Journal of Intellectual Property and Information Technology Law (JIPIT)
ISSN: 2788-6727 – Volume 5, Issue 1 (2025) p.59-103
DOI: https:doi.org/10.52907/jipit.v5i1.623

Reinterpreting the Identifiability of Personal Data in the Age of Artificial Intelligence

Josphat I. Ayamunda[*]

ABSTRACT

This paper analyzes a significant legal issue pertaining to the application of Artificial Intelligence (AI) to personal data, namely the re-identification of natural persons to their personal data. Since the development and use of AI technologies heavily rely on analyzing large amounts of data and identifying links among them, AI may be used to retrace and de-anonymize data about natural persons, creating new personal data protection risks. Using the permissive theory of data protection law, this paper analyzes the identifiability element of the notion of personal data within the data protection laws of Kenya and the European Union (EU) with a view to explaining how it can be interpreted so as to avoid an overly restrictive data protection regime that could hinder beneficial AI innovations and/or deprive individual data subjects of the protection that the legislator intended them to have. This paper proposes restricting identifiability to the time of data processing at issue, and limiting the assessment of identifiability to the controller or processor, and persons who are likely to receive the information, rather than the public and the broader community of users. That way, the study offers an option that narrows the definition of personal data to reasonable limits that facilitate responsible AI innovation while maintaining adequate and effective protection of data subjects.

Keywords: Artificial Intelligence, Facilitative Objective of Data Protection, Identifiability of Individuals, 'Means Reasonably Likely to Be Used' Test, Permissive Interpretation, Personal Data

* LLB (UoN), MLitt in Law (Oxon), PhD Law Researcher (UDSM). Advocate of the High Court of Kenya and Lecturer in intellectual property and information technology law at Moi University, School of Law, Kenya. Email: ayamunda@mu.ac.ke

TABLE OF CONTENTS

I. INTRODUCTION

The problem this study addresses is the difficulty of interpreting the concept of personal data in Kenya and the European Union (EU) in ways that promote societally beneficial artificial intelligence (AI) technology without exposing data subjects to detrimental effects. In both jurisdictions, the term personal data is defined in pertinent data protection laws (DPLs) (which are the reference of interpretation) as 'any information relating to an identified or identifiable individual' (Kenyan Data Protection Act (DPA), s. 2; European Union, (GDPR), 2016, art. 4.1). These jurisdictions also share the same problem regarding the scope of personal data. The problem lies in the fact that, on balance, EU interpreters go along with the broad notion of personal data because of the fear that a somewhat restrictive interpretation might leave individuals deprived of protection, while their Kenyan counterparts generally prefer a narrow interpretation, despite the risk of leaving individuals under-protected (*Peter Nowak v. Data Protection Commissioner* (2017); *Allen Waiyaki Gichuhi and Charles Wamae v. Florence Mathenge and Ambrose Waigwa* (2022)). It is therefore not entirely clear how personal data should, in fact, be interpreted to avoid the undesirable outcomes on either end of the interpretative spectrum.

Neither the EU GDPR nor the Kenyan DPA contains specific provisions for AI. However, it is important to consider the connection between the personal nature of information and technological development. In connection with DPL's definition of personal data, the primary question is whether and to what extent technological developments, such as fully trained AI systems, permit the identifiability of natural persons now and in the future. In this regard, AI raises two key issues in particular: '(i) the re-personalization of anonymous data, namely the re-identification of the individuals to which such data are related; (ii) and the inference of further personal information from personal data that are already available' (European Union, 2020, p. 36).

From the perspective of a permissive theory of data protection law, personal data and the use of AI typifies the inherent conflict between the dual objectives pursued by DPL: the protection of personal rights on the one hand and the facilitation of an effective and competitive data economy on the other (Paal, 2022). By analyzing large amounts of data and identifying links among them, AI may be used to de-anonymize personal data.

Interventions through DPLs based on the prohibitive theory of data protection law aim at enforcing a high standard of personal data protection that can limit the free flow of data, which is at the heart of the development of AI technologies. Conversely, permissive DPL interventions ensure protection of personal data while allowing its usage to foster innovation by framing data protection as a tool of fairness and transparency that defines how personal data is processed within a set framework of checks and balances. However, since self-learning mechanisms primarily drive AI technology, even system programmers might not always know which types of data were processed in what ways and which inferences were drawn from which data correlations (Poscher, 2022). This self-adaptive behavior of some types of AI systems might lead to a lack of transparency and, therefore, run counter to the transparency principle, a core feature of permissive theory-based DPLs. As such, some scholars suggest abandoning the traditional concept of personal data protection (Poscher, 2022).

While this paper acknowledges the transparency challenge posed by AI, it is not entirely clear whether and how AI's lack of transparency rules out narrowing down personal data as an appropriate option for addressing emerging AI technologies. As detailed in section B of part III below, it is granted that although the approach of the Court of Justice of the EU (CJEU) to delimiting the notion of personal data in such cases as *Patrick Breyer v. Bundesrepublik Deutschland* (2016) is somewhat narrower than the 'distinguishing approach' taken by the Article 29 Working Party (A29WP), the CJEU'S approach still favors a

broad meaning of identification. On the one hand, the CJEU's approach is narrow, in that it reduces identification to lookup identification (i.e., by means of identifiers connecting a person to their real-world identity). On the other hand, it is broad in recognizing the role of additional auxiliary information in making the data subject individually distinct from the group to which the information is merely relatable.

The CJEU's approach, which favors broad interpretations of the concepts of identifiability and the relational link between information and the individual, must be seen in light of the overarching regulatory objective of EU DPL: to ensure effective and complete protection of data subjects. The CJEU first introduced the principle of effective and complete protection in the *Google Spain SL, Good Inc v Agencia Espanola de Proteccion de Datos and Mario Costeja Gonzalez* Case (2014). This principle requires that the notion of personal data be construed in such a way that it remedies any restrictive effects on the scope of legal protection for the individual. In other words, it is believed that a narrow interpretation of the concept of personal data would go against the aim of affording a data subject effective and complete protection. However, as many scholars have argued, it is highly doubtful whether a broad concept of personal data necessarily attains effective and complete protection of the data subject (Lynskey, 2023; Vergnolle, 2022).

Against that background, this article analyzes some significant difficulties with the identifiability component of the notion of personal data and offers suggestions for overcoming them. As illustrated in part III of this paper, the interpretation of what constitutes 'a means reasonably likely to identify a person' can lead to the over-inflation of the identifiability requirement. In that regard, this work seeks to test the hypothesis that there may be other interpretations or identifiability tests that could yield a relatively high identifiability threshold, thereby narrowing personal data to a reasonable dimension. The current identifiability test seems to be contextual. Thus, eventuating in an overly broad

concept of personal data. Although this outcome might appear to contribute to the complete and effective protection of individual data subjects, it risks stifling societally beneficial AI innovations. The fact that the wording of 'the means reasonably likely to be used' test is without a time horizon logically makes all data personal, as there is no way of telling whether something will always be anonymous or not. Therefore, questions might be raised about the appropriateness of a notion of 'identifiability' that is not time-bound. This is not an area that has been thoroughly explored by existing academic literature. To this extent, this paper examines the possibility of putting a time limit on identifiability.

This paper argues that because the permissive theory of data protection emphasizes the facilitative objective of the right to protection of personal data, it offers a powerful departure point for delimiting the identifiability element of the notion of personal data. It proposes that the idea of personal data (in Kenya) can, and should, be interpreted using a method that considers the underlying permissive nature of data protection law, in order to ensure that it is flexible enough not to deprive individuals of their right to data protection, while remaining sufficiently narrow as not to cover all data all the time.

This paper proceeds by presenting a brief outline of the fundamental theories of data protection and how they guide and manifest themselves in Part II. This is followed by Part III, which examines the meaning of the identifiability element of personal data. Next, Part IV of the paper traces the historical development of 'the means reasonably likely to be used' test and its subsequent interpretation by courts and data protection authorities. The paper then makes suggestions for a permissive interpretation of the identifiability element of the notion of personal data in Part V.

II. THEORETICAL FRAMEWORK ON DATA PROTECTION

Within the data protection academic community, there are two theories linked to the understanding of the right to personal data protection as either a permissive or a prohibitive right. Although these conceptualizations differ, 'both start from the premise that the overarching purpose of the right is to counter the power and knowledge asymmetries that emerge, in an increasingly digitalized society, between the controllers and data subjects' (Christofi and Verdoodt, 2019).

On the one hand, the prohibitive theory of the right to the protection of personal data conceives the right as proscriptive, in the sense that it primarily prohibits the processing of personal data (Fuster and Gurtwirth, 2013). Plainly, according to the prohibitive theory, the right to personal data protection in principle forbids data processing. Under this theory, data protection is closely linked to opacity and to shielding individuals against the use of or interference with their personal information, with a view to protecting the rights and interests of the individual data subject.

