They who must not be identified—distinguishing personal from non-personal data under the GDPR

Michèle Finck* and Frank Pallas**

Key Points

- In this article, we examine the concept of non-personal data from a law and computer science perspective.
- The delineation between personal data and non-personal data is of paramount importance to determine the GDPR's scope of application. This exercise is, however, fraught with difficulty, also when it comes to de-personalized data—that is to say data that once was personal data but has been manipulated with the goal of turning it into anonymous data.
- This article charts that the legal definition of anonymous data is subject to uncertainty. Indeed, the definitions adopted in the GDPR, by the Article 29 Working Party and by national supervisory authorities diverge significantly. Whereas the GDPR admits that there can be a remaining risk of identification even in relation to anonymous data, others have insisted that no such risk is acceptable.
- After a review of the technical underpinnings of anonymization that is subsequently applied to two concrete case studies involving personal data used on blockchains, we conclude that there always remains a residual risk when anonymization is used. The concluding section links this conclusion to the more general notion of risk in the GDPR.

One year after the European Union's (EU) General Data Protection Regulation (hereafter 'GDPR' or 'the Regulation') became binding, uncertainty continues to surround the definition of some of its core concepts, including that of personal data. Drawing a dividing line between personal data and non-personal data is, however, paramount to determine the scope of application of European data protection law. Whereas personal (including 'pseudonymous') data is subject to the Regulation, non-personal data is not. Determining whether a given data item qualifies as personal data is thus crucial, and increasingly burdensome as more data are being generated and shared.

Notwithstanding the pivotal importance of the distinction between personal and non-personal data, it can, in practice, be extremely burdensome to differentiate between both categories. This difficulty is anchored in both technical and legal factors. From a technical perspective, the increasing availability of data points as well as the continuing sophistication of data analysis algorithms and performant hardware makes it easier to link datasets and infer personal information from ostensibly non-personal data. From a legal perspective, it is at present not obvious what the correct legal test is that should be applied to categorize data under the GDPR.

Recital 26 GDPR announces that data is anonymous if it is 'reasonably likely' that it cannot be linked to an identified or identifiable natural person. National supervisory authorities and the Article 29 Working Party (the 'A29WP' which is now the European Data Protection Board—'EDPB') have, however, provided interpretations of the concept that conflict with this legislative text. It will indeed be seen below that whereas Recital 26 GDPR embodies a test based on the respective risk of identification, the Working Party has developed a parallel test that considers that there can be no remaining risk of identification for data to qualify as anonymous data. Notwithstanding, anonymization is an important

[*] Michèle Finck, Max Planck Institute for Innovation and Competition, 80539 Munich, Germany. Email: Michele.finck@ip.mpg.de. The authors would like to express their gratitude to the anonymous reviewer for very helpful comments as well as to Kai Ebert for exemplary research assistance. Thanks for fruitful discussions on technical details also go to Jacob Eberhardt.

[**] Frank Pallas, Technische Universität Berlin, Information Systems Engineering Research Group (ISE), 10587 Berlin, Germany

concept from the perspective of other notions and requirements in European data protection law, such as that of data minimization. The difficult determination of what constitutes a 'reasonable likelihood' of identification further burdens practitioners' work. Beyond, the explicit inclusion of the new concept of pseudonymous data in the Regulation has confused some observers.

This article charts the resulting entanglements from an interdisciplinary perspective. It evaluates the GDPR's definition of personal and non-personal data from a computer science and legal perspective by proceeding as follows. First, the legal concepts of personal and non-personal data are introduced through an analysis of the legislative text and its interpretation by different supervisory authorities. Secondly, we introduce the technical perspective on modifying personal data to remove person-relatedness. The third section applies the preceding insights in looking at practical examples of blockchain use cases. A concluding section thereafter builds on previous insights in engaging with the risk-management nature of the GDPR. It will be seen that, contrary to what has been maintained by some, perfect anonymization is impossible and that the legal definition thereof needs to embrace the remaining risk.

The legal definition of personal data under the GDPR

The GDPR only applies to personal data, meaning that non-personal data falls outside its scope of application. The definition of personal data is hence an element of primordial significance as it determines whether an entity processing data is subject to the various obligations that the Regulation imposes on data controllers. This underlines that the definition of personal data is far from merely being of theoretical interest. Rather, the contours of the concepts of personal and non-personal data are of central practical significance to almost anyone processing data. Notwithstanding, '[w]hat constitutes personal data is one of the central causes of doubt' in the current data protection regime.[1]

The Regulation adopts a binary approach that differentiates between personal data and non-personal data and subjects only the former to its scope of application.[2] In contrast with this binary legal perspective, reality operates on a spectrum between data that is clearly personal, data that is clearly anonymous and anything in between.[3] Today, much economic value is derived from data that is not personal on its face but can be rendered personal if sufficient effort is put in place. Beyond, there is an ongoing debate as to whether and if so under which circumstances personal data can be manipulated to become anonymous. Indeed, whereas some data can be anonymous data from the beginning (such as climatic sensor data with no link to natural persons), other data may at some point be personal data but then be successfully manipulated to no longer relate to an identified or identifiable natural person. This underscores that the classification of personal data is dynamic. Depending on context, the same data point can be personal or non-personal and hence be subject to the Regulation or not.

This section introduces three causes of the bewildered definition of personal data. First, there is doubt regarding the appropriate legal test to be applied. Secondly, technical developments are further complicating this definitional exercise. Thirdly, the introduction of an explicit legal category of 'pseudonymous' data in the GDPR has induced confusion.

Personal data

Article 4(1) GDPR defines personal data 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.[4]

Personal data is hence data that directly or indirectly relates to an identified or identifiable natural person. The Article 29 Working Party has issued guidance on how the four constituent elements of the test in Article 4(1) GDPR—'any information', 'relating to', 'an identified or identifiable', and 'natural person'—ought to be interpreted.[5]

Information ought to be construed broadly, and includes objective information (such as a name or the presence of a given substance in one's blood) as well as subjective analysis such as information,

[1] Lilian Edwards, 'Data Protection I: Enter the GDPR' in Lilian Edwards (ed), *Law, Policy and the Internet* (Hart 2018) 84.

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, Recital 26 (GDPR).

[3] Note however the argument that in the future all data may be personal data, Nadezhda Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10 Law, Innovation and Technology 40.

[4] Art 4(1) GDPR.

[5] Article 29 Working Party, Opinion 04/2007 on the Concept of Personal Data (WP 136) 01248/07/EN, 6.

opinions, and assessments.[6] The European Court of Justice (ECJ) has, however, clarified in the meantime that whereas information contained in the application for a residence permit and data contained in legal analysis qualify as personal data, related legal analysis (the assessment) does not.[7] Personal data can moreover take any form and be alphabetical or numerical data, videos, and pictures.[8] Note, moreover, that Article 4(1) GDPR refers to 'information' rather than just data, indicating that the data appears to require some informational value. Of course, the distinction between information and data is not always easy to draw.

Data is considered to be *relating to* a data subject 'when it is about that individual'.[9] This includes information that is in an individual's file but also vehicle data that reveals information about the data subject.[10] An individual is considered to be *identified or identifiable* where they can be 'distinguished' from others.[11] This does not require that the individual can be identified by a name, rather she could also be identified through alternative means such as a telephone number.[12] This underlines that the concept of personal data ought to be interpreted broadly, a stance that has been embraced by the Court time and time again. It held in *Nowak* that the expression 'any information' reflects 'the aim of the EU legislature to assign a wide scope to that concept, which is not restricted to information that is sensitive or private, but potentially encompasses all kinds of information, not only objective but also subjective'.[13] In *Digital Rights Ireland*, the ECJ established that metadata (such as location data or IP addresses combined with log files on retrieved web pages) which only allows for the indirect identification of the data subject can nonetheless be personal data as it makes it possible 'to know the identity of the person with whom a subscriber or registered user has communicated and by what means, and to identify the time of the communication as well as the place from which that communication took place'.[14] Finally, Article 4(1) GDPR underlines that personal data must relate to *a natural person*. The GDPR does not apply to legal persons or the deceased.[15]

The above overview has underlined that the concept of personal data ought to be interpreted broadly. Yet, not all data constitute personal data.

Differentiating between personal data and non-personal data

The European data protection framework acknowledges two categories of data: personal and non-personal data. There is data that is always non-personal (because it never related to an identified or identifiable natural person) and there is also data that once was personal but no longer is (as linkage to a natural person has been removed). The legal test to differentiate between personal and non-personal data is embodied in Recital 26 GDPR according to which:

> [p]ersonal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.

Data not caught by this test falls outside the scope of European data protection law. Indeed, Recital 26 GDPR goes on to state that:

> The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.

Pursuant to the GDPR data is hence personal when the controller or another person is able to identify the data subject by using the 'means reasonably likely to be used'. Where personal data never related to a natural person or is no longer reasonably likely to be attributed to a natural person, it qualifies as 'anonymous

[6] Ibid.

[7] Joined Cases C-141/12 and C-372/12 *YS* [2014] EU:C:2014:2081.

[8] A29WP on the concept of personal data (n 5) 7; Case C-345/17 *Sergejs Buivids* [2019] EU:C:2019:122, para 31.

[9] A29WP on the concept of personal data, ibid 9.

[10] Ibid 10.

[11] Ibid 12.

[12] Ibid 14; Case C-101/01 *Bodil Lindqvist* [2003] EU:C:2003:596, para 27.

[13] Case C-434/16 *Peter Nowak* [2017] EU:C:2017:582, para 34.

[14] Cases C-293/12 and C-594/12 *Digital Rights Ireland* [2014] EU:C:2014:238, para 26.

[15] Bart van der Sloot, 'Do Privacy and Data Protection Rules Apply to Legal Persons and Should They? A Proposal for a Two-tiered System' (2015) 31 Computer Law and Security Review 26.

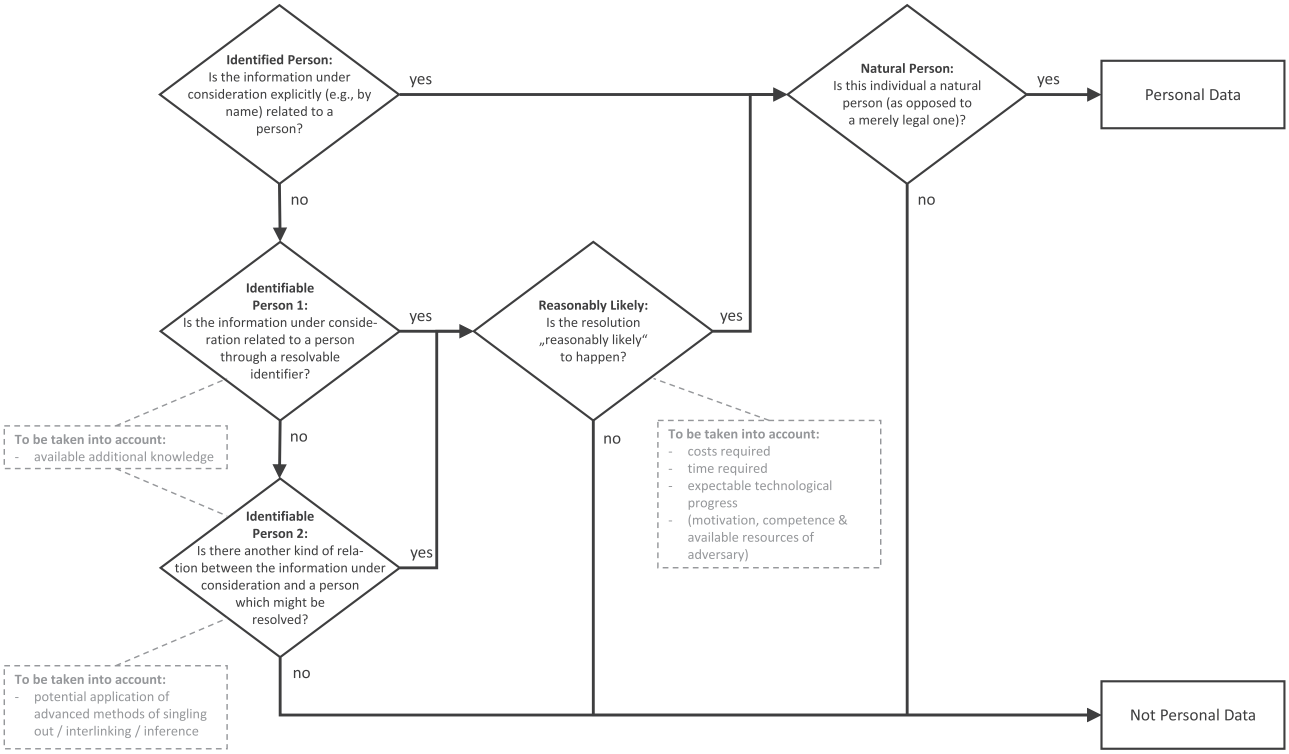
Identified Person:
Is the information under consideration explicitly (e.g., by name) related to a person?
yes
no
Natural Person:
Is this individual a natural person (as opposed to a merely legal one)?
yes
no
Personal Data
Identifiable Person 1:
Is the information under conside-ration related to a person through a resolvable identifier?
yes
no
Reasonably Likely:
Is the resolution „reasonably likely" to happen?
yes
no
To be taken into account:
- available additional knowledge
To be taken into account:
- costs required
- time required
- expectable technological progress
- (motivation, competence & available resources of adversary)
Identifiable Person 2:
Is there another kind of rela-tion between the information under consideration and a person which might be resolved?
yes
no
To be taken into account:
- potential application of advanced methods of singling out / interlinking / inference
Not Personal Data

Figure 1 Assessment scheme for person-relatedness of data under the GDPR

information' and eschews the Regulation's scope of application. Figure 1 depicts the test to be applied to determine whether information constitutes personal data.

