Personal Data, Algorithms and Profiling in the EU: Overcoming the Binary Notion of Personal Data through Quantum Mechanics

Alessandro El Khoury*

Abstract

In this paper I propose to analyse the binary notion of personal data and highlight its limits, in order to propose a different conception of personal data. From a risk regulation perspective, the binary notion of personal data is not particularly fit for purpose, considering that data collection and information flows are tremendously big and complex. As a result, the use of a binary system to determine the applicability of EU data protection law may be a simplistic approach. In an effort of bringing physics and law together, certain principles elaborated within the quantum theory are surprisingly applicable to data protection law, and can be used as guidance to shed light on many of today's data complexities. Lastly, I will discuss the implications and the effects that certain processing operations may have on the possibility of qualifying certain data as personal. In other terms, how the chances to identify certain data as personal is dependent upon the processing operations that a data controller might put in place.

1 Introduction

Personal data is any information related to an identified or identifiable natural person.[1] Sometimes it is obvious which information constitutes personal data; some other times the exercise becomes complex and may lead to unexpected results. The paramount principle upon which EU data protection law is based is the possibility of qualifying certain information as *personal data*. Whenever the piece of information carried by data can be separated from the physical person to whom that information refers, the rules and safeguards stemming from EU data protection law become inapplicable. In this sense, data is conceived as binary: it is either personal or not.

The possibility of identifying, directly or indirectly, a person through a number of pieces of information – individual or combined – highlights the complexities that data protection experts are currently experiencing when dealing with technologies and techniques such as Big Data,[2] Cloud Computing,[3] data mining and collection of information through the Internet of Things (IoT).[4] Devices of all sorts around us are constantly collecting information to provide services, yet not all data collected falls within the category of personal data strictly speaking. This amount of non-personal information can, however, quickly lead to the identification of a physical person and reveal very personal aspects such as political orientation or sexual preferences.

In this article, I propose to analyse the binary notion of personal data and highlight its limits in the current EU General Data Protection Regulation (GDPR).[5] *Breyer* v. *Deutschland*[6] shows that from a risk-regulation perspective, the binary notion is not particularly fit for purpose, considering that data collection and information flows are complex processes. This calls for a different conception of personal data, which should go beyond its binary definition, and instead, focus on its inherent, relative nature. Data could indeed be personal and non-personal at the same time: the relevant distinction can be made only in a specific moment, while putting the data in the context of processing operations carried out around it. This article, therefore, purports to show that the use of a binary system to determine the applicability of EU data protection law may be too simplistic an approach.

For this purpose, it employs quantum mechanics as a guide to shed more light on the matter. At the beginning of 1900, when certain observations on matter could not be described through classical physics, physicists began

[*] Alessandro El Khoury, LLM, Legal and Policy Officer, DG Health & Food Safety, European Commission. The information and views set out in this article are those of the author and do not necessarily reflect the official opinion of the European Commission.

[1] The definition we adopt is based on EU data protection law. *See* after the third section.

[2] Big Data has been defined as a data set whose size is beyond the ability of typical database software tools to capture, store, manage and analyse. *See* J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh & A.H. Byers. *Big Data: The Next Frontier for Innovation, Competition, and Productivity* (2011).

[3] Cloud Computing has to be understood as a methodology through which a vast measure of pooled and virtualised resources can be accessed. *See* A. El Khoury, 'Data Protection and Risk Regulation. Cloud Computing: A Case Study' (LLM thesis on file at LUISS School of Governance, Rome).

[4] With the term 'Internet of Things', we refer to a global network infrastructure linking uniquely identified physical and virtual objects, things and devices through the exploitation of data capture, communication and actuation capabilities. *See* A. Guimarães Pereira, A. Benessia and P. Curvelo, *Agency in the Internet of Things*, Publications Office of the European Union (2013), at 7.

[5] European Parliament and Council Regulation 2016/679, OJ 2016 L 119/1.

[6] Case C-582/14, *Patrick Breyer* v. *Bundesrepublik Deutschland*, ECLI:EU:C:2016:779.

doi: 10.5553/ELR.000114 - ELR december 2018 | No. 3

thinking differently, and quantum mechanics arose as a new branch of physics. Interestingly, it seems that certain principles elaborated within quantum theory may be appropriate for describing data. By drawing inspiration from quantum mechanics, this article aims to ultimately overcome the binary notion of personal data and find a right balance in the application of EU data protection law. This conclusion is also supported by the fact that today it has become rather easy to identify a data subject due to the increasing affordability of certain processing operations and the tools to perform them.

2 Setting the Scene: We Live in a World of Data

The idea behind the quote 'Data is the new oil'[7] is elementary: oil was – and most likely still is – the basis of the world's economy during the twentieth century. Refined to produce plastics, fuel and many other materials, oil can be converted into many different commodities. A legitimate question would be, what does oil have to do with data? Both commodities – oil and data – can be traded and their trade volumes and prices can affect stock markets in different ways. It was demonstrated that changes in oil prices could predict stock market return worldwide,[8] whereas the impact that data can have on stock markets is tied to the reliability that companies feeding off the data can project on the general public.[9]

Moreover, oil is not a self-sufficient commodity: once refined and transformed its sub-products cannot be reverted to oil. This concept was well summarised by Scaruffi:

> [T]he difference between oil and data is that the product of oil does not generate more oil (unfortunately), whereas the product of data (self-driving cars, drones, wearables, etc.) will generate more data (where do you normally drive, how fast/well to drive, who is with you etc.).[10]

Differently from oil, not all data is equal. In this sense, data is more comparable to rocks: there are common and inexpensive, and rare and expensive ones. When a piece of information refers to a human being, it becomes personal data. Not all personal data has the same economic value: there are different values, different pieces of information linked to that data which can make it more or less attractive for business operators according to the type of business they are running.[11] For an advertisement company, geographical data on potential customers is valuable: the company might use that information to target its advertisements and promptly show offers from restaurants to nearby potential customers. This geographical data (technically called 'geotag') needs to be placed in the context of activities that a potential customer is carrying out in a determined time and space. Knowing the potential customer is located close to a restaurant whose advertisement can be shown by the advertisement company is valuable data. If the potential customer is hiking in a forest, however, knowing his specific location does not bring any advertising potential, because there are no restaurants nearby to advertise.[12]

Data is to be understood in broad terms, and according to Ackoff, is raw and does not have a meaning in itself.[13] In the case of geographical data, latitude and longitude are just numbers, coordinates on a map; when matched with a physical person, they become a geotag, an information conveying that a person is physically located somewhere. Therefore, information is data that has been given a meaning by way of relational connection with other data.[14] The meaningfulness of this information has a different degree of appreciation for the subject making use of it.

Another difference between data and oil is that the latter is a scarce resource, whereas the former is virtually infinite, self-sustainable and self-replicable. To understand these concepts we can imagine a timeline, a sequence of events starting at *time 0* and ending at *time 10*. The actual length between *0* and *10* is not relevant. A barrel of oil will always be a barrel of oil throughout the timeline, or until it is transformed into something else. On the contrary, data and the information held within it changes according to its intended use and with time . For example, a person's name is likely to remain unchanged, but if we consider body temperature, its variation throughout a timeline might reveal other pieces of information, such as that the person has a cold or is performing physical activity. This different degrees of information provided by data allows for an interesting observation: data has both an intrinsic and extrinsic value. The intrinsic value is by virtue of the piece of

[7] The quote is often attributed to different people. *See* M. Kuneva, European Consumer Commissioner in a 2006's Speech http://europa.eu/rapid/press-release_SPEECH-09-156_en.htm (last visited 24 June 2018); G. Rometty, IBM CEO in a Speech to the Council of Foreign Relations in 2013 https://siliconangle.com/blog/2013/03/11/ibms-ceo-says-big-data-is-like-oil-enterprises-need-help-extracting-the-value/(last visited 24 June 2018).