Consequently, the rationale for the right to data protection is conceived as providing individuals with control over their personal information. Put another way, under the prohibitive theory, data protection is essentially a right intended to reduce information and power asymmetries between data controllers and data subjects by giving the latter control over their personal data and its processing. That way, it is hoped that the individual will be effectively protected from the harms that might arise from the processing of their personal data.

On the other hand, the permissive theory of the right to the protection of personal data conceives the right as consisting of rules that regulate and limit the processing of personal data, but do not prohibit it. This permissive conception 'assumes that personal data in principle may and will be processed, but asserts that such processing should be fair' (Fuster and Gurtwirth, 2013, p. 533). In the words of Dalla Corte (2020):

> 'Data protection is meant to allow information sharing: there would be no need for it if there were a general prohibition of personal data disclosure, and the law very seldom prohibits the processing of personal data, rather mandating the requirements to be respected to make it lawful' (p. 148).

Thus, DPLs are *not* barriers to the processing of personal data but *permissive* laws to *enable* the use of personal data within a set of standards and safeguards designed to protect data subjects. Put another way, the right to data protection enshrines a claim grounded in fairness, aimed at safeguarding the processing of personal data. This theory is based on the premise that 'data protection is pragmatic and that allowing data processing by public and private entities is desirable—or even necessary in modern society' (Christofi and Verdoodt, 2019).

The essence of the prohibitive approach is control, whereas the permissive approach places a premium on fairness and checks and balances as the essence of rights. As a permissive right, personal data protection is designed to provide safeguards to individuals whenever their personal data is processed, rather than to prevent data processing per se. Data protection, therefore, allows 'personal data processing if and to the extent that it is subject to a certain set of rules, *rather than prohibiting it unless specific requirements are satisfied* (emphasis added)' (Dalla Corte, 2020, p. 150).

As a permissive regime, data protection is designed to facilitate the flow of personal data (as opposed to curtailing it) by providing safeguards for individuals whenever their personal data is processed. As an inherently permissive regime, data protection is meant to channel personal data processing through a set of rules rather than prohibit it *tout court* (Dalla Corte, 2020). It regulates the processing (not by prohibiting it) but by setting the rules for doing it.

By framing data protection as a permissive right, a tool of transparency that defines how personal data is processed, this paper emphasizes the facilitative objective of data protection rather than the protective one. Accordingly, the paper proposes

an interpretation of the identifiability element of personal data whose essence is framed as substantive fairness rather than individual control.

It is observed that the two theories reflect the dual objectives of data protection: protecting individual rights and promoting economic growth by facilitating the free flow of data. The current interpretation of personal data places significant emphasis on the prohibitive theory and the protective rationale that follows. It highlights effective protection of the individual as the most important objective of DPL; however, it underestimates the other equally important objective of enabling the free movement of personal data. This eventuates in a very high threshold for identifiability and sets an impossibly high standard for anonymization (Information Commissioner's Office (n.d.)).

While that approach to interpreting identifiability is designed to provide effective and complete protection for data subjects, it hinders the development of AI innovations in the following manner. The development of practical AI innovations requires massive amounts of data. However, DPLs limit the data needed for training AI systems by designating certain data as personal, to safeguard the rights and interests of individuals. Ordinarily, such laws require a lawful basis (for instance, the data subject's consent) for the processing of personal data. Where the data controller is unable to establish a lawful basis, they can anonymize the personal data in question in order to process it without being bound by a DPL. However, a high threshold of identifiability creates an unattainable ideal of complete anonymity that may impede the practical and beneficial use of data for AI innovation.

Thus, narrowing personal data in line with the permissive theory and the resultant facilitative rationale could prove useful for data controllers developing AI in two ways. First, it would enlarge the pool of non-personal data that they can use without the constraints of DPL. Second, and perhaps most importantly, it would make it easier for them to anonymize personal data.

This is because the suggested permissive interpretation emphasizes reasonableness and assesses identifiability at the time of processing or within a specified time limit.

III. MEANING OF IDENTIFICATION

The idea of identifiability is a 'possibility of' identification. However, it is not entirely clear what identification in DPL means. As Purtova (2022) argues, it makes little sense to argue if a natural person is 'identifiable' when it is not clear when a natural person would be 'identified' and what it means to identify somebody.

While the DPA of Kenya does not define what identification is, it defines an identifiable natural person as a person who can be identified directly or indirectly, by reference to an identifier (DPA, s. 2). It also specifically includes the term 'online identifiers' within the definition of what constitutes personal data (DPA, s. 2). The European Union's Article 29 Working Party (A29WP)/ European Data Protection Board (EDPB)[1] gave identification a broad interpretation to mean distinguishing one in a group—the 'distinguished from' approach. It says that a person is considered to be identified or identifiable where they can be 'distinguished' from others (Article 29 Data Protection Working Party, 2007). It distinguishes between direct identifiability, when information immediately singles out a specific individual, and indirect identifiability, when a person is singled out through a unique combination of data.

In reference to anonymized data, A29WP put forth three criteria that should 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

[1] The A29WP is the predecessor of the European Data Protection Board (EDPB) established by the GDPR as an independent advisory body, providing recommendations and opinions to the European Commission on matters relating to data protection. It was dissolved in 2018, and it did not create any binding rules.

to an individual, and (iii) information concerning an individual can still be inferred (Article 29 Data Protection Working Party, 2014). 'Singling out' was not mentioned in the Data Protection Directive (DPD). Still, it can be linked back to the Working Party's opinion on the concept of personal data, in which identifiability is equated with the possibility of singling someone out (Finck and Pallas, 2020). While Recital 26 of the GDPR now explicitly references 'singling out', inference and linkability are elements considered by the Working Party but not expressly mentioned in the GDPR. Although A29WP emphasized that meeting these three thresholds (i.e., singling out, linkability, and inference) is very difficult, it underlined the need to consider all the means reasonably likely to be used to identify an individual (Article 29 Data Protection Working Party, 2007).

Regarding the means to identify, A29WP, which is now the European Data Protection Board (EDPB), stated that a mere hypothetical possibility to single out the individual is not enough to consider the person as identifiable (Article 29 Data Protection Working Party, 2007). Similarly, if that possibility is negligible, the person should not be considered as identifiable, and the information would not be regarded as 'personal data'. The criterion of 'all the means likely reasonably to be used either by the controller or by any other person' should, in particular, take into account all the factors at stake (Article 29 Data Protection Working Party, 2007).

Identifiability can therefore be defined as the possibility of distinguishing one from a group. Based on A29WP's broad interpretation of identification, various scholars have attempted to define identification comprehensively (Leenes, 2008). For instance, Ronald Leenes (2008) bases his argument on the idea that identification is the act of being distinguished from, or 'individualized', within a group of subjects, the identifiability set. He distinguishes four types of identification: look up, recognition, classification, and session.

The term ' lookup identification' refers to the process of identifying a named individual using an identifier, such as a name, phone number, passport number, or even an IP address, where there is a registry, directory, or table that links that identifier to the named individual (i.e., their civil identity) (Leenes, 2008). An individual can be 'looked up' using an identifier in the real world, hence the name. The term 'real world' is used here in contradistinction to the term 'online world'. The former refers to the physical environment and the tangible interactions within it, while the latter is a digital space accessed through the Internet, characterized by virtual interactions.

Recognition identification requires a person's physical presence or activity; it refers to the identification of a person without reference to that person's civil identity (Leenes, 2008). A person is recognized as an identifiable person if they have previously been known to that person or if they provide attributes, such as a description of their physical appearance, to the entity performing the identification. Under classification identification, a person is identified as a member of a specific group or category. The goal is to identify a person as a member of one or more organizations, not to establish their civil identity or to acknowledge them (Leenes, 2008). Lastly, session identification tracks a person during a certain interaction, and the duration of the session identifier is constrained to that period of interaction. Session cookies are an example of how an online store can tailor a customer's buying experience by ensuring that the website remembers the goods in a customer's shopping cart (Purtova, 2022).

This understanding of identification brings many data-driven practices within the protective scope of DPL that might involve personal data processing but are not tied to an individual's real-world identity, such as online behavioral advertising when the data processed does not include real-world identifiers (Purtova, 2022). As such, the deconstruction of the concept of identifiability into more specific subcategories accords precisely with the prohibitive approach to data protection and promotes

the protective rationale of the right to personal data protection. However, from the perspective of the permissive nature of the right to personal data protection, it could be argued that identification means real-world identification only, since the concept of identifiability is a 'possibility of' identification, and identifiability must include the possibility of this real-world identification. In this light, real-world identification means distinguishing an individual by establishing their civil identity.