The test devised by Recital 26 GDPR essentially embraces a risk-based approach to qualify information.[16] Where there is a reasonable risk of identification, data ought to be treated as personal data. Where that risk is merely negligent, data can be treated as non-personal data, and this even though identification cannot be excluded with absolute certainty. A closer look reveals, however, that some of the elements of this test suffer from a lack of clarity, resulting in particular from contrasting interpretations by various supervisory authorities.

Making sense of the various elements of Recital 26 GDPR

Although Recital 26 GDPR appears to embrace a straightforward approach to distinguish between personal and non-personal data, in practice, it has often proven difficult to implement. This becomes obvious when dividing the overall test embodied in the GDPR into its various constituent elements.

What risk? The uncertain standard of identifiability

Recital 26 GDPR formulates a risk-based approach to determine whether data is personal in nature or not. Where identification is 'reasonably likely' to occur, personal data is in play, where this is not the case the information in question is non-personal. Some national supervisory authorities have embraced interpretations of the GDPR that largely appear in line with this risk-based approach. The UK Information Commissioner s Office (ICO), for instance, adopts a relativist understanding of Recital 26 GDPR, stressing that the relevant criterion is 'the identification or likely identification' of a data subject.[17] This acknowledges that 'the risk of re-identification through data linkage is essentially unpredictable because it can never be assessed with certainty what data is already available or what data may be released in the future'.[18] The Irish Data Protection Authority (DPA) deems that it is not 'necessary to prove

[16] This risk-based approach is introduced in further detail in the paper's final section.

[17] Information Commissioner's Office, 'Anonymisation: Managing Data Protection Risk Code of Practice' (November 2012) 16 <https://ico.org.uk/media/1061/anonymisation-code.pdf> accessed 9 January 2020.

[18] Ibid.

that it is impossible for the data subject to be identified in order for an anonymisation technique to be successful. Rather, if it can be shown that it is unlikely that a data subject will be identified given the circumstances of the individual case and the state of technology, the data can be considered anonymous'.[19]

In its 2014 guidelines on anonymization and pseudonymization, the Article 29 Working Party, however, adopted a different stance. On the one hand, the Working Party acknowledges the Regulation's risk-based approach.[20] On the other hand, it, however, appears to devise its own independent zero-risk test. Its guidelines announce that 'anonymisation results from processing personal data in order to irreversibly prevent identification'.[21] Similarly, the guidance document announces that 'anonymisation is a technique applied to personal data in order to achieve irreversible de-identification'.[22] This strict position is in line with earlier guidance from 2007 according to which anonymized data is data 'that previously referred to an identifiable person, but where that identification is *no longer possible*'.[23] This means that 'the outcome of anonymisation as a technique applied to personal data should be, in the current state of technology, *as permanent as erasure, i.e. making it impossible to process personal data*'.[24]

What is more, the Working Party considers that 'when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data'.[25] This has been criticized as there may well be scenarios where a controller wants to release anonymous data while needing to keep the original dataset, as would be the case where a hospital makes available anonymized data for research purposes while retaining the original data for patient care.[26] This in itself is a rejection of the risk-based approach as it considers the risk stemming from keeping the initial data to be intolerable. As Stalla-Bourdillon and Knight explain, the combination of A29WP emphasis on original dataset and wording of Recital 26 GDPR is 'problematic since as the definition of pseudonymisation refers to both identified and identifiable data subjects the risk remains that data will be considered pseudonymised as long as the raw dataset has not been destroyed, even if the route of anonymisation through aggregation has been chosen'.[27] Beyond, the opinion also uses expressions that are difficult to make sense of such as 'identification has become reasonably impossible'—although it is unclear what reasonably (a qualified term) impossible (an absolute term) could mean.[28]

Compared to the risk-based approach of the GDPR, the Working Party thus appears to consider that no amount of risk can be tolerated. Indeed, the concepts of irreversibility, permanence, and impossibility stand for a much stricter approach than that formulated by the legislative text itself. Whereas Recital 26 acknowledges that anonymization can never be absolute (such as where technology changes over time), the Working Party's absolutist stance indicates that anonymization ought to be permanent. These diverging interpretations have prevented legal certainty as to what test ought to be applied in practice.[29]

The tension between the A29WP's no-risk stance and the risk-based approach embraced by Recital 26 GDPR can also be identified in guidance released by national authorities. To the French Commission Nationale de l'Informatique et des Libertés (CNIL), anonymization consists in making 'identification practically impossible'. It deems that anonymization 'seeks to be irreversible' so as to no longer permit the processing of personal data.[30] This reference to impossibility is more helpful as it clarifies that impossibility is the goal, yet also recognizes that it can be difficult to achieve in practice. To the French Conseil d'État, the highest national administrative court, data can, however, only be considered anonymous if the direct or indirect identification of the person becomes 'impossible', and this notwithstanding whether evaluated from the perspective of the data controller or a third person.[31] In

[19] Data Protection Commission, 'Guidance on Anonymisation and Pseudonymisation' (June 2019) 5 <https://www.dataprotection.ie/sites/default/files/uploads/2019-06/190614%20Anonymisation%20and%20Pseudonymisation.pdf> accessed 9 January 2020.

[20] Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques (WP 216) 0829/14/EN, 11–12, 23–25.

[21] Ibid 3.

[22] Ibid 7.

[23] Ibid 21 (emphasis added).

[24] Ibid 6 (emphasis added).

[25] Ibid 9.

[26] Khaled El Emam and Cecilia Álvarez, 'A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques' (2015) 5 International Data Privacy Law 73, 81–82.

[27] Sophie Stalla-Bourdillon and Alison Knight, 'Anonymous Data v. Personal Data - A False Debate: An EU Perspective on Anonymisation, Pseudonymisation and Personal Data' (2017) 34 Wisconsin International Law Journal 284, 301.

[28] A29WP on Anonymisation Techniques (n 20) 8.

[29] El Emam and Álvarez (n 26) 75.

[30] Commission Nationale de l'Informatique et des Libertés, 'Comment prévenir les risques et organiser la sécurité de vos données ?' (16 April 2019) <https://www.cnil.fr/fr/comment-prevenir-les-risques-et-organiser-la-securite-de-vos-donnees> accessed 9 January 2020.

[31] Conseil d'État, 10ème – 9ème ch. réunies, décision du 8 février 2017, N° 393714 (citing art 2 of the Law of 6 January 1978) ('une telle donnée ne peut être regardée comme rendue anonyme que lorsque l'identification de la personne concernée, directement ou indirectement, devient impossible que ce soit par le responsable du traitement ou par un tiers').

contrast, the highest French civil court, the cour de cassation concluded that IP addresses are personal data without justifying why this is the case.[32] The Finnish Social Science Data Archive similarly considers that for the data to count as anonymous 'anonymisation must be irreversible'.[33]

What elements ought to be taken into account to determine whether anonymization has occurred?

Pursuant to Recital 26 GDPR, the relevant criterion to assess whether data is pseudonymous or anonymous is identifiability.[34] To determine whether an individual can be identified consideration ought to be given to 'all means reasonably likely to be used'. This includes 'all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments'.[35]

In addition, the A29WP considers that three criteria ought to be considered to determine whether de-identification has occurred namely if (i) it is still possible to single out an individual; (ii) it is still possible to link records relating to an individual, and (iii) information concerning an individual can still be inferred.[36] Where the answer to these three questions is negative, data can be considered anonymous. It should be noted that while Recital 26 GDPR now makes explicit reference to 'singling out', inference and linkability are elements considered by the Working Party but not explicitly mentioned in the GDPR.

Singling out refers to 'the possibility to isolate some or all records which identify an individual in the dataset'.[37] Linkability denotes the risk generated where at least two data sets contain information about the same data subject. If in such circumstances an 'attacker can establish (e.g. by means of correlation analysis) that two records are assigned to a same group of individuals but cannot single out individuals in this group', then the used technique only provides resistance against singling out but not against linkability.[38] Finally, inference may still be possible even where singling out and linkability are not. Inference has been defined by the Working Party as 'the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes'.[39] The Working Party underlined that meeting these three thresholds is very difficult.[40] This is confirmed by its own analysis of the most commonly used 'anonymisation' techniques, which revealed that each method leaves a residual risk of identification so that, if at all, only a combination of different approaches can successfully de-personalize data.[41]

What is the relevant time scale?

Recital 26 requires that the 'means' to be taken into account are not just those that are presently available, but also 'technological developments'. It is, however, far from obvious what timescale ought to be considered in this respect. Recital 26 does not reveal whether one ought to account for ongoing technological changes (such as a new technique that has been rolled out across many sectors but not yet to the specific data controller or processor) or whether developments currently just explored in research should also be given consideration. To provide a concrete example, it is not obvious whether the still uncertain prospect of quantum computing should be factored in when determining whether a certain encryption technique could transform personal data into anonymous data.[42]

The A29WP indicated that one should consider both 'the state of the art in technology at the time of the processing' as well as 'the possibilities for development during the period for which the data will be processed'. In respect to the second scenario, the lifetime of the data is a key factor. Indeed, where data is to be kept for a decade, the data controller 'should consider that possibility of identification that may occur also within the ninth year of their lifetime, and which may make them personal data at that moment'.[43] This indicates that the data in question only becomes personal information in the ninth year, yet from the beginning the controller must be aware of, and prepare for, that possibility. This highlights the GDPR's nature as a risk-management framework, which is further explored in the concluding section.

Pursuant to the Working Party, the relevant system 'should be able to adapt to these developments as they

[32] Cour de cassation, chambre civile 1, arrêt du 3 novembre 2016, N° 1184 (15-22.595).

[33] 'Anonymisation and Personal Data' (*Finnish Social Science Data Archive*, 24 June 2019) <https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html> accessed 9 January 2020.

[34] Recital 26 GDPR.

[35] Ibid.

[36] A29WP on Anonymisation Techniques (n 20) 3.

[37] Ibid 11.

[38] Ibid 11.

[39] Ibid 12.

[40] Ibid 4.

[41] Ibid 3.

[42] 'Quantum computers will break the encryption that protects the internet' *The Economist* (London, 20 October 2018) <https://www.economist.com/science-and-technology/2018/10/20/quantum-computers-will-break-the-encryption-that-protects-the-internet> accessed 9 January 2020.

[43] A29WP on the concept of personal data (n 5) 15 (emphasis added).

happen, and to incorporate the appropriate technical and organisational measures in due course'.[44] Indeed, the assumption is that data becomes personal data at the moment identification becomes possible. The relevant question appears to be whether a given dataset can be matched with other datasets from the perspective of availability rather than technical possibility. The risk that the entity in possession of the dataset may in the future acquire (access to) additional information that, in combination, may enable identification is accordingly not considered to legally qualify data in the present. This has been criticized as 'the characterisation of anonymised data should also be dependent upon an ongoing monitoring on the part of the initial data controller of the data environment of the dataset that has undergone anonymisation'.[45] In fact, in times where data generation continues to accelerate, an entity may have access to a dataset that on its face is anonymous but might then, purposefully or not, subsequently gain access to a dataset containing information that enables re-identification. The resulting data protection risks are considerable. Yet, it is questionable how a test addressing this *ex ante* could be fashioned as there is often little way of predicting what data may be generated or acquired in the future. As such it might be more realistic to acknowledge data's dynamic nature and that anonymous data becomes personal data as soon as identification becomes possible. In any event, data controllers have a monitoring obligation and must adopt technical and organizational measures in due course.

Personal data to whom?

To determine whether information constitutes personal data, it is important to know from whose perspective the likelihood of identification ought to be assessed. Recital 26 provides that to determine identifiability 'account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller *or by another person* to identify the natural person directly or indirectly'. This formulation appears to indicate that it is not sufficient to evaluate identifiability from the perspective of the controller but potentially also any other third party.

