circular economy, including through facilitating the maintenance and repair of the connected products in question'.

9 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) SWD(2022) 34 final (23.2.2022) 9, 15. See also rec 2 Data Act.

10 Apart from addressing data semantic interoperability in the context of switching data processing service providers.

11 SWD(2022) 34 final (23.2.2022) 133.

12 rec 20 Data Act.

13 COM(2022) 68 final, rec 17: 'Such data should include data in the form and format in which they are generated by the product, but not pertain to data resulting from any software process that calculates derivative data from such data as such software process may be subject to intellectual property rights.'

14 ibid.

15 ibid rec 14.

16 References to 'raw' data are made in the context of the impact of the Data Act on the database protection sui generis. SWD(2022) 34 final 132, 138.

17 Drexl J and others, 'Position Statement of the Max Planck Institute for Innovation and Competition of 25 May 2022 on the Commission's Proposal of 23 February 2022 for a Regulation on harmonised rules on fair access to and use of data (Data Act)' < https://pure.mpg.de/rest/items/item_3388757_4/component/file_3395639/content > para 333 ff; Podszun R, *Der EU Data Act und der Zugang zu Sekundärmärkten am Beispiel des Handwerks* (Nomos 2023) 41

4 Something must have prompted the Council of the EU to introduce within its negotiation mandate[18] a technically dense Recital 14(a) that specifies the technical state of data covered by the Data Act, along with the notion of 'metadata that is necessary to interpret and use [data]' as part of the data holders' obligations.[19] These proposals made their way into the final version of the Data Act, while the reference to data 'in the form and format' that is generated by a product was omitted. Ostensibly, the EU legislature must have recognised that the latter would not suffice for unlocking the value of data through its use.

5 In the following, we take a close look at data usability, which is both a legal parameter delineating the scope of rights and obligations introduced by the Data Act and a technical precondition for harnessing the value of data, as aspired by the legislature. By doing so, we aim to make an original contribution to the existing analysis of the Data Act.[20] The analysis is structured as follows: Part II explains the key aspects of data usability that are relevant for understanding the technical state of data falling within the ambit of the Data Act. Part III examines the notions of 'pre-processed data', 'readily available data', 'inferred or derived data', 'metadata' and the related qualitative criteria – 'significant investment', 'simple operations', 'disproportionate effort' – that are applied to determine the scope of data covered by the Data Act. It identifies interpretative difficulties presented by these notions and criteria, introducing uncertainty in delineating the scope of new data-sharing obligations. In Part IV, we consider how the Data Act treats the technical state of data in view of the policy objectives, and contemplate an alternative approach where 'readily available data' and 'pre-processed data' would not be restricted by the criteria of 'a simple operation', 'disproportionate effort', and 'significant investment'. In conclusion, we submit that, while the final version of the Data Act represents an improvement over the initial proposal in terms of data usability, the imposition of the limiting criteria on the scope of 'readily available data' and 'pre-processed' data is not justified, whether viewed from the perspective of technical necessity, legal certainty, or a balance of interests.

B. Why does the technical state of data matter?

6 The value of data can be realised only when its technical state allows for processing in a particular use case. This section explains the concept of data usability within the context of data generated through the use of connected products and related services, which is a focus of the Data Act.

I. Data usability as a purpose-oriented concept

7 Neither a commonly agreed-upon definition of the usability of sensor-generated data nor a universal taxonomy of data processing exists.[21] In essence, the usability of sensor-generated data is a characteristic of the technical state of data, indicating its suitability relative to the intended purpose, whether it be sharing, record-keeping, display, status tracking,

ff; Kerber W, 'Governance of IoT Data: Why the EU Data Act Will Not Fulfill Its Objectives' (2023) 72 GRUR International 120, 126 ff.

18 Council of the European Union, 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). Mandate for negotiations with the European Parliament (17 March 2023) 2022/0047(COD) <https://data.consilium.europa.eu/doc/document/ST-7413-2023-INIT/en/pdf>.

19 ibid arts 3(1), 4(1), and 5(1).

20 Eckardt M and Kerber W, 'Property Rights Theory, Bundles of Rights on IoT Data, and the EU Data Act' (2024) European Journal of Law and Economics, https://doi.org/10.1007/s10657-023-09791-8; Kerber W, 'EU Data Act: Will New User Access and Sharing Rights on IoT Data Help Competition and Innovation?' (2024) Journal of Antitrust Enforcement, 10.1093/jaenfo/jnae011; Chiarella ML and Borgese M, 'Data Act: New Rules about Fair Access to and Use of Data' (2024) 10 Athens JL 47; Stuhldreier MA, 'Fostering Innovation by Utilising Big Data: The Data Act and the Risk of Quasi-Exclusivity Reinforcing Data Lockups' in Nadia Naim (ed), *Developments in Intellectual Property Strategy* (Springer 2024); Colangelo G and Borgogno O, 'Shaping Interoperability for the Internet of Things: The Case for Ecosystem-Tailored Standardisation' (2024) 15 European Journal of Risk Regulation 137; Hennemann M and others, *Data Act: An Introduction* (1. Auflage, Nomos 2024); Picht PG, 'Caught in the Acts: Framing Mandatory Data Access Transactions under the Data Act, Further EU Digital Regulation Acts, and Competition Law' (2023) 14 Journal of European Competition Law & Practice 67; Leistner M and Antoine L, 'IP Law and Policy for the Data Economy in the EU' (2023) 17 Economics 1; Schweitzer H, Metzger A, 'Data Access under the Draft Data Act, Competition Law and the DMA: Opening the data treasures for competition and innovation? (2023) GRUR Int. 337; Metzger A, Schweitzer H, 'Shaping Markets: A critical evaluation of the draft Data Act' (2023) 1 ZEuP 42; Paal F, 'Access to Data in the Data Act Proposal' (2023) ZfDR 249; Kerber (n 17); Podszun (n 17); Drexl J and others (n 17).

21 Different qualities of data have been discussed as the components of data usability in technical, managerial, and economic literature. See eg Chen B, 'What is Data Usability? Definition, Examples, and Best Practices' (*Metaplane*, 29 May 2023) <https://www.metaplane.dev/blog/data-usability-definition-examples>.

machine learning, business analytics and decision-making, or other applications. Data usability is enhanced as a dataset[22] is processed within the data value chain, progressing from raw sensor data to a state more closely aligned with the pursued objective. Given that data usability is defined and assessed relative to the purpose of data processing, it is not a fixed characteristic that can be universally defined.[23]

8 The purpose of each data processing step within the data value chain is to improve data usability qualitatively and/or quantitatively. The results of each processing phase can be assessed in terms of qualitative and quantitative benchmarks, such as 'accuracy' and 'precision'. Table 1 (annex) presents a non-exhaustive list of major types of processing[24] sensor-generated data: value calibration, data value de-noising, missing data value imputation, data selection, and data extraction.[25] It also illustrates the respective contributions of these steps to data usability with respect to the assumed objectives.

