14. Identity, Profiles and Pseudonyms in the Digital Environment

Miranda Mourby[1] *& Elaine Mackey*

Abstract
The boundaries of personal data are determined by the concept of 'identity'. Personal data, as defined under the GDPR, is information relating to an identified or identifiable natural person. In this chapter, we argue that the informational 'identity' of an identified/identifiable person is characterised by the potential for privacy impact. Our informational identity is, in essence, the sum of all the information which can impact our rights. We use profiles and pseudonyms as an illustration of this definition. Profiles permit scrutiny of an individual – and thus 'identify' them through the intrinsic privacy impact of this evaluation. Pseudonyms alone do not allow individuals to be evaluated, which is why they are not, in and of themselves, personal data.

Keywords: identity; pseudonymisation; profiling; anonymisation; personal data

1. Introduction

What is an identification? Some information is deemed sufficiently 'us' to warrant legal protection, but this category of information shifts all the time, and the logic underpinning these shifting parameters is far from explicit. The idea of 'identity' determines the scope of data protection law in the EU, which safeguards the rights of 'identified' and 'identifiable' individuals. Without understanding when a person is – or might be – 'identified', we cannot be sure when these rights arise.

1 Miranda Mourby would like to acknowledge support from the EU-STANDS4PM consortium (www.eustands4pm.eu) that was funded by the European Union Horizon2020 framework programme of the European Commission under Grant Agreement #825843. She is also grateful to the School of Law at the University of Sheffield, whose funding supported this work in part.

Van der Sloot & van Schendel. *The Boundaries of Data*. Amsterdam: Amsterdam University Press, 2024. DOI: 10.5117/9789463729192_CH14

This chapter clarifies the concept of 'identity' in EU and associated national data protection law by flipping conventional wisdom on its head. It is often asserted that privacy and data protection rights arise when an individual is or can be identified. But without a clear understanding of what it means to be 'identified', this statement is not particularly meaningful. As the growth of the online infosphere increasingly detaches identity from traditional 'real-world' signifiers, the time may have come to recognise that an individual is instead 'identified' when information engages their rights to privacy and/or data protection. As profiling is thought to engage privacy and data protection rights and is proliferating within the Big Data environment (de Hert & Lammerant, 2016), it is a useful touchstone in understanding identification in digital information.

This chapter therefore attempts to delineate the contours of 'identity' in data protection law by exploring two associated concepts: profiling and pseudonymisation. We have selected these concepts because they are respectively associated with *direct* and *indirect* identification. We suggest that the parameters of 'direct' identification – information that is, in and of itself, an identification with nothing further required – help to reveal the nature of an identity in data protection law. The UK is used as a particular case study because it has, in its post-Brexit modification of the EU General Data Protection Regulation (GDPR), introduced the concepts of direct and indirect identification into a statutory definition of identifiable individuals, which adds precision to the definition that can be inferred at EU level.

The concepts of pseudonymisation and profiling under the GDPR are therefore worth unpacking because they help illustrate the circumstances in which identification takes place in the online infosphere. In the absence of a definition of 'identified' or 'identifiable' individuals in the EU Regulation itself, these subsidiary concepts provide contrasting definitions of a directly identifying 'profile' (which engages an individual's rights through evaluation of their personal characteristics) with a 'pseudonym' (which also uniquely represents people but does not permit analysis or scrutiny of them as individual subjects without further information). The 'unique' nature of the pseudonym may only be a particular variation in a hashing code; it does not signify any immediately discernible personal information. Put simply, therefore: if a profile alone is an identification, and a pseudonym alone is not, the contrast between the two helps us explain what is and is not an identity in online information.

Ultimately, we suggest that the defining feature of 'identity' in data is the capacity of information to interfere with individuals' privacy and data protection rights. As profiling data permit scrutiny of individuals in a way that pseudonymised data should not, this distinction between the two

concepts provides a useful illustration of the difference this capacity of interference makes in practice.