The GDPR is a fundamental rights framework and the ECJ has time and time again emphasized the need to provide an interpretation thereof capable of ensuring the complete and effective protection of data subjects. From this perspective, it matters little from whose perspective data qualifies as personal data—anyone should protect the data subject's rights. In the academic literature, there has long been a debate as to whether there is a need to only focus on the data controller (a relative approach) or any third party (an absolute approach).[46] Some have criticized the absolute approach, highlighting that the reference to 'another person' eliminates 'the need for any risk management because it compels the data controller to always make the worst possible assumptions even if they are not relevant to the specific context'.[47]

Some supervisory authorities appear to have embraced a half-way test between the absolute and relative approach. The ICO formulated the 'motivated intruder' test whereby companies should determine whether an intruder could achieve re-identification if motivated to attempt this.[48] The motivated intruder is assumed to be 'reasonably competent' and with access to resources such as the Internet, libraries, or all public documents but should not be assumed to have specialist knowledge such as hacking skills or to have access to 'specialist equipment'.[49]

In the European Courts, *Breyer* is the leading case on this matter.[50] Mr Breyer had accessed several websites of the German federal government that stored information regarding access operations in logfiles.[51] This information included the visitor's dynamic IP address (an IP address that changes with every new connection to the Internet. Breyer argued that the storage of the IP address was in violation of his rights. The ECJ had already decided in *Scarlet Extended* that static IP addresses 'are protected personal data because they allow those users to be precisely identified'.[52] Recital 30 GDPR also considers IP addresses as online identifiers.

The Court noted that the collection and identification of IP addresses was carried out by the Internet Service Provider (ISP), whereas in the case at issue the collection and identification of the IP address was carried out by the German federal government, which 'registers IP addresses of the users of a website that it makes accessible to the public, without having the additional data necessary in order to identify those users'.[53] A dynamic IP address is not data related to an identified natural personal but can be considered to make log

[44] Ibid 15 (emphasis added).

[45] Stalla-Bourdillon and Knight (n 27) 288.

[46] For a brief overview of the relative and absolute approaches, see Gerald Spindler and Phillip Schmechel, 'Personal Data and Encryption in the European General Data Protection Regulation' (2016) 7 JIPITEC 163.

[47] El Emam and Álvarez (n 26) 83.

[48] Information Commissioner's Office (n 17) 22.

[49] Ibid

[50] Case C-582/14 *Patrick Breyer* [2016] EU:C:2016:779.

[51] Ibid.

[52] Case C-70/10 *Scarlet Extended* [2011] EU:C:2011:771, para 51.

[53] *Breyer* (n 50) para 35.

entries that relate to an identifiable person where the necessary additional data are held by the ISP.[54] The dynamic IP address accordingly qualified as personal data even though the data to identify Mr Breyer was not held by German authorities but by the ISP.[55]

In isolation, this would imply that the nature of data ought not just to be evaluated from the perspective of the controller (German authorities; the relative approach) but also from the perspective of third parties (the ISP; the absolute approach). Indeed, 'there is no requirement that all the information enabling the identification of the data subject must be in the hands of one person'.[56] However, that finding may have been warranted by the specific facts at issue. The Court stressed that whereas it is in principle prohibited under German law for the ISP to transmit such data to website operators, the government has the power to compel ISPs to do so in the event of a cyberattack. This is also an interesting statement as it implies that cyberattacks are events that are 'reasonably likely'—arguably highlighting that the standard of reasonable likelihood to be applied is not very strict. *Breyer* hence confirms the risk-based approach in Recital 26 GDPR as the Court indeed evaluates the actual risk of identification.

The *Breyer* ruling also begs an additional question. Indeed, the Court's emphasis on the legality (for the government only) of compelling ISPs to reveal the data necessary to re-personalize a de-personalized dataset was key to its conclusion. This makes us wonder, on the one hand, whether the illegality of an act that enables identification means that it should always be considered as reasonably unlikely.

This relativist approach to identifiability has been endorsed in other contexts as well. Some have argued in relation to cloud computing that 'to the person encrypting personal data, such as a cloud user with the decryption key, the data remain "personal data"'.[57] In the Safe Harbor agreement, the Commission considered that the transfer of key-coded data to the USA is not a personal data export where the key was not revealed or transferred alongside the data.[58] Recently, an English court embraced a cautionary approach to *Breyer*, arguing that whereas the ECJ's ruling depended 'on specific factual aspects of the German legal system', it should not be held that the mere fact that a party can use the law to gain access to data to 'identify a natural person' would make that procedure a 'means reasonably likely to be used'.[59]

The above would indicate that the perspective from which identifiability ought to be assessed is that of the data controller. In *Breyer*, Advocate General Campos Sánchez-Bordona warned that if the contrary perspective were adopted, it would never be possible to rule out with absolute certainty 'that there is no third party in possession of additional data which may be combined with that information and are, therefore, capable of revealing a person's identity'.[60]

As a consequence, there is currently 'a very significant grey area, where a data controller may believe a dataset is anonymised, but a motivated third party will still be able to identify at least some of the individuals from the information released'.[61] Research has moreover pointed out that where a data controller implements strategies to burden the re-identification of data, this does not mean that adversaries will be incapable of identifying the data, particularly since they might have a higher tolerance for inaccuracy as well as access to additional (possibly illegal) databases.[62] On the other hand, adopting an absolute approach could effectively rule out the existence of anonymous data as ultimately there will always be parties able to combine a dataset with additional information that may re-identify it.

An objective or subjective approach?

It is furthermore unclear from whose perspective the risk of identification ought to be evaluated. Recital 26 foresees that a 'reasonable' investment of time and financial resources should be considered to determine whether a specified natural person can be identified. There is, however, an argument to be made that what is a 'reasonable' depends heavily on context. The characterization of data is context-dependent, so that personalization 'should not be seen as a property of the data but as a property of the environment of the data'.[63] It is indeed fair to assume that reasonableness ought to be evaluated differently depending on whether the entity

[54] Ibid, para 39.

[55] Ibid, para 49.

[56] Ibid, para 43; *Nowak* (n 13) para 31.

[57] Kuan Hon and others, 'The Problem of "Personal Data" in Cloud Computing: What Information Is Regulated? - The Cloud of Unknowing' (2011) 1 International Data Privacy Law 211, 219.

[58] Commission Decision 2000/520/EC of 26 July 2000 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequacy of the protection provided by the safe harbour privacy principles and related frequently askedquestions issued by the US Department of Commerce [2000] OJ L215/7, 24.

[59] *Mircom International Content Management & Consulting Ltd v Virgin Media Ltd* [2019] EWHC 1827 (Ch) [27].

[60] *Breyer* (n 50), Opinion of AG Campos Sánchez-Bordona, para 65.

[61] Article 29 Working Party, Opinion 03/2013 on Purpose Limitation (WP 203) 00569/13/EN, 31.

[62] Michael Veale, Reuben Binns and Jef Ausloos, 'When Data Protection by Design and Data Subject Rights Clash' (2018) 8 International Data Privacy Law 105, 107.

[63] Stalla-Bourdillon and Knight (n 27) 311–12.

concerned is a private person or a law enforcement agency or a major online platform. Whereas a case-by-case basis is in any event required, it is not obvious what standard of reasonableness ought to be applied, specifically whether the specific capacities of a relevant actor need to be factored in or not. Moreover, it is not clear whether an objective or subjective approach ought to be adopted. A subjective approach would require consideration of all factors within one's knowledge—specifically who has access to relevant data that enables identification. An objective approach would, however, require a broader evaluation, including who has access to information in the present and who might gain access to relevant data in the future.

The Irish DPA suggested that it should first be considered who the potential intruder might be before determining what the methods reasonably likely to be used are. Furthermore, organizations 'should also consider the sensitivity of the personal data, as well as its value to a potential intruder or any 3rd party who may gain access to the data'.[64] Indeed, when anonymized data is shared with the world at large, there is a higher burden to ensure effective anonymization as it is virtually impossible to retract publication once it becomes apparent that identification is possible.[65] With this in mind, it should be evaluated what other data these controllers have access to (such as public registers but also data available only to a particular individual or organization).[66] This appears to imply that all (known) data controllers need to be considered to determine the person-relatedness of a dataset. Restricting this exercise to known data controllers seems reasonable as it would be impossible for any party to exclude with absolute certainty that there is not another party able to identify allegedly anonymous data.

The UK ICO similarly considers that when anonymized data is released publicly, it is not only important to determine whether it is really anonymous data from the perspective of the controller releasing the data but also whether there are third parties that are likely to use prior knowledge to facilitate re-identification (such as a doctor that knows that an anonymized case study relates to one of her patients).[67] This indicates that what needs to be accounted for is the knowledge of third parties that could reasonably be expected to attempt to identify data, the subjective approach. Indeed, an absolute objective approach would present challenges as it would require much better knowledge of the wider world than a data controller typically has. A hospital releasing data that is 'anonymised' from its own perspective (such as for research purposes) cannot reasonably evaluate whether any other party in the world may have additional information allowing for identification. This is a particular challenge in open data contexts. Although those releasing the data may be confident that it is anonymous they cannot exclude with certainty whether other parties may be able to identify data subjects on the basis of additional knowledge they hold. An important open question that remains in this domain is thus from whose perspective the quality of data ought to be assessed: from the perspective of any third party or only of those third parties reasonably likely to make use of the additional information they have to proceed to re-personalize a de-personalized dataset. If the latter is the case, then the follow-on question becomes what parties can be considered reasonably likely to make use of such information and, moreover, whether presumed intent ought to be considered as a relevant factor here.

The purposes of data use

Finally, the A29WP stressed that when determining the nature of personal data, it is crucial to evaluate the 'purpose pursued by the data controller in the data processing'.[68] Indeed, 'to argue that individuals are not identifiable, where the purpose of processing is precisely to identify them, would be a sheer contradiction in terms'.[69] In the same vein, the French supervisory authority held that the accumulation of data held by Google that enables it to individualize persons is personal data as 'the sole objective pursued by the company is to gather a maximum of details about individualised persons in an effort to boost the value of their profiles for advertising purposes'.[70] In line with this reasoning, public keys or other sorts of identifiers used to identify a natural person constitute personal data.

After having introduced the general uncertainties regarding the taxonomy of personal and anonymous data, it will now be seen that ongoing technical developments further burden the legal qualification of data.

[64] Data Protection Commission (n 19) 8.

[65] Ibid.

[66] Ibid.

[67] Information Commissioner's Office (n 17) 25–26.

[68] A29WP on the Concept of Personal Data (n 5) 16.

[69] Ibid.

[70] Commission Nationale de l'Informatique et des Libertés, 'Délibération No. 2013-420 of the Sanctions Committee of CNIL, Imposing a Financial Penalty Against Google Inc' (8 January 2015) <www.cnil.fr/fileadmin/documents/en/D2013-420_Google_Inc_ENG.pdf> accessed 13 May 2019.

Technical developments and the definition of personal data

With the advent of ever more performant data analysis techniques and hardware as well as the heightened availability of data points, it is becoming increasingly straightforward to relate data to natural persons. Some have observed that data protection law may become the 'law of everything' as in the near future all data may be personal data and thus subject to the GDPR.[71] This is so as 'technology is rapidly moving towards perfect identifiability of information; datafication and advances in data analytics make everything (contain) information; and in increasingly 'smart' environments any information is likely to relate to a person in purpose or effect'.[72] The A29WP warned in the same vein that 'anonymisation is increasingly difficult to achieve with the advance of modern computer technology and the ubiquitous availability of information'.[73]

In light of these technical advancements, establishing the risk of re-identification can be difficult 'where complex statistical methods may be used to match various pieces of anonymised data'.[74] Indeed 'the possibility of linking several anonymised datasets to the same individual can be a precursor to identification'.[75] A particular difficulty here resides in the fact that it is often not known what datasets a given actor has access to, or might have access to in the future. The A29WP's approach to anonymization has accordingly been criticized as 'idealistic and impractical'.[76]

Research has amply confirmed the difficulties of achieving anonymization, such as where an 'anonymised' profile can still be used to single out a specific individual.[77] Big data moreover facilitates the de-anonymization of data through the combination of various datasets.[78] It is accordingly often easy to identify data subjects on the basis of purportedly anonymized data.[79] Some computer scientists have even warned that the de-identification of personal data is an 'unattainable goal'.[80] Recent research has confirmed that allegedly anonymous datasets often allow for the identification of specific natural persons as long as the dataset contains the person's date of birth, gender, and postal code.[81]

The language of anonymous data has been criticized as 'the very use of a terminology that creates the illusion of a definitive and permanent contour that clearly delineates the scope of data protection laws'.[82] This, however, is not the case where even data that is anonymous on its face may be subsequently matched with other data points. Early examples for such re-personalization of datasets thought to be anonymous include the re-identification of publicly released health data using public voter lists[83] or the re-personalization of publicly released 'anonymous' data from a video streaming platform through inference with other data from a public online film review database.[84] More recent research suggests that 99.98 per cent of the population of a US state could be uniquely re-identified within a dataset consisting of 15 demographic factors.[85]

In light of the above, it might be argued that the risk-based approach to anonymization enshrined in Recital 26 GDPR is the only sensible approach to distinguishing between personal and non-personal data. Indeed, in today's complex data ecosystems, it can never be assumed that the anonymization of data is 'as permanent as erasure'. Data circulates and is traded, new data sets are created, and third parties may be in possession of information allowing linkage, which the original data controller has no knowledge of. There are accordingly considerable complications in drawing the boundaries between personal and non-personal data. The GDPR now recognizes that when data is modified to decrease linkability, this does not necessarily result in anonymous but rather in pseudonymous data.