Real-world identification occurs when a lookup identifier is associated with a named individual. In that case, the possibility of singling an individual out in other ways would not suffice. This permissive approach would not only promote the facilitative objective of DPL but also enhance substantive fairness by affording the individual protection where they have been distinguished from a group (rather than where there is merely a hypothetical or negligible possibility of singling out).

IV. THE 'MEANS REASONABLY LIKELY TO BE USED' TEST

In this section, focused on the concept of 'means reasonably likely to be used', the analysis is organized into seven key themes. These themes are primarily based on the list of factors typically included in DPLs for determining which means are reasonably likely to be used to identify the concerned natural person. As indicated in the introductory part of this paper, in both Kenya and the EU, the term personal data is ordinarily defined as 'any information relating to an identified or identifiable individual' (DPA, s. 2; GDPR, art. 4.1). Therefore, in both jurisdictions, an identifiable person is one who can be identified directly or indirectly.

According to DPLs of the EU, to determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used (European Union, 2016 (GDPR), recital 26). In order to establish whether means are

reasonably likely to be used to identify the data subject, one is required to take into account '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' (European Union, 2016 (GDPR), recital 26).

Kenyan DPLs are largely modelled on the GDPR, and the conventional EU-influenced definition of personal data is adopted verbatim by the DPA (African Declaration on Internet Rights and Freedoms Coalition, 2021; *Nubian Rights Forum v. Attorney General*, 2020, para 758). Notably, the DPA does not expressly provide for the 'all means reasonably likely to be used' test. This might be viewed as a manifestation of the legislator's intention to oust the test from Kenyan DPLs. Yet, in their day-to-day practice, data protection stakeholders in Kenya refer to the EU sources (even though the EU legal framework is not binding on them). For instance, in *Nubian Rights Forum v. Attorney General* (2020), the High Court of Kenya analyzed the extent to which the DPA complied with internationally accepted data protection standards. It decided that it would be guided by the principles developed by the Organization for Economic Cooperation and Development (OECD)—namely, the OECD Privacy Principles (1980)—which, in its view, was a more comprehensive and internationally recognized data protection framework deeming it most appropriate for its purposes (*Nubian Rights Forum v. Attorney General*, 2020, para 843). However, the High Court proceeded to refer to, and was guided by, the GDPR throughout the analysis of the issues in question. It heavily relied on definitions of personal data in the GDPR, decisions of the CJEU, and advisory opinions of EU bodies on personal data, arguing the concepts of identifiability in the DPLs of Kenya and the EU are similar (*Nubian Rights Forum v. Attorney General*, 2020, paras 758-760).

A review of the database of the rulings by the data commissioner, for a Kenyan perspective of identifiability, does not show any complaint that raised issues requiring the application

of 'the means reasonably likely to be used' test. This study finds that the probable reason is that the complaints in question concerned cases in which the data subjects were identified rather than identifiable. As such, where the data subject is identified rather than identifiable, 'the means reasonably likely to be used' test becomes redundant. In other words, where the data subject is identified, there is no need to apply 'the means reasonably likely to be used' test.

The 'means reasonably likely to be used' test under the GDPR applies to both direct and indirect identifiability. As established, the test is only necessary to determine if a person is identifiable. When data contains unique identifiers, the test has little or no practical relevance, typically because the data subject is already identified (i.e., an individual's identity is immediately obvious from the data itself). As such, there is no need to analyze whether it is reasonably likely for someone to identify the data subject. This appears to be the norm for the complaints determined and made available in the public database of the Office of the Data Protection Commissioner (ODPC) as of August 2025. Although not formally binding, the 'all means reasonably likely to be used' test, as developed in EU DPLs, has undeniable persuasive authority in Kenya and provides the most comprehensive guidelines for data controllers on how to apply the notion of personal data in practice.

A. Historical Development of the Means Reasonably Likely to be Used Test

Rossi (2020) argues that 'it was decided very early on that 'personal data' would mean any data, regardless of the sensitivity of its contents with regards to 'privacy', and, in the overwhelming majority of cases, relating to an individual, and not a group' (p. 38). The reason for that is explained as follows:

> 'This difficulty in determining once and for all, *erga omnes*, for everyone, what informational contents would be 'private' and which one would be excluded from 'data protection' led to the adoption of a broad definition

> that was agnostic to the content. *Any data, whatever its content, would be 'personal' if there was a link between the said data and a 'person"* (emphasis added) (Rossi, 2020, p. 40).

Similarly, the French Law adopted in 1978 applied to nominative information. It stated that 'nominative data' (*donnée nominative*) was any data 'directly or indirectly' related to a person (Rossi, 2020) or as data allowing the identification of the person concerned by them (Fuster, 2014). The French approach 'did not circumscribe the purpose of regulating the processing of data related to individuals to the single objective of protecting privacy or private life' (Fuster, 2014, p. 64).

Significantly, and in contrast to the German approach (which viewed personal data processing as a prohibitive notion), the French Law envisaged a permissive approach to data protection (Christofi and Verdoodt, 2019). In line with the permissive theory (and, therefore, the facilitative rationale of DPL), the French Law assumed that data processing was to take place and, therefore, the law's core objective should be the description of conditions that should govern such processing (Fuster and Gurtwirth, 2013). The only exception to this general rule that processing should be permitted was reserved with regard to sensitive data (Fuster and Gurtwirth, 2013).

At the international level, the OECD Guidelines (Ministerial Council of the OECD, 1980) and Convention 108 (Council of Europe, 1981a) referred to an 'identified or identifiable' person. The OECD Guidelines defined personal data as 'any information relating to an identified or identifiable individual (data subject)' (Ministerial Council of the OECD, 1980, Art. 1(b)). Article 2(a) of Convention 108, the first binding international legal instrument, has the exact definition as the OECD Guidelines. However, as indicated below, in later legal instruments (and related revisions), this definition was expanded. Despite the embroilment between data protection and privacy in the international legal instruments, both Convention 108 and the OECD Guidelines emphasized the idea that, in personal data processing, it is nec-

essary to balance the protection of individuals with the free flow of personal data. Not only does this reveal the long-established dual objectives of DPL, but it is also reminiscent of the dualistic prohibitive/facilitative conceptualization of the right to data protection.

The DPD (the predecessor of the General Data Protection Regulation (GDPR)), expanded on the core definition of personal data in the Council of Europe Convention No. 108 for the protection of individuals with regard to automatic processing of personal data by combining two elements: 'identified or identifiable' and 'directly or indirectly' and specifically clarifying when individuals should be deemed to be 'identifiable' (by 'anyone'). However, it should be noted that the broadening of the concept of 'personal data', which demarcates the material scope of the data protection instruments, did not begin in the 1980s. As van de Sloot (2017) aptly demonstrates, 'the two Resolutions for data processing from 1973 and 1974 simply defined 'personal information' as information relating to individuals (physical persons)' (p. 5).

This expansion continued in the 1980s, when the subsequent Convention defined 'personal data' as any information relating to an identified or identifiable individual (Council of Europe, 1981b). The explanatory report stressed that an 'identifiable person', a new element in the definition, meant a person *who could be easily identified from the data* (Sloot, 2017, p. 5). It did not cover the identification of individuals by means of very sophisticated methods. In the DPD, *the definition of personal data was expanded to include* an identifiable person who can be identified, directly or indirectly, by reference to an identification number or to one or more factors specific to their physical, physiological, mental, economic, cultural, or social identity. This is viewed as introducing an extensive, non-exhaustive list of possible identifying factors and as inserting 'indirect' identifiable data (Sloot, 2017, p. 5).

The definition of personal data is not at all modified in the 'modernized' Convention 108 of 2018, though an 'extensive gloss has been added in the Explanatory Memorandum to the Modernized Convention' (Korff and Georges, 2019, p. 97). Finally, in the GDPR, the term 'personal data' is defined in an even slightly broader manner. For instance, in Article 4(1), it further expands the notion of personal data, by clarifying that a person can also be identifiable by means of an 'online identifier'. Even though the GDPR expanded the list of identifiers, the constituent elements of personal data remained the same (Tetiana, 2021). In this respect, some of the significant points of controversy raised and thrashed out regarding the delimitation of the notion of personal data in the GDPR are examined immediately below.

During discussions on the GDPR, the EU Commission's initial proposal did not change the fundamental elements of the definition of 'personal data' but did reorder the presentation of the concepts of 'personal data' and 'data subject' by tying the former to the latter. The definition in Article 2 of the DPD reads as follows:

> (a) 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

That definition was inverted to read as follows in Article 4 of the GDPR:

> (i) 'data subject' means an identified natural person or a natural person who can be identified, directly or indirectly, *by means reasonably likely to be used by the controller or by any other natural or legal person*, in particular by reference to an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person (emphasis added);
>
> (ii) 'personal data' means any information relating to a data subject.