2. Identity in Data Protection Law

As Sullivan (2011) emphasises, it is important to discern the meaning attributed to the concept of 'identity' in a particular legal context:

> Identity has traditionally been a nebulous notion and in referring to 'identity' without defining it, much of the legal literature in this area lacks precision. It gives the impression that 'identity is identity' whereas the constitution, function and nature of identity depends on context … it is important to differentiate the 'purely legal relations' from other non-legal conceptions. (p. 6)

In order to delineate the meaning of identity in the context of data protection law, it is necessary to grapple with the GDPR's usage of the terms 'identified', 'identifiable' and 'identifier.' These occur in the definition of personal data in the GDPR:

> 'personal data' means 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. (GDPR, Article 4[1])

It is easy to lose one's bearings within a definition so densely packed with the terms identified, identifiable, identifier and identity. Interestingly, while the term 'identifiable' is elaborated upon as meaning someone who 'can be identified', the word 'identified' itself is not explained, leaving an ultimate ambiguity as to what 'identity' means for the purposes of the GDPR. The list of 'identifiers' is perhaps a clue, but these pieces of information appear only to refer to *means* of identification and not identification itself. As the UK Information Commissioner's Office (2020) clarifies, 'whether any potential identifier actually identifies an individual depends on the context.' For example, 'a person who enjoys the theatre' may be an aspect of cultural identity, but without further information to link this no doubt scintillating insight into one particular person, it is no more identification than it is trope, fiction or hypothesis.

We have suggested that two types of personal data can be established within the GDPR:

– information that is, in and of itself, identification (relating to an 'identified' individual);
– information that can be linked indirectly to an identified individual, including pseudonymised data, which is information on an 'identifiable' individual.

In order to answer the question of what identification is, we are principally concerned with the first type of personal data – information that is *in itself* an identification. The latter category is essentially a secondary subset of personal data, caught by the regulation if they can be linked to information that either in combination or in itself constitutes identification. The core question, therefore, is what quality or qualities of data render information an identification.

We will answer this question of 'what is identification?' by relating to direct identification, i.e. information relating to *identified* individuals. Within privacy and data protection, data that are characterised as 'personal' – and therefore as linking to individuals' 'identity' – tend to be information with sufficiently close association to an individual to justify their 'stake' in the information. As Laurie states in the context of genetic data, 'individuals have an interest in this information because it relates to them and can affect their lives' (Laurie, 2002).

2.1. Facial Images as Direct Identification

A UK case that illustrates this association with identification and the idea of a personal stake in information is the High Court judgment in *Bridges v. South Wales Police*, which was believed to be the first time any court in the world had considered the use of automated facial recognition software (AFR). The claim for, inter alia, infringement of data protection legislation was brought by Edward Bridges with the support of the campaigning organisation Liberty.

In brief, *Bridges* concerned the collection of facial images by police at rugby matches for the purposes of AFR. It was argued in submissions that the police would require further powers to match the facial images to individuals in order for them to constitute personal data (per *Breyer*). In other words, the images were not an identification in and of themselves, and 'identifiability' would only be triggered with the presence of an additional means reasonably likely to be used to identify people.

The Court rejected this argument, however, on the basis that the images were an identification in and of themselves:

> Where the data in issue is biometric facial data, we see no need for the analysis adopted by the CJEU in Breyer (in the context of information comprising dynamic IP addresses). Whether or not such information is personal data may be open to debate, as is apparent from the judgment in Vidal-Hall [2016] QB 1003. However, the biometric facial data in issue in this case is *qualitatively different* and clearly does comprise personal data because, *per se*, it permits immediate identification of a person. (R. [on the application of Bridges], 2020; emphasis added)

The phrase 'immediate identification' makes it clear that an image of a face is an identification in and of itself, having the 'quality' of being identity per se. This is reminiscent of Sullivan's description (cited above) of the 'identity is identity' mentality. Although the reasons for this are not elaborated upon, it seems overwhelmingly contextually likely that the Court bore the civil liberty implications mentioned above in mind, meaning that the location of the information within the regulatory framework of privacy and data protection was a pressing concern in this determination. The risks revealed by the evolution of AFR thus make a compelling argument for consideration of images of faces as an identification, and thus an identification in the eyes of data protection law.