II. Data pre-processing

9 Calibration[26] and de-noising are foundational data processing steps that are crucial for data interpretability and usability. Usually performed early in the data value cycle, these steps are generic in nature compared to purpose-specific data transformations and enhance the results of the follow-on steps. These generic steps can be considered as data pre-processing and are briefly explained below, given their relevance to the scope of the Data Act.[27]

1. Calibration and data accuracy

10 As sensors interact with the physical environment, they generate electrical signals, which are digitised into raw data. For example, a temperature sensor generates signals that are converted into raw data, not direct temperature values. However, the link between this raw data and understandable units like degrees Celsius can be unclear. To determine this relationship, a formula[28] is required to convert the raw sensor data into a form with an interpretable unit of measurement. This formula can be obtained through a process called calibration, a procedure of comparing the raw sensor data with that of a calibration standard[29]. This process typically involves placing the sensor in a controlled environment with stable temperatures at selected levels, measuring the actual temperature values with the standard, and recording the raw sensor data to establish a relationship and derive a calibration formula.

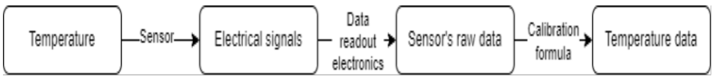

Figure 1: A schematic view of the conversion process from physical temperature to temperature data

11 The outcome of the conversion is characterised in terms of the accuracy of data, a quantitative measure of the difference between raw data values and their true values. Accuracy serves as a quantitative measure of data usability – improved accuracy denotes higher usability. Such a difference is known as a systematic error and, therefore, a lower accuracy value indicates better accuracy.[30] Several

22 A dataset can include data from different sources, as well as metadata.

23 For example, if A's goal is to sell raw temperature sensor data to B, who needs it for data analytics aimed at product improvement, the usability of such data would be higher for A than for B.

24 These steps can be, but do not have to be, performed consecutively. While calibration and de-noising are almost a must-have for sensor data, other steps are optional and some steps might need to be iterated.

25 Some may categorise de-noising, missing value imputation, and selection into data cleaning/cleansing as they detect and correct or remove corrupt or inaccurate data values. On the other hand, extraction and other techniques, including discretisation and normalisation, can be referred to as 'data transformation'.

26 Yeong DJ et al,. 'Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review' (2021) 21(6) Sensors 2140, https://doi.org/10.3390/s21062140.

27 Below at C.I. While technical literature uses the term 'data pre-processing', there is no fixed catalogue of operations falling within this category. In this paper, we apply the term 'data processing' as encompassing any data processing activity required to achieve the goal and refer to certain generic operations – typically necessary to enable purpose-specific use of data, such as calibration and de-noising – as 'pre-processing'. As discussed in part III, the Data Act is not explicit on the types of data processing considered as 'pre-processing'.

28 The formula can consist of one or more equations, taking raw sensor data as input and providing an output with an interpretable unit of measurement (e.g., degrees Celsius). This formula may also be visually represented in a graph, featuring a curve that illustrates the correspondence between the raw sensor data value and the standard's data value.

29 Fraden J, *Handbook of Modern Sensors: Physics, Designs, and Applications* (5th edn, Springer 2016) 24-26.

30 In this context, accuracy is, counterintuitively, defined as a measure of error rather than a positive feature. It is typically expressed either as an absolute term (e.g. ±5 for temperature data) or equivalently as a percentage of the sensor's full scale (e.g. ±5% if the full scale is 100). Fraden

factors can influence the accuracy of calibrated data, including the accuracy of the calibration standard, the accuracy of the calibration formula, and the sensor's sensitivity to environmental changes, such as temperature variations. While there is no universal standard for the minimum acceptable accuracy, it is determined relative to a specific objective. For instance, if calibrated data is utilised only to indicate outdoor temperatures, worse accuracy might be more tolerable compared to situations where the data is employed to monitor temperature-sensitive plants in a laboratory environment.

2. De-noising and data precision

12 Noise, also known as random or stochastic error, is a type of error distinct from the systematic error as the above-described measure of accuracy. Noise is unavoidable[31] and uncorrelated with the physical phenomenon being measured. Since a sensor first produces electrical signals, any environmental factor that interferes with the sensor or the supporting electronics can induce noise in the signal[32], and consequently, in the sensor's digitised raw readings.[33] Given that noise is uncorrelated with the physical phenomenon, it cannot be calibrated away, and thus, it remains in the calibrated data.

13 The level of noise is measured in terms of precision.[34] Without noise, the data value should stay constant if the physical phenomenon being measured is also unchanged. However, noise causes the data value to fluctuate around that constant level. Precision measures the amount of fluctuation in the sensor data (either raw or calibrated, given that noise passes freely without reduction due to the conversion of raw data to calibrated data). Thus, the more fluctuation, the lower the precision.

14 Calibration and de-noising are the foundational steps within the sensor data processing chain. Figure 2 illustrates a typical data processing workflow using temperature sensor data as an example that can be extrapolated to other types of sensor-generated data, considering their measurement specifics.

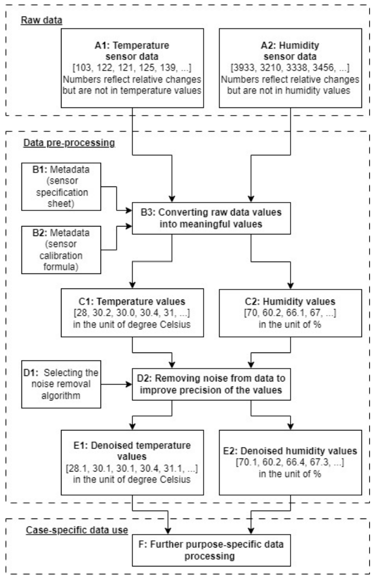

Figure 2 Data processing workflow exemplified by temperature and humidity sensor data[35]

III. An optimal technical state of acquired data

15 Given that data is frequently acquired elsewhere, the question arises about the optimal state in which data should be obtained to allow for its meaningful processing in a given use case. The answer depends on technical and practical considerations within a specific context. Essentially, the choice is between obtaining raw data or data that has undergone generic processing steps (that is, calibration and de-noising in the case of sensor-generated data). In principle, raw data can be usable if accompanied by sufficient metadata. Raw or generically processed data possesses greater potential for fulfilling various purposes and producing diverse outcomes. In some cases, sharing data in a pre-processed form can be both commercially and technically suitable for both the data holder and the data user. While obtaining purpose-specific processed data can be an option when purposes align, even minor differences may lead the data recipient to prefer conducting pre-processing themselves. Thus, there should not be a bias that the more data is processed, the greater its usability. In reality, the data user knows its own needs best and would be better off with data that allows for the most flexibility and diversifiable results.

(n 29) 39-42.

31 ibid 243-244.

32 ibid 237-238.

33 Some sources of noise include electromagnetic interference from a power converter that is connected to the circuit board hosting the sensors, and random vibrational movements of electrons (the carriers of the sensor's signal) which are proportional to temperature and thus called the 'thermal noise'. Apart from factors related to the electronics, natural noise can be introduced, for instance, by turbulent flow around a pressure sensor during air pressure measurement, or by ambient noise from pedestrians and cars when measuring sound levels by using an audio receiver.

34 Sometimes a related but distinct term 'reproducibility' is used as a measure of noise in the sensory context.

35 In this scheme, Bs and Ds are processing steps, while As, Cs and Es are the data states.

16 In summary, this part underscores that data usability is a characteristic of data defined and assessed in relation to a specific purpose. Two foundational pre-processing steps of sensor-generated data explained above – calibration and de-noising – have specific benchmarks and measures associated with data usability, namely accuracy and precision. These attributes denote continuous qualities that can vary in degree, while the acceptable level can be determined in relation to the intended purpose of data usage.