This inversion raised a lot of opposition, and so did the words 'by means reasonably likely to be used by the controller or by any other person', which were removed from the article containing the definitions, but kept in recital 26 (Rossi, 2020, p. 43). This arrangement is a manifestation of the legislator's intention to endorse 'an absolute concept of 'personal data' while leaving some flexibility for interpretation by including the word 'reasonably' (Rossi, 2020, p. 43).

The absolute concept of 'personal data' states that to determine whether information constitutes personal data, the likelihood of identification of the individual ought to be assessed from the perspective of any third party. This is in contrast to the relative approach, which requires that, to determine whether information constitutes personal data, the likelihood of identification of the data subject be assessed from the perspective of the data controller (Finck and Pallas, 2020). The absolute concept of personal data is strongly influenced by and closely connected to the prohibitive theory and seeks to ensure the complete protection of the data subject. This stands in contrast to the relative approach, which follows the permissive theory and aims at enabling the free flow of personal data.

It is worth noting that the words 'by means reasonably likely to be used by the controller or by any other person' have their origin in the legislative process of the DPD. The original proposal for the DPD contained the notion of 'depersonalization' (Commission of the European Communities, 1990). This concept was understood as the alteration of personal data so that it could no longer be associated with an identifiable natural person (Sloot et al., 2022). The explanatory memorandum stated that the data could be regarded as depersonalized, even if, in theory, it could be repersonalized with disproportionate technical and financial resources. It defined depersonalization as 'modify[ing] personal data in such a way that the information they contain can no longer be associated with a specific individual or an individual capable of being determined except *at the price of an excessive*

effort (emphasis added)' (Commission of the European Communities, 1990, p. 50). Although that suggestion was considered and rejected, the same type of formulation later reappeared in one of the DPD's recitals under the term anonymization. The terminology used in the recital that was adopted referred to in Recital 26 of the DPD as 'all the means likely reasonably to be used either by the controller or by any other person'.

Two other major controversial issues regarding the meaning of identification are pseudonymization and singling out. The negotiations on these issues are well documented by Rossi (2020), thus:

> 'Several interest groups from the industrial coalition lobbied for the inclusion of a definition of a concept of 'pseudonymous data' in the GDPR. …The industrial coalition was successful in including a concept of 'pseudonymization' in article 4 of the GDPR, but the coalition of privacy advocates managed to get it to be phrased in a way that ensured that pseudonymous data would explicitly stay contained within the scope of 'personal data': … Thus, pseudonymization in the GDPR is now a useful – and sometimes necessary – extra safeguard, but not something that can be used to be exempted from certain data protection principles or data subject rights. Its inclusion did not move data protection law in the EU away from an absolute definition of 'personal data' (p. 43).

In that way, EU DPL negotiators/lawmakers arguably followed a prohibitive approach. The same approach was taken in reference to the proposed amendments by some privacy advocates and non-governmental organizations (NGOs) to add the words 'single out' or 'singling out' to the definition of 'personal data'. While 'singling out' never made it to the final version of the definition, it has been included in recital 26 of the GDPR, which states that '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.' As Rossi (2020) puts it:

> 'During the writing process of the GDPR, it appears that the new definition [of personal data], compared to that of Directive 95/46, includes new elements in favor of an absolute interpretation of the notion. While

> a concept of 'pseudonymization' was included, pseudonymous data was explicitly included in the scope of 'personal data' and there was no serious attempt to restrict the application of data protection principles to data that relates to the 'private life', 'privacy' or 'intimacy' of the data subject' (p. 45).

Ultimately, while the DPD referred to the right to privacy thirteen times, the GDPR, which replaced the DPD, deleted all references to privacy entirely.

B. Interpretation and Application of 'the Means Reasonably Likely to be Used' Test

It is evident from the legislative history of the concept of 'all means reasonably likely to be used' that the initial proposal was for the phrase 'at the price of an excessive effort' (Sloot et al, 2022, p. 79). The reference to 'all means likely' seems more open, contextual, and fluid than the term 'excessive effort'. The term 'excessive effort' is more objective and less contextual. It also has a higher threshold. As such, the 'excessive effort' test is more restrictive than the 'means reasonably likely to be used'. However, this proposal was rejected because the EU legislator did not intend to make the test for indirect identifiability objective or objectifiable (Sloot et al., 2022). It was felt that 'excessive effort' should be deleted, for a processing task requiring excessive effort today may require no effort at all next year (Sloot et al., 2022, p. 78). While this concern remains valid today, one might question how far into the future the competent interpreter should look to properly evaluate excessive effort.

One of the controversial issues in reference to the words 'by means reasonably likely to be used by the controller or by any other person' as used in Recital 26 of GDPR regards the phrase 'by the controller or by any other person'. The challenge the phrase raises is that, despite being a manifestation of the EU legislator's intention to endorse an absolute concept of personal data, it is too open-ended to be helpful to data controllers and

data processors. This is because there is no predetermined limit on the parties that might be involved in the identification process. For instance, is it any person who may reasonably access the data or any conceivable third party globally?

In other words, the absolute concept of 'personal data' requires that to ascertain whether any given information is personal data, one should assess the likelihood of identification of the individual from the perspective of any third party (and not just the perspective of the data controller) (Finck and Pallas, 2020). The difficulty with that approach is that one can never rule out with certainty the possibility that a third party will disclose the identity of the person in question (Groos and van Veen, 2020).

By contrast, the relative approach provides that, to establish whether information constitutes personal data, the likelihood of the individual's identification ought to be assessed from the perspective of the data controller or data processor (Finck and Pallas, 2020). Whereas the absolute concept of personal data is based on the prohibitive theory and seeks to ensure a high level of protection of the individual, the relative approach—which rests on the permissive theory—is better placed to not only facilitate the free flow of information but also improve the effectiveness of the personal data protection regime. This is because the relative approach limits the assessment of identifiability to the controller or processor and to persons likely to receive the information, rather than to anyone in the world, thereby easing the compliance burden. The identifiability test might be made more objective as discussed immediately below.

1. Direct and Indirect Identifiability

Both the DPA and GDPR define an identifiable natural person as a person who can be identified directly or indirectly, 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. The definition recognizes that a per-

son may be identified or identifiable directly or indirectly. Identifiability refers to the possibility of being identified in the future (Article 29 Working Party, 2007). A person is considered directly identified—distinguished from a group—by name or another identifier that is obtained, and where no additional information is necessary (Purtova, 2022).

As a corollary, a person is directly identifiable when such a unique identifier is not yet available but is reasonably likely to be obtained. Conversely, a person is considered indirectly identified when distinguished by a unique combination of non-unique identifiers that is complete and requires no additional information. Consequently, a person is indirectly identifiable when a unique combination of non-unique identifiers forms an incomplete chain and further information is necessary and is reasonably likely to be obtained (Purtova, 2022).

Purtova (2022), while discussing the meaning of identification under the GDPR, underscores that a data subject is directly identifiable when a unique identifier is not obtained yet, but it is reasonably likely to be obtained. This is in contrast to the case of a directly identified person, which entails that a data subject is distinguished from a group, by name or another unique identifier, which is obtained, and where no additional information is necessary.

The quest to categorically define what constitutes a unique identifier has not been easy. It is instructive to observe that A29WP illustrated a number of unique identifiers (Article 29 Working Party, 2007). These include names, file numbers, identification numbers, and IP addresses. Inasmuch as efforts are underway to obtain a list of unique identifiers, there are concerns that even the name, regarded as the most distinctive unique identifier, may sometimes fail to distinguish an individual from a group. This uncertainty on the state of unique identifiers was canvassed in the case of *Patrick Breyer v. Bundesrepublik Deutschland* (2016). In this case, Breyer had accessed several websites of the German Federal Government that stored

information regarding access operations in logfiles. This information included the visitor's dynamic Internet Protocol (IP) address (i.e., an IP address that changes with every new connection to the Internet). He argued that storing his IP address violated his rights, especially his right to data protection. The court, in rendering its decision, held that a dynamic IP address is not data relating to an identified natural person but can be considered in combination with other entries to make a person identifiable. In essence, the court ruled against the consensus that an IP address is a unique identifier, as established in the *Scarlet v SABAM* (2011) case, in which the court held that internet users' IP addresses were protected personal data because they allow users to be precisely and uniquely identified.