2.2. IP Addresses as (In)Direct Identification

IP addresses, on the other hand, are not as straightforward a proposition. An IP address alone is not necessarily an identification because it does not create sufficient potential for consequence for, or inference about, the user of the related device, but an IP address combined with browsing history data across a number of websites *is* generally held to be an identification because it creates a profile. Evidence for this argument can be found in Recital 30 GDPR:

> Natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags.
>
> This may 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. (emphasis added)

This recital seems to draw a reasonably clear distinction between potential identifiers (such as an IP address) and the combination of information that

profiles an individual, adding up to enough usable information to constitute an actual identification.

Further illustration of how IP addresses can fail to meet the standard of direct identification comes from the 2016 judgment Case C-582/14 of the Court of Justice of the European Union in *Patrick Breyer v Bundesrepublik Deutschland*, which we will refer to as the *Breyer* judgment.

In the *Breyer* case, the German government collected information in case its websites came under attack and it was necessary to identify the perpetrators:

> With the aim of preventing attacks and making it possible to prosecute 'pirates', most of those websites store information on all access operations in logfiles. The information retained in the logfiles after those sites have been accessed include the name of the web page or file to which access was sought, the terms entered in the search fields, the time of access, the quantity of data transferred, an indication of whether access was successful, and the IP address of the computer from which access was sought. (CJEU, 2016, para. 14)

These retained IP addresses had no immediate privacy consequences for the associated individuals unless the German government took additional steps to build a picture of these people. It was confirmed at paragraph 38 of the judgment that the dynamic IP addresses were not personal data in and of themselves:

> In that connection, it must be noted, first of all, that it is common ground that a dynamic IP address does not constitute information relating to an 'identified natural person', since such an address does not directly reveal the identity of the natural person who owns the computer from which a website was accessed, or that of another person who might use that computer. (CJEU, 206, para. 38)

3. Pseudonyms and Profiles

The terms 'pseudonyms' and 'profiles' are used in this chapter to refer to the end products of GDPR pseudonymisation and profiling respectively. While these terms may, in other contexts, both refer to representations of individuals that fall short of an identification (e.g. a psychological 'profile' of a criminal suspect who sends letters under a 'pseudonym' but has yet to be

identified), in the context of EU data protection law, they denote different levels of identifiability.

A 'pseudonym' is traditionally defined as an alternative to one's 'real' identity, for example as a 'false or fictitious name, esp. one assumed by an author; an alias' (Oxford University Press, 2007). In the context of the GDPR, personal data that have undergone pseudonymisation are associated with an 'alias' or something falling short of an actual identification. The data thus requiring additional information to be linked back to the 'real' identity of the natural person:

> 'pseudonymisation' means 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. (Article 4[5] GDPR)

A profile, by contrast, permits the evaluation of personal characteristics under its definition in Article 4(4) GDPR:

> 'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.

This automated evaluation of personal characteristics is, we suggest, sufficient intrusion into privacy and data protection rights to constitute an identification in and of itself, even if there are no other consequences for the data subject. For example, a profile of an individual's online behaviour is likely to involve novel inferences about that person, which are of value for commercial exploitation, which then steps over the boundary of anonymous, unobserved browsing even before any attempt to 'reach' or affect the individual is made. The use of profiling in the digital environment therefore illustrates the underlying logic of identification: where there is intrusion, there is identification, even if the digital profile bears questionable resemblance to someone's 'real' identity.

Table 14.1 attempts a summary of how we distinguish the GDPR terms 'profiling' and 'pseudonymisation':