C. How does the Data Act account for data usability?

17 The key insight from the preceding section is that mere data accessibility does not ensure the realisation of its value in a given use case. Equally important is the technical state of the data, enabling its further processing. In the following, we analyse how the Data Act factors in this aspect.

I. 'Pre-processed data'

1. Definition

18 Recital 15 clarifies that the scope of the Data Act covers both:

19 data 'which are not substantially modified, meaning data in *raw* form, also known as source or primary data which refer to data points that are automatically generated without any further form of processing', and

20 'data which have been *pre-processed* for the purpose of making them *understandable* and *useable* prior to subsequent processing and analysis' (emphasis added).

21 The latter category 'includes data collected from a single sensor or a connected group of sensors for the purpose of making the collected data *comprehensible for wider use-cases* by determining a physical quantity or quality or the change in a physical quantity, such as temperature, pressure, flow rate, audio, pH value, liquid level, position, acceleration or speed' (emphasis added). For those not tech-savvy, this might require an explanation. Recall that the Data Act defines data as a 'digital representation of acts, facts or information'.[36] In the case of sensor-generated data, such representations result from the conversion[37] of an analogue signal to a digital signal[38] taking place within a converter that can be located in a device or on a server. Raw sensor data – data resulting from the conversion of an analogue signal to digital – is indeed not comprehensible or usable because such data does not represent the physical values/quantities. For that, data should be calibrated,[39] which corresponds to the wording of Recital 15: 'determining a physical quantity or quality or the change in a physical quantity'. If we look at Figure 2 and try to locate the type of data pre-processing described therein, it would be step B1 – converting raw values to meaningful values.

22 If calibration of data values only exemplifies data pre-processing, as signalled by the wording 'includes', what other technical operations on data can count as 'pre-processing'? Such operations would, in effect, delineate the scope of the rights and obligations under the Data Act as far as the technical state of data is concerned. As explained in Part II, data processing entails a sequence of operations that progressively enhance data usability, bringing it closer to the technical state aligned with the intended purpose. Where exactly did the legislature intend to delimit the scope of the Data Act when introducing the notion of 'pre-processed' data? The concretisation of making data 'comprehensible for wider use-cases' in Recital 15 presupposes data-processing steps *generic* in nature, as opposed to purpose-specific data processing. Besides calibration, this could potentially include de-noising.

2. Insubstantial investment

23 While Recital 15 does not provide other examples of pre-processing operations that improve data usability or comprehensibility, it does place a constraint on data pre-processing: such pre-processing 'should not be interpreted in such a manner as to impose an obligation on the data holder to make substantial investments in cleaning and transforming the data'. Thus, theoretically, it may also include data transformation beyond calibration, such as 'cleaning' (step D2 in Figure 2),[40] as long as

36 '…and any compilation of such acts, facts or information, including in the form of sound, visual or audio-visual recording' (art 2(1) Data Act).

37 While the Data Act does not define the terms 'generate', 'obtain', and 'collect' (data), all these activities should be interpreted – in line with the definition under art 2(1) Data Act – as acts of transforming real acts and facts into their digital representation, such as by converting an analogue signal into a digital signal in the case of sensor-generated data.

38 See Figure 1 and the accompanying explanation.

39 For explanation, see above at B.II.1.

40 As mentioned earlier, data cleaning/cleansing can be understood to encompass processes that detect, correct,

this would not entail 'substantial investment'.

24 If these criteria were to be applied to delineate the scope of the data holder's obligation to make data available, certain aspects require clarification. First, the characteristics of the technical state of data (usable/understandable) and the data holder's investment in data processing (substantial) denote continuous qualities that vary by degree, which prompts the question of the applicable threshold. Second, such criteria are relative – what constitutes comprehensible or usable data, or substantial investment, depends on a perspective or a point of reference. For data usability, the point of reference is the purpose of data processing. By which standard is the substantiality of investment to be determined, and by whom? Furthermore, how do these criteria correlate? Since it cannot be generally presumed that making data understandable and usable always requires an insubstantial investment, how should tension be resolved if making the data usable, as deemed by the data user, requires an investment deemed substantial by the data holder? The greater the misalignment between the criteria of data usability and the insubstantiality of investment, the greater the legal uncertainty regarding the scope of obligations for making data available, and the greater the potential for disputes between the data holder and the product/service user.

25 To explore this potential, let us first consider the practical aspect: How significant are the expenses associated with data pre-processing? The most straightforward case is providing product or service data in a 'commonly used format'[41] which would typically entail trivial costs.[42] Concerning calibration, the tendency is also rather towards an insubstantial cost. Sensor and device manufacturers routinely verify their product's sensors for performance, including for quality assurance.[43] Therefore, it is assumed that data holders should be able to provide calibrated data without substantial additional – i.e. discounting necessary equipment expenses – costs. However, it is worth noting that the cost of calibration can vary depending on calibration quality, which in turn impacts data accuracy and usability. For instance, data accuracy may suffer if calibration is done by a layperson in a poorly controlled environment and with a subpar calibration standard. In contrast, device or sensor manufacturers would usually be in a position to achieve superior results due to better standards, equipment, and a better-controlled environment at their disposal.

26 The question may further arise about the expenditures that are relevant for evaluating the substantiality of investment. Would the costs incurred by a device- or sensor manufacturer to purchase calibration equipment count? For instance, inertial sensors like an accelerometer or a gyroscope can be calibrated with or without precision equipment. While calibration can be performed in both cases, the cost for precision equipment is undoubtedly higher, resulting in better accuracy. Furthermore, some cases might require sensor re-calibration to ensure accuracy throughout the product's lifetime.[44]

27 In the case of de-noising, a device's circuit board could be designed to reduce the level of noise from within the circuit. However, additional de-noising software can deal with noise from unpredictable sources. The factors impacting the cost of de-noising include the choice of the de-noising methods, as well as the complexity and number of de-noising algorithms. The quality and its acceptable level may vary depending on the purpose, influencing the cost of de-noising.[45] Thus, if a device or sensor

or remove corrupt or inaccurate data values, such as de-noising, imputation of missing values, and selection. See above at B.II.2.

41 Which formats are 'commonly used' can vary depending on the context and purpose, and it can be interpreted within the relevant industries or technical communities. The guidance on this term, which is also employed in the General Data Protection Regulation, may provide further insights. See Article 29 Data Protection Working Party, 'Guidelines on the Right to Data Portability, 16/EN WP 242' <https://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp242_en_40852.pdf> 13 (clarifying that 'the GDPR does not impose specific recommendations on the format of the personal data to be provided' and emphasising the purpose-bound approach to interpretation).

42 The term 'format' in this context refers to structures such as Excel (xlsx, xls), CSV, SQL, Parquet, JSON, and XML, each of which has own standard, at a minimum, indicating how the data should be stored and read.

43 Sensors are usually sold with product specification sheets detailing calibration results.

44 While it is impractical to re-calibrate typical personal-use products such as refrigerators, watches, and phones, in the case of industrial equipment – especially where accuracy is crucial for safety and/or where the product's sensors may shift significantly over time – re-calibration is necessary.