[8] G. Driesprong, B. Jacobsen & B. Maat, 'Striking Oil: Another Puzzle?', 89 *Journal of Financial Economics* 307 (2008).

[9] *See* 'Facebook Stock is in the Red for the Year After the FTC Confirms Investigation', http://fortune.com/2018/03/26/facebook-stock-ftc-investigation-cambridge-analytica/(last visited 24 June 2018).

[10] P. Scaruffi, *Humankind 2.0* (2016), available at https://www.scaruffi.com/singular/bigdata.html (last visited 18 November 2018).

[11] *See also* I.N. Cofone and A.Z. Robertson, 'Privacy Harms', 69 *Hastings Law Journal* 1039, at 1049-1053 (2018) where the concept of the Privacy Bell is discussed. Despite the authors refer to privacy, and not to data protection, the same theory could be used to describe the degree of information on a data subject that data could provide.

[12] For an in-depth analysis on the use of geotags and Big Data, *see* J.W. Crampton, M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M.W. Wilson & M. Zook, 'Beyond the Geotag: Situating 'Big Data' and Leveraging the Potential of the Geoweb', 40 *Cartography and Geographic Information Science* 130 (2013).

[13] R.L. Ackoff, 'From Data to Wisdom', 16 *Journal of Applied Systems Analysis* 3 (1989).

[14] *Ibid.*

information carried by the data. In the previous example, it would be the fact that the body has a certain temperature in a specific moment. When that information is, however, put in correlation with the same data from a different moment of the timeline, it allows inferring a new information.[15]

For companies it makes sense to collect, aggregate and analyse any kind of data, even the one that, prima facie, does not seem to identify a person or highlight a pattern. The reason is that this data could prove useful when put in correlation with other data sets: it could show trends and correlations in those data sets where people are identified, thus transforming into personal data the first data set as well.[16]

So far, this article has focused on the more theoretical aspects of data and information. Now it is time to apply those aspects to concrete cases. The world these days is populated by smart devices capable of collecting and sharing any type of data, by IoT, Cloud Computing and Big Data, which are at the basis of services not even imaginable few years ago. All these technologies are the equivalent of the tools used to extract and refine oil. Tiny sensors collect data, which is shared and processed in the Cloud and ultimately stored in Big Data. Cloud, Big Data and IoT are three different perspectives of complex data processings: IoT *gathers*, Cloud *processes* and Big Data *stores* data. This picture portrays data as a commodity – the fuel running a complex mechanism of systems. Thus, the fundamental question is, how is this commodity regulated? According to EU law, data as such is not regulated, but it becomes strictly regulated when it can be qualified as personal. This leads to another question: are the boundaries of personal data and non-personal data so well defined to justify such a binary approach to data regulation?[17]

3 The Current Notion of Personal Data: From the GDPR to the Case Law of the ECJ

The definition of personal data in Article 4 of the GDPR[18] largely draws from and overlaps with the old definition enshrined in Article 2 of Directive 95/46/CE[19] (Data Protection Directive, or DPD hereinafter), which the GDPR aimed at replacing and updating. The main difference between the two is in the use of the concept of *identifier*: it is used implicitly in the GDPR and explicitly in the DPD. Identifiers are not defined in the GDPR, but they have to be understood as a piece of information holding a particularly privileged and close relationship with the data subject, such as cookies or internet protocol addresses.[20] Recital 30 of the GDPR explains that identifiers are important as they may leave traces of the data subject in a particular environment, which, once combined with other identifiers and information, may be used to create profiles of the data subjects and to identify them.[21]

3.1 The Practical Issues of Identifiers

Some of the issues revolving around identifiers could be understood by analysing the *Cambridge Analytica* scandal, which called into question how Facebook collects and shares personal data.[22] After the scandal became public, Mark Zuckerberg (CEO of Facebook) was summoned before the United States Congress and the European Parliament to answer on how and when Facebook collects and shares data. A recurrent question concerned the so-called *shadow profiles*.[23]

[15] Cofone and Robertson, above n. 11. The Privacy Bell shows mathematically how the degree of privacy changes according to the degree of plausible assumptions that can be made on a person: more plausible assumption, less privacy. The same concept is applicable to data protection.

[16] This shows why, in academia, some researchers call the debate between anonymous data and personal data a false debate. *See* S. Stalla-Bourdillon and A. Knight, '*Anonymous data* v. *personal data* – a False Debate: an EU Perspective on Anonymization, Pseudonymization and Personal Data', 34 *Wisconsin International Law Journal* 284 (2017) and S.Y. Esayas, 'The Role of Anonymisation and Pseudonymisation Under the EU Data Privacy Rules: Beyond the 'All Or Nothing' Approach', 6 *European Journal of Law and Technology* 1 (2015).

[17] Some courts in the United States shared the same perplexity. *See Sanders* v. *ABC*, 978 P.2d 67 (Cal. 1999), where 'privacy, for purposes of the intrusion tort, is not a binary, all-or-nothing characteristic. There are degrees and nuances to societal recognition of our expectations of privacy'. If the case deals with privacy, the same reasoning is valid for data protection if we consider that the presence or absence of privacy is logically linked to the fact that data is personal or not, although the right to privacy and the right to data protection have fundamental differences in their scopes and limitations. *See e.g.* Case C-28/08 *P, Commission/Bavarian Lager*, [2010] ECR I-6055, para. 60, and J. Kokott and C. Sobotta, 'The Distinction Between Privacy and Data Protection in the Jurisprudence of the CJEU and the ECtHR', 3 *International Data Privacy Law* 222 (2013).

[18] Personal data is defined as 'any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, 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', Art. 4, GDPR, above n. 5.

[19] Art. 2(a) of Directive 95/46 defines personal data as 'any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity'. European Parliament and Council Directive 95/46/CE, OJ 1995 L 281/31.

[20] Compare with Art. 29 Data Protection Working Party, *Opinion 4/2007 on the concept of personal data* (2007) at 14.

[21] *See also* R.E. Leenes, 'Do You Know Me? Decomposing Identifiability', *Tilburg University Legal Studies Working Paper No. 001/2008*, where the identifiability is divided in four subcategories: L-, R-, C- and S-identifiability. L-identifiability allows individuals to be targeted in the real world on the basis of the identifier, whereas this is not the case for the other three. In fact, R-identifiability can be further decomposed into the S-type, which is a technical kludge, and C-type, which relates to the classification of individuals as members of some set.

[22] The *Cambridge Analytica* scandal concerned the collection of personal data of around 84 million Facebook users by British political consulting Cambridge, which used it to steer the US presidential elections of 2017.

[23] Shadow profiles are an aggregation of information concerning a particular data subject who has not yet been formally identified. *See,* in par-

Shadow profiles are based off a basic function of the Internet: most websites collect information on visitors to tailor services such as advertisements or store users' preferences to provide a better browsing experience. Facebook's peculiarity is that whenever one of its features (as simple as a *like and share* button) is embedded in a website, it sends data about its use to Facebook, even if the user's activity on the webpage was only limited to browsing.[24] This data is full of identifiers such as cookies, Internet Protocol (IP) addresses and many others.[25] Facebook counts around 2.2 billion users monthly,[26] and it is not difficult to understand why most websites today embed features from it, thus allowing a large collection of identifiers. A sufficient amount of identifiers can be used to infer information about a virtually unknown person (technically, a not-yet-identified data subject).