<table>
<tr><th></th><th>Pseudonymisation</th><th>Profiling</th></tr>
<tr><td>Definition under Article 4 GDPR</td><td>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.</td><td>Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements.</td></tr>
<tr><td>Information associated with the process</td><td>A pseudonym plus a string of information</td><td>A string of information that allows an evaluation of personal characteristics</td></tr>
<tr><td>Example</td><td><table><tr><td>Masked IP address</td><td>Age in bands of ten years</td><td>Gender</td><td>Employed [y/n/]</td><td>Industry of employ-ment</td><td>Number of times X website visited last 12 months</td></tr></table></td><td><table><tr><td>IP address</td><td>Website address</td><td>Date visited</td><td>Time + Duration of visit</td><td>Products viewed</td><td>Products bought</td></tr></table><table><tr><td>Masked IP address</td><td>Website address</td><td>Number of times Website visited last 12 months</td><td>Most popular product viewed last 12 months</td><td>Most popular product bought in the last 12 months</td></tr></table></td></tr>
<tr><td>How the processes differ</td><td>– A process to mask identity
– A process that reduces detail in the personal information, leaving the substance of the data untouched
– Aims to prevent direct identification</td><td>– A process to create an identity (may not match real world identity)
– A process to create new information, 'a profile', from the underlying original data
– Aims to evaluate the individual</td></tr>
<tr><td>Purpose of the process</td><td>To dissociate individuals from information. It enables the exploration of patterns within data whilst preventing direct identification.</td><td>To associate individuals with information. It enables the targeting of natural people in pursuit of a service, product or message with the potential to 'reach' the individual.</td></tr>
</table>

Table 14.1

3.1. Profiles as Direct Identification: IAB Europe

An important example of profiling is the 'Transparency and Consent', or 'TC String', generated by consent management platforms to record the consent preferences of visitors to websites regarding the use of their data.

This 'TC String' was considered in the judgment of Case DOS-2019-01377 before the Litigation Chamber of the Belgian Data Protection Authority (the APD) in a case we will refer to as the 'IAB Europe decision' (APD, 2022).

The APD handed down a decision in February 2022 as the lead supervisory authority under the 'one-stop-shop' mechanism of Article 56 GDPR. Its judgment was reviewed and approved by a number of Concerned Supervisory Authorities representing the Netherlands, Latvia, Italy, Sweden, Slovenia, Norway, Hungary, Poland, Portugal, Denmark, France, Finland, Greece, Spain, Luxemburg, Czech Republic, Austria, Croatia, Cyprus, Germany (Berlin, Rhineland-Palatinate, North Rhine, Westphalia, Saarland, Lower Saxony, Brandenburg, Mecklenburg-Western Pomerania and Bavaria) and Ireland.

This was not a judgment of the Court of Justice of the European Union, or indeed any other European court. Nonetheless, the breadth of data protection authorities represented – and the consequent scale of the litigation – makes the decision an important precedent within Europe, particularly within the world of online behavioural profiling.

Interactive Advertising Bureau Europe (IAB) is a federation of approximately 5,000 companies across Europe. IAB developed a Transparency and Consent framework as a best practice standard so that real-time bidding could be conducted in compliance with the GDPR (in theory).

Real-time bidding (RTB) was deemed sufficiently complex that it required introduction at the outset of the decision, with diagrammatic representation of the interactions. A distinction was drawn with 'traditional' advertising, in which the advert is negotiated manually between business and publisher. Instead, the machinations of RTB take place 'behind the scenes', with data subjects unaware of the identity of actors involved or even necessarily aware that their information is being automatically auctioned for the opportunity of advertising to them.

The profiling involved in RTB was deemed to be a key element of the processing that IAB had facilitated. There was no controversy that the data used for and generated by this profiling were personal data. This is interesting, as the information used for RTB was very heterogenous, potentially including:

> *URL of the visited site ▪ category or subject of the site ▪ operating system of the device ▪ browser software and version ▪ manufacturer and model of the device ▪ mobile operator ▪ screen dimensions ▪ unique user identification set by vendor and/or buyer. ▪ unique person identifier from the Ad Exchange, often derived from the Ad Exchange's cookie. ▪ the user identification of a DSP, often derived from the Ad Exchange's cookie that is synchronised with a cookie from the DSP's domain. ▪ year of birth ▪ gender ▪ interests ▪ metadata reporting on consent given ▪ geography ▪ longitude and latitude ▪ post code*

While some data included in the RTB processing are what would conventionally be deemed an identifier (gender, post code, year of birth), others are more device-orientated and not 'personal' in the conventional sense (e.g. screen dimensions, browsing software, etc.).