[71] Purtova (n 3) 40.

[72] Ibid 40.

[73] A29WP on Purpose Limitation (n 61) 31.

[74] Information Commissioner's Office (n 17) 21.

[75] Ibid.

[76] Stalla-Bourdillon and Knight (n 27) 298.

[77] Akiva Miller, 'What Do We Worry about When We Worry about Price Discrimination? The Law and Ethics of Using Personal Information for Pricing' (2014) 19 Journal of Technology Law & Policy 41.

[78] Paul Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' (2010) 57 UCL Law Review 1701; Michael Veale and others (n 62) 113.

[79] Latanya Sweeney, 'Simple Demographics Often Identify People Uniquely' (2000) Data Privacy Working Paper 3, Pittsburgh <https://dataprivacylab.org/projects/identifiability/paper1.pdf> accessed 9 January 2020.

[80] Arvind Narayanan and Vitaly Shmatikov, 'Myths and Fallacies of Personally Identifiable Information' (2010) 53 Communications of the ACM 24, 26.

[81] Luc Rocher, Julien M Hendrickx and Yves-Alexandre de Montjoye, 'Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models' (2019) 10 Nature Communications 3069.

[82] Stalla-Bourdillon and Knight (n 27) 287.

[83] Latanya Sweeney, 'k-anonymity: A Model for Protecting Privacy' (2002) 10 International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 557.

[84] Arvind Narayanan and Vitaly Shmatikov, 'Robust De-anonymization of Large Sparse Datasets' in *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (IEEE Computer Society 2008) 111–25.

[85] Rocher, Hendrickx and de Montjoye (n 81).

The concept of pseudonymous data under the GDPR

Article 4(5) GDPR introduces pseudonymization as the

> processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.[86]

The concept of pseudonymization is one of the novelties of the GDPR compared to the 1995 Data Protection Directive. There is an ongoing debate regarding the implications of Article 4(5), in particular, whether the provision gives rise to a third category of data beyond those of personal and anonymous data. A literal interpretation reveals, however, that Article 4(5) GDPR deals with a method, not an outcome of data processing.[87] Pseudonymization is the 'processing' of personal data in such a way that data can only be attributed to a data subject with the help of additional information. This underlines that pseudonymized data remains personal data, in line with the Working Party's finding that 'pseudonymisation is not a method of anonymisation. It merely reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure.'[88] Thus pseudonymous data is still 'explicitly and importantly, personal data, but its processing is seen as presenting less risk to data subjects, and as such is given certain privileges designed to incentivise its use'.[89] The Irish supervisory authority concurs that pseudonymization 'should never be considered an effective means of anonymisation'.[90]

The GDPR explicitly encourages pseudonymization as a risk-management measure. Pseudonymization can serve as evidence of compliance with the controller's security obligation under Article 5(f) GDPR and confirm that the data protection by design and by default requirements has been duly considered.[91] Recital 28 further provides that '[t]he application of pseudonymisation to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations.'[92] According to Recital 29, pseudonymization is possible 'within the same controller' when that controller has taken appropriate technical and organizational measures. It is interesting to note that Recital 29 explicitly facilitates this in order to 'create incentives to apply pseudonymisation when processing personal data'.

Pseudonymized data can, however, still be linked to natural persons. Recital 30 recalls that data subjects may be 'associated with online identifiers provided by their devices, applications, tools and protocols, such as Internet protocol addresses, cookie identifiers or other identifiers'.[93] These enable identification when they leave traces which 'in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them'.[94]

It is worth stressing that even though pseudonymized data may fall short of qualifying as anonymized data, it may be caught by Article 11 GDPR, pursuant to which the controller is not obliged to maintain, acquire, or process additional information to identify the data subject in order to comply with the Regulation.[95] In such scenarios, the controller does not need to comply with Articles 15–20 GDPR unless the data subject provides additional information enabling their identification for the purposes of exercising their GDPR rights.[96]

There is thus ample recognition that whereas pseudonymization serves as a valuable risk-minimization approach, it falls short of being an anonymization technique. Before the revision of data protection law through the GDPR, there was some confusion regarding the legal distinction between pseudonymization and anonymization. Some supervisory authorities considered that pseudonymization can produce anonymous data.[97] It has been suggested that this confusion may be rooted in the fact that in computer science pseudonyms are understood as 'nameless' identifiers and thus not necessarily anonymous data.[98] In any event, the GDPR is now unequivocal that

[86] Art 4(5) GDPR.

[87] Miranda Mourby and others, 'Are "Pseudonymised" Data Always Personal Data? Implications of the GDPR for Administrative Data Research in the UK' (2018) 34 Computer Law & Security Review 222, 223.

[88] A29WP on Anonymisation Techniques (n 20) 3.

[89] Edwards (n 1) 88. See also Mourby and others (n 87) 222.

[90] Data Protection Commission (n 19) 13.

[91] Arti 25 GDPR. See further Edwards (n 1) 88.

[92] Recital 28 GDPR.

[93] Recital 30 GDPR.

[94] Recital 30 GDPR.

[95] Art 11(1) GDPR.

[96] Art 11(2) GDPR.

[97] Information Commissioner's Office (n 17) 21 ('[t]his does not mean, though, that effective anonymisation through pseudonymisation becomes impossible').

[98] Frederik Zuiderveen Borgesius, 'Singling Out People Without Knowing Their Names - Behavioural Targeting, Pseudonymous Data, and the New Data Protection Regulation' (2016) 32 Computer Law & Security Review 256, 258.

pseudonymized data is still personal data. Interestingly, however, the GDPR only looks towards one specific method of identifier replacement—referred as 'traditional pseudonymisation' below—that uses additional, separately kept information to re-personalize pseudonymised data. The A29WP uses a different definition for pseudonymization,[99] only increasing terminological confusion around 'pseudonymisation'.

Moreover, it is also worth noting that Article 4(5) GDPR may be read as considering that whenever there is additional data available (with the same controller?) that allows for the personalization of a de-personalized dataset, then this always amounts to personal data. Stated otherwise, a data controller is unable to anonymize data by separating a dataset that is de-personalized from a dataset that would enable re-personalization, even where the adoption of technical and organizational measures makes re-personalization reasonably unlikely. Given the pronounced practical relevance of that question, the adoption of regulatory guidance to specify whether this is in fact the case would be helpful.

Having laid out the legal foundations for determining whether a certain piece of data is to be considered personal data or non-personal data under the GDPR, we now move on to the technical dimension of anonymization and pseudonymization.

Technical approaches to identifier replacement

Different technical approaches can be used to remove explicit links to natural persons from data that differ regarding the possibilities of re-personalization and, in particular, the additional knowledge and resources (in the form of computational power) necessary to achieve re-personalization. They also differ with regard to the linkability of single data points within a dataset or across different datasets. We therefore present different established patterns of replacing explicit identifiers in datasets. Given the importance of re-identification to legally qualify data, we discuss these patterns particularly with respect to the means reasonably likely to be used test.

There are two different starting points for re-personalization, namely (1) re-identification starting from clear-text information, eg when we have a person's ID and want to find all data points related to this ID from a set of de-personalized data and (2) re-identification starting from a de-personalized dataset, eg when we want to know the identities behind (all or some) data points matching certain criteria. We moreover have to distinguish between (i) identifier-based re-identification (learn the relation between a clear-text identifier and its obfuscated counterpart) and (ii) content-based re-identification (learn which person is behind an obfuscated ID based on the content—like motion profiles—linked to this obfuscated ID).

To illustrate, imagine a scenario where the transfer of goods and respective payments among different actors be tracked without revealing the parties' identities. For this case, the four possible approaches for re-identification can be depicted as follows:

	1. Start from known person	2. Start from content
A. ID-based re-identification	'Find all transactions that John Smith was involved in, based on his known ID'	'Find the persons involved in transaction *X*, based on known IDs of all persons to be considered'
B. Content-based re-identification	'Find all transactions that John Smith was involved in through matching transaction data with his known bank account history'	'Find the persons involved in transaction *X* through matching transaction data with bank account histories of all persons to be considered'

Depending on the de-personalization pattern, the likelihood of a successful re-personalization can vary strongly between those four approaches. We therefore briefly introduce several established de-personalization patterns and discuss respective likelihoods for each of them.

Pattern 1: traditional pseudonymization

This is the traditional way of achieving pseudonymization. It essentially consists in replacing those elements

[99] 'replacing one attribute . . . in a record by another', A29WP on Anonymisation Techniques (n 20) 20.

of a data point that represent explicit identifiers (ID numbers or a combination of first and last name with date of birth) with a random number[100] and creating a separate table that matches this random number to the explicit identifiers. This results in the original dataset being split into two separate datasets which can be stored and handled differently. For instance, a data point like would be separated into three data points stored in two different datasets:

Sender	Receiver	Date	Type of good	Price
John Smith	Jane Miller	12 February 2019	Notebook	1375.12

Sender	Receiver	Date	Type of good	Price
2342	1337	12 February 2019	Notebook	1375.12

Pseudonym	Name
1337	Jane Miller
2342	John Smith

Without access to the pseudonym table, data from the transaction dataset do not, in itself, allow to directly identify data subjects. Data in the transaction table is pseudonymized as, in line with Article 4(5) GDPR, it 'can no longer be attributed to a specific data subject without the use of additional information'.

In such scenarios, re-identification must often explicitly remain possible in some cases, such as for fraud prevention. The pseudonym table is typically held separately and only accessed in such cases. Consequently, any party having access to this table can re-personalize the transaction data, meaning that it can likely not qualify as anonymous data at least from the perspective of the data controller having access to said data. What is interesting about Article 4(5) GDPR is that it appears to assume that as long as there is additional information, this is pseudonymous—not anonymous data, hence not envisaging the possibility that there can be sufficient technical or organizational safeguards making identification reasonably unlikely.

For identifier-based re-identification, re-identifiability does not significantly depend on whether we start with a clear-text ID ('John Smith") and want to find all transactions this ID was involved in or if we start with a particular transaction and want to know the clear-text identities of the involved parties. In both cases, what is needed is access to the pseudonym mapping table. With access to this table, no further knowledge is needed for re-identification. At least to those having access to this mapping table, pseudonymized data is personal data.

Parties without access to this table might pursue content-based re-identification. A party with access to Jane Smith's payment history and the pseudonymized transaction table could easily learn what she bought by simply matching these two datasets and finding the transaction that was made some days before 15 February 2019 for exactly €1357.12. Similarly, a party wanting to know who bought the Notebook for €1357.12 on 12 February 2019 and having access to a sufficient amount of potentially relevant bank account or credit card histories could try to identify the person by scanning these histories for transfer entries matching the price and date.[101] In practice, both situations could emerge in case of an online retailer handing over pseudonymized transaction data to banks or credit card companies.

Besides access to the pseudonym table, person-relatedness can also arise from access to other data that helps re-identify pseudonymized transactions through content-matching (starting from a transaction or person). The possibilities for content-matching include more sophisticated approaches of automated and sometimes fuzzy pattern-matching across multiple datasets, where machine-learning algorithms identify patterns that were not explicitly searched for. For instance, an algorithm could find a congruence pattern between a particular cookbook being ordered to a given postal code and a specific combination of vegetables being ordered to the same postal code from a different online shop. This might then be used to re-identify—with some certainty—a person from a de-personalized dataset.[102] The

[100] In so doing, it is ensured that random numbers are not allocated twice to avoid conflicts.

[101] The pseudonym may also be used to ascertain the correctness of re-identification by matching other transactions of the same pseudonym to the same bank account history.

[102] The possibilities go thus beyond those mentioned by the A29WP on Anonymisation Techniques (n 20) 21.

likelihood of this approach being successful can hardly be estimated in advance, as it depends on the available data. This raises an interesting question of broader relevance for the correlation methods used by techniques such as machine- and deep learning, namely what percentage of accuracy needs to be achieved for an identification technique to be considered reasonably likely. While this is an important question in this specific context, it is also pivotal for the relation between the GDPR and Artificial Intelligence more generally.

Pattern 2—hash-based ID replacement[103]

A distinct way of removing explicit links to data subjects from data is to replace those parts of the data that represent explicit identifiers by the hash of these data. A hash is the result of a hash function, which is a well-defined method for mapping a piece of data of arbitrary size onto a hash value.[104] A given hash function (such as SHA3 or the outdating SHA2) always produces the same hash value for a given piece of input data.

- The SHA3-Hash[105] of 'John Smith' will always be '900085141454845707708206b78720a71d908aaaacd571f054efdccbd7e6c7da'.
- The SHA3-Hash of 'Jane Miller' will always be '799eb189fd71f62a2cbb044286dba9ff778ce84920663ef01d6fce6687af4b26'.