What is the use of this aggregated data? In the case of Facebook, when a person registers to it, the platform associates the shadow profile with that person. Without shadow profiles, the database containing personal data of a new user should be empty. Any collection of personal data should begin only at the moment of registration and, in any case, after the user has given explicit consent to it. However, when shadow profiles are used, correlations are done automatically by Facebook, and the already performed data collection and analysis are associated with that data subject. In turn, the platform can immediately offer enhanced services such as suggesting a friend list or displaying advertisements of interest for that user in a surprisingly (or worryingly) accurate fashion.

Identifiers as such do not do have to be understood as personal data: they hold a privileged relationship with the data subject because they can describe certain of his characteristics.[27] Yet, they have the potential to become personal data, at later stages. All the more so when an identifier not conceived to collect personal data could be re-engineered into an identifier carrying a high degree of personal identifiability.[28]

The practical issue of identifiers and personal data has been presented to better understand the impact of the reasoning followed by the European Court of Justice.

3.2 *Breyer* v. *Bundesrepublik Deutschland*

In the landmark judgement delivered on 19 October 2016 in *Patrick Breyer* v. *Bundesrepublik Deutschland*[29] (*Breyer* hereinafter), the Court of Justice of the European Union (ECJ hereinafter) determined that dynamic IP addresses constitute personal data in relation to a certain provider, where it has the legal means that would enable it to identify the data subject through additional data held by another provider.

The case originated from a request for preliminary ruling from the German Federal Court of Justice (FCJ), in relation to an action brought by Mr. Breyer – a former member of the parliament in Schleswig-Holstein – against the Federal Republic of Germany, concerning the registration and storage by the latter of the IP address allocated to him, alongside the date when he accessed several websites run by German federal institutions, the terms entered in the search fields and the quantity of data transferred. Data retained by the German federal institution, no matter how specific, did not allow the identification of Mr. Breyer, thus falling outside the notion of personal data and the protection of the DPD. However, such identification would have been possible if the Internet Service Provider (ISP) had revealed sufficient information to identify the person operating behind a dynamic IP address.[30]

Whether static IP addresses should be considered personal data or not was already answered by the ECJ in 2011. In *Scarlet Extended*,[31] the ECJ concluded that static IP addresses should be considered personal data because they allow the precise identification of the user.[32] According to the Court, there are two elements to consider: one technical and one legal. Technically, the ISP assigns an IP address to a device, and this IP is always the same (static IP); legally, the underlying contract for the Internet service provisioning will be undertaken between the ISP and a natural or legal person, under whose responsibility the connected device is operated. This is why in *Scarlet Extended* the ECJ based its conclusions on the fact that an injunction by a court

ticular, the question asked by New Mexico Representative Ben Lujan (full transcript available at https://techcrunch.com/2018/04/11/facebook-shadow-profiles-hearing-lujan-zuckerberg/?guccounter=1 (last visited 7 July 2018) and by MEP Syed Kamall (*see* Facebook's written answers available at http://www.europarl.europa.eu/resources/library/media/20180524RES04208/20180524RES04208.pdf> (last visited 7 July 2018).

[24] *Ibid*.

[25] *See* M.D. Ayenson, D.J. Wambach, A. Soltani, N. Good, and C.J. Hoofnagle, 'Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning', Available at SSRN: https://ssrn.com/abstract=1898390; D. Barth-Jones, 'The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now', Available at SSRN: https://ssrn.com/abstract=2076397 and F.J. Zuiderveen Borgesius, 'Singling Out People Without Knowing Their Names – Behavioural Targeting, Pseudonymous Data, and the New Data Protection Regulation', 32 *Computer Law & Security Review* 256 (2016).

[26] Statistics are referred to the second quarter of 2018, https://www.forbes.com/sites/dantedisparte/2018/07/28/facebook-and-the-tyranny-of-monthly-active-users/#383c9c8f6aea (last visited 4 November 2018).

[27] It is the case for keystroke dynamics applied for personal authentication, which relies on the fundamental assumption that keystroke dynamics (i.e. how a certain person types on a keyboard) is almost unique for each person. *See* G. Gabla, 'Applying Keystroke Dynamics for Personal Authentication' Available at SSRN: https://ssrn.com/abstract=2508480.

[28] *See* Barth-Jones, above n. 25; and Zuiderveen Borgesius, above n. 25.

[29] *Breyer*, above n. 6. *Breyer* was ruled under the DPD. Differences with the GDPR will be marked throughout the analysis.

[30] An IP address is a logical numeric address assigned to every device connected to a network to identify it. These addresses are assigned by an ISP to a host in a fixed or dynamic fashion. In the former case, a device will always use the same IP address, whereas in the latter case, the IP address is assigned each time the device connects to the network. IP addresses exist to identify a specific device, but they are not necessarily meant to identify the person operating it in a given moment, all the more so when the IP address is a dynamic one. *See* S. Feit, *TCP/IP: Architecture, Protocols, and Implementation with IPv6 and IP Security* (1996).

[31] Case C-70/10, *Scarlet Extended*, ECLI:EU:C:2011:77.

[32] *Ibid*. para. 51.

to an ISP to install technical means to analyse the traffic generated by a certain IP address, in order to monitor the use of peer-to-peer software[33] used to infringe intellectual property rights, was precluded by EU data protection law.[34]

If the nature and function of static IP addresses are clear, dynamic IP addresses are trickier. The difference in the identifiability features of static and dynamic IP addresses can be understood through an example: if we picture IP addresses as coats of different colours used to identify doctors in a hospital, using a static IP means that each doctor will always wear the same coat; on the contrary, a dynamic IP address entails that each time a doctor enters the hospital's premises, he will be assigned one random coat from the ones available. This latter concept is known in information technology as Dynamic Host Configuration Protocol (DHCP), and it prevents two devices from being assigned the same IP address and thus causing a conflict in the network architecture. That coat alone, however, does not bring sufficient information to enable an identification of the doctor wearing it: it holds a privileged relationship with the data subject, but alone it does not allow its identification.

The DPD enshrines in Recital 26 a key principle to ascertain whether an identifier actually allows for the identification of a data subject: the means likely reasonably to be used by the controller or by any other person to identify the data subject.[35] The GDPR provides the same principle in a same-numbered recital,[36] but it adds to it that to consider those means as 'likely reasonably to be used', account should be given of all objective factors, such as costs and the amount of time required for the identification, taking into consideration the available technology at the time of the processing and the technological developments. In a nutshell: *feasibility* and *capability*. A technical means is likely reasonably to be used according to the *feasibility* of its use and the *capability* of a data controller to use it, which include technical implementation, time and the costs and benefits of doing so. Technical implementation, economic cost and time need to be put in relation to the potential economic benefit of the operation for the data controller.

The concept of *means likely reasonably to be used* generated a large debate in German academia, which polarised around a *subjective* and an *objective* criterion.[37] According to the *objective criterion*, a person can be identified when, regardless of the capability of a certain data controller to identify him, the identification is *feasible* by combining data from different sources. The *subjective criterion* relies on the concrete *capability* of a certain data controller to make use of its means to identify the data subject. The main difference between the two criteria lies in the relevance given to the data controller. For instance, the sheer size of means available might make all the difference in understanding whether certain data is personal or not for that specific controller. Relativity at its best!