45 Different de-noising methods are described in literature. See eg Buades A, Coll B and Morel JM, 'A Review of Image Denoising Algorithms, with a New One' (2005) 4(2) Multiscale Modeling & Simulation 490, https://doi.org/10.1137/040616024; Banos O and others, 'On the Use of Sensor Fusion to Reduce the Impact of Rotational and Additive Noise in Human Activity Recognition' (2012) 12(6) Sensors 8039, https://doi.org/10.3390/s120608039; Du J, Gerdtman C and Lindén M, 'Signal Quality Improvement Algorithms for MEMS Gyroscope-based Human Motion Analysis Systems: A systematic review' (2018) 18(4) Sensors 1123, https://doi.org/10.3390/s18041123.

manufacturer de-noises data for their purposes, the quality level may or may not align with the data user's needs.

28 Accordingly, while it would be desirable for 'pre-processed' data to include calibrated and de-noised data, the limitation that pre-processing can only involve 'insubstantial investment' might be suboptimal from a data usability perspective. Alternatively, if the device manufacturer provides raw sensor data along with the relevant metadata[46] – information necessary for leveraging techniques such as sensor fusion for de-noising – such data can, in principle, be converted into calibrated and de-noised data. Nevertheless, it would be advantageous for data users if the device manufacturer, with a better understanding of the device and access to a larger sensor network for sensor fusion, could provide de-noised data.

29 In summary, it is not entirely clear how the criteria of insubstantial investment and usable/understandable data introduced by Recital 15 align and should be cumulatively applied to delineate the scope of the Data Act. The minimal prerequisites for data usability – calibration and de-noising – already suggest that the notion of pre-processed data may involve a trade-off between data usability and the compliance with the yet-to-be-clarified requirement of 'insubstantial investment'.

30 The question arises as to whether the statement in Recital 15, stipulating that both raw and pre-processed data 'fall within the scope of this Regulation', implies that the latter necessarily falls within the scope of the obligations to make data available, as considered next.

II. 'Readily available data'

1. The definition

31 While the term 'pre-processed data' appears only in Recital 15 Data Act, the data holder's obligations to make data available under Articles 4 and 5 refer to 'readily available data'.[47] The latter is defined as 'product data and related service data that a data holder lawfully obtains or can lawfully obtain from the connected product or related service, without disproportionate effort going beyond a simple operation'.[48] On the surface, this definition does not specify the technical state of such data – whether 'readily available data' is confined to raw data or can/must encompass pre-processed data. This question directly bears on the scope of the data holder's obligations. An indication that the fulfilment of this obligation can involve data processing is found in Recital 47, which explains that the cost of making data available includes technical costs, comprising 'the costs for processing, necessary to make data available, including costs associated with the formatting of data'.

2. Can data be processed before it is obtained from a product or service?

32 To understand the technical state in which data should be made available, let us consider what 'obtaining' data by the data holder refers to, bearing in mind that only 'simple operations' would count. The act of 'obtaining' data technically refers to the transmission of data from a device to the data holder's server. For related services, the data resides on either the service provider's server or the server operating the service. In which state does a data holder typically obtain data *from* a connected product or related service? And can any type of data (pre-)processing take place within the device at all before data is obtained from a product through transmission to a server? The decision-making of relevant entities in this regard can be influenced by different technical and practical considerations. As explained earlier, the conversion from an analogue to a digital signal typically takes place within the device. Subsequent data processing on a server allows the data holder to make changes to the data processing chain at any time.[49] Processing within a product offers benefits of offline use, cost savings on server computation, and pre-aggregation of data to reduce network traffic fees. However, if the product allows operation offline, then all steps relevant to the product's offline functionality have to occur within the product.[50]

46 On this option, see below at C.4.

47 art 4 Data Act. This notion was first introduced in the Council's version (n 18). Notably, in the Council's negotiation mandate, 'readily available data' was also in Article 3(1), which lays down an obligation to design products or provide services in a way to make product data and related service, in the wording of the final version, data 'directly accessible to the user'.

48 art 2(17) Data Act.

49 For instance, if the product manufacturer/service provider intends to implement a new function or improve an existing function of a product/service.

50 For example, the data processing chain of a sports watch may span over three computational entities – the watch, a mobile phone connected to the watch via Bluetooth, and a remote server connected to the mobile phone via the Internet. Since the watch is designed to work in standalone mode, it processes sensor data to support all its functions, such as calculating and displaying the heartbeat rate. The

33 Thus, in principle, data that can be obtained from a connected product or a related service is not confined to raw data but can extend to data that has undergone any transformations performed in-device or on a server.

34 To define which data falls within the meaning of 'readily available data', two cumulative criteria need to be further considered: the obtaining of such data should (i) be lawful, and (ii) should not involve 'disproportionate effort going beyond a simple operation'. Let us address each in turn.

3. In which technical state is product and service data 'lawfully obtained'?

35 Of relevance to this inquiry is whether the conditions of lawfully obtaining connected product or related service data explicitly or implicitly suggest any particular technical state of data or impose any restrictions thereon.

36 The sources of 'lawful obtaining' of data are exemplified in Recital 20: 'such as by means of the connected product design, the data holder's contract with the user for the provision of related services, and its technical means of data access'. Thus, both technical/factual means (via product design)[51] and a contractual basis for obtaining data would fulfil the condition of data being lawfully obtained, given that 'such as' indicates non-cumulativeness of conditions. Before the Data Act, the initial allocation of rights in sensor-generated data had not been statutorily prescribed, at least not at the EU level, leading to the frequent confusion between de facto exclusive control over data by device manufacturers and legal ownership of data.[52] In this context, Article 3 Data Act can be viewed as the first attempt at the EU level to statutorily allocate access and usage rights to users of connected products or related services. Furthermore, the Data Act appears to strengthen[53] the user's position by mandating that 'a data holder shall only use any readily available data that is non-personal data on the basis of a contract with the user'.[54] However, this limitation would not apply to data processing occurring *within* the product or service, i.e. before data is obtained *from* a product or service, which is the reference point of the definition of 'readily available data'.

4. Which operations should be deemed as 'disproportionate' and 'going beyond simple'?

37 The qualifiers 'disproportionate' and 'simple' serving as the delineators for 'readily available data' – consequently, the obligation to make data available – necessitate clarification. Given their relative character, questions inevitably arise concerning the threshold for simplicity and the point of reference for proportionality. For instance, if conversion from an analogue to a digital signal already constitutes a simple operation, should it be sufficient for the data holder to deny a claim for making available data in any (pre-)processed form? As discussed in Part II, every subsequent data-processing operation can vary in terms of both technical complexity and costs involved. Where is the line meant to be drawn? One could suggest that the rule of thumb would apply in a given situation, in light of its circumstances. However, this may jeopardise the objectivity of assessment and legal certainty. Furthermore, questions arise as to whether the criteria of 'disproportionate effort' and 'a simple operation' pertain solely to the act of obtaining data *from* the product or service, or if they are also applicable to data processing operations occurring *within* the product or service. Either way,

mobile phone, equipped with the watch's application, may process heartbeat rate data to display a performance review with historical data as one of the application's offline functions. However, certain functions, such as exercise recommendations, may require an internet connection to the remote server for aggregating and processing the watch user's and other users' historical data. Such 'division of labour' in the data processing chain is determined by product design – whether a function should work online and/or offline – and variations in computational and data storage capabilities among these three entities.

51 Notably, rec 20 explicitly states that a manufacturer's control over the generation of and access to data through the product technical design does not confer legal rights to such data in a manufacturer. In the wording of rec 20: 'In many sectors, manufacturers are able to determine, through their control of the technical design of the connected products or related services, what data are generated and how they can be accessed, despite having no legal right to those data.' Thus, while obtaining data by way of a product's technical design is deemed to be lawful, it does not translate into legal rights over such data.