This apparent confusion in jurisprudence is somewhat ameliorated by legal doctrine. For instance, as Purtova (2022) correctly argues, '*Breyer'* effectively reduces the meaning of identification and 'identified' to lookup identification, through identifiers that connect a person to their real-world identity (p. 178). While this narrows down the idea of identification, it takes out of the protective scope of DPL many data-driven practices which have long been assumed to involve personal data processing and thus fall under DPL, but which are not tied to a real-world identity of an individual by his phone number, address, passport number, a name or similar identifiers (Purtova, 2022, p. 178). Examples of such practices are:

> 'online behavioral advertising when the data processed does not include real-world identifiers and a person is 'reached' through the online identifiers alone, facial recognition when the facial templates are not associated with the real-world identity, [and] individual profiling targeted at a person by means other than offline identifiers' (Purtova, 2022, p. 178).

Accordingly, under EU DPL, the *Breyer* decision must be read in light of the DPL's aim to ensure effective and complete protection of data subjects (Purtova, 2022). This means that identification should be interpreted broadly as distinguishing from a group. This position (i.e., interpreting identification as distinguishing from a group) may also apply in Kenya. Although

DPA does not define what identification is, it defines an identifiable natural person as a person who can be identified directly or indirectly by reference to an identifier. Specifically, it includes the term 'online identifiers' within the definition of what constitutes personal data.

The common factor between the two jurisdictions is that interpreting identification as distinguishing from a group reflects data protection as an individual right, meant to protect individuals rather than the groups to which they belong. It protects the person from the moment they become individually distinct from the group of individuals they belong to.

i. Direct Identifiability

Unique identifiers are defined as pieces of information that hold privileged and close relationships with the person (Article 29 Working Party, 2007). However, A29WP did not explicate the criteria for determining what is privileged and close. To avoid such a lacuna, Purtova (2022) has provided a reliable definition of an identifier as a piece of information that, alone or in combination with other identifiers, distinguishes a person from a group. For instance, it is generally appreciated that each human being has a voice that is distinct and different from everyone else's (McGettigan and Lavan, 2017). GDPR has taken the position that voice recognition is an example of a physical or physiological biometric identification technique (Panvier, 2021). A similar approach is taken under section 2 of the DPA. Voice discloses an abundance of personal information about the speaker.

Kroger (2022) argues that benign voice recordings can reveal sensitive attributes such as a person's geographical origin, health status, personality, gender, and age. However, in practice, voice has proved challenging to identify directly or to make data subjects directly identifiable. Authentication is a huge problem. This is true as ear-witness testimonies, or hearsay, have been established to be notoriously unreliable and inaccurate (Kroger, 2022). Therefore, while each person possesses a unique voice, it

is not apparent how functional that uniqueness is to make them directly identifiable.

A29WP underlined that a purely hypothetical possibility of identification is insufficient to meet the standard of means reasonably likely to be used (Borgesius, 2016). Instead, all relevant factors should be considered to assess this possibility (Article 29 Working Party, 2007). Such factors include: the cost of identification; intended explicit or implied purposes; risk of organizational dysfunctions and technical failures; state of the art in technology at the time of processing, including possible developments in the future, within the lifetime of processing.

ii. Indirect Identifiability

It is instructive to note that an identifier, such as an image, can, in some situations, directly identify a data subject without the addition of other necessary information. This was appreciated in *R v. The Chief Constable of South Wales Police and Secretary of State* (2019). The case concerned a police facial recognition system in which closed-circuit television (CCTV) cameras captured facial images of passers-by within their range, first distinguishing human faces and then distinguishing one face from another, enabling matching the images with biometric templates on watch lists. The plaintiff was within the cameras' range on two occasions and filed a suit claiming that his personal data was being unlawfully processed, even though he was not matched with the watch list on either occasion. The court ruled that processing of the plaintiff's image constituted processing of personal data even before the matching of facial images and possible recognition. The court reasoned that at this instance, the plaintiff was directly identifiable on the basis that the information recorded by the facial recognition system individuates him from all others, through singling him out and distinguishing him from all others. As such, the plaintiff was already directly identifiable without the addition of further information, such as by matching his image against others.

2. Risk of Organizational and Technical Failures

The standard for the relevant possibility of identification as discussed above is whether or not the means of identification are 'reasonably likely to be used'. One factor to consider is the risk of organizational dysfunctions and technical failures, including data breaches.

In *Patrick Breyer v. Bundesrepublik Deutschland* (2016), the court was faced with the issue of whether the possibility to combine a dynamic IP address with additional data held by the internet service provider constitutes a 'means likely reasonably to be used to identify' the data subject. The court found the website provider to have the means reasonably likely to be used to identify the website visitors. This was on the premise that, should there be a cyber-attack, the website providers could contact the competent authority, which could then obtain information from internet service providers to initiate criminal proceedings. This decision shows an appreciation of the risk of organizational dysfunctions and technical failures as factors reasonably likely to be used as means of identification.

Remarkably, the intended explicit or implicit purpose of processing has also been identified as a factor to consider (Purtova, 2018). Processing only makes sense if it allows identification of specific individuals and treats them in a certain way (Article 29 Working Party, 2007). The Court in *Patrick Breyer v. Bundesrepublik Deutschland* (2016) considered and accepted the Advocate General's argument that 'the possibility of combining IP address with additional data would not be reasonably likely if it was prohibited by law or practically impossible due to a disproportionate effort in terms of time, cost and manpower, so that the risk of identification appears in reality to be insignificant'. This argument by the Advocate General is particularly relevant to this paper, so far as the factor of illegality of the means used in identification is concerned. That factor is discussed in the fourth subsection below.

Whether or not the risk of organizational dysfunctions and technical failures (including data breaches) is reasonably likely to occur, the answer should be in the affirmative. For instance, there have been several cases of data breaches in Kenya. Thus, one might question the adequacy and effectiveness of the law of Kenya in overcoming the evidential difficulty of establishing liability for leakage of personal data. A case in point is *Joshua Kiprop Kisorio v. Safaricom Plc & 4 others* (2021). In that case, the court was clear that it had been established that there was a leakage of the Petitioner's information to third parties. As a result, the Petitioner lost out on a business venture worth over three million United States Dollars. However, it was not clear whether the leakage was done with the acquiescence or knowledge of the first Respondent (i.e., the data controller). It is almost impossible to safeguard data against organizational dysfunctions and technical failures, including data breaches. Due to the ubiquitous human involvement in or interaction with data processing, it is likely that there will always be a party that could leak the data without leaving any traces.

Perhaps the crucial question here is whether the leakage of identifiers affords the data controller, or a third party, a means reasonably likely to be used for identification. Notably, as the CJEU states, means reasonably likely to be used for identification would not be the case if the identification of the data subject was prohibited by law or practically impossible (*Patrick Breyer v. Bundesrepublik Deutschland* (2016)). In that regard, one might argue that if identifiers are unlawfully made public, they afford a data controller a means reasonably likely to be used, because such identification using publicly available information is neither prohibited nor would it involve disproportionate effort in terms of time, cost, or manpower.

However, this paper suggests that the risk of the data controller taking that opportunity to link the leaked identifiers to the data in its possession and identify the individual concerned appears insignificant. That is because, while it is theoretically

possible to do so, the data controller is unlikely to intend to apply the identifiers to identify individuals due to factors such as legal or contractual restrictions, internal policies, or technical and organizational measures (Clifford Chance, 2022).

3. The Influence of the Current and Foreseeable Technological Developments

This theme concerns the technology available at the time of processing and future technological developments. Determining the means likely reasonably to be used for identification requires consideration of the current state of technology and the constant increase in computing power, as well as the know-how and tools available.

For this paper, technological developments relating to AI are of utmost importance. AI significantly impacts the identifiability element of the concept of personal data, blurring lines and creating new challenges for data protection. The ever-growing ability of AI to analyze vast datasets, including personal information, and draw inferences, can make it easier to identify individuals, even from seemingly anonymized data (Poscher, 2022). In particular, increased identifiability can lead to adverse impacts, including the following: First, AI algorithms can identify individuals from seemingly anonymized data by uncovering patterns and correlations that are not obvious to humans (European Union, 2020). Second, AI can re-identify individuals from datasets previously considered de-identified, particularly when combined with other available data sources (Paal, 2022). Third, this ability to identify individuals from seemingly innocuous data raises concerns about the effectiveness of anonymization techniques.

To mitigate the harshness of this technological reality, DPL increasingly assumes that greater protection should be granted to safeguard individuals, as is evidenced, for instance, by Recital 6 of the GDPR. This happens, for example, by expanding the notion of personal data, as articulated in the *Google Spain case* (2014), to attain effective and complete data protection for data

subjects via a broad interpretative approach. Although the case is best known for its recognition of a 'right to be forgotten' in EU DPL, it is also significant for introducing the principle of 'effective and complete protection' by stating that the aim of DPL is to ensure effective and complete protection of data subjects (*Google Spain case* (2014), para 34).