In *Breyer*, the two criteria applied as follows: for the *subjective criterion*, IP addresses become personal data only when there is the concrete capacity of a provider who has access to that information to use his own resources to identify the data subject (*e.g.* by performing more correlation with other data sets or even collecting additional data); on less theoretical grounds, by applying the *objective criterion,* IP addresses become personal data only when a data subject can be concretely identified, regardless of the abilities and the means of a provider to do so.[38]

The choice between the two criteria has a fundamental meaning when dealing with dynamic IP addresses. In that case, the means *likely reasonably to be used* to identify the data subject are allegedly more complex, expensive and time consuming to implement. Thus, if theoretically an identification is possible (subjective criterion), it does not mean that this could happen in practice (objective criterion). The question referred to the ECJ by the German FCJ, however, has a remarkable subjective element. What the FCJ fundamentally asks is if a dynamic IP address stored by an online media provider (*i.e.* the owner of a website) has to be considered already personal data for that provider, in the case where only a third party has the additional information necessary to identify the data subject[39] which accessed the online media through that dynamic IP address in a specific moment in time.[40]

The groundbreaking element of *Breyer* does not consist in the ECJ's ruling that, under certain conditions, dynamic IP addresses are personal data, but rather in the legal reasoning followed to reach those conclusions – that same reasoning is applicable *mutatis mutandis* to similar categories of data and subjects. This reasoning is based on three key elements.

[33] Peer-to-peer networking is a distributed computing architecture allowing the partitioning of tasks between different devices (peers) connected to a network, thus allowing a substantial degree of anonymity when sharing files of considerable size. *See also* R. Ambrosek, *Shawn Fanning: the Founder of Napster* (2006) where the facts behind the very first peer-to-peer software called 'Napster' are re-construed.

[34] Notably, the ECJ ruled, 'Directives 2000/31, 2001/29, 2004/48, 95/46 and 2002/58, read together and construed in the light of the requirements stemming from the protection of the applicable fundamental rights, must be interpreted as precluding an injunction made against an ISP which requires it to install the contested filtering system'. *Scarlet Extended*, above n. 31, para. 55.

[35] Recital 26, DPD, above n. 19.

[36] Recital 26, GDPR, above n. 5.

[37] *See* M. Schreibauer, '§ 11 Telemediengesetz (4 to 10)', in M. Esser, P. Kramer & K. von Lewinski (eds.), *Kommentar zum Bundesdatenschutzgesetz. Nebengesetze* (2014); J. Nink and J. Pohle. 'Die Bestimmbarkeit des Personenbezugs. Von der IP-Adresse zum Anwendungsbereich der Datenschutzgesetze', in *Multimedia und Recht* (9/2015), at 563-67. J. Heidrich and C. Wegener, 'Rechtliche und technische Anforderungen an die Protokollierung von IT-Daten. Problemfall Logging', 8 *Multimedia und Recht* 487 (2015). H. Leisterer, 'Die neuen Pflichten zur Netz– und Informationssicherheit und die Verarbeitung personenbezogener Daten zur Gefahrenabwehr', 10 *Computer und Recht* 665 (2015).

[38] See *Breyer*, above n. 6, paras. 52-54.

[39] *Breyer*, above n. 6, para. 31.

[40] Notably, dynamic IP addresses change at every connection; thus, the reasoning has to be strictly bound to the possibility of identifying a data subject in a specific moment of a timeline.

First, from a technical perspective, dynamic IP addresses belong to the general category of metadata[41]: metadata is not personal data, but may contain data about personal data. For instance, typical metadata applied to personal data would be the date when the personal data *surname* has been changed in a system. If we strictly apply the binary approach adopted in the DPD or in the GDPR, metadata stays outside the protection provided by EU data protection law, meaning that every processing operation on that metadata is possible, including transfers outside the EU and recombination with other data.

The second element stays in the nature of power used by the German federal institutions to obtain information from the ISP. Public entities can act either in their public capacity, representing the public interest (*cum imperio*), or as any other legal entity (*sine imperio*).[42] When acting *cum imperio*, a public administration does not act as a peer towards its counterparts – it exercises a public power with an outreach not possible for private operators – whereas, acting *sine imperio* does not entail an exercise of public power and the outreach is the same as any other private operator. In *Breyer*, the German federal institution acted *sine imperio*.[43]

The third element concerns the outreach of an action *sine imperio,* which, according to the ECJ, consists of any possible channel not prohibited by law to achieve the desired result.[44] These channels could be, for instance, contractual clauses foreseeing the trading of metadata between two entities acting *sine imperio* one against the other.[45] Such clauses could be very easily inserted in a service provisioning agreement between different service providers in a contract for Cloud Computing services and,[46] concerning mere metadata, none of the guarantees foreseen by EU data protection law could prevent such trading.[47]

The three aforementioned key elements have to be tested within the framework of 'means likely reasonably to be used' provided by Recital 26 of the DPD and GDPR. Earlier we used the terms *feasibility* and *capability*, but what the ECJ concluded in a much more complex manner is that the possibility for a data controller to obtain further data from a third party to identify a data subject has to be understood within its capability to do so. In fact, 'that would not be the case if the identification of the data subject was prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort'.[48]

Proportionality is another element that has to be accounted for. It entails at least two sub-elements: an effort and a subject performing it. Lifting a hundred kilograms is a remarkable effort for a human, but is a negligible effort for a crane. On those same lines, imposing a certain contractual clause where metadata has to be transferred to a data controller might be a negligible effort, if that data controller is someone the size of Google or Facebook.[49] Moreover, in the proportionality check, a significant role is also played by the reward that those efforts bring.

The conclusion of *Breyer* is that dynamic IP addresses are not personal data per se, but they can become so for a data controller if it has lawful means to obtain any further data that would allow the identification of the data subject. The same reasoning is applicable to any kind of metadata, which brings two questions: I any data potentially personal data? Is the binary notion of personal data adequate to respond to the challenges posed by the complex world of Big Data?

4 Big Data, Anonymisation, Pseudonymisation and Data Analysis

Breyer shows how data is subject to a double relativity. One relativity aspect concerns the very nature of the data (personal or not) against the means that a controller can put in place to reconstruct that data as personal; in this case, the controller performs an identification. The other relativity (hence double relativity) concerns the effort needed to reconstruct non-personal data as personal, which is not relative to the means used, but to the data controller performing it and to its capacity to do so. To put it in different words, at the beginning of this article I used the example of the timeline, from *time 0* to *time 10*. What *Breyer* shows is that non-personal data located at *time 0* could become personal data in another moment of the timeline, depending on the subjects having access directly or indirectly to it. Moreover, the possibility of non-personal data to mutate its nature depends on the theoretical means that a controller can potentially put in place to do so (if we opt for the subjective criterion), or the means that it actually puts in place, only when it makes use of them (if we opt for the objective criterion).

To add another layer of complexity to this reasoning, we should also take into account the issue of data anonymisation. This practice has been described as the process

[41] Metadata has to be understood as data about other data. *See* J. Pomerantz, *Metadata* (2015), at 16.

[42] *See* E. Casetta, *Manuale di Diritto Amministrativo* (2008), at 300.

[43] M. Reimann and R. Zimmermann, *The Oxford Handbook of Comparative Law* (2007), at 1274.