The element of controversy, however, lay in the TC string. The TC string is 'a character string consisting of a combination of letters, numbers and other characters' (para. 41). At paragraph 95 of the judgment, the APD (2022) found that:

> the generation of the TC String in itself constitutes, without any doubt, processing of personal data. The issue at hand is the automated creation, by a CMP registered with the TCF, of a unique and linked set of characters intended to capture a specific user's preferences regarding permitted data exchanges with advertisers. (emphasis added).

The ultimate determination by the APD that the unique set of characters capturing a user's preferences constituted personal data was transformative for the digital economy, acknowledging a whole new link in the chain of information as personal data in and of itself.

The APD's decision is congruent with the logic of this chapter. Although the relevant combination of numbers, letters and characters may not resemble the person in question in a way we would see them with human eyes, in an automated context, this string represents an actionable personal characteristic: their preferences regarding data exchange. It constitutes information that could impact upon the privacy of the person's internet browsing and is therefore, understandably, an identification.

It is important to remember that an identity for the sake of data protection law may be very different from the social, 'real-world' ways we recognise and differentiate people. Identification does not need to include a name or the capacity to physically locate the individual in the real world but could reveal enough information about them to provide an interface to affect

them. McMahon and others illustrate this with the scenario of a woman who miscarries but then continues to receive ads targeted to her perceived pregnancy; a digital profile does not need to correlate accurately with a lived reality to have an impact on her (McMahon et al., 2020). Accurate or not, it would therefore make sense for this profile to be a protected digital identity in order to protect the natural living individual who will be impacted by it.

In this sense, it would not matter if the digital profile correlated poorly with the 'real-world' or 'offline' identity of the individual. Writing for the BBC, Carl Miller conducted a number of subject access requests and uncovered a strange array of inferential judgments made about him based on his browsing history, including that he was a woman trying to conceive, a 'love aspirer' and a disengaged worker with little perceived interest in reading (Miller, 2019). Even if the digital profile of an individual bears little relation to the individual's social or physiological identity, or their own subjective sense of self, it could nonetheless have consequences for them at least in terms of personalised advertisements and (as in the case of misidentification) may have all the more consequences for being wrong. When inaccurate information impacts upon individuals, there is no need to have recourse to the concept of 'fake privacy' (Burgess, 2018) if the digital identity is understood as the clusters of data that can impact a natural, living person.

The IAB Europe case illustrates the increasing penetration of the internet into our daily lives and the consequent expansion of online activity among the digitally connected majority of Europeans, meaning that many of us have an increasing proliferation of 'virtual identities' (Wachter, 2018). Any attempt to rationally delineate those virtual identities that are sufficiently connected with us to constitute a 'profile', and those sufficiently detached to be a 'pseudonym', reveals the lack of attention generally given to the question at the heart of the scope of data protection law: what is an identity in information?

If privacy and data protection are inherently connected to the 'integrity of information constituting one's identity', we cannot understand the boundary of personal data without a common agreement on what information *is* our identity. The general complacency on this issue stems from an apparent assumption that it must be obvious, that 'identity is identity' (Sullivan, 2011). The Spanish AEPD and the European Data Protection Supervisor recently collaborated to address common misunderstandings relating to anonymisation, but the ensuing guidance still falls into the 'identity is identity' trap, stating 'direct identifiers are somewhat trivial to find, indirect identifiers, on the other side, are not always obvious' (AEPD, 2021).

Our exploration of profiling versus pseudonymisation in this chapter shows that direct identifiers are *not* always trivial to define. The evolution of case law since 2016 has shown an expansion of what is considered direct identification in an online environment due to increasing recognition of the power of online profiles – even those that cannot be attributed to the 'real-world' identities of named, gendered, geographically located individuals.