52 Drexl J and others, 'Data Ownership and Access to Data – Position Statement of the Max Planck Institute for Innovation and Competition of 16 August 2016 on the Current European Debate', <https://www.ip.mpg.de/fileadmin/ipmpg/content/stellungnahmen/positionspaper-data-eng-2016_08_16-def.pdf>; Kim D, 'No One's Ownership as the Status Quo and a Possible Way Forward: A Note on the Public Consultation on Building a European Data Economy' (2018) 13 Journal of Intellectual Property Law & Practice 154.

53 But see Kerber (n 17) (assuming that the users would 'agree in this initial contract that the manufacturers or data holders get all rights to use and commercialize this non-personal data for the entire lifetime of the IoT device').

54 art 4(13) Data Act.

what would be the consequences if the data holder considers the efforts or operations involved as going beyond 'simple' and 'proportionate'? Could this potentially serve as a backdoor to deny access to data, given that there is no obligation for products or services to be designed in such a way that 'readily available data' only involves 'simple operations' and 'proportionate efforts'?

38 The notions of 'disproportionate effort' and 'simple operations' within the definition of 'readily available data' may invoke 'significant investment' as a delineating criterion of 'pre-processed' data falling within the scope of the Data Act, according to Recital 15. While there is no explicit linkage between Articles 4(1) and 5(1) and Recital 15 Data Act, an interpretation in light of the explanations in the Recital suggests that the data holder's obligations to make data available can encompass data in a calibrated or further (pre-)processed form, to the extent that such processing does not involve 'substantial investment', supposedly aligned with the notions of 'beyond a simple operation' and 'disproportionate effort'. As noted above, the relative nature of these qualifiers introduces some indeterminacy in interpreting the scope of data-sharing obligations.

39 To summarise, on the surface, data-sharing obligations under the Data Act do not explicitly require data holders to make available data in any 'pre-processed' form. The conversion from an analogue to a digital signal alone – i.e. the provision of raw data – can be argued to suffice for complying with the definition of 'readily available data'. The relevance of the reference to 'pre-processed' data laid down in Recital 15 for the obligations of data holders under Articles 4 and 5 remains open to interpretation.

III. 'Inferred and derived' data and information

40 The notion of 'readily available data' is contrasted with information and data 'inferred' or 'derived' from connected product or related service data, which 'should not be considered to fall within the scope' of the Data Act.[55] Notably, the rationale behind this delineation is based on the involvement of 'additional' investment and 'proprietary' algorithms and software. As articulated in Recital 15, inferred or derived information/data constitute 'the outcome of additional investments into assigning values or insights from the data, in particular by means of proprietary, complex algorithms, including those that are a part of proprietary software'. Situations to which Recital 15 refers would typically involve data analytics, usually performed on aggregated data, including through sensor fusion.[56] By 'assigning values', it hints at the use of data as input for developing machine learning (ML) models, while 'insights' may refer to predictions generated by ML models that enable the functionality of ML-based systems and applications.

41 References to 'additional investment' in data analytics, 'proprietary' algorithms, and 'proprietary' software indicate an intention to safeguard the economic interests of the data holders. This rationale aligns with the conventional logic of intellectual property (IP), where restricting third-party access to and usage of the 'fruits' borne by investment is assumed to incentivise innovation, which in this context may translate into innovation in the field of data analytics and ML. While this cannot be read as conferring any exclusive rights in derived/inferred data, it is notable that they are treated as 'untouchable' by default due to the very reason of being derived through (potentially) 'proprietary' algorithms and software – the mere fact that inferred/derived data can result from 'proprietary'[57] algorithms and software is deemed sufficient to limit restrict access to such information/data.

42 Furthermore, inferred or derived 'data could include, in particular, information derived through sensor fusion, which infers or derives data from multiple sensors, collected in the connected product, using proprietary, complex algorithms and which could be subject to intellectual property rights'.[58] The clause 'which could be subject to intellectual property rights' logically refers to 'data' or 'information', even though it grammatically correlates with 'sensor fusion' (which, as such, cannot be 'subject to' IP rights). One may wonder what kind of data or information resulting from sensor fusion could be protectable by IP rights. A plausible candidate might be an ML model as part of a patentable invention, but a model is not 'information'. Trade secrets do not come into question because they are not considered IP 'rights'.[59] While the linkage to IP is not articulated,

55 rec 15 Data Act.

56 For an explanation, see Table Annex.

57 The source of this proprietary status of algorithms is not quite clear, given that, as such, they cannot be protected by copyright or patents. Recital 15 also uses more cautious wording stating that 'algorithms' can be 'part of proprietary software'.

58 rec 15 Data Act.

59 rec 16 of Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, OJ L 157, 15.6.2016, p. 1–18. See also Proposal for a Directive of the European Parliament and of the Council on the protection of undisclosed know-how and business information (trade secrets) against their unlawful

the legislature seems to have presumed - bluntly and pre-emptively - its limiting effect on access to data.

43 In summary, dichotomies between substantial and insubstantial, simple and complex are applied to delineate the scope of the Data Act: raw data is defined as data that is 'not substantially modified', simple operations are a criterion of 'readily available data', '(in)substantial investment' is a criterion of 'pre-processed data', and derived/inferred data or information is that which results from 'complex' algorithms and additional (i.e. beyond insubstantial) investment. The challenge is that these criteria exist along a continuum with some range of legal uncertainty in between where it can be unclear whether a process might be rather simple or complex, or whether the associated investment or effort might be more or less substantial. If the motivation behind excluding substantial investment from the scope of the data-sharing obligation stems from protecting economic interests, a relevant reference point would be the definition of investment under the Database Directive, which includes 'the deployment of financial resources and/or the expending of time, effort and energy'.[60] The question may still arise regarding the investment that should be deemed relevant in this context, such as whether the expenditure associated with developing a data-processing algorithm would fall within this category.

IV. Metadata

44 Another latecomer to the Data Act, motivated by data usability considerations, was the notion of 'metadata' as part of access and usage rights and respective obligations, first introduced by the Council of the EU.[61] Defined as 'a structured description of the contents or the use of data facilitating the discovery or use of that data',[62] metadata should include inter alia 'basic context and timestamp, to make the data usable, combined with other data'.[63]

45 Notably, in the case of the obligation to make product data and related service data directly accessible to the user by design, metadata is supposed to be *included* in the connected product or related service data.[64] In contrast, in the case of the obligations to make data available to the user or third parties, metadata should be provided *in addition* to the 'readily available data'.[65] For metadata to be literally and technically 'included' in the connected product or related service data to be made directly accessible by product or service design, such metadata first needs to be placed within the same file[66] as product or related service data, located either in a product,[67] or on a remote server.

46 Metadata is an umbrella term – an exhaustive categorisation of information and data falling within this notion in all possible use scenarios is unfeasible. The Data Act adopts a purpose-based approach to determining the relevant metadata subject to data-sharing obligations when it emphasises that the 'relevant' metadata is data 'necessary' for interpreting and utilising the connected product or related service data for further purposes.[68]

47 The question may arise whether the Data Act imposes any constraints on the scope of metadata subject to the data holder's obligation to make such data either

acquisition, use and disclosure COM(2013) 813 final (28.11.2013) 3 (noting that trade secrets are 'not protected as a classical [intellectual property right]'). See also art 49(e) and (f) Data Act, distinguishing between the impact on intellectual property rights and on trade secrets as part of an evaluation of the Data Act.