[44] *Breyer*, above n. 6, para. 47.

[45] C.J. Hoofnagle, 'Big Brother's Little Helpers: How Choicepoint and Other Commercial Data Brokers Collect, Process, and Package Your Data for Law Enforcement', 29 *N.C.J. Int'l L. & Com. Reg* 595 (2003).

[46] *See* C. Reed, 'Information "Ownership" in the Cloud', *Queen Mary School of Law Legal Studies Research Paper No. 45/2010*.

[47] Which explains why the data processing put in place by Facebook to perform shadow profiling, despite being despicable, is perfectly compatible with EU data protection rules.

[48] *Breyer*, above n. 6, para. 46.

[49] *See* S. Bradshaw, C. Millard & I. Walden, 'Contracts for Clouds: Comparison and Analysis of the Terms and Conditions of Cloud Computing Services', 19 *International Journal of Law and Information Technology* 187 (2011) where the authors refer to Terms and Conditions offered by Cloud computing providers in business-to-business contracts.

through which a data controller manipulates data sets in a database in order to make it difficult to identify data subjects.[50] Data anonymisation is also often referred to as 'de-identification'.[51] There are several techniques through which data anonymisation can be achieved, and the difference lies in the cost, complexity, ease of use and robustness.[52] In this sense, we can apply the same proportionality check described for the transformation of metadata in personal data: there will be an initial effort to anonymise personal data, and the anonymisation will be as strong as the effort put in place by the data controller to anonymise that data. Therefore, the robustness of an anonymisation processing is directly proportional to the effort put in place by the data controller, which is also logically impacted by three factors: the degree of robustness that the data controller wants to achieve for those categories of personal data subject to anonymisation; the means likely reasonably to be used to that end and the costs and benefits balance of the anonymisation processing.

Today, the possibility of using virtually unlimited computing power resources, thanks to Cloud Computing[53] and accessing data from tremendously big databases called Big Data, is not reserved for big corporations or governments. The very basis of Cloud Computing is its capability of providing enterprise-like services for any kind of user who can afford the price: the more powerful the service, the higher the price.[54] Data anonymisation is surely a privacy-enhancing technology, but it is also a threatening technology for data protection due to the binary notion of personal data and the so-called accretion problem.[55] The accretion problem postulates that once an adversary has linked two anonymised databases together, he can add the newly linked data to his collection of outside information and use it to help unlock other anonymised databases.[56] Theoretically, the risk increases exponentially for each further database correlated and, as we emphasised earlier, data protection rules are applicable only as long as we are dealing with personal data. If the *personal* element disappears, there are no safeguards for that data. *Pas gráve*, one may argue: if anonymised information suffers a data breach, nobody's rights to data protection or privacy will be violated. From a logical perspective, this is true. The amount of data and metadata present in Big Data, and the simplicity with which they can be computed in a Cloud system by anyone, however, poses a serious risk of reidentification.[57]

There is another interesting debate about anonymisation, and it concerns the 'pseudonymisation' technique. Pseudonymisation involves substituting the real identifying information with a code number or a nickname. Article 29 Data Protection Working Party has described it as 'the process of distinguishing identities'. Such a process aims at collecting additional data related to the same individual without having to know his identity.[58] The problem with pseudonymisation is that it gives the false hope of creating a safe harbour from data protection obligations,[59] thus legitimating high-risk processing operations (such as profiling) under the impression that any claim for damages of unlawful processing could be prevented.[60] Also, the GDPR in Article 6(4)(e) provides that pseudonymisation is an appropriate safeguard,[61] at the same level as encryption.[62] In reality, the means *likely reasonably to be used* are becoming more and more affordable and common thanks to the technologies described earlier. Thus, a correct risk assessment should conclude that re-identification of a data subject is more likely to happen than to retaining a permanent de-identification (or pseudonymisation).

It has been argued that current anonymisation techniques do not favour the data subject's right to self-determination, meaning that the degree of freedom that a data subject can exercise on its personal data is very limited. For instance, when personal data is anonymised, a data subject is faced with difficulty already at the stage of identifying that personal data is being processed. Thus, the data subject cannot verify whether its

[50] P. Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization', 57 *UCLA Law Review* 1701, at 1707 (2010).

[51] S. Latanya, 'Weaving Technology and Policy Together to Maintain Confidentiality', 25 *Journal of Law, Medicine & Ethics* 98, at 100 (1997): 'The term anonymous implies that the data cannot be manipulated or linked to identify an individual'.

[52] *See*, for instance, the basic guides to data anonymisation published by the Personal Data Protection Commission of Singapore, *Guide to Basic Data Anonymisation Techniques* (2018); and the European Data Protection Supervisor, *Opinion 3/2018 – EDPS Opinion on Online Manipulation and Personal Data* (2018).

[53] S. Chen, H. Lee & K. Moinzadeh, 'Pricing Schemes in Cloud Computing: Utilization-Based versus Reservation-Based', *Production and Operations Management* (2018).

[54] For a more detailed overview of Cloud contracts *see* Bradshaw, Millard & Walden, above n. 49.

[55] A. Narayanan and V. Shmatikov. 'Robust de-anonymization of large sparse datasets', 111 *IEEE Symposium on Security and Privacy* (2008).

[56] *See e.g.* B. Krishnamurthy and C.E. Wills, 'On the Leakage of Personally Identifiable Information Via Online Social Networks', 7 *WOSN '09 Proceedings of the 2nd ACM workshop on online social networks* (2009).

[57] *See e.g.* D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.A. de Montjoye & A. Bourka, 'Privacy by Design in Big Data: An Overview of Privacy Enhancing Technologies in the Era of Big Data Analytics', *ENISA: European Union Agency for Network and Information Security* (2015).

[58] Art. 29 Data Protection Working Party, above n. 20, at 18.

[59] Esayas, above n. 16, at 6-8.

[60] It is the case of the shadow profiling operations performed by Facebook through the placement of cookies, which was fined an incremental penalty of 250,000 EUR per calendar day of non-compliance by the Court of First Instance of Brussels in a judgment of 16 February 2016. See also the joint declaration of the French, Spanish, Belgian and Dutch Data Protection Authorities of 4th December 2015 https://www.cnil.fr/sites/default/files/typo/document/Declaration_commune_Groupe_de_contact_Facebook.pdf (last visited 6 July 2018).

[61] The same choice is made in Art. 25(1), where pseudonymisation is presented as a privacy by design measure, Art. 32(1)(a) considering pseudonymisation as adequate safeguard for the security of processing and Art. 40(2)(d), where pseudonymisation becomes a key element of the codes of conducts of enterprises.