In this *Google Spain* case, the CJEU adopted a broad approach to the meaning of controllership in reference to search engine providers and personal data, arguing that a broad interpretation of data protection concepts is in line with the principle of 'effective and complete protection' (*Google Spain case* (2014), para 34). In that regard, the concept of personal data should be construed so as not to affect a broader understanding of identification as distinguishing from a group. In the EU in particular, 'all means reasonably likely' as the standard for determining the possibility of identification is in keeping with the protective rationale of DPL and the principle of effective and complete data protection of data subjects.

Although the principle of 'effective and complete protection' has, since its first use in the *Google Spain case* (2014), been applied in several subsequent cases, it is argued that attempts to make the protection offered by DPL more complete risk jeopardizing its practical effectiveness (Lynskey, 2023; Vergnolle, 2022). Considering the growing amounts of information gathered from 'smart' environments capable of assessing or affecting individuals, some legal scholars hold that it is safer to assume that any information is likely to relate to a person (Purtova, 2018). However, other data protection scholars, such as Dalla Corte (2019), are reluctant to adopt that view and instead argue that DPL and the related doctrine and jurisprudence contain elements that enable the avoidance of an overly extensive interpretation of the notion of personal data. Accordingly, they emphasize the permissive nature of the right to personal data protection and the facilitative objective of innovation that benefits from narrowing personal data. This approach is also echoed by other scholars who,

while cognizant of the protective objective of DPL, argue that 'in order to better protect individuals, we should aim at better delimiting – actually, at narrowing down – the scope of application of data protection law, as opposed to deliberately or inadvertently…expanding it' (Fuster, 2023, p. 383).

To ascertain whether means are reasonably likely to be used, account should be taken of all objective factors, such as the costs and the time required for identification. This includes considering the available technology at the time of processing and ongoing technological development. However, questions might be raised about the appropriateness of a notion of 'identifiability' that is not time-bound. For instance, the fact that the wording of 'the means reasonably likely to be used' test is without a time horizon logically makes all data personal, as there is no way of telling whether something will always be anonymous or not. This issue is explored in the seventh subsection below.

4. Legality of the Means Considered

The DPA does not explicitly address the legality of recordings between individuals. One can, however, draw inferences from other laws relevant to data protection to determine the legality of identifying individuals in the context of AI applications to personal data. Article 31 of the Constitution of Kenya ('the Constitution') enshrines the right to privacy in Kenya. It stipulates that every individual has the right not to have their privacy in their communications infringed. Drawing from this provision, it can be inferred that one can face a civil lawsuit for unjustifiably recording and disseminating a private conversation. For instance, if a data controller secretly recorded a conversation in a public office (e.g., a police station), a question would arise as to whether the recorded audio falls within the scope of personal data protection requirements. The framers of the DPA intended to protect data subjects from unnecessary processing of personal data without their consent or another lawful basis. It is immaterial where the data was collected, whether in a public place or a

private place, provided the information contains personal data. The processing of that data should comply with the DPA.

Part IV of the DPA sets out the principles and obligations for the protection of personal data. Section 28 contains provisions for the collection of personal data. It stipulates that the data controller or data processor shall collect data only from the data subject. Section 29 states that a data controller or data processor shall, before collecting personal data, inform the data subject of various facts, including that personal data is being collected and the purpose for which it is being collected. Section 30 proceeds to provide for the lawful processing of personal data. It stipulates that a data controller or processor shall not process personal data, unless the data subject consents to the processing for one or more specified purposes or where the processing is necessary for at least one of the eight purposes stated therein. Section 32 stipulates that a data controller or processor shall bear the burden of proof for establishing a data subject's consent.

In *Shakunt Rajnikant Shah v. Bhupendra Motichand Shah t/a John Cumming & Company & another* (2020) the issue was whether a secretly recorded audio of a family meeting, without the consent of another party, could be admitted as evidence. The proceedings of the family meeting were disputed, and the audio could provide clarity about what happened. While citing *S v. Ismail & Others* (2004), the court observed that, in South Africa, a participant in a conversation could secretly record it without committing an offence, depending on the circumstances. The court, however, took cognizance of section 29(b) of the DPA, which stipulates that a data subject must be informed that their personal data is being collected. It concluded that regardless of who is collecting the data, whether a participant or a third party, it will not be quick to exempt potential evidence from compliance with the law unless sufficient reasons are provided.

As indicated under the second theme in this section, the Court in *Breyer* observed that the possibility of combining one set of information with additional auxiliary information would

not be considered as reasonably likely if it were prohibited by law. For this analysis, the crucial question is why illegality is ruled out as a means available to the controller or to another person, in reference to identifiability. This inquiry is critical, given that people may still be harmed when information is illegally obtained from third parties, or when third parties illegally obtain the auxiliary information necessary to tie data to a specific natural person from data controllers or processors. The legal rule in question appears to be premised on the mistaken assumption that all data controllers and processors always act lawfully, or that third-party activities involving personal data processing may not constitute illegalities. Perhaps the relevant legal question is: 'all means reasonably likely to be used' by whom? By the controller or by a third party? The answer is that it is either the controller or another third party (e.g., a competent authority or a motivated attacker) who can single out a natural person, directly or indirectly, through all means likely to be reasonably used. This logically leads to a low threshold below which a person is considered legally identifiable.

While it is theoretically possible for a data controller (or any other person, e.g., a motivated intruder) to link one set of data with another to make individuals identifiable, an assessment of whether that is probable will be based on the actual capabilities, both in terms of cost, skill, and technological ability, of the data controller in question. However, the distinction between actual and theoretical circumstances is likely to be fluid, given the rapid technological advancements in the processing of personal data, such as AI systems.

While an order from a competent authority will properly be a 'means likely reasonably' to be used to identify the natural person to which the information relates, the conduct of a motivated attacker will not. This is because the attacker's conduct will not be deemed to be 'reasonable'. Yet the conduct of a motivated attacker remains a means likely to be used to identify the natural person to whom the information relates (and thus might cause

harm to the individual). It is, therefore, hard to see how, even though identification can also result from illegal acts (e.g., hacking), the law appears to disregard that mode as a 'means likely reasonably' to be used to identify the natural person to which the information relates. In other words, does the term 'reasonable' reference the likelihood of illegal means being used or the reasonableness of using illegal means? According to A29WP, the likelihood of identification under Recital 26 must be 'reasonable'. In reference to the theme at hand, the term 'reasonable' references the likelihood of illegal means being used.

Purtova's (2018) analysis offers some guidance on how to approach this concern. Purtova (2018) points out that by going along with the opinion of the Advocate General (AG) that identification would not be reasonably likely if prohibited by law, the court's view is in direct contradiction to A29WP's opinion on the concept of personal data. According to the court in *Breyer*, the legality factor, by itself, would not have any restrictive effect on the scope of personal data. In Purtova's (2018) words, 'a legal possibility for the controllers to obtain additional information that enables them to identify data subjects is merely one of 'all the factors at stake' that should be taken into consideration to assess the possibility of identification' according to A29WP. (p. 64).

While A29WP named the possibility of organizational dysfunction (including data security breaches resulting from illegal acts) among the relevant factors to be assessed, the restrictive potential of *Breyer* lies in its starting point: what would be unreasonable. Reflecting on the court's argument, Purtova (2018) argues that 'a more nuanced reasoning would be that a legal prohibition to combine data for identification would make the means of identification 'less reasonably likely to be used', rather than 'not reasonably likely' (pp. 64-65).

Although that is a plausible argument, it still results in a low threshold for when a person is considered legally identifiable. This is because, rather than narrowing the list of factors

to be included in determining which means should be deemed reasonably likely to be used, it somewhat expands it. In that regard, this paper supposes that it is reasonable to argue that the phrase 'less reasonably likely to be used, rather than not reasonably likely' is meant to equate illegality with a negligible level of probability.

Dalla Corte (2019) provides further cues for interpreting the legality factor, noting that 'Recital 26 of the GDPR[2] models the potential attacker as the data controller plus 'another person" (p. 7). Viewed that way, it appears that the legal rule in question presupposes that the motivated attacker (and not the data controller) is likely to engage in illegal means of identification. Yet that does not provide much help in narrowing the list of factors to be included in determining which means should be deemed reasonably likely to be used, or in raising the threshold for considering an individual identifiable. That is so because anyone (other than the data controller) could be 'the other person'—the potential attacker or the subject—attempting the re-identification of de-identified or anonymized data, or the one possessing the auxiliary information necessary for that purpose. In other words, that interpretation does not limit assessment of identifiability to the controller or processor. Instead, it leaves it open to anyone in the world, thereby making it unduly broad.