3.2. Pseudonyms as 'Indirect' Identification

It is potentially confusing that a 'pseudonym' can superficially appear the same as a profile, which is also a string of letters and characters. The reason why pseudonymised data are not, however, a direct identification is that they should not permit scrutiny or other action vis-à-vis an individual (e.g. authorising the sharing of their data, in the above example). The French Data Protection Authority (the CNIL) provides the following example:

> an economics researcher has entered into a partnership with a family allowance fund (CAF) which has databases containing the names, dates of birth and addresses of applicants for housing allowance in 2019, as well as the amounts of allowances received and the number of people in the household.
>
> In order to carry out this research and meet data protection requirements, the researcher and CAF have agreed that the latter works on pseudonymised data. For this, the CAF will replace the names and dates of birth with a unique identifier (instead of deleting the columns) and will replace the complete addresses with only the municipalities.
>
> It will thus be possible for the researcher to compare identifiers between databases to find common recipients, without being able to know their identity directly. (CNIL, 2022; emphasis added)

In the above example, the researcher is crucially concerned with *trends across a dataset* rather than scrutinising or making decisions about any individual within it. As such, even if the 'unique identifier' pseudonym was similar in composition to the TC string, its presence within pseudonymised data as opposed to profiling data means that it does not immediately reveal anything about an individual that interferes with their privacy. It is only the risk of 'indirect' identification through combination with other information

that makes this information personal data: it is not an identification in and of itself, as it does not directly impinge on privacy.

3.3. Direct and Indirect Identification

In the above examples, the distinction between 'direct' and 'indirect' identification is key. Direct identification requires no further information and therefore means that the data in question are a legally protected identity without the risk of further attribution. As we have seen above, the French CNIL has referred to pseudonymisation as representing a risk of 'indirect identification', and the UK Parliament has undertaken to go a step further by placing this distinction into law, in proposed updates to its Data Protection Act 2022:

> (3A) An individual is identifiable from information 'directly' if the individual can be identified without the use of additional information.
>
> (3B) An individual is identifiable from information 'indirectly' if the individual can be identified only with the use of additional information. (UK Parliament, 2022, p. 2)

The UK has even gone as far as to propose its own definition of pseudonymisation to clarify that which was set out in the GDPR, indicating that '"pseudonymisation" means the processing of personal data in such a manner that it becomes information relating to a living individual who is only indirectly identifiable' (UK Parliament, 2022, p. 3). While this is only one national interpretation of the GDPR, it does chime with the logic of the CNIL's pseudonymisation scenario, cited above. This helps to reinforce the idea that a pseudonym falls short of a direct identification because it is not immediately revelatory about an individual in a way that will interfere with their rights.

In all EU jurisdictions, the definition of identity will also establish the parameters of data protection law, which protects identified and identifiable people. The scope of this law should be understood with reference to its central purpose: the safeguarding of individual rights within a free market of digital information. Where these rights are engaged by the collection, construction or inference of information, the data should be considered an identification. The difference between pseudonymisation and profiling illustrates this acid test of intrusion in practice.

Data that have undergone GDPR pseudonymisation should not permit evaluation of personal characteristics; they should only reveal trends across individuals. Where reasonable likelihood of attribution back to particular people is removed (though control of the data environment), it may be possible for such pseudonymised personal data to be rendered anonymous. However, careful consideration should be given as to whether the same information could permit profiling in a different context; through combination with other information, or through automated scrutiny with advanced algorithms. These are among the risks of identification that must be excluded by any means reasonably likely to be used for the information to be considered anonymous, per Recital 26 GDPR.

Clarifying the digital identity as distinct from a 'pseudonym' is not just an academic exercise: our privacy and data protection rights are bound up in this concept. We therefore use profiling as a case study of intrusion and impact, which illustrates when information is of such intrinsic value that it constitutes an aspect of identity, thus warranting legal protection.

4. Profiles, Pseudonyms and Anonymity

We have previously written a paper in which we explored the introduction of the 'pseudonymisation' to data protection law within the GDPR. We argued that the data 'environment' (which includes other data, people, the presence or absence of information governance controls and infrastructure) can be managed to render such unattributed information functionally anonymous in the hands of a third party who has no access to the identifiers (Mourby et al., 2018). The controversy surrounding this question continues. Our argument drew on the concept of 'functional anonymisation' and appears to align with the UK Information Commissioner's Office draft updates to their anonymisation guidance post-GDPR (Elliot et al., 2016), but the 'bigger picture' from the European Data Protection Board (EDPB) is still outstanding, as the EU-wide board of regulators is still reviewing the 2014 European guidance on anonymisation (EDPB, 2021).