[62] Encryption can be applied to provide pseudonymisation, but the two processing are logically distinct operations. There is a general understanding that key-coded data may not even be considered personal data so far as there are appropriate measures to exclude re-identification, such as a strong encryption algorithm, a strong encryption key and a secure key. *See* W.K. Hon, C. Millard & I. Walden, 'The Problem of 'Personal Data' in Cloud Computing: What Information is Regulated? – the Cloud of Unknowing', 1 *International Data Privacy Law* 211.

records are getting adequate protection. This kind of dispute is, however, substantially unfounded. Whenever personal information is anonymised, it ceases to be *personal*. Thus, the data subject does not have any legal right over it. It is for this very reason that the real emphasis should be on the moment right before the anonymisation and on the process of anonymisation itself. Once data is anonymised, it can be transferred without boundaries, and as the European Commission stated in 2009, this is not even considered a data transfer in the legal sense.[63] Moreover, other than giving technical advice and guidance on which anonymisation logic exist and what are some of their risks and advantages, and providing examples on their use, public regulatory bodies, such as national data protection authorities, cannot do much more, as anonymisation relies on complex algorithms that are often subject to intellectual property rights.[64]

From a conceptual perspective the distinction between personal and non-personal data is neat; yet, we underlined that this binary approach does not bring a real added value when data protection has to be implemented practically, because the possibilities of identifying a data subject are not the same for every data controller and change according to the circumstances as well. Yet, the legal definition of personal data remains a purely binary one.[65]

If, until now, we were able to substantiate our reasoning without the need to dig into Big Data's technicalities, the next set of issues inevitably demands so. Notably, another set of problems strictly linked to the technical aspects of Big Data – conceptually distinct from data anonymisation and very close to data reidentifiability – are those of data mining and predictive analysis.

Data mining is commonly defined as a set of automated techniques used to extract buried or previously unknown pieces of information from large databases. Data mining makes it possible to unearth patterns and relationships, and then use this new information to make proactive, knowledge-driven business decisions.[66]

From a data protection perspective, data mining is a processing operation and is neutral: the same data mining techniques can be applied to different databases, whether they contain personal data or not. Business operators are increasingly relying on data mining as it allows them to understand the market better and make better decisions.[67] Moreover, thanks to Cloud Computing, the costs of computing services powerful enough to run data mining algorithms are considerably low.[68] The main issues with data mining are that by mining Big Data, the algorithm can find patterns among data sets, thus unveiling further information that was not originally included in those data sets, and de-anonymise personal data that was previously anonymised.[69]

Predictive analysis is a particular type of statistical analysis that can provide, with a certain degree of certainty, answers to certain questions.[70] For instance, by analysing a set of anonymised information, the predictive analysis could tell whether a certain buyer of a product is a man or a woman or if it is a reliable debtor.[71] Once one anonymised information is de-anonymised (remember the accretion problem, and the proportional effort), all the other anonymised information about that (now) identified data subject is immediately correlated to him or her: this is what technically happens behind the curtains of Facebook's shadow profiling.

The relativity of personal data, and the ease with which the virtual border between personal and non-personal data can be disregarded, calls for a different approach, a different conception of personal data – one more attuned with the reality of data processing taking place in today's world – a notion of personal data that draws from quantum mechanics.

5 Overcoming the Notion of Personal Data through Schrödinger's Cat: Quantum Superposition and Quantum Entanglement of Personal Data

Quantum mechanics is a branch of physics developed in the early twentieth century by brilliant minds such as Erwin Schrödinger, Max Planck, Neils Bohr, Albert Einstein and Werner Heisenberg following a series of educated guesses inspired by a thorough knowledge of physics.[72] Quantum theory aimed to describe and explain the behaviour of matter at an atomic and subatomic level, which could not be explained by classical physics, in order to answer very practical questions such as why hot objects glow at a different colour depending

[63] European Commission, *Frequently Asked Questions Relating to Transfers of Personal Data from the EU/EEA to Third Countries* (FAQ B.1.9) (2009), available at http://ec.europa.eu/justice/policies/privacy/docs/international_transfers_faq/international_transfers_faq.pdf (last visited 18 November 2018).

[64] *See* Art. 29 Data Protection Working Party, *Opinion 5/2014 on Anonymisation Techniques* (2014), at 11 where the analysis revolves around the logic behind certain anonymisation techniques, but it refrains from referring to specific commercial solutions.

[65] *See also* Hon, Millard & Walden, above n. 62.

[66] A. Cavoukian, *Data Mining: Staking a Claim on Your Privacy* (1998), at 4.

[67] J.P. Bigus, *Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support* (1996), at 9. I.N. Cofone, Ignacio & A. Robertson, 'Consumer Privacy in a Behavioral World', 69 *Hastings Law Journal* 1471 (2018).

[68] P. Ruxandra-Stefania, 'Data Mining in Cloud Computing', 3 *Database Systems Journal* 67 (2012).

[69] Art. 29 Data Protection Working Party, above n. 64, at 5.

[70] *See* E. Siegel, *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* (2016).

[71] On the problem of credit scoring *see* D.K. Citron and F.A. Pasquale, 'The Scored Society: Due Process for Automated Predictions', 89 *Washington Law Review* 1, at 16 (2014).

[72] S.M. Barnett, J. Jeffers & J.D. Cresser, 'From Measurements to Quantum Friction', 18 *Journal of Physics: Condensed Matter* S401 (2006).

on their temperature. This article does not claim to redefine quantum physics or enrich its postulations but humbly aims at borrowing specific observations and applying them to data protection law to check whether they could be of help in ultimately providing a more fit-for-purpose definition of personal data.

The basic intuition is that the issue of describing (or measuring) the nature of data as personal or non-personal is very similar to the problems that the illustrious minds behind quantum theory tried to resolve. Bohr wrote, '[A] measurement to a certain degree deprives the information given by a previous measurement of its significance for predicting the future course of the phenomena. Obviously, these facts not only set a limit to the extent of the information obtainable by measurements, but they also set a limit to the meaning which we may attribute to such information'.[73] If we consider a single data in today's interconnected and complex world, its size and velocity of transmission are negligible. Any data (personal or not), regardless of its ability to provide descriptive details of a data subject, is shared between systems at very high speed similar to what happens to protons and electrons in a subatomic system. For the same reason, Bohr concluded that what matters is the unambiguous description of the matter's behaviour, rather than its measurement in a given moment.[74]

Our starting point is the conclusion reached by the ECJ in *Breyer*: data can be personal or non-personal sometimes, according to certain criteria. This emphasises the need to have a notion of personal data capable of providing an unambiguous description, rather than a measurement. The same matter can be better understood through Schrödinger's famous cat experiment.

Schrödinger's cat is a thought experiment imagined in 1935 by the physicist Erwin Schrödinger[75] and used to describe two fundamental principles of quantum mechanics: quantum superposition and quantum entanglement. Specifically, the experiment involves a cat in a sealed box with a bottle of poison, a Geiger counter and a radioactive source. The radioactive source has a 50 per cent chance of decaying. As soon as the Geiger counter detects the decay, a mechanism breaks the bottle of poison in the box, killing the cat. It is not possible to know if the cat is dead or alive before opening the box. Thus, the cat, in the timeline of the experiment, is both dead and alive at the same time. This state of matter is described in quantum mechanics as quantum superposition, and it entails that any two or more quantum states (the cat is dead or alive) can be added together (hence the name superposition) and the result will be another valid quantum state (for the cat, that status would be the cat being dead and alive at the same moment).[76] The main difference with binary systems is that in those, the result can only be true or false, *1* or *0*, but never both together, whereas in quantum mechanics the result can be *1*, *0* or a combination of the two.

Quantum superposition could also be understood through the famous *heads or tail*, where a coin is flipped in the air and the players have to guess on which side the coin is going to land. In a timeline that goes from *0* to *10*, where *0* is the moment just before the coin is flipped and *10* is the moment when the coin lands showing one of the two faces, in any moment between *0* and *10* the coin is potentially showing both *heads* and *tail*.