Therefore, viewing this issue from a permissive or facilitative perspective, this paper aligns with Breyer's restrictive potential but adds that the phrase 'less reasonably likely to be used, rather than not reasonably likely' is meant to equate illegality with a negligible level of probability. When approached in that manner based on the permissive theory, the scope of personal data would be limited to either of two situations, namely (i) where the information is identifiable by the controller by reasonable means or (ii) where the controller ought to know that another person will likely obtain the information as a result of the processing and the

[2] Similar to what Recital 26 of the Data Protection Directive (DPD).

individual will likely be identifiable by that person by reasonable means.

5. Risk of Identification

The risk of identification refers to the possibility that an individual can be identified, directly or indirectly, from data processed or handled by an entity (i.e., data controller or processor) (Finck and Pallas, 2020). It arises when the data being processed, even if it may not explicitly mention the individual's name or other identifying information, can still be used to identify the individual through different means or by combining it with other data (Finck and Pallas, 2020). Under DPLs, organizations or entities that handle personal data are required to take appropriate measures to minimize the risk of identification. This may include anonymization or pseudonymization of data, encryption, data aggregation, or other techniques to protect individuals' identities.

This interpretation accords precisely with the jurisprudence and related doctrine on this issue. For instance, in *Breyer,* the CJEU ruled that the means to identify a data subject would not be considered as reasonably likely to be used if the identification of the data subject was practically impossible because it requires a disproportionate effort in terms of time, cost, and manpower, so that the risk of identification appears in reality to be insignificant. In that regard, this paper agrees with the data protection scholars who emphasize the permissive nature of the right to personal data protection, arguing that in *Breyer's case,* 'the Court did not mean to equate impossibility with zero probability, but with a negligible level of the latter' (Dalla Corte, 2019, p. 25).

6. Amount of Time Required for and Cost of Identification

In reference to GDPR, to determine when a data subject is identifiable, consideration ought to be given to 'all means reasonably likely to be used'. This includes all objective factors, such as the costs and time required for identification, taking into account the available technology at the time of processing and technolog-

ical developments. A proper scrutiny of Recital 26 of the GDPR shows that it is neither decisive nor precise about the timeline for when data may be presumed not to be identifiable. It vaguely states that the 'means' to be considered are not just those that are presently available, but also 'technological developments' (Finck and Pallas, 2020, p. 16). However, A29WP seemed to give a clearer understanding by indicating 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' (Article 29 Working Party, 2007, p. 15). This means that the data controller will have to consider the contemporary time and the last period for which they will process the available data.

The future possibility of identification is based on a risk-management framework. For example, where data is to be kept for a decade, the data controller 'should consider the possibility of identification that may occur also within the ninth year of their lifetime, and which may make them personal data at that moment' (Finck and Pallas, 2020, p.18). This indicates that the data in question becomes personal information only in the ninth year, yet from the beginning, the controller must be aware of, and prepare for, that possibility.

Lastly, and most importantly, in determining a 'reasonable' amount of time and financial resources to identify an individual, it is unclear whether a subjective or objective test should be applied. 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, however, would require a broader evaluation, including who currently has access to information and who might gain access to relevant data in the future (Finck and Pallas, 2020, p. 19). Nevertheless, it can be inferred that the objective test conforms with A29WP's opinion and GDPR that stipulates all means reasonably likely to be used should include consideration of 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' (Finck and Pallas, 2020, p.19).

Despite that, a recurring challenge in the literature on the scope of the notion of personal data is the impediments to data-driven innovation and the burdens that organizations associate with DPLs, particularly with regard to the broad concept of identifiability. Data controllers are especially adamant about this challenge. They feel that identification for the purposes of data protection should not be expanded beyond information that individuates the person, in the sense that they are singled out and distinguished from all others by the controller or processor, at the time of the processing.

It is extremely difficult to achieve anonymization in the eyes of data protection law because data controllers and processors need to concern themselves with future identifiability by anyone in the world. Thus, data controllers and processors take the view that the fact that DPLs set no specific horizon or time limit on the evaluation of the means reasonably likely to be used makes the test less objectifiable and, hence, leads to a broad notion of personal data. This means that, keeping an eye on the technological developments, it is highly probable that at some point in time, the data will be considered personal data again.

These challenges pose obstacles to innovation and business by stifling the free flow of information in the guise of ensuring a high level of protection of individuals, while underestimating the equally important facilitative objective of data protection that the permissive theory conveys. It is to the consideration of how this difficulty might be overcome that the next theme in this section now turns.

7. Time Horizon on the Evaluation of the Means Reasonably Likely to be Used

As already pointed out, GDPR establishes 'the means reasonably likely to be used to identify' as a test in determining when a person is identifiable. To ascertain whether means are

reasonably likely to be used, account should be taken of all objective factors. This is followed by the appreciation that a piece of data can be anonymous at the time of collection, but turn into personal information later, just sitting there, simply by virtue of technological progress (Purtova, 2018). Therefore, storage is a crucial element for understanding the concept of identifiability. In other words, more extended data storage increases the likelihood of identification. This is because longer storage provides more opportunities to link the data in question with other publicly available information over time, thereby satisfying the cumulative criteria in the definition of personal data. It is partly for that reason that the principle of storage limitation is a fundamental feature of DPLs. This principle requires that personal data be kept in a form that permits identification of the data subject for no longer than is necessary for the purpose for which it was processed.

However, there is one difficulty with this approach to the identifiability test. The period within which the identification can be done is too open-ended to be helpful to data controllers and processors. There is no time horizon. Data controllers and processors want certainty so they can arrange their affairs to avoid liability. The lack of any time limit on when data can be re-identified or de-anonymized means that it is highly likely that, at some point, the data will be linked to a natural person and thus become personal data again (Sloot, van Schendel, and López, 2022). This is particularly so given the nature of technological developments, such as those in AI. In this regard, some legal scholars and a number of technical experts conclude the following:

> 'It is almost always highly likely that, in 20 years' time, data that are anonymous now can be de-anonymized. Under the current legal regime, when data are stored for that long, such means reasonably likely to be used must be taken into account when determining whether the data protection regime applies, while it is next to impossible to foresee how the technological landscape and the availability of data will evolve in the next 20 years' (Sloot, van Schendel and López, 2022).

In relation to this theme, the findings of this study suggest the need for a fixed point in time in the future at which the re-identification or de-anonymization of data, or the general loss of identifiability, can be assumed to end. The suggestion to narrow down the notion of identifiability by adding a time limit on re-identification is a valid point for enhancing the protection of individuals while facilitating responsible AI innovations. Although absolute anonymity is often unattainable, defining a specific horizon for potential re-identification helps manage data protection risks more effectively. This approach acknowledges that the probability of re-identification can change over time due to evolving technologies and data availability, and it allows for a more practical and nuanced approach to data protection.

The concept of a time limit on identifiability in data de-identification is not a universally standardized calculation but rather a consideration of how long data remains susceptible to re-identification, given technological advancements and the availability of auxiliary information. There is no single, definitive formula for calculating this time limit. Instead, it involves assessing the evolving risk of re-identification over time. Current assessments suggest twenty years to be the maximum period of time within which anonymity might be attainable (Sloot, van Schendel, and López, 2022). Beyond that time, re-identification or de-anonymization of data becomes inevitable. The twenty-year time period starts to run from the moment the data controller or processor receives personal data without identifiers (and they have no right to gain access to those identifiers); or from the moment the data controller or processor anonymizes the data in question.

It is acknowledged that this suggestion might not solve the problem of re-identification, given that more enabling technologies may still facilitate re-identification in the next twenty years. However, it is proposed as a common-sense perspective designed to make it easier for information outside of the definition of personal data to be available for scientific research purposes (such as training AI models), as long as appropriate safeguards are in

place to prevent re-identification of individuals. It will also make it easier for data controllers to anonymize data, facilitating AI development.

V. CONCLUSION

This article sets out and analyzes the identifiability requirement of the notion of personal data in light of technological developments in AI in Kenya and the EU. It highlights interpretations or identifiability tests that could lead to a relatively high identifiability threshold, and hence, contribute to narrowing down personal data to a reasonable dimension that is facilitative of AI innovations. While the meaning of identification can be interpreted narrowly, that should not mean excluding certain potentially harmful practices and their effects from the scope of DPL, or depriving the people affected by them of the protection the legislator intended for them.