The preceding sections have shed light on the distinction between profiles and pseudonyms, which forms a central question of this chapter. We can perhaps summarise how this distinction maps onto the personal-anonymous data boundary in Text Box 14.1:

Profiles, Pseudonyms and Anonymity

Profiles: a collection of information with the potential to impact the rights to privacy and data protection of one or more natural persons through automated evaluation of personal characteristics. Profiles thus relate to an 'identified' individual and do not need any further attribution to constitute personal data.

Pseudonyms: information that has undergone GDPR pseudonymisation will still be personal if it can be attributed back to individuals through means reasonably likely to be used (rendering them identifiable per *Breyer*).

Text Box 14.1

To anonymise information, therefore, it is necessary to eliminate:

– Reasonable means of attributing information to individuals through management of the data environment (to prevent the subject becoming *identifiable*).
– The capacity of the information itself to allow individuals to be profiled and thus *identified*.

It is worth noting that longitudinal data that show an individual's behaviour over time (e.g. from a tracking cookie) will be much more difficult (if not impossible) to anonymise than a list of 'hits' on a website. Even if both types of information involve hashed or masked IP addresses, the former is far more likely to enable profiling and therefore remain personal data.

The GDPR could be described as a missed opportunity to provide a clear definition of anonymity versus pseudonymity, and indeed to address the underlying definition of what constitutes 'identification'. As it stands, however, the reader must parse an implicit definition from Recital 26:

> The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal 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.
>
> 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. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.

Elsewhere we have outlined at length how definitions of anonymisation and pseudonymisation can be gleaned from this recital (Mourby et al., 2018). In essence, data that can be attributed to a natural person by means reasonably likely to be used are *indirectly* identifying and are thus pseudonymous personal data. Anonymous data are data for which identification by any means reasonably likely to be used is considered remote. The length of Recital 26 alone illustrates the complexity of demarcating personal and anonymous data in a way that is both logically consistent and consistent with the terminology of the GDPR. This was not unavoidable, however. When reviewing a draft of the GDPR, the Committee on Civil Liberties, Justice and Home Affairs of the European Parliament recommended clarification of these concepts back in October 2012:

> In order to reach the best level of data protection and enable new business models, we need to encourage the pseudonymous and anonymous use of services. Clearly defining 'anonymity' should also help data controllers understand when they are outside the scope of the Regulation. For the use of pseudonymous data, in sense of the data controller is able to single out individual persons by a pseudonym, there could be alleviations with regard to obligations for the data controller. (LIBE, 2012)

To reconcile this paragraph with our working definitions of profiles and pseudonyms, the mere 'singling out' of a person by reference to a pseudonym could be seen as falling a step short of evaluating their personal characteristics in a privacy-intrusive way. As such, it remains logical to see pseudonyms as indirectly identifying personal data, even when they

permit singling out. This appears to have been borne out by the trends we have identified in regarding pseudonymised data as indirect identification.

In short, as pseudonymised data are only personal because of the risk of further attribution, they can be anonymised by eliminating reasonable risk of connection with additional information. Profiling data, however, are directly identifying and cannot be anonymised unless they are modified to the point that they no longer permit the immediate evaluation of personal characteristics.

5. Conclusion

This chapter has suggested that 'identity' in data protection law should be understood not in the psychological sense of how we perceive ourselves but in the 'digital' sense of information with sufficient potential impact on us individually that it should be recognised as a legally protected aspect of self. Although we have focused on profiling as an intrusion into privacy that thus constitutes an identification, the engagement of other fundamental rights could also justify treating the data as personal. For example, where the automated evaluation is of personal characteristics protected under equality laws, identification due to the engagement of the right to non-discrimination should also be considered.

The question of whether information constitutes an identification can thus be considered in two stages:

- Does the information, in and of itself, provide enough detail about the individual that they can be profiled, scrutinised, judged or otherwise experience (even without their knowledge) consequences from this information? If so, they have been 'identified' by the information.
- Can it be combined with other information – either already in the hands of the controller, or which they can obtain through means reasonably likely to be used – in such a way to achieve identification? If so, the individual is 'identifiable'.