60 Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases OJ L 77, 27.3.1996, p. 20–28, rec 40.

61 Namely, rec 14a and 56; art 3, 4, 5, 14, 17, and 19 of the version of the Council of the EU (n 18). The Commission's proposal referred only to metadata generated by the customer's use of a service which should be portable according to the provisions on switching between data processing services.

62 art 2(2) Data Act.

63 rec 15 Data Act.

64 art 3(1) ('Connected products shall be designed and manufactured, and related services shall be designed and provided, in such a manner that product data and related service data, including the relevant metadata necessary to interpret and use those data, are, by default, easily, securely, free of charge, in a comprehensive, structured, commonly used and machine-readable format, and, where relevant and technically feasible, directly accessible to the user.').

65 Both arts 4(1) and 5(1) Data Act state that 'the data holder shall make available readily available data, *as well as* the relevant metadata' (emphasis added).

66 Timestamps – an example of metadata mentioned in Recital 15 – are usually placed side-by-side with sensor values in one data file. The decision of whether to store metadata in the same file as the data depends on technical and practical factors. Opting for separate files for data and metadata allows for avoiding redundant metadata duplication, enhancing memory efficiency, and maintaining metadata consistency and currency.

67 It might not be even feasible to make all relevant metadata 'directly accessible' from on-device data storage or from a remote server at any point in time, already for the reason that the product manufacturer or service provider may not know all purposes for which users might need metadata for the subsequent data uses to fulfil the obligation under art 3(1) Data Act. See also below (n 78).

68 rec 15 and 20; art 3(1), 4(1), 5(1) Data Act.

directly accessible[69] or readily available.[70] While the provisions referring to metadata do not contain any direct, specific indication regarding the scope of metadata,[71] one could suggest that the legislature might not have deemed such limitations as necessary because it had already included safeguards for trade secrets, potentially embedded within metadata, to protect the interests of trade secret holders, who may or may not be data holders. Indeed, the protection of trade secrets is factored into the data access and usage rights.[72] While data-sharing obligations extend to trade secrets, they presuppose only inter partes disclosure,[73] subject to contractual and technical measures agreed upon with the trade secret holder.[74] This concerns sharing product and service data, along with metadata, with product/service users, as well as third parties.[75] Furthermore, a trade secret holder can, under some conditions, withhold, suspend, or refuse to share trade secrets.[76] It is worth noting that the mandatory sharing of trade secrets – even when subject to safeguarding measures to protect confidentiality – does constitute a limitation on the trade secret holder's rights, in the sense that it restricts their discretion in deciding with whom to share trade secrets and whether to share them at all.[77]

48 Furthermore, the question arises: What if the data holder simply does not have metadata – or does not have *all*[78] of 'the relevant metadata necessary to interpret and use those data'?[79] This issue is seemingly not regulated under the Data Act. Considering that data usability is a relative concept, the metadata at the disposal of the data holder might make product or service data more usable but not ideal from the prospective data user's perspective. Should a dispute between the data holder and the user arise in this regard, the user can contest the fulfilment of the obligations before a dispute settlement body or 'seek an effective remedy' before a Member State's court or tribunal.[80]

V. An interim conclusion

49 The overall approach taken by the Data Act regarding data usability can be characterised as establishing minimum conditions for data utilisation. From a technical perspective, even if only raw sensor-generated data is made available, the inclusion of all 'relevant' metadata should enable its utilisation. The practicality, feasibility, and efficiency of this approach would depend on the specifics of the scenario and the technical and economic capabilities of the data user. From a legal perspective, the technical state of shareable data – hence, the scope of data-sharing obligations – are challenging to delineate due to the ambiguous legal criteria examined in this part. This ambiguity introduces the potential for disputes if such limiting criteria are interpreted in a way jeopardising data utilisation. Considering that the latter is the very purpose of the Data Act, data usability may and should carry significant weight in the legal assessment in contested cases.

69 art 3(1) Data Act.

70 arts 4(1) and 5(1) Data Act.

71 Apart from an exemplifying reference to the data's 'basic context and timestamp' (rec 15). From a technical perspective, contextual information should encompass the sensor's location, which is particularly useful in cases where multiple sensors detect the same physical phenomenon, as well as the sensor's specifications, typically including details such as calibration accuracy, sensor precision, etc.

72 rec 31; arts 4(6)-(8) and 5(9)-(11) Data Act.

73 rec 31 Data Act: 'While this Regulation requires data holders to disclose certain data to users, or third parties of a user's choice, even when such data qualify for protection as trade secrets, it should be interpreted in such a manner as to preserve the protection afforded to trade secrets under Directive (EU) 2016/943.'

74 arts 4(6) and 5(9) Data Act. In particular, such agreed measures directed at the preservation of the 'confidentiality of data considered to be trade secrets' include 'model contractual terms, confidentiality agreements, strict access protocols, technical standards and the application of codes of conduct' (rec 31 Data Act).

75 arts 4(6)-(8); 5(9)-(11); 6(2)(c), (g); 8(6); Data Act.

76 arts 4(6)-(8) and 5(9)-(11) Data Act.

77 This follows from the trade secret holder's (voluntary) consent being the condition for the lawful acquisition, use, and disclosure of trade secrets (art 4 of Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, OJ L 157, 15.6.2016, p. 1–18).

78 In practice, manufacturers may not have at their disposal all the metadata relevant to the needs of the prospective data users, as the assessment of the relevance of certain metadata can differ between a data recipient and a manufacturer. For instance, if a manufacturer utilises a temperature sensor solely to generate an on-off signal, indicating whether the temperature exceeds a specific threshold, the manufacturer may not have the metadata, e.g. concerning the sensor's accuracy and calibration outside the temperature range of interest. However, this incomplete information may become an issue of missing metadata if a data recipient decides to use the sensor data for recording temperatures beyond the manufacturer's range of interest.

79 recs 15 and 20; arts 3(1), 4(1), and 5(1) Data Act.

80 art 10(13) Data Act.

D. Normative considerations

50 The Data Act serves as a regulatory instrument aiming to 'maximise the value of data in the economy and society'.[81] In light of its instrumental nature, the validity of the Data Act hinges on how well it aligns with the intended objectives. Furthermore, its legitimacy is contingent upon its adherence to the balance of interests as a fundamental principle of policymaking.

I. Uncertainty within the 'means-ends' relationship

51 According to the intervention logic outlined by the Commission in its ex-ante impact assessment, the Data Act should maximise the value of data, particularly by increasing the availability of data for innovation.[82] In this logic, the new access and use rights, along with the corresponding obligations to make data available, specifically target 'legal uncertainty for consumers and businesses concerning data access and use' and 'abuse of contractual imbalances with regard to data access' in the B2B and B2C context.[83]

52 As discussed, data usability was not envisaged in the initial proposal but was addressed at a relatively late stage in the legislative process. While several provisions of the Data Act bear on data usability, the overall impression is that it lacks thorough consideration. In an attempt to remedy the shortcomings of the original proposal, a number of concepts were introduced – 'source or primary' data, 'data in raw form [...] which are not substantially modified' distinguished from 'pre-processed data' which does not involve 'substantial investment' in processing, contrasted with 'readily available data' delineated by 'a simple operation' and 'disproportionate effort', yet distinct from 'derived' or 'inferred' data or information defined by 'additional investment' and the complexity of an algorithm. This terminology appears convoluted, lacks coherence and clarity, and undermines legal certainty in defining the scope of data falling within the obligation to make data available. Furthermore, comparing the notion of 'readily available data' under Articles 4(1) and 5(1) with making data 'directly accessible' under Article 3(1) Data Act, the criteria of simplicity of operations or proportionality of effort, applicable to the former type, might lead to discrimination between the scope and technical states of data 'directly accessible' vs. made 'readily available' to users.