Generally speaking, current hash functions can in most cases be assumed to produce (significantly) different results for (even slightly) different input values. Most importantly, however, the original input data cannot be recalculated from the hash value. Hash functions are thus non-revertible or one-way functions. Applying this pattern, we get an obfuscated dataset like the following:

Sender	Receiver	Date	Type of good	Price
900085141...	799eb189f...	12 February 2019	Notebook	1375.12

In contrast to the first pattern, there is no pseudonym mapping table here. Instead, the mapping between a clear-text ID and the replacement is solely defined by the hash function. This has important implications for re-identifiability. ID-based re-identification of obfuscated data requires no additional knowledge (except knowing which hash function was used—which can be guessed or tried out) and the effort is negligible. Any party can easily recalculate the hash value of 'John Smith' using freely available software libraries or even online tools and—with access to the obfuscated transaction data—thereby easily identify all transactions John was involved in. ID-based re-identification starting from a known person thus requires very little effort.

Due to the one-way characteristic of hash functions and the lack of a mapping table, directly re-identifying participant IDs from obfuscated transaction data (ID-based re-identification starting with given content) is impossible. The only possible ways to learn that John Smith is the data subject behind the stored hash value are (i) to try out all possible identities that might be the original clear-text value, apply the hashing function, and then check whether the result is a match ('brute-forcing') and (ii) to identify John Smith based on inference with other data, for instance, based on depersonalized transactions in one database and personalized transactions with similar amounts in another database.

Leaving aside inference-based re-identification for the moment, the effort necessary for a successful re-identification through brute-forcing depends on how many different candidates for clear-text identities exist. If there are 20, calculating the hashes for each and comparing them to the obfuscated transaction data are a negligible effort.[106] With an increasing amount of possible clear-text values, the effort increases. If half of the inhabitants of the EU (around 250 million people) were possible candidates, re-identifying one of them from a given obfuscated dataset would on average require the calculation and comparison of 125 million hashes.

[103] Different from the A29WP on Anonymisation Techniques (n 20) 20ff, we avoid a terminology of 'hash pseudonyms' here as it would unnecessarily provoke confusion about implications with regard to the legal concept of pseudonymization.

[104] For a more extensive explanation of hashing techniques, see also Agencia Española de Protección de Datos and European Data Protection Supervisor, 'Introduction to the Hash Function as a Personal Data Pseudonymisation Technique' (October 2019) <https://edps.europa.eu/sites/edp/files/publication/19-10-30_aepd-edps_paper_hash_final_en.pdf> accessed 9 January 2020.

[105] Using SHA3-256 as example, calculated via <https://emn178.github.io/online-tools/sha3_256.html> 9 January 2020.

[106] See generally Ed Felten, 'Does Hashing Make Data "Anonymous"?' (*Federal Trade Commission*, 22 April 2012) <https://www.ftc.gov/news-events/blogs/techftc/2012/04/does-hashing-make-data-anonymous> accessed 9 January 2020.

While the Working Party assumes—in line with its absolute approach—that hash-based de-personalization does not render data anonymous because of the mere possibility of brute-forcing,[107] the relative approach would require an assessment of whether such a brute-force ID-based re-identification is 'reasonably likely'. This, in turn, depends on how many attempts can actually be made in what timeframe and at what cost, which requires that we know whether an objective or subjective approach to identification be adopted.

Even though scenario-specific factors play a role and technology constantly becomes more powerful, having reference points regarding respective 'hash-rates'—at least in the sense of orders of magnitude—would be highly valuable. Technologies and hash-rates here significantly differ between different hash-algorithms. Highly-optimized single-purpose devices specifically built for hashing-intensive Bitcoin mining[108] provide a hash rate of more than 20 trillion (20 * 10^{12}) SHA-2 (256 bit) hashes per second for less than 2000 Euro.[109] Leaving aside additional energy costs, the required 125 million hashes could be brute-forced within nanoseconds with just one such device. For the more advanced SHA-3 hash function, the highest performance is currently achieved on graphics cards (GPUs). Just one card costing less than 1000 € achieves more than 30 million hashes per second[110] and could thus brute-force our example in around 4 seconds. When using hash functions explicitly designed to resist brute-forcing[111] such as PBKDF2 or bcrypt, in turn, similar graphics cards still achieve more than 500[112]/12 thousands of hashes per second and card and would thus require 250 seconds/174 minutes, respectively. Even when additional technical factors limit the speed achievable in practice by a factor of 10, a hash-obfuscated data point could still be easily re-personalized without additional knowledge except the base population. Even for a minimally motivated party with moderate technical competencies, such a re-identification of hash-based ID replacements can be considered as 'reasonably likely' as soon as the 'base population' employed as hashing input is well-known.[113]

Where possible input values are not known or do not directly denominate a data subject (eg when ID-card numbers are used instead of names), content-based re-identification might also be applied. This approach is essentially the same as for traditional pseudonyms. However, it is eased by the fact that hash-based obfuscation produces the same hash for the same input data such as 'State-ID-K8484556547128B' across organizational or contextual boundaries, allowing to easily interlink data across different datasets (eg the—yet unknown—person that bought good X according to dataset A also was at place Y yesterday morning, according to dataset B).[114] With cross dataset interlinkability, content-based re-personalization becomes more likely as it can be applied to a rich, cross-domain information base, allowing to apply more capable approaches for singling out and inference and also to propagate a successful content-based re-personalization to other datasets interlinked via the same hash. Even though still hard to estimate, the likeliness of content-based re-personalization is considered higher in the case of simple, hash-based ID replacement than it is in the case of traditional pseudonyms.

To avoid brute-forcing across a known base population and the inherent hash-based interlinkability across different datasets, two additional practices, called salting and peppering, are broadly used.

Pattern 2a——hash-based ID replacement with salted and peppered hashes

Salting and peppering are two techniques broadly used in hash-based password storage. In both cases, additional data is added to the clear-text data before the hash function is applied, but the added data differs between contexts so that resulting hashes also differ. For instance,

- The SHA3-Hash of 'password1-abc' is '56a95c8615cb8ebc4d838de840719abb18fc00cfefe0bfc304539ca3be5714cb', while

[107] See A29WP on Anonymisation Techniques (n 20) 20.

[108] Even though we will come back to blockchain-related scenarios below, the focus is only on the hash rates achievable with dedicated hardware in general.

[109] Example: Antminer S15 <https://www.antminerdistribution.com/antminer-s15/> accessed 9 January 2020.

[110] See, eg Steven Walton, 'Ethereum Mining GPU Benchmark' (*Techspot*, 29 June 2017) <https://www.techspot.com/article/1438-ethereum-mining-gpu-benchmark/> accessed 9 January 2020. This hash rate is a conservative estimate since it is based on Ethereum's modified hashing algorithm which adds additional overhead to avoid broad use of single-purpose hashing devices. With multiple graphics cards operated in parallel and without intentional overhead, hundreds of millions of hashes per second are achievable at moderate effort and cost.

[111] The existence of such hash functions explicitly designed for cases like the one discussed here was obviously out of scope for the A29WP when stating that 'Hash functions are usually designed to be relatively fast to compute' (A29WP on Anonymisation Techniques (n 20) 20).

[112] In less secure modes of operation, 4 million hashes per second and card can also be achieved.

[113] Insofar, we are in line with the A29WP on Anonymisation Techniques (n 20) 20, considering simple hash-based ID replacements as pseudonyms.

[114] See also A29WP on Anonymisation Techniques (n 20) 21: 'Linkability will still be trivial between records using the same pseudonymised attribute to refer to the same individual.'

- the SHA3-Hash of ‘password1-xyz’ is ‘13c296717a3fdcd3aea25354f5f74cd53bdc9134a4a29c06549d992e3046bfab’.

However, salting and peppering significantly differ regarding the additional data and its storage.[115] Peppering uses one additional, secretly held piece of data for every hash in a particular context (such as a password database). In the above example, the pepper ‘abc’ is appended to any password before hashing it for storage in database A. As long as the pepper is chosen randomly and held secret, the same input data (the password) does not result in the same hash across databases. More importantly, it also hinders brute-force attacks on leaked databases of hashed passwords based on a list (‘base-population’) of popular passwords as any attacker would have to try out all possible peppers for any guessed password. A randomly chosen pepper of 4 bytes (32 bits)—the minimum length suggested by the NIST—increases the necessary guessing effort by the factor of 2^{32} (more than 4 billion) as long as the pepper remains secret.

Salting, in turn, uses different additional data for every entry and the data attached to the password of user 1 thus differs from that one used for the password of user 2. When the resulting hashes are leaked together with these user-specific salts, the salts do not increase the brute-force complexity but still ensure that two users with the same password have different password hashes, thus avoiding that identical passwords used by different users can be identified based on the stored hashes alone.[116]

Both techniques are broadly used (ideally in combination) in the context of secure password storage but may also be employed for the purpose of de-personalization. This approach provides benefits over the plain hash-based approach, as illustrated below.

With 32 bits of pepper, every item from the base population would (leaving aside possible optimizations in brute-forcing) have to be tried out more than 4 trillion times, statistically necessitating more than half a quintillion ($0.5 * 10^{18}$) tries until the first successful re-identification in our 250 million examples. For the comparably brute-force-friendly SHA-2, this would require approximately 26,800 seconds (less than eight hours) on optimized hardware. Insofar, the A29WP’s statement that peppered (or ‘keyed’[117]) hash functions lead to a brute-forcing effort ‘sufficiently large to be impractical’ seems implausible. Rather, actual numbers should be carefully examined.

For SHA-3 and one current GPU, it would take more than 200,000 days (or 20 days with 10,000 current GPUs)—an effort that might still be deemed ‘reasonably likely’ for certain edge-cases with high-interest adversaries today (if a subjective approach ought to be adopted) and for even more cases with ongoing technological progress. With the even more resistant bcrypt-hashes, however, 32 bit of pepper would lead to more than 140 years with 10,000 current GPUs, clearly contradicting the A29WP’s implicit assumption that there are no significant divergences between different hash functions.[118] Even though these numbers only represent rough estimates, they clearly demonstrate that the employed hash function and pepper length significantly influence the likelihood of successful re-identification. Calculating the expectable number of tries for successful re-identification and setting them into relation with practically achievable hash-rates thus provides valuable guidance for assessing the likelihood of re-identification.

If peppering is carried out with actually secret, random, and sufficiently sized peppers and using an appropriate hash algorithm, the effort of brute-force ID-based re-personalization with regard to a known base population thus rises significantly. Also, the relation between a peppered hash and a real ID inferred through content-based re-identification for database A would have no impact on the re-personalizability of datasets from database B if different peppers are used. For both cases, a successful re-personalization should thus not be considered reasonably likely. For the party conducting such peppered hash-based ID replacement and thus holding the otherwise secret pepper (and any other party possibly getting hold of it), the likelihood of successful re-personalization is not affected.

Salting, in turn, could be applied to hash-based ID replacement to ensure that the same clear-text ID gets mapped to different substitutes within one dataset. This would lead to non-interlinkable entries in the dataset (eg it is not obvious that two transactions involved the same actor). In use cases where such interlinking is not necessary but where it should still be possible to verify that a particular party was subject to a given transaction, salting could provide a benefit over replacing IDs with

[115] With regard to salting and peppering, terminology is inconsistent in the literature. The NIST, for example, refers to both approaches presented in the following as ‘salting’ (see, eg NIST, ‘Special Publication 800-63B – Digital Identity Guidelines’ (June 2017) 5.1.1.2 <https://pages.nist.gov/800-63-3/sp800-63b.html#memsecretver> accessed 9 January 2020).

[116] See, eg Dan Kaminsky, ‘Salt The Fries: Some Notes on Password Complexity’ (*Dan Kaminsky’s Blog*, 5 January 2012) <https://dankaminsky.com/2012/01/05/salt-the-fries-some-notes-on-password-complexity/> accessed 9 January 2020.

[117] A29WP on Anonymisation Techniques (n 20) 20 refers to what is herein called peppering as ‘keyed hash functions with stored key’.

[118] A29WP on Anonymisation Techniques (n 20) 20: ‘Hash functions are usually designed to be relatively fast to compute.’

pure random numbers or deleting them completely. Salting, however, also requires the employed salt to be stored together with each data point. For instance, the data points

Sender	Receiver	Date	Type of good	Price
John Smith	Jane Miller	12 February 2019	Notebook	1375.00
John Smith	Ken Wolfe	13 February 2019	. . .	. . .

would be replaced by

Sender	Used Salt sender	Receiver	Used salt receiver	. . .	Price
b4c71ab30 . . .	abc	2f187454. . .	xyz	. . .	1375.00
99a1c090f . . .	a12	c352772. . .	A51	. . .	. . .