Although the GDPR does not explicitly link the definition of profiling with that of personal data, the decisions we have reviewed have placed interference with individual rights at the heart of the concept of identification. As such, profiling provides an important illustration as to when information is sufficiently intrusive into fundamental rights in and of itself that can

justifiably be called an identification. This has been contrasted with pseudonymisation, in which case the question of identification is less certain.

We have therefore considered the theoretical underpinning of the concept of identity in data protection law but also provided some practical guidance. In particular, our analysis highlights that longitudinal data that show individual behaviour over time (e.g. from a cookie) will be much more difficult to anonymise than a logfile of website visitors that only provides a single snapshot in time. Ultimately, however, our central contribution has been to show that it may now be helpful to determine the scope of identity in data protection law with reference to fundamental rights, and not (as is often suggested) the other way around. For all that the category of 'identity' shifts as technology evolves, the underlying benchmarks of privacy and non-discrimination rights are sufficiently stable to provide a reliable sense of who we are as we navigate the digital environment.

References

Agencia Española Protección Data & European Data Protection Supervisor (AEPD). (2021). *10 misunderstandings related to anonymisation*. https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf

Authorité de protection de données (APD). (2022). Litigation Chamber, Case DOS-2019-01377, Concerning: Complaint relating to Transparency & Consent Framework, Decision on the merits 21/2022 of 2 February 2022

Burgess, M. (2018). *The law is nowhere near ready for the rise of AI-generated fake porn*. Wired. https://www.wired.co.uk/article/deepfake-app-ai-porn-fake-reddit

CNIL. (2022). *Scientific research* (*excluding health*)*: Challenges and advantages of anonymization and pseudonymization*. https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation

Committee on Civil Liberties, Justice and Home Affairs of the European Parliament (LIBE). (2012). Working Document 2 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). https://www.europarl.europa.eu/doceo/document/LIBE-DT-497802_EN.pdf?redirect.

de Hert, P., & Lammerant, H. (2016). Predictive profiling and its legal limits: Effectiveness gone forever? In B. van der Sloot, D. Broeders, & E. Schrijvers (Eds.), *Exploring the boundaries of Big Data* (pp. 145–167). Amsterdam University Press.

Elliot, M., Mackey, E., & O'Hara, K. (2016). *The Anonymisation Decision-Making Framework*, 2nd ed. UKAN Publications. https://ukanon.net/framework/

European Data Protection Board (EDPB). (2021). *EDPB letter to the European institutions on the privacy and data protection aspects of a possible digital euro*. https://edpb.europa.eu/system/files/2021-07/edpb_letter_out_2021_0112-digitaleuro-toep_en.pdf

Information Commissioner's Office. (2020). *What is personal data?* https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data/

Information Commissioner's Office. (2022). *Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance*. https://ico.org.uk/media/about-the-ico/consultations/4019579/chapter-3-anonymisation-guidance.pdf

Laurie, G. (2002). *Genetic privacy: A challenge to medico-legal norms*. Cambridge University Press.

McMahon, A., Buyx, A., & Prainsack, B. (2020). Big data governance needs more collective responsibility: The role of harm mitigation in the governance of data use in medicine and beyond. *Medical Law Review*, *28*(1), 155–182.

Miller, C. (2019). *Would you recognise yourself from your data?* BBC. https://www.bbc.co.uk/news/technology-48434175

Mourby, M., et al. (2018). Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. *Computer Law and Security Review*, *34*(2), 222–233.

Oxford University Press. (2007). *OED*. https://www.oed.com/

R. (on the application of Bridges) v Chief Constable of South Wales. (2020). EWCA Civ 1058.

Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/ (General Data Protection Regulation). (2016). OJ L119/1.

Sullivan, C. (2011). Digital identity: An emergent legal concept. University of Adelaide Press.

UK Parliament. (2022). Data Protection and Digital Information Bill. https://publications.parliament.uk/pa/bills/cbill/58-03/0143/220143.pdf

Wachter, S. (2018). Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR. *Computer Law and Security Review*, *34*(3), 436–449.