53 Given the relative nature of the legal concepts involving relative qualifiers 'substantial', 'simple', and 'disproportionate', a certain middle ground appears inevitable, which introduces uncertainty. While courts may eventually need to establish a threshold and develop a corresponding test, having guidance clarifying the criteria regarding the technical state of data subject to the obligation of making data available could have streamlined data access. The absence of a specific[84] or general[85] mandate vested by the Data Act in the European Commission or the European Data Innovation Board suggests that the legislature had not anticipated uncertainty regarding the technical aspects of data usability. The European Commission could proactively address this issue by developing guidance clarifying these criteria and what exactly they imply for the technical state of data subject to the obligation of making data available. To the extent that ambiguity surrounding the applicable threshold can be leveraged to interpret data-sharing obligations narrowly, compromise data usability, or give rise to disagreements over the technical state of data between the data holder and the user or third-party data recipients, these qualitative criteria may jeopardise the benefits anticipated from the Data Act.

II. An alternative approach?

54 The Data Act has already faced criticism for the overall design of its data-sharing mechanism, being deemed cumbersome in practice, lacking a sound economic justification, and suboptimal for fostering the data economy.[86] Even though this framework is not going to be changed in the near future, we would like to contemplate an alternative approach: What if the qualitative criteria of 'a simple operation' and 'disproportionate effort' were eliminated from the definition of 'readily available data' – along with eliminating substantial investment as a criterion of 'pre-processed data' – in view of their potential to diminish the scope and technical state of data, and, consequently, data utility? In other words, what if data were subject to the data-sharing obligations in the same technical state and scope as it is obtained from a product or service, including pre-processing that takes place within that product or

81 SWD(2022) 34 final 26-28.

82 ibid.

83 ibid.

84 Such as the development and adoption of interoperability standards in the context of common European data spaces and data processing services.

85 Akin to Article 47 of the Digital Market Act (laying the basis for the Commission to 'adopt guidelines on any of the aspects of this Regulation in order to facilitate its effective implementation and enforcement').

86 (n 17).

service to ensure its functionality?[87] Assuming all other parameters of the data-sharing regime stay the same, how would eliminating such constraints impact the equilibrium of interests, relative to the baseline established by the Data Act?

55 From the data usability perspective, removing the qualitative constraints on shareable data would be beneficial. In principle, even if 'readily available data' turns out to be data in its raw form,[88] it would allow the data user to extract value through purpose-specific processing if supplemented with the relevant metadata. As noted earlier, raw or generically processed data holds the highest potential for generating diverse outcomes and serving various use cases. In the case of sensor-generated data, it would be advantageous in terms of data usability if in-device processing of connected product data included calibration and de-noising, as the resulting level of accuracy and precision is typically sufficient to ensure product functionality. Provided that the relevant metadata is made available, raw or generically processed data can serve both primary purposes (i.e., ensuring product functionality, including product maintenance and repair) and secondary purposes, where data serves as input in new product or service development, often involving data aggregation.

56 From a legal perspective, omitting the criteria of 'simple operation' and 'disproportionate effort' from the definition of the 'readily available data' would reduce legal uncertainty concerning the determination of an elusive threshold of simplicity and proportionality, especially considering that the point of reference (proportionate to what?) is unclear.

57 From a balance-of-interests perspective, removing constraints on 'readily available data' – to the extent this could enhance data usability – would benefit prospective data users, both product/service users and third parties of their choice. For users, this would not entail additional costs, given that data should be made available to them free of charge to them (while the corresponding cost would be calculated within the market price of the product or service). For third-party recipients, this is a matter of compensation which they have to pay for data anyway.[89] Given that data can be made available to third-party data recipients under fair, reasonable, and non-discriminatory (FRAND) terms and conditions,[90] these terms can reflect the difference in the technical state of the data, i.e., either reduced to 'simple operations' or involving processing beyond this level. Hence, they can be adjusted to reflect the cost of data processing.[91] In this view, it is unclear why shareable data should be constrained by the 'simplicity' of operations, 'proportionality' of efforts, or 'substantiality' of investment.

58 For data holders, the current constraints within the definition of 'readily available data' might appear as a safeguard for their economic incentives and, hence, one would conjecture negative consequences ensuing if they were removed. Limitations on the scope and the technical state of shareable data[92] under the Data Act might be read as a precaution to prevent data-sharing obligations from becoming 'too burdensome' for data holders. Some could view this as the legislature's attempt to strike a fair balance between enabling broader access to and meaningful utilisation of data across a broad spectrum of use cases while avoiding imposing onerous requirements on parties under data-sharing obligations. However, such a restrictive approach to data sharing, tiptoeing around the data holders, might also be viewed as overly favouring their interests, without a sound justification.[93]

59 In principle, the requirement to share data in the technical state as it is obtained from a product or service would not interfere with the economic calculus underlying the current data-sharing obligations under the Data Act, particularly by imposing additional costs on data holder. By requiring data to be made directly accessible by the

87 While the technical state of data is determined by the product or service design, there is still some room for variability. For instance, the product can be designed to transmit data states A, B, C, D, and E. By default, the 'related service' may only necessitate states A and B, resulting in only A and B being transmitted. However, C can also be transmitted to the user if necessary.

88 As argued earlier, the conversion of an analogue to a digital signal can already be argued to satisfy the definitional criteria of 'readily available data' under art 2(17) Data Act.

89 art 8 Data Act.

90 This is not to idealise the FRAND system, the shortcomings of which have been discussed elsewhere. See eg Drexl J and others (n 17) para 99 ff; Picht PG, 'Caught in the Acts: Framing Mandatory Data Access Transactions under the Data Act, Further EU Digital Regulation Acts, and Competition Law' (2023) 14 Journal of European Competition Law & Practice 67, 26 ff; Kerber (n 17) 126. To clarify, here we are only comparing the option of removing the restrictions on the accessible and shareable data versus the existing baseline adopted in the Data Act, without challenging the latter.

91 art 9 Data Act.

92 This manifests in excluding the following categories of data from the scope of the Data Act: cleansed or transformed data requiring 'substantial investment', inferred or derived data or information due to 'additional investment', and readily available data if it requires 'disproportionate effort going beyond a simple operation' (rec 15; arts 2(17), 4(1), and 5(1) Data Act).