In our case, the cat or the coin represents data. The fact that the cat is dead or alive or the coin flips on one face or the other represents the fact that data is measured as personal or not. Theoretically, from an observer standpoint, every data not yet identified as personal behaves in the same manner: it is non-personal as long as a data controller does not perform a processing operation suitable of correlating that non-personal data with personal data or an individual, thus converting its nature from non-personal to personal. What puts data in the superposition state is the availability of the *means likely reasonably to be used* by a data controller to identify a data subject from that data. This is why we used the adverb 'theoretically'. Theoretically, we could envisage a set of non-personal data that is kept isolated from any processing operation capable of putting it in correlation with other databases. This is possible either because that non-personal data is collected and stored in a way to be inaccessible or non-compatible with any other data set (thus preventing reidentification) or just because it is swiftly deleted after having achieved its purpose. It was noted, however, that these cases are an exception rather than the rule.[77]

Observing that data is in the quantum superposition state also entails another logical conclusion. Quantum superposition as such is a neutral state: it comprises the case where data becomes personal, but also the opposite, where personal data is anonymised and loses its *identification* properties.[78] This observation is significant for understanding another concept described by quantum mechanics: quantum entanglement.

Quantum entanglement is a very particular quantum mechanical, physical phenomenon[79] in which two particles are so deeply linked that they share the same existence, no matter their physical distance. Once two particles are entangled, even if they are in the superposition status, their measurement will bring the same result.[80] To resume our example of *heads or tails*, if we flip two coins, and these are entangled, any measurement taken during their spin would lead to the same result: the two coins showing the same face. In the case of data, the entanglement consists in the possibility of linking together information from different data sets and pro-

[73] A. Plotnitsky, *Niels Bohr and Complementarity. An Introduction* (2012), at 68.

[74] J.A. Wheeler and W.H. Zurek, *Quantum Theory and Measurement* (2014), at 5.

[75] E. Schrödinger, 'Die gegenwärtige Situation in der Quantenmechanik', 23 *Naturwissenschaften* 807 (1935).

[76] P.A.M. Dirac, *The Principles of Quantum Mechanics* (1947), at 1-18.

[77] Art. 29 Data Protection Working Party, above n. 64, at 5.

[78] Alternatively, pseudonymised with all the *caveats* highlighted before.

[79] A. Einstein, B. Podolsky & N. Rosen, 'Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?', 47 *Physics Review* 777 (1935).

[80] Wheeler and Zurek, above n. 74, at 422-5.

cessing it in a specific time. At the time of processing, if one data becomes personal, then all the other data from different data sets linked to it exhibit quantum entanglement, and they become personal data too. The bond shared by the different data is their possibility of providing a piece of information sufficient to identify differently the data subject. Then, the fact that this data can be put in correlation provides the entanglement that changes the nature of data that was not personal, in a domino-effect fashion.

Going back to *Breyer*, the collection of dynamic IP address and log files by the online media provider consists of specific data on a particular subject's factual circumstances.[81] The data subject has yet to be identified as Mr. Breyer, and the identification becomes possible only when the ISP reveals information sufficient to achieve the identification. The data held by the ISP and the data held by the online media provider are entangled. They are physically distant, because they are stored in two different systems that are not linked physically or logically to one another; once superposition is triggered by the *means likely reasonably to be used* by the online media provider, the data in the two systems exhibit entanglement and can ultimately be *measured* as personal data. In other words, the entanglement among all the data present in the two data sets allows for an immediate measurement as personal data when a bridge is built between them: this bridge involves the possibility of putting in correlation one non-personal data from a data set with one personal data from another data set. This operation instantly exposes the entanglement (due to the correlations already made within each database), and all data suddenly becomes personal.

Notably, the entanglement – this intangible link or, to use the words of Einstein, this 'spooky action at a distance'[82] – involves the fact that certain data are inherently capable of describing an action, a property or a fact of a data subject. Through this description, data can directly or indirectly contribute to the identification of the data subject. Therefore, the intangible link consists in the fact that all data originates from the same data subject.

6 Quantum Theory and the GDPR

This long and complicated reasoning explained in the previous sections leads to two important conclusions. First, a correct approach to the notion of personal data should aim at providing an unambiguous description of it, rather than a predetermined measurement. In practice, this means taking into account the fact that data is in the quantum superposition state and could exhibit quantum entanglement.

A binary approach fails at grasping these complexities and, above all, fails at describing the true nature of personal data in a world of Big Data and infinite possible processing operations. Quantum superposition and quantum entanglement are a great aid in describing the reality of what can happen to data and personal data when placed in the context of the free-flow of information, and where practically any data controller has access to technical or legal means likely reasonably to be used to achieve the identification of the data subject. Second, rather than measuring the nature of data in a given moment and anchoring to it the applicability of EU data protection law, the focus should be on the processing operations triggering quantum superposition and what surrounds them – meaning that the focus should be on those *means likely reasonably to be used* to transform non-personal data into personal data. The status (personal or not) of data cannot be measured with sufficient certainty or, better, cannot describe the nature of data unambiguously because that status might change in the future depending on the data controller attempting the identification and its available means. If we assume that most data can potentially become personal, from a risk-regulation perspective, it is safer to assume that data is in the superposition status. The focus then shifts on the means used to entangle data and on the safeguards that should apply to those processing operations. In fact, due to those processing operations data exhibits entanglement and can be measured as personal.

On applying quantum superposition to the notion of personal data, the result necessarily moves away from a binary approach and three statuses of data can be observed: personal, non-personal and potentially personal. *Personal data* is data that has already identified (directly or indirectly) a data subject; *non-personal data* is data that does not and cannot (even theoretically) identify a data subject; finally, *potentially personal data* is a residual category, a grey zone, for which identification has not occurred yet, but it has not been excluded either.

If we apply the notion of entanglement to these three new categories, the focus becomes the processing operation that forces data to exhibit quantum entanglement. As a consequence, despite not having measured data as personal in a specific moment, some provisions of the GDPR should be applicable. The rationale behind this consequence lies in the fact that the GDPR foresees obligations and safeguards that could be respected by data controllers without identifying a data subject. This core group of obligations, I argue, would represent a standard for best practice that should be capable of shielding the data controller from liabilities, and possible data subject from damages.[83] Moreover, this conclu-

[81] *Breyer*, above n. 6, paras. 23-24.

[82] A. Einstein, 'Reply to Criticism: Remarks Concerning the Essays Brought Together in This Co-Operative Volume', in P.A. Schilpp (ed.), *Albert Einstein: Philosopher-Scientist* (1949) 665.

[83] The data subject is considered as eventual because its identification has not happened yet, but could happen at a later stage, when the damage has already been caused. It will be always the case, for instance, for the transfer of personal data in a third country that does not provide an adequate level of safeguards according to Chapter V of the GDPR. In

sion allows avoiding any measurement *a posteriori* of data as being personal, similarl to the conclusions of the ECJ in *Breyer*. Finally, a solution of this kind would be desirable in a legal framework where administrative fines for unlawful processing of personal data could have significant economic consequences, such as reaching €10,000,000 or 2 per cent of the annual turnover of a company.[84]

We mentioned earlier this *core* group of provisions of the GDPR which should be applicable regardless of the measurement of data as personal in a specific moment. To conclude this section it seems worth presenting a table (Table 1) including these provisions and the rationale behind their applicability.

In conclusion, the core group of provisions listed in Table 1 should provide a fair balance between meeting the need of data controllers to carry out their businesses in a profitable manner without excessive burdens and preventing them from harming data subjects involuntarily. The listed provisions deal with the correct management of data flow in a company and should already be in place for other reasons, mostly linked to the monitoring of the business activities, their profitability and the development of new processing operations.