Knowing the used salt for a particular data point, it is easy to verify, eg a claim by John Smith that he actually was part of a transaction—simply by applying the hash function to the name combined with the salt. Similarly, the known salt could also be used in brute-force de-personalization with similar efforts for a single record as for the plain (unsalted and unpeppered) hash example above.[119] With regard to the necessary effort for re-personalization (and, thus, for the assessment of likeliness), the only difference is that it would have to be taken for any single data point to be re-personalized, while in the plain case, a successful re-personalization immediately applies to all occurrences of the respective hashed ID. In the end, however, the use of salting will provide only limited benefit over plain hash-based replacements regarding the likeliness of re-personalization.

Pattern 3: content hashing

Even where identifiers, the explicit links between data and data subjects are removed (or sufficiently obfuscated) from datasets, natural persons may still be identified on the basis of content data. This becomes obvious when we consider location data. A continuous location history indicating a certain, unchanged overnight position and a rather constant position during workdays could be attributed to an individual with additional knowledge about home and work addresses of the assumed base population. Similarly, starting with a known home and work address of a given person, the complete motion profile may be discovered. Such 'content-based re-identification' can also be applied in significantly more sophisticated ways by interlacing multiple datasets and advanced methods of data analytics.

To avoid content-based re-identification, different technical approaches can be used, ranging from data aggregation (which may guarantee numerical levels of 'anonymity' like k-anonymity,[120] l-diversity,[121] or t-closeness[122]), over data coarsening to differential privacy,[123] which allows statistically meaningful analysis without revealing concise data about individuals in a dataset. Given the vast amount of possible approaches[124] and ongoing developments, we abstain from detailed analyses. One approach shall, however, briefly be introduced.

Content hashing is widely used whenever a checksum functionality is needed to ensure data integrity such as in

[119] Again, the A29WP on Anonymisation Techniques (n 20) 20 is rather unspecific here, stating that salted hash functions 'can reduce the likelihood [of re-personalization while it] may still be feasible with reasonable means'. With salts being stored together with the hashes, it is questionable what reduction it refers to. In case it assumes salts not to be remembered, in turn, the statement lacks explanation why it is considered more likely than in the case of peppered ('keyed') hashes.

[120] See Sweeney (n 83) 557.

[121] See Ashwin Machanavajjhala and others, 'L-diversity: Privacy Beyond k-anonymity' (22nd International Conference on Data Engineering, Atlanta, April 2006) <https://ieeexplore.ieee.org/abstract/document/1617392> accessed 9 January 2020.

[122] See Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian, 't-Closeness: Privacy Beyond k-Anonymity and l-Diversity' (IEEE 23rd International Conference on Data Engineering, Istanbul, April 2007) <https://ieeexplore.ieee.org/abstract/document/4221659> accessed 9 January 2020.

[123] For a vivid introduction, see Christine Task, 'An Illustrated Primer in Differential Privacy' (2013) 20 XRDS 53.

[124] For a first overview to respective approaches from the rather legal viewpoint, see, eg A29WP on Anonymisation Techniques (n 20) 31ff.

the context of digital signatures. A chosen hash algorithm like SHA-3 is applied to a piece of data of arbitrary size (such as a word processing file). This hash can be used to verify the integrity of the document later on as any change would lead to a different hash. Besides the application in digital signature schemes, content-hashing can prove that a certain dataset exists without revealing the data itself. If, for instance, our trading system should allow participants to prove that they participated in a transaction, it could—following a simple commitment scheme[125]—take the whole transaction data:

- Sender: 'John Smith'; Receiver: 'Jane Miller'; Date: 12 February 2009; Type: 'Notebook'; Price: 1375.12

and create a hash thereof ('bef4ee0b2 ...') that could then be released publicly. John Smith, being in possession of the whole transaction data could then hand over this transaction data to a third party which could also hash it and compare the result with the published hash, proving that exactly these transaction data are also present in the transaction system.

Due to the one-way characteristic of hashes, it is impossible to directly recreate the transaction content from such a content hash. Also, the published hash does not contain any IDs of the parties involved in the original transaction. The only way to uncover the content of the transaction behind the published hash resembles brute-forcing: trying out all possible content combinations for all possible identities, hashing each of them, and comparing the hash with the one made publicly available until the hashes match and, thus, the original transaction data are unveiled.

To estimate the necessary effort for a successful recreation of the original content—and, thus, to decide whether it is reasonably likely—we have to identify the available parameter space of transactions. In most reasonable scenarios, however, this parameter space will be sufficiently large to make it unlikely.[126] To avoid brute-forcing, it is common practice to add additional random values—a blinding factor[127]—to the clear-text before hashing it. This is comparable to peppering and salting but here consists of an individual random value per content to be hashed which is not stored along with the hash result. This can be assumed as a viable way for making successful brute-forcing reasonably unlikely even for small parameter spaces. We now proceed to demonstrate how the above insights can be applied in practical blockchain use cases.

Personal data on blockchains

In recent times, there has been ample discussion in the literature and in policy circles whether data conventionally stored on blockchains qualify as personal data. Many of the currently discussed use cases for blockchain involve personal data. Over the past few years, the points of tension between blockchains and the GDPR have been amply discussed—including questions of when and under which circumstances on-chain data qualifies as personal data.[128] Actors interested in using DLT and worried about GDPR compliance will seek to avoid the processing of personal data to start with. Our analysis however confirms that this is far from straightforward as much of the data conventionally assumed to be non-personal as a matter of fact qualifies as personal data. We will discuss in particular (i) what categories of data conventionally stored on blockchains are likely be personal data, (ii) the impact of different approaches of de-personalization, and (iii) related implications for the design and implementation of blockchain applications.

Scenarios

Our analysis does not seek to provide a comprehensive technical description of blockchains and blockchain data, which have been explained elsewhere.[129] Rather, our

[125] For the idea of commitment schemes in general, see, for instance, Torben Pryds Pedersen, 'Non-Interactive and Information-Theoretic Secure Verifiable Secret Sharing' in Joan Feigenbaum (ed.) *Advances in Cryptology - CRYPTO '91* (Springer 1992). For one of the initial publications on the general approach, see also Manuel Blum, 'Coin Flipping by Telephone a Protocol for Solving Impossible Problems' (1983) 15 ACM SIGACT News 23. For a simplified explanation, see, eg 'Pedersen Commitment' (*Beam*) (n.d.) <https://www.beam.mw/beampedia-item/pedersen-commitment> accessed 9 January 2020.

[126] If 100 known persons are on the transaction system, 50 different types of goods are traded, the transaction took place during the last 365 days, and prices range from 0.01 to 10,000.00 (in 0.01 steps, summing up to 1 million possible price points) there are around 180.5 trillion possible combinations. Without sorting out unlikely combinations and assuming all combinations to be similarly likely instead (and using the same numbers as above), it would statistically require ca 90 trillion tries until success, taking less than an hour for SHA-3 but more than 2000 days on a single GPU for PBKDF2 and more than 200 years for bcrypt. In real scenarios, unlikely combinations would be excluded, leading to less combinations to be tested. Nonetheless, other parameters will typically have a larger parameter space (the day might, for instance, be replaced by a millisecond-timestamp), which will in most cases outweigh this reduction. Nonetheless the estimation must be made separately for the specific case, of course.

[127] 'Pedersen Commitment' (*Beam*) (n 125).

[128] See, for instance, Michèle Finck, *Blockchain Regulation and Governance in Europe* (CUP 2019).

[129] Jean Bacon and others, 'Blockchain Demystified' (2017) Queen Mary School of Law Legal Studies Research Paper <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3091218> accessed 9 January 2020; Arvind Narayanan and others, *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction* (Princeton University Press 2016); Andreas Antonopoulos, *Bitcoin & Blockchain - Grundlagen und Programmierung* (2nd edn, O'Reilly 2018).

focus lies on two exemplary and sufficiently differing scenarios for blockchain usage that involve personal data.

Scenario 1—monetary transactions

Two friends, John Smith and Jane Miller repeatedly meet in a café to spend time together. Typically, both pay their bills using a cryptocurrency based on a public blockchain. Accordingly, each transaction is recorded in a public ledger, whereas their names are not. John uses the same cryptocurrency and wallet to pay subscription fees for a video streaming service and the food delivery service he uses occasionally. Jane's primary usage of her cryptocurrency wallet consists of voluntary contributions to a community-run makerspace she frequently visits, which maintains an internal 'contributors leaderboard' that links her name with the transaction's origin. She used the delivery service suggested by John Smith only once. Since having done so, however, she repeatedly receives advertisements from it suggesting she 'have a pizza with John instead of just coffee'.

Analysis. To determine whether blockchain transaction data are personal data (and to whom), we assume comparably plain givens, which reflect a rather typical usage of such transactions. All transactions between Jane, John, the cafe, the video service, the food delivery service and the makerspace are carried out through a public blockchain. Here, balances are ascribed to (and, thus, held in) addresses. An address can be considered as the public key belonging to a private–public keypair randomly generated by a particular user. Users create and manage their addresses (there can be and typically are more than one per user) in wallets, which might be a wallet app on a smartphone or a hardware device. A real-world analogy would be a purse (the wallet) containing multiple credit cards with particular numbers (addresses)—albeit with the extension that users can generate new credit cards and numbers on their own. Leaving out several details like transaction fees, transactions publicly stored and confirmed on the blockchain specify transfers between addresses, implying that users can also transfer amounts between different addresses held by one person.[130] This can be done in various ways.

Simplistic case: unaltered addresses. If John holds an address in his wallet with a balance of 0.001 BTC and wants to pay a coffee for 0.00005 BTC, he can transfer this amount from his address A to the address key of the cafe B and sign this transaction with the private key corresponding to A. In blockchains using proof-of-work, miners can then validate this transaction based on the public key A and the publicly known balance. With the same address A now holding a reduced balance of 0.00095 BTC, he can transfer the necessary amount to the address C of food service for delivering a pizza to him as well as his monthly fee to the address D of the video streaming service. Similarly, Jane can use an unaltered address E for transferring the funds to the cafe (address B), the makerspace (address F), and the delivery service (address C).

The downside of this approach is that everyone able to link an address to a natural person can re-personalize all other transactions of that address. The food delivery service is able to link address A to John and address E to Jane based on their order. Following the approach of ID-based re-personalization, it can also identify all other transactions these persons made with the same address with minimal effort. As soon as it also knows that address B belongs to the cafe (which can particularly be assumed for addresses explicitly intended for receiving transactions from many individuals), it can determine that Jane and John likely know another.

In our example, the transaction data are not explicitly related to a natural person but to an identifier (the addresses). These identifiers are pseudonyms in the traditional sense: quasi-random numbers initially linked to users' identities only in their wallet apps (thus resembling the secretly held pseudonym table, echoing the definition of Article 4(5) GDPR). To decide whether the transaction data on the public blockchain are personal data, we therefore have to assess whether these addresses are reasonably likely to be resolved. This assessment leads to different results for different parties. With the additional order information (especially the delivery address), the food delivery service is able to resolve the pseudonym with minimal (and, thus, reasonable) effort, also rendering all other mentioned transaction data personal data from the perspective of this service.[131] For the maker space, the situation is similar regarding all of Jane's transactions, given that it maintains the relation between Jane and her address for implementing the leaderboard. The streaming service could operate without disclosing John's real identity to the service operator so re-personalization would only be considered reasonably likely if any further information is available.

[130] For an intelligible introduction to Blockchain transactions, see also 'How Does Bitcoin work?' (*Bitcoin*) (n.d.) <https://bitcoin.org/en/how-it-works> accessed 9 January 2020, referring to Bitcoin but basically also reflecting monetary transactions on other blockchains like Ethereum.

[131] See also Danny Yang, Jack Gavigan and Zooko Wilcox-O'Hearn, 'Survey of Confidentiality and Privacy Preserving Technologies for Blockchains' (*R3*, 14 November 2016) 8 <https://www.r3.com/wp-content/uploads/2017/06/survey_confidentiality_privacy_R3.pdf> accessed 9 January 2020.

Whether and how this could be the case would depend on further factors of service implementation (eg whether accounts are used and how).

Common case: one-time addresses. To counteract easy re-personalization, a different usage scheme is used by most wallet applications. It builds on the fact that new addresses can be generated at will and in arbitrary number.[132]

Furthermore, transactions can have multiple input and output addresses, allowing John to specify, for instance, that 'all 0.001 BTC currently assigned to address A are withdrawn, 0.00005 BTC of which are transferred to address B and the remaining 0.00095 BTC (the "change") are to be transferred to address X', with X being a new address generated and controlled by John through his wallet. This scheme is used in most current wallet applications to implement non-persistent addresses by default. The newly created 'change' or 'shadow address' X is then used as the source for subsequent transactions.[133] Monetary transactions made by John do thus not originate from one and the same address and the relation between the source and the newly created change address is not discoverable by other means (eg calculating X from A or 'back-calculating' A from X is not possible, at least for other parties than John himself). From the publicly available transaction data, it is not directly visible whether B or X is the newly created change address controlled by John. For the parties receiving cryptocurrency from John, the first change address (X) can be attributed to him, but for subsequent transactions, they face the same challenge of not being able to distinguish target and change addresses directly. Additional approaches like the chaff coins called 'mixins' used in Monero,[134] the so-called 'CoinJoin'[135] transactions, or dedicated mixes[136] can also be used to further obfuscate the relation between a given transaction and its participants.