Therefore, this paper proposes that, first, instead of using the highly contextual 'all means reasonably likely to be used' test, the interpreter should work with the 'at the price of an excessive effort' test. Second, the notion of identifiability should be narrowed down by adding a specific horizon or time limit on when data can be re-identified or de-anonymized, to wit: twenty years. Third, the definition of 'personal data' should be amended to limit the scope of personal data to circumstances where the controller can themselves identify an individual from the information they are processing, by using reasonable means; or where the controller ought to know that another person will likely obtain the information as a result of the processing and the individual will likely be identifiable by that person by reasonable means at the time of the processing.

REFERENCES

African Declaration on Internet Rights and Freedoms Coalition. (2021). *Privacy and personal data protection in Africa: A rights-based survey of legislation in eight countries*. Retrieved June 10, 2023, from https://africaninternetrights.org.

Allen Waiyaki Gichuhi and Charles Wamae v. Florence Mathenge and Ambrose Waigwa, ODPC Complaint No. 677 of 2022, Final Determination.

Article 29 Data Protection Working Party. (2007, June 20). *Opinion 4/2007 on the concept of personal data.*

Article 29 Working Party. (2014). *Opinion 05/2014 on anonymisation techniques (WP 216) 0829/14/*EN.

Borgesius, F. J. Z. (2016). Singling out people without knowing their names – Behavioral targeting, pseudonymous data, and the new Data Protection Regulation. *Computer Law and Security Review*, 32, 256–271. https://doi.org/10.1016/j.clsr.2015.12.004

Christofi, A., & Verdoodt, V. (2019). *Exploring the essence of the right to data protection in a smart city context: A report in the framework of the SPECTRE research project*. Retrieved July 25, 2022, from https://spectreproject.be/

Clifford Chance. (2022). The UK's data protection and digital information bill: Further reform on the horizon. Retrieved August 28, 2025, from https://www.clifford chance.com

Commission of the European Communities. (1990). COM(90) 314 final - SYN 287 and 288. Brussels, 13 September 1990.

Council of Europe. (1981a). Convention no. 108 for the protection of individuals with regard to automatic processing of personal data, as updated in 2018 by Protocol CETS No. 223 ('Convention 108+').

Council of Europe. (1981b). Convention for the protection of individuals with regard to automatic processing of personal data. Council of Europe, European Treaty Series No. 108.

Dalla Corte, L. (2019). Scoping personal data: Towards a nuanced interpretation of the material scope of EU data protection law. European Journal of Law and Technology, 10(1). https://doi.org/10.5555/ejlt.v10i1.000

Dalla Corte, L. (2020). *Safeguarding data protection in an open data world: On the idea of balancing open data and data protection in the development of the smart city environment* (Ph.D Thesis). Tilburg University, Tilburg.

Data Protection Act (Kenya), Act No. 24 of 2019 (DPA).

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (DPD). *Official Journal of the European Communities*, L 281/31, 24 October 1995.

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons

with regard to the processing of personal data and on the free movement of such data *(General Data Protection Regulation). Official Journal of the European Union*, L 119/1, 4 May 2016.

European Union. (2020). *The impact of the General Data Protection Regulation on artificial intelligence*. Brussels: European Union.

Explanatory Memorandum to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data. (1981). Strasbourg, 28 January 1981.

Finck, M., & Pallas, F. (2020). They who must not be identified: Distinguishing personal from non-personal data under the GDPR. *International Data Privacy Law*, *10*(1), 11-36. https://doi.org/10.1093/idpl/ipz026

Fuster, G. G. (2014). *The emergence of personal data protection as a fundamental right of the EU*. London: Springer.

Fuster, G. G. (2023). Book review of *L'effetivité de la protection des personnes par le droit des données à caractère personnel* (by Suzanne Vergnolle). *European Data Protection Law Review, 9(3),* 383–385.https://doi.org/10.21552/edpl/2023/3/10

Fuster, G. G., & Gurtwirth, S. (2013). Opening up personal data: A conceptual controversy. *Computer Law & Security Review, 29*

Google Spain SL, Good Inc v Agencia Española de Protección de Datos and Mario Costeja González, Case C-131/12 [2014] ECLI:EU:C:2014:317.

Groos, D. & van Veen, B. (2020). Anonymised data and the rule of law. *European Data Protection Law, 4*, 498-508.

https://doi.org/10.1007/978-94-6265-198-2_1

https://doi.org/10.31045/2413-6360.2021.93.5

Information Commissioner's Office. (n.d.). How do we ensure anonymization is effective? Retrieved August 29, 2025, from, https://ico.org.uk

Joshua Kiprop Kisorio v. Safaricom Plc & 4 others; Abdinajib Adan Muhumed (Interested Party) [2021] eKLR.

Korff, D., & Georges, M. (2019). *The DPO handbook: Guidance for data protection officers in the public and quasi-public sectors on how to ensure compliance with the European Union General Data Protection Regulation* (p. 97). Available at <https://ssrn.com/abstract=3428957>.

Kroger, J., et al. (2022). Personal information inference from voice recordings: User awareness and privacy concerns. *Proceedings on Privacy Enhancing Technologies, 2022(1)*

Leenes, R. (2008). Do they know me? Deconstructing identifiability. *University of Ottawa Law & Technology Journal, 4(1),* 135–161. Retrieved April 12, 2023, from <https://pure.uvt.nl/ws/portalfiles/portal/1310856/>.

Lynskey, O. (2023). Complete and effective data protection. *Current Legal Problems, 76,* 297–343. https://doi.org/

McGettigan, C., & Lavan, N. (2017). Human voices are unique but we're not that

good at recognizing them. *Scientific American*. Retrieved from https://www.scientificamerican.com/article/human-voices-are-unique-but-were-not-that-good-at-recognizing-them.

Ministerial Council of the OECD. (1980). Recommendation of the Council concerning guidelines governing the protection of privacy and transborder flows of personal data. *OECD Doc. C(80)58/FINAL (Sept. 23, 1980)*. Retrieved from https://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm.

Nubian Rights Forum &2 others v. Attorney General & 6 others, Child Welfare Society & 9 others (Interested Parties) [2020] eKLR.

OECD. (1980). Ministerial Council of the Organisation for Economic CooperationCooperation & Development, Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, OECD Doc. C(80)58/FINAL (Sept. 23, 1980).

Paal, B. P. (2022). Artificial intelligence as a challenge for data protection law. In S. Voeneky (Ed.), *The Cambridge Handbook of Responsible Artificial Intelligence* . Cambridge: Cambridge University Press. [DOI: 10.1017/9781108657092].

Panvier, J. (2021). When is voice (a special category of) personal data under GDPR? Retrieved from https://www.linkedin.com/pulse/when-voice-special-category-personal-data-under-gdpr-janvier-parewyck?trk=public_profile_article_view.

Patrick Breyer v. Bundesrepublik Deutschland. (2016). Case C-582/14. [ECLI:EU:C:2016:779]. *European Court of Justice*. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX-%3A62014CJ0582>.

Peter Nowak v. Data Protection Commissioner, Case C-434/16, ECLI:EU:C:2017.

Poscher, R. (2022). Artificial intelligence and the right to data protection. In S. Voeneky (Ed.), *The Cambridge Handbook of Responsible Artificial Intelligence*. Cambridge: Cambridge University Press.

Purtova, N. (2018). The law of everything: Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology, 10*(1), 40. https://doi.org/10.1080/17579961.2018.1452176

Purtova, N. (2022). From knowing by name to targeting: The meaning of identification under the GDPR. *International Data Privacy Law, 12*(3), 163-183. https://doi.org/10.1093/idpl/ipac007

R (on the application of Edward Bridges) v. The Chief Constable of South Wales Police and Secretary of State for the Home Department (2019).

Rossi, J. (2020). *Data protection and right to privacy: Investigating the contested notion of personal data* (English summary). [PhD Thesis, Université de Technologie de Compiègne].

S v. Ismail & Others [2004] ZAWCHC 39.

Scarlet v. SABAM. (2011). Case C-70/10. [ECLI:EU:C:2011:771].

Shakunt Rajnikant Shah v. Bhupendra Motichand Shah t/a John Cumming & Company & another, Petition no. 14 of 2020, [2021] eKLR.

Sloot, B. (2017). Legal fundamentalism: Is data protection really a fundamental right? In Leenes, R., van Brakel, R., Gutwirth, S., & De Hert, P. (Eds.), *Data protection and privacy: (In)visibilities and infrastructure* (pp. 3–30). Springer.

Sloot, B., van Schendel, S., & López, C. A. F. (2022). *The influence of (technical) developments on the concept of personal data in relation to the GDPR*. Tilburg Institute for Law, Technology and Society of Tilburg University.

Tetiana, L. (2021). The concept of personal data: From academic perspective to practical implications. *Young Scientist, 5*(93), 111–112.

Vergnolle, S. (2022). *L'effectivité de la protection des personnes par le droit des données à caractère personnel*. Bruylant.