7 Concluding Remarks

The purpose of this article has been to demonstrate that, despite the best intentions to regulate personal data in a stringent manner, its legal notion has very practical implications. We live in a world dominated by data exchanges, where the saying 'If you are not paying for it, then you are the product' is dramatically fitting. The *Cambridge Analytica* scandal showed that the possibility of transforming data into personal data is very real, and *Breyer* demonstrated its legal implications. In both cases, the binary notion of personal data seemed to be a weak tool to determine the applicability of EU data protection law.

Similar to the problems that physicists had to solve when quantum theory was developed, the notion of personal data has to describe unambiguously the behaviour of personal data in a real-world scenario. The consequences of not doing so are to be mistaken by the measurement of data as personal (or not) in a specific moment, with the certainty that such a result could be reversed at a later stage. This is all the more so when the whole applicability of EU data protection rules depends on that measurement.

The article shows that quantum theory may provide a better point of view, thus enabling the selection of a number of core provisions of the GDPR to avoid the detriment of data subjects, who could suffer damages, and of data controllers, which will have to pay for those damages.

this case, (not yet personal) data can be legally transferred; yet, when that data is used for the identification of the data subject or to enrich a profiling operation that has already taken place, the ultimate result is that the data subject is damaged, but has no legal claim over the data controller that performed the transfer.

[84] Art. 83(4) of the GDPR. That amount can be doubled easily according to paras. 5 and 6 of the same article, in case where a company bases its core business on processing data, which only afterwards reveals to be processing of personal data. In fact, paras. 5 and 6 deal with specific cases where either the processing operation went too far and the data subject is irremediably damaged by this or the data controller does not comply with an order of the supervisory authority. In the case where the processing of data is based on the wrong assumption that the data processed is not personal, it is very common to have data transfers towards third countries outside the guarantees of Chapter V. Thus, the processing operations are also engineered on that wrong assumption, and redesigning them is a process that necessarily takes a certain amount of time, during which the company can easily be put out of business.

Tabel 1 Core group of GDPR provisions

Points (b), (c), (d) and (f) of Article 5(1):
Article 5 deals with the principles related to the processing of personal data. In particular, the principles of purpose limitation, data minimisation, accuracy and integrity and confidentiality should be applicable. In turn, those provisions that are strictly related to the presence of a data subject (or the possibility of identifying it) have been excluded.[85] The reasoning is that the data controller might deal with data for which he does not have means likely reasonably to be used to identify the data subject, and may be completely unaware of the fact that that data could lead to the identification of a data subject.[86] On the contrary, the principles we identified as applicable are related to the design of the processing operations and prevent reckless processing of data.
Point (f) of paragraph 1 **and paragraph** 4 **of Article 6:**
Article 6 deals with the lawfulness of processing. Although we deemed as not necessary the provision under point (a) of Article 5(1), the lawfulness referred to in point (f) of Article 6(1) refers to the legitimate interests pursued by the data controller, for instance, its freedom to conduct business. Article 6(4) enriches Article 6(1) and sets further limitations to the processing operations, which include an assessment of the compatibility of the reasons for the further processing, of the need to use encryption or pseudonymisation and an evaluation of the type of data that is being further processed.
Point (f) of Article 14(2):
While Article 14 entails the existence of an identified data subject, the overall goal of the article can be understood from Recital 30 of the GDPR. The idea is that the data controller has to keep track of the personal data it processes. If we apply quantum superposition, and we accept the conclusion that data could turn into personal data at some point, then the data controller should always keep track of where it gets data, where it sends it for further processing, from how long that data is kept and if it transfers it outside the EU.
Section I of Chapter IV, Articles 24 to 31:
Section I of Chapter IV, Articles 24 to 31 establish the obligations between data controllers and data processors. The relationship between the two is fundamental to establishing a good model of governance for the processing operation because although the data processor processes data on behalf of the data controller, it might have a certain degree of flexibility in how certain operations are technically performed.
Article 33:
Article 33 on the notification of data breach towards authorities should be applicable every time a data controller is not able to demonstrate that the data processed under its responsibility is *non-personal data* according to the notion we provided earlier, meaning that the data breach notification should be performed every time the controller has not taken steps to ensure that the data processed is *non-personal* data. Data protection authorities should be put in the position of knowing whether a breach of data that is *potentially personal* could lead different entities, such as cybercriminals, to use the breached data sets with other data sets and ultimately identify data subjects.[87]

[85] In particular, the principle of lawfulness, fairness and transparency and the principle of storage limitation entail obligations that are determined by the data subject. For instance, those two principles will be applied in a very different manner if the data subject is an adult or a minor.

[86] If we consider that data is in the superposition status, and the data controller did not take any measure to make sure that data falls in the *non-personal* data category, then it is legitimate to conclude that another data controller might get access to that data in superposition and make use of its means to combine it with other personal data and ultimately make the data in superposition exhibit entanglement, thus transforming it into personal data.

[87] The accretion problem as such is a neutral process and can be used for legitimate or illegitimate purposes.

Article 37:

The designation of the data protection officer should become the rule where data is processed on a large scale. The designation should be based on the exception provided for in paragraph 4.

Chapter V, Articles 44 to 50:

Transfers of personal data to third countries or international organisations are risky operations by nature because data is transferred to a different jurisdiction with different (or no) safeguards. For this reason, the GDPR allows such transfers only in very limited circumstances and only where the data controller or processor have adopted appropriate safeguards. Therefore, considering the quantum superposition of data, this whole Chapter should be applicable in all cases where the data controller did not put in place mechanisms to ensure that data falls in the *non-personal data* category. The reason for such a stringent conclusion is that once data is transferred outside the EU, it does not matter if it becomes personal: it will still be outside the reach of EU data protection safeguards. All the more so in the case where economic operators amass vast amounts of *potentially personal* data (*e.g.* dynamic IP addresses) and perform the reidentification of subjects outside the EU, in countries where there are no safeguards for personal data and operations like mass-profiling for surveillance reasons are common.[88] The result of that identification can facilitate the use of data mining and predictive analytics techniques, which would ultimately unveil even more personal data on the data subject, with the final goal of using this aggressive profiling on that data subject in the EU.

[88] It is the case for the very recent Social Credit System developed by China. According to this, nothing prevents the fact that China amasses a large amount of *potentially personal* data and performs the identification of tourists or foreigners visiting China, at the border, where biometric data is collected from pictures. *See, for instance,* G. Sgueo, 'Tetris, La Cina e la gamification dei servizi pubblici', available at http://www.forumpa.it/citta-e-territorio/tetris-la-cina-e-la-gamification-dei-servizi-pubblici (last visited 8 November 2018), and A. Cagaan, 'China's Social Credit System raises privacy concerns over surveillance', available at https://www.veridiumid.com/blog/chinas-social-credit-system-raises-privacy-concerns-surveillance/(last visited 8 November 2018). It was also the case for the Prism programme run in the United States by the National Security Agency, which was the main driver behind the ECJ judgment in Case C-362/14 *Maximilian Schrems* v. *Data Protection Commissioner*, where the court stated that the *EU-US Safe Harbour Agreement* was not a legitimate tool for the transfer of personal data from the EU to the United States . *See* A. El Khoury, 'The Safe Harbour Is Not A Legitimate Tool Anymore. What Lies In the Future of EU-USA Data Transfers?', 6 *European Journal of Risk Regulation* 659 (2015).