With such mechanisms, plain ID-based re-personalization cannot be carried out beyond those transactions for which the relation is known. The café cannot recognize John and Jane as jointly recurring customers, the delivery service cannot identify further transactions of its known customers, and the makerspace and the streaming service cannot relate incoming transactions to Jane and John based on the addresses involved in publicly visible transactions.[137] Concerning pure ID-based re-personalization, only transactions explicitly linked to Jane and John are resolvable and only from the perspective of the party having access to this linkage like the food delivery service for the delivery payment transaction.

Content-based re-identification may, however, reveal another non-ID-based relation between the publicly visible transactions and a natural person which might be resolved with reasonable likelihood. Here, a substantial body of literature exists, proposing a multitude of different approaches for content analysis of blockchain transactions through transaction flow analysis,[138] identification of specific transaction patterns,[139] or time matching.[140] Research demonstrates that address obfuscation through one-time addresses and comparable approaches can, through content analysis of transactions, be 'reverted' so that different addresses can be identified as being controlled by the same person with reasonable certainty (although without actually 'proving' this fact). Based on such content-based address clustering, in turn, a party being able to re-personalize only one of the clustered addresses can,

[132] See Satoshi Nakamoto, 'Bitcoin: A Peer-to-peer Electronic Cash System' (2008) 6 <https://bitcoin.org/bitcoin.pdf> accessed 9 January 2020: 'a new key pair should be used for each transaction to keep them from being linked to a common owner.'

[133] Yang, Gavigan and Wilcox-O'Hearn (n 131) call the underlying concept 'one-time-address' and also introduce the related concept of 'stealth addresses'. Both approaches differ technically but with only marginal implications for the question of resolvability discussed here.

[134] Malte Möser and others 'An Empirical Analysis of Traceability in the Monero Blockchain' (2017) <https://arxiv.org/pdf/1704.04299/> accessed 9 January 2020.

[135] Felix Konstantin Maurer and others, 'Anonymous CoinJoin Transactions with Arbitrary Values' (2017 IEEE Trustcom/BigDataSE/ICESS, Sydney, August 2017) <https://ieeexplore.ieee.org/abstract/document/8029483> accessed 9 January 2020.

[136] See Malte Möser and others, 'An Inquiry into Money Laundering Tools in the Bitcoin Ecosystem' (APWG eCrime Researchers Summit, San Francisco, September 2013) <https://ieeexplore.ieee.org/abstract/document/6805780> accessed 9 January 2020, introducing obfuscation approaches for Bitcoin transactions typically used in money laundering.

[137] For cases nonetheless requiring the attribution to a given person, like the makerspace's leaderboard, this attribution must be made separately. For this, different approaches like comments in the transaction can be used—these shall, however, not be discussed further herein.

[138] Fergal Reid and Martin Harrigan, 'An Analysis of Anonymity in the Bitcoin System' in Yaniv Altshuler and others (eds), *Security and Privacy in Social Networks* (Springer 2013) 197–223.

[139] For transactions with multiple input addresses, it can usually be assumed that these input addresses belong to the same person, 'collecting' deposits from multiple addresses into a single one. See, eg Elli Androulaki and others, 'Evaluating User Privacy in Bitcoin' in Ahmad-Reza Sadeghi (ed), *Financial Cryptography and Data Security* (Springer 2013). Similar approaches were also used by Sarah Meiklejohn and others, 'A Fistful of Bitcoins: Characterizing Payments among Men with No Names' (Internet Measurement Conference, Barcelona, October 2013) <https://link.springer.com/chapter/10.1007/978-3-642-39884-1_4> accessed 9 January 2020, and many others.

[140] Möser and others (n 134). A comprehensive and recent survey of existing approaches for re-identification of blockchain transactions is given by Merve Can Kus Khalilov and Albert Levi, 'A Survey on Anonymity and Privacy in Bitcoin-like Digital Cash Systems' (2018) 20 IEEE Communications Surveys & Tutorials 2543.

with some certainty, also consider the other addresses from a cluster to belong to the same person. The food delivery service could conduct such an analysis and on that basis assign the one-time addresses used for paying in the café to its known customers. It could then deduce that John and Jane know each other. Similarly, it could identify that John repeatedly made transactions to the video rental service (assuming that the target address is publicly known) as well as the makerspace could re-identify Jane's transactions with the food delivery service.

Such content-based analytics are however always heuristic and therefore only provide probable but not certain or even proven associations between different addresses. Furthermore, the whole field of blockchain-related transaction analytics is highly dynamic on both sides: new clustering approaches are continuously developed, regularly followed by the introduction of new countermeasures. These countermeasures typically only work from the point in time they are established, leaving past (and undeletable) transactions open to re-identification. This underlines the broader challenge of accounting for technical developments in the sense of Recital 26 GDPR.

The relation to a natural person can (with mentioned exceptions) not be established based on identifiers alone in the case of one-time addresses and comparable approaches being used. Our test then suggests to consider non-ID-based possibilities for re-identification. It is possible to re-identify transactions through a combination of content-based clustering and ID-based re-identification for any party able to match one of the clustered addresses to an identity. More sophisticated clustering approaches could, however, allow to re-personalize past transactions. Whether such re-identification is to be considered reasonably likely mainly depends on the necessary effort for executing the analysis and the availability of required additional knowledge. Research has demonstrated that necessary efforts and resources for conducting analyses are not excessive.[141] Necessary additional knowledge is typically collected from public sources only and average data analysts would have the required competencies. This re-identification of transactions with non-static addresses is thus to be considered reasonably likely for any party able to link one such address to a person's identity, even though it is always subject to some level of uncertainty.

Scenario 2—ID-based notarization of diplomas

Jane Miller successfully completed her MSc at the University of Blockchain, which provides tamper-proof electronic notarization of academic degrees. The notarization concept is designed to allow potential future employers to easily verify applicants' credentials. Participating universities register all degrees in a public, blockchain-based catalogue. Any entry comprises the field of study, the degree awarded, the courses attended and grades, the final grade, a reference to the person (the type of which shall for the moment be left open), and a signature verifying the identity of the awarding university. When applying for a job, Jane Miller is asked to prove that she actually holds an MSc. She provides her potential employer with the reference information necessary to identify 'her' entry in the catalogue.

Analysis. This use case relies on the advantages of blockchains for data management. These data can originally be of a personal nature but this does not necessarily imply that the data stored on the blockchain also is personal data—this depends on the actual implementation of the use case.

Simplistic case: clear-text data. For the sake of completeness, we assume that all diploma-related data are stored on a public blockchain in clear-text (although this is not usually done in practice).

Diploma ID: <ID number of this diploma>
Name: „Jane Miller"
Citizen ID: 558091684
Issuer: „Blockchain University"
Issuer ID: <University's publicly known ID number>
Degree title: „MSc."
Year of completion
Course title A: <Title A>
Course grade A: <Grade A>
. . .
Course title P: <Title P>
Course grade P: <Grade P>
Final grade: <Final grade>
Issuer signature: <87db523a92f . . .>

This dataset thus contains data identifying Jane and the university, her degree, all 16 courses and grades, and her final grade. To certify the authenticity of the diploma,

[141] The approach presented by Möser and others (n 134) requires a graph database of only 11.5 GB and there is no indication that a particularly noteworthy compute infrastructure was used for analyses. Some further examples are provided by Khalilov and Levi (n 140) 2561, mentioning expenses of 2500 USD or less for executing different analytical approaches.

the university adds a digital signature to the dataset, so that the integrity of the dataset as well as its creation by the university can be proven. When applying for a job, Jane would hand over her diploma ID to the prospective employer who could then retrieve the diploma dataset from the public blockchain, and verify that it originates from the university and was not adulterated.

Applying our test scheme to this case is quite straightforward: these data are explicitly linked to Jane Miller through her name (and her Citizen ID which might be resolvable for several parties). These data are also publicly available on the blockchain. It is therefore personal data from the perspective of any arbitrary party.

Advanced case 1: hash-based pseudonyms. Hash-based pseudonyms might be used instead of clear-text identifiers. In a simplified delineation, this would result in the hash values of 'Jane Miller' and her citizen ID being included in the dataset:

Diploma ID: <ID number of this diploma>
Name-Hash: 799eb189f...
Citizen-ID-Hash: 894640a71...
Issuer: „Blockchain University"
...

As delineated in our initial explanation of hash-based pseudonyms, writing such datasets to a public blockchain would eliminate the direct linkage between the diploma dataset and Jane Miller. However, there still is an indirect link through the hash-based pseudonym. To determine whether the data is personal data, we therefore have to consider the question whether re-identification through the hashed name or citizen ID is reasonably likely. Like above, we here again have to distinguish between re-identification starting from a known ID (eg 'What are the grades of Jane Miller?') and re-identification starting from known content (eg 'Whom does this diploma dataset belong to?').

With all diploma datasets being accessible on a public blockchain, finding Jane's diploma (starting from a known identifier) only requires to hash her name or citizen ID and search for the respective dataset. Any party interested in Jane's diploma could do so with minimal effort. For finding out the person behind a given diploma dataset (ID-based re-identification starting with given content), any party could hash the names or citizen-IDs of all potential diploma holders and compare these to the hashed name or ID. Where the base population is 250 million, such a re-identification must also be considered reasonably likely even for weakly motivated adversaries. From both perspectives, re-personalization of diploma data must thus be considered reasonably likely for plain hash-based pseudonymization.

It is interesting to ponder whether adding salt and pepper to the hashing would alter that conclusion. Peppering would require that the university adds a random, secretly held piece of data to the name and citizen ID of every single diploma before hashing them, whereas this piece of data is the same for all issued diplomas. This would prevent diploma validation without revealing the employed secret piece of data. It is thus not a reasonable strategy here.

Salting would be a reasonable strategy when salt is not publicly known but kept secret and only selectively disclosed, thus resembling the blinding factor from the pattern of content-hashing above: assume that the university adds 4 random bytes to the name and the citizen ID before hashing them and lets Jane know these 4 random bytes. When applying for a job, Jane could then hand over these bytes to the prospective employer together with the ID of her diploma. The employer could then easily validate the dataset identified with the ID. For parties not knowing the secret blinding factor, however, the efforts for brute-forcing through all existing diploma datasets and, for each of them, through all possible blinding factors to find the diploma for a given ID would tremendously increase necessary effort: instead of just one try per available diploma, this would now require more than four billion (2^{32}) tries per diploma. When using proper hash algorithms, this already results in significant effort.[142] Especially with increasing the blinding factor's size to, eg 8 bytes (increasing efforts by the factor of another four billion), this leads to a point where ID-based re-identification (either starting with a dataset or a given ID) is not to be considered reasonably likely for those parties not knowing the blinding factor. The question whether diploma data is to be considered personal data, therefore, depends on the access to the blinding factor.

Content-based re-identification must also be considered. Jane's dataset might be identified based on the

[142] Assuming just 1000 diplomas to be issued und publicly stored, identifying Jane's from these would thus statistically require 500 * 2^{32} = 2,147,483,648,000 (more than 2 trillion) tries. With the above-mentioned 30 million SHA-3 hashes per second for a current GPU, this would result in 20 hours of brute-forcing. For the more brute-force-resistant hashing algorithms PBKDF2 and bcrypt, respectively, it would take roughly 50 days/5.5 years. Finding the right person for a given diploma dataset, in turn, would require to brute-force through the whole available population, increasing necessary effort even further.

combination of courses she attended: let us assume the university offers 40 master's courses from which 16 must be taken, allowing for ca 63 billion different course combinations.[143] For the sake of simplicity, we assume no interdependencies between different courses to exist. If we know just 5 out of the 40 available courses to have been taken by Jane, we have one out of ca 658,000[144] possible five-course combinations.

When we assume 1000 diplomas to be issued per year and know Jane's year of graduation, this allows us to identify Jane's dataset with reasonable certainty—the probability that one out of 999 fellow graduates chose the same five courses is reasonably low (ca 1/658). As soon as Jane's particular five-course combination actually appears in the dataset only once, we have likely identified her diploma dataset. Knowing this dataset, in turn, also provides us with information about the other courses she took and her respective grades. Content-based re-identification starting with a known person and some courses she took would thus not be prevented by the ID-replacement approaches considered above. As soon as we consider it reasonably likely that a party knows or is able to find out a combination of courses, content-based re-identification is reasonably likely. Other possibilities for content-based re-identification based on inference with other datasets would only add to this.