93 Kerber (n 17).

user free of charge,[94] the legislature must assume that the relevant costs, including building data-sharing infrastructure, will be passed on to the consumer, i.e. factored into the price of the product or service. Otherwise, this requirement would not be rational or economically viable. While these costs can be calculated within the market price of a product or service, data holders can also charge additional compensation for making data available to third-party data recipients. Here we do not question the economic logic of this model. Our point is that removing constraints on 'readily available data' would not impose on data holders additional costs relative to what is already required under the Data Act. Neither would this interpretation require the data holder to provide additional data processing beyond what already occurs within the product or service to ensure its functionality. In this view, it is unclear how removing constraints on readily available data – i.e. data generated and pre-processed to the point at which it is obtained from a connected product or related service – could jeopardise the economic incentives of data holders. If the restrictive criteria – 'simplicity' of operations, 'proportionality' of efforts, and 'substantiality' of investment – enable data holders to further maximise their profits at the expense of diminished data usability, one can question the current 'balance of interests' established by the Data Act.[95]

60 More broadly, protection of investment, incentives, and competitive advantage surfaces in several instances, such as when prohibiting using shared data for developing competing – interchangeable or substitutable – products;[96] when providing for the possibility for the data holders to request reasonable compensation for making data available in the context of B2B relations to 'promote continued investment in generating and making available valuable data, including investments in relevant technical tools';[97] when emphasising the importance 'to preserve incentives to invest in products with functionalities based on the use of data from sensors built into those products';[98] and when pointing to 'the lack of predictability of economic returns from investing in the curation and making available of datasets or data products' as a 'substantial hurdle to data sharing by businesses'.[99]

61 Of all these concerns, confining 'readily available data' by criteria of 'simple operations', 'disproportionate effort', and 'insubstantial investment' appears most relevant for incentives for data curation. However, it is questionable whether mandatory sharing of data puts at risk the incentives for data curation if such curation is confined to in-device or on-server data processing as part of *ensuring product functionality*, and given that the cost of processing can be factored within the product/service price, as well as the compensation for making data available. Given that the Data Act provides limited grounds for refusing an access request,[100] the restrictive criteria of 'simple operations' and 'disproportionate effort' cannot be invoked to substantiate a refusal to make data available altogether. Instead, the data holder may attempt to rely on these constraints to limit the readily available data in terms of its scope and technical state. However, from a practical perspective, it might be more feasible and beneficial for the data holder to make data available in the technical state it is obtained from a product or service and factor the related cost into the amount of 'fair compensation', rather than splitting data flows into two tracks – one with data in its 'natural' condition and the other one satisfying the restrictive qualitative criteria of 'readily available data'.

62 In summary, all other things being equal, removing constraints on the shareable data could have been more net-positive. Recognising that amending the Data Act remains a distant prospect, this consideration could be incorporated into dispute resolution and judicial practices, as well as future sectoral legislation. This could involve either removing the above-discussed constraints on the scope of shareable data or applying a stricter standard for defining what qualifies as 'disproportionate effort' or 'substantial investment'. To emphasise, this paper does not delve into the analysis of whether and to what extent the compromise reached within the Data Act is economically sound and balanced from a broader perspective of innovation incentives, including beyond those of data holders. Instead, we consider the existing deal as a baseline and explore the option of omitting constraints from the definitions of 'readily available data' and 'pre-processed' data, relative to this baseline. At the same time, it is worth noting that concerns have been raised about whether the baseline is optimal and justified from an incentives perspective, whether the compensation is needed to 'promote continued investment in generating' data,[101] and whether the latter is at risk at all.[102]

E. Conclusion

94 art 3(1) Data Act.

95 For a critical perspective on the overemphasis on the protection of incentives for data holders, see Kerber (n 17).

96 recs 32, 39, and 57; arts 4(10) and 6(2)(e) Data Act.

97 rec 46 Data Act.

98 rec 30 Data Act.

99 rec 26 Data Act.

100 Namely based on security reasons and trade secrets protection (art 4(2) and (8) and art 5(11)).

101 rec 46.

102 Kerber (n 17).

63 From the outset, the Data Act was conceived as a horizontal instrument, leaving the door open for further legislation to accommodate sectoral specifics, provided that sector-specific rules align with the Data Act.[103] Despite the Commission's engagement with stakeholders during the preparatory stage, the adopted horizontal, top-down approach had to maintain a generic – agnostic to the specific requirements of individual sectors or use cases – stance regarding the rules. The limitations of this 'access-in-the-abstract' strategy became evident during the late stage of the legislative process when it became apparent that some vital technical details had been overlooked. The late attempt to pivot and align the Data Act with the technical practicalities of data-sharing and usage resulted in populating the statutory text with ambiguous and hardly practical notions, including 'readily available data', 'disproportionate efforts', 'simple operation', 'pre-processed data', and 'significant investment'. This initiated a cycle of perpetual clarification, wherein the introduction of 'clarifying' terms necessitates further clarification.

64 In this paper, we examined how the Data Act addresses the need to enable data usability, apart from data accessibility, both of which are equally important for the maximisation of the value of data. As shown, the definition of the technical state of data constitutes a parameter of data access and usage rights, directly bearing on the scope of data subject to data-sharing obligations under the Data Act. However, the limiting criteria applicable to 'readily available data' pose a challenge in delineating this scope and might offset data usability. As an alternative approach, we have considered omitting such criteria from the definition of readily available data and argued that this holds the potential to yield a more positive overall outcome in terms of technical usability, legal certainty, and a balance of interests.

103 SWD(2022) 34 final 7. However, considering that subsequent rules should align with the Data Act, the concern is that the Data Act might pre-emptively limit the flexibility of these rules to accommodate for the specifics of the sector or use cases.

1 Table Annex

Some data pre-processing steps and their contribution to and dependence on data usability.

The type of data transformation	Changes to data	Possible contributions to the usability of the transformed data in future steps	Possible dependence on the usability of the data being transformed	Cost considerations
Calibration	Converts raw sensor data (unitless signal strength) to calibrated data with known accuracy and an interpretable unit of measurement such as degree Celsius for temperature.	Usually an early step, any future step that builds upon a well-calibrated dataset will benefit from the better accuracy so acquired. With interpretable data, relevant physical laws might be applied to treat the data in a future data pre-processing step.	Calibration requires sensor data to be available so that comparisons can be made between the sensor's readings and the standard values being calibrated.	Calibration equipment cost or calibration service charge.
De-noising	Reduces the fluctuation in data caused by noise to increase the signal-to-noise ratio.	Usually, an early step, as any future step that builds upon a dataset with minimal noise will benefit from the better precision so acquired. Revealing the signals helps discover patterns in the extraction pre-processing step. Imprecise data is bad for many machine learning algorithms.	Missing values can degrade the performance of de-noising algorithms that rely on aggregating existing data values.	Labour cost in research and development, involving examination of the characteristics of the data being treated, as well as selecting and configuring the best-performing approach through experimentation with various possible approaches.
Missing value imputation	Fills the values that are missing due to reasons such as sensor or device downtime, communication loss, or data corruption.	Increase the percentage of available data, which is important for statistically based machine learning algorithms; Many machine-learning algorithms cannot deal with missing values.	Many imputation algorithms make use of existing values (from any co-working sensors) to estimate the missing ones. Therefore, inaccurate and/or imprecise existing values will result in poor estimations.	
Selection (including techniques such as outliers detection, feature selection, data reduction, and instance selection)	Filters out unusable data such as irrelevant data, or data samples with outlier values or too many missing values.	Removing unhelpful data may improve the performance of a machine-learning model.	Inaccurate or imprecise data might lead to wrong decisions.	
Extraction (including techniques such as feature engineering and data fusion)	Creates new data from the existing dataset, e.g., temperature and relative humidity can be combined to get the amount of water vapour in the air.	New data, which is a strong indicator of the variable being predicted by a machine learning model, can boost the model's performance. Aggregation of data can reduce network traffic.	Garbage-in, garbage-out: the quality of the selected and extracted data depends on the quality of the data being transformed.	