Advanced case 2: off-chain content certified on-chain. Even though consciously implemented hash-based pseudonymization of diploma data might thus avoid ID-based re-identification, content-based re-identification is still a problem. To avoid this, finally, blockchain-based diploma notarization can also be implemented following the content-hashing pattern laid out above. Jane would receive the initially sketched clear-text dataset from her university. In addition, the university would create a hash of the entire dataset and sign this hash with its private key so that it can be decrypted with its public key, resulting in the hash of the dataset again. Only this signature is, together with the diploma ID stored on the public blockchain.[145] A prospective employer then receives the diploma dataset from Jane, retrieves the signature from the blockchain, and decrypts it to original hash.[146] It can then build the hash of the data it got from Jane and compare this to the university-signed hash retrieved from the blockchain.

This model is preferable over previous solutions. There is no data with meaningful content written to the public blockchain but only a diploma ID and a university-signed hash of the content. As explained for content-hashing in general above, the only possible way of learning something about diploma holders would be to brute-force all possible contents of diploma datasets, hash each possible occurrence, and compare respective hashes to those stored on the blockchain—when such a hash-match occurs, the diploma content currently tried out actually exists.

This approach would particularly prove problematic if the diploma dataset only contained the Name, the degree, and the final grade. Trying out all possible grades for a given name would be trivial and consequently, even a hash would have to be considered personal data in itself as it allows to deviate information about a natural person just by brute-forcing all possible combinations of valid content. With more content being included and thus increasing the possible parameter space (eg all attended courses and respective grades), this problem attenuates but still does not disappear.[147] When—as also laid out as being common practice above—adding some random blinding factor to the diploma dataset before initially hashing it, however, this risk can easily be minimized to the point where the likeliness will hardly be considered reasonable anymore.

In the case of only hashes of sufficiently blinded diploma data being stored on the blockchain to allow prospective employers to verify data provided to them off-chain, the hashes stored on the blockchain are thus likely not to be considered personal data. However, this

[143] 40!/24! = ca 1.3 * 10^{24} ordered combinations, 40!/24!/16! = ca 63 billion combinations without ordering.

[144] 40!/35! = ca 79 million ordered combinations, 40!/35!/5! = ca 658,000 combinations without ordering.

[145] A comparable approach for notarizing the time-stamped existence of arbitrary documents, albeit without signatures etc, is, for instance, provided at <https://notary.bitcoin.com/> accessed 9 January 2020. For an overview also including other approaches to blockchain-based notary services, see KC Tam, 'Notarization in Blockchain (Part 2)' (*Medium*, 28 August 2018) <https://link.medium.com/yzb4akr1OY> accessed 9 January 2020.

[146] A comparable approach is, for instance, implemented in the MIT MediaLab's Digital Certificates Project. See <https://certificates.media.mit.edu/> 9 January 2020. An alternative, yet more sophisticated approach could also employ DHT-based off-chain storage integrated with on-chain mechanisms like the one proposed by Guy Zyskind and others, 'Decentralizing Privacy: Using Blockchain to Protect Personal Data' (IEEE Security and Privacy Workshops, San Jose, May 2015) <https://ieeexplore.ieee.org/abstract/document/7163223> accessed 9 January 2020.

[147] Recall, for example, the ca 63 billion possible combinations for 16 out of 40 courses. As an order of magnitude, the hashing of all these combinations would on a current GPU statistically require ca 17.5 hours for PBKDF2 or more than 30 days for bcrypt until a match is found.

is only the case when the possible parameter space of the original data is sufficiently large and/or sufficient blinding was actually carried out.

Conclusion: anonymization as risk management

The above analysis has shown that Recital 26 GDPR embraces a risk-based approach to determine whether data qualifies as personal data. If data can be matched to a natural person with reasonable likelihood, it qualifies as personal data and falls within the GDPR's scope of application. If de-personalization has been sufficiently strong so that identification is no longer reasonably likely, this is non-personal data and accordingly falls outside the Regulation's scope of application. It has, however, also been seen that other elements of European data protection law echo a stance that at least partially conflicts with this risk-based spirit. Indeed, the A29WP appears to have embraced a parallel test according to which no risk of identification can be tolerated. Further statements by various courts and supervisory authorities fall somewhere on the spectrum between both approaches, clearly highlighting the lack of consensus regarding the legal test to be applied, hence threatening the homogenous application of data protection law across the EU. The preceding analysis has moreover illustrated related technical uncertainties and an analysis of two specific blockchain use-cases has confirmed that it can hardly ever be excluded that data which is ostensibly anonymous is transformed into personal data. This final section takes these difficulties and divergences as its starting point and makes the argument that the only realistic test to be applied to anonymization is the risk-based approach, as made clear by the text of the GDPR.

The determination and management of risk are important notions in data protection law. As Kuner and others note, data protection 'has long relied on risk management as a critical tool for ensuring that data are processed appropriately and that the fundamental rights of individuals are protected effectively'.[148] This becomes clear from a mere reading of the legislative text. There are many references to risk in the GDPR, too many to all be named here. For example, the Regulation encourages pseudonymization in order to 'reduce the risks to the data subjects concerned'.[149] Sensitive data deserves special protection in light of the 'significant risks to the fundamental rights and freedoms' that emerge where it is processed.[150] Where a data breach occurs, the controller is exempted from its notification obligations where it can demonstrate that the breach is 'unlikely to result in a risk to the rights and freedoms of natural persons'.[151] Codes of conduct are to calibrate the controllers' obligations on the basis of the risk resulting from their processing operations.[152] The controller has to implement technical and organizational measures that account for the specificities of processing as well as the risks for the rights and freedoms of natural persons.[153] Similarly, data protection by design and by default measures ought to be calibrated on the basis of risk.[154] Data protection impact assessments are required where processing is likely to result 'in a high risk to the rights and freedoms of natural persons'.[155]

These references relate to different contexts and the meaning of risk may differ in each of them. Nonetheless, the legislative text's insistence on risk underlines that ultimately data protection law is a form of risk management. Where personal data is processed, a whole range of different data protection risks emerge. What the law does not do is prohibit related processing to exclude that any risks materialize. Instead, it rather formulates requirements that ought to be followed to minimize risk. This is the logical consequence of a number of factors inherent to data protection law. First, fundamental rights are not absolute but must rather be balanced against the rights and freedoms of others.[156] To achieve absolute protection, any processing of personal data would have to be outlawed. Secondly, the GDPR itself underlines that whereas data protection is an (and arguably the most important) objective it pursues, it also pursues another goal, namely the 'strengthening and the convergence of the economies of the internal market'.[157] The pursuit of the Digital Single Market presupposes that data is processed, albeit subject to the GDPR's risk-management measures.

Below, it is argued, first, that a risk-based approach to anonymization would be in line with the GDPR's overall risk-based approach, and secondly, that an alternative approach would lead to the impossibility of applying this legal framework, and relatedly a profound revision of how data protection law currently operates.

[148] Christopher Kuner and others, 'Risk Management in Data Protection' (2015) 5 IDPL 95.

[149] Recital 28 GDPR.

[150] Recital 51 GDPR.

[151] Recital 85 GDPR.

[152] Recital 98 GDPR.

[153] Art 24 GDPR.

[154] Art 25 GDPR.

[155] Art 35 GDPR.

[156] See Art 52(1) and 52(2) of the EU Charter of Fundamental Rights.

[157] Recital 2 GDPR.

The risk-based approach to anonymization and data protection as risk management

Risk is a core concept of European data protection law. While the notion is used in different fashions in different contexts, the repeated reference thereto underlines that the legal obligations stemming from the Regulation are in many ways formulas for dealing with the risk to the rights and freedoms of data subjects that arise where personal data is processed.

At the same time, risk is not explicitly mentioned in the provisions engaging with anonymization and pseudonymization. Notwithstanding, the case can be made that the notion ought to guide the overall interpretation of the GDPR. Beyond, risk is also mentioned in contexts closely connected to anonymization. Article 25 GDPR is concerned with data protection by design and by default, both methods that seek to address the risks of personal data processing. It is well known that anonymization is an important data protection by design method.[158] Others agree that there is a link between anonymization and the GDPR's overall risk-based approach. The UK ICO's guidance document on anonymization is entitled 'Anonymisation: Managing Data Protection Risk'.[159] The supervisory authority moreover considers that anonymization does not need to be 'completely risk free'—rather what is required is that the risk of identification is mitigated 'until it is remote'.[160]

Seen through this prism, anonymization can be fashioned as a means of reducing the risks that data processing generates. Furthermore, the notion of risk can inform how we think of anonymization and its (lack of) absoluteness. It has been observed above that there can never be an absolute form of anonymization. Rather, a residual risk of identification always remains. This very fact is echoed by the risk concept as 'it is impossible to reduce risks to a zero level'.[161] This is particularly the case with technological risk as 'particularly when dealing with new technologies and activities, scientific uncertainty is due to the fact that risks relate to future outcomes of action which is inherently unpredictable'.[162]

Risk is commonly framed as a two-part concept: it first requires a forecasting of the future and second that decisions are made on the basis of that forecast.[163] The same two-step approach is inherent to the qualification of data under Recital 26 GDPR. Whereas in other domains, such as environmental law, there are established theories of risks and practices of risk management this is not yet the case in relation to data protection in general or anonymization in particular. To some extent, general guidance documents such as the ISO guidelines on risk management could be helpful.[164]

To incorporate the concept of risk in the decision whether data is personal data requires us to determine—at least in orders of magnitude—the time and money necessary for successful re-identification. Well-founded estimations based on at least rough calculations mapped to sound technical givens (such as possible hash rates) like the ones above will therefore be indispensable for making explicit, conscious, and justifiable decisions in this regard. They should thus be conducted more often and may also make their way into institutionalized risk-related processes of data protection law like data protection impact assessments where those are required.

Risk is also well suited to address technical developments in data protection law. For instance, the notion has been helpful to make sense of core concepts such as data minimization and purpose limitation in big data analysis.[165] As such, the concept could also serve to address the changing risks to identification stemming from technological advancements. This underlines that risk is a useful criterion to determine whether data qualifies as personal data. It is also worth considering the alternative and ponder what would happen to data protection law if the absolutist approach to identification were adhered to.

The alternative to the risk-based approach: system change

Accepting that there always remains a residual risk of identification even where data is anonymized appears to be the only realistic option in light of contemporary developments. Research has amply highlighted that anonymization is never absolute. If the law were to insist that it must be, the only logical conclusion would be

[158] Information Commissioner's Office (n 17) 7.

[159] Ibid.

[160] Ibid.

[161] Raphael Gellert, 'Understanding the Notion of Risk in the General Data Protection Regulation' (2018) 34 Computer Law & Security Review 279, 280.

[162] Elizabeth Fisher, *Risk Regulation and Administrative Constitutionalism* (Hart 2007) 7.

[163] Peter Bernstein, *Against the Gods - The Remarkable Story of Risk* (Wiley 1998) 3.

[164] International Organization for Standardization, 'ISO 31000:2018 Risk management – Guidelines' (February 2018) <https://www.iso.org/standard/65694.html> accessed 9 January 2020.

[165] Raphael Gellert, 'We Have Always Managed Risks in Data Protection Law: Understanding the Similarities and Differences Between the Rights-based and the Risk-based Approaches to Data Protection' (2016) 4 European Data Protection Law Review 481, 482.

that data that once was personal data can only ever be pseudonymized but never anonymized. This would not only reject the spirit of Recital 26 GDPR in favour of an absolute approach but also radically change the nature and status of data protection law. Indeed, one would then need to rethink the very distinction between personal and non-personal data on which EU law is currently based.

If it could never be taken for granted that personal data has been successfully transformed into non-personal data, then any information that was ever in the scope of the GDPR would need to be presumed to forever remain within that scope. This was indeed the claim made by Ohm in a seminal 2004 article on the broken promises of privacy.[166] Ohm underlined that in light of recent developments in ICT, perfect anonymization had become impossible as there are always theoretical or real limitations to anonymization. As a consequence, he called for the abolishment of the distinction between personal and non-personal data in data protection and privacy laws.

If one were to transpose that reasoning to the EU data protection law regime, the very concepts of pseudonymous and anonymous information, affirmed by the GDPR, would need to be abolished, which would effectively result in a profound modification of the core of data protection law. Beyond, it is also questionable whether this would generate desirable practical effects. On the one hand, it might be argued that all data that once was personal remains personal and hence subject to the protections of the GDPR. On the other hand, however, there would be no more incentives for data controllers to transform personal data into anonymous data, which would be detrimental to data protection. Indeed, anonymizing data, even with a small remaining risk of re-identification, can be a more effective means of protecting the rights and interests of data subjects compared to leaving this data in its initial state yet applying controllers' duties and data subjects' rights to such data. In one scenario, information about data subjects would in all likely circumstances never be revealed, in another, it would be but data subjects would have rights over such data they are however unlikely to enforce in practice. The incentivizing function of anonymization should hence not be neglected as it can be a powerful tool encouraging data controllers to behave in a data protection-friendly manner.

doi:10.1093/idpl/ipz026
Advance Access Publication 10 March 2020

[166] Ohm (n 78).