IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 August 2023, accepted 22 September 2023, date of publication 2 October 2023, date of current version 10 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321578

RESEARCH ARTICLE

Using Differential Privacy to Define Personal, Anonymous, and Pseudonymous Data

TAO HUANG[1] AND SHUYUAN ZHENG[2], (Member, IEEE)
[1]School of Law, City University of Hong Kong, Hong Kong, China
[2]Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan

Corresponding author: Shuyuan Zheng (zheng@ist.osaka-u.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science KAKENHI (JSPS KAKENHI) under Grant 21J23090 and Grant 21K19767, and in part by the Japan Science and Technology Agency SICORP (JST SICORP) under Grant JPMJSC2107.

ABSTRACT Defining personal, anonymous, and pseudonymous data is a vital issue for data protection law. Current approaches adopted by legal regimes are either too absolute to be practical or too vague to be manageable. Differential privacy (DP), as a newly emergent technical tool, can help define the different categories of data by quantifiably measuring identification risks of databases. Through the selection of a privacy budget in advance, data controllers can delineate the boundaries among personal, anonymous, and pseudonymous data in an auditable and reviewable manner, as well as incorporate these definitions into the broader practice of data risk management. This article offers concrete steps for applying this approach in practice and argues that such an approach not only enhances certainty, consistency, and transparency, but also inspires a new model of interaction between law and technology. Recognizing that this approach is not perfect, the article then discusses some challenges and directions for future research.

INDEX TERMS Anonymization, differential privacy, personal data, pseudonymization.

I. INTRODUCTION

Personal, anonymous, and pseudonymous data are three common concepts of data adopted by data protection regimes in the world. This tripartite classification plays an important role in the regulatory system of data protection: for data controllers or processors, it determines their compliance duties and burdens; for data subjects, it influences what rights they will enjoy; for government officials who enforce data protection laws, it prescribes the standards of enforcement they will need to follow. However, the dividing lines that separate the three categories of data are either impractical or too ambiguous. The legal mandate that builds anonymization upon zero risk of reidentification is based on an illusory conception of data risk and its relationship with data utility. Recently issued guidelines from the EU and the UK have rightly emphasized the contextual nature of anonymization and pseudonymization. However, by delegating the determination of data categories to the prudential considerations of data controllers, these guidelines leave them with few processes, standards, or institutions to guide their practices. At best, they have listed some contextual factors for data controllers' reference; but the endeavor to define data categories is still far more of subjective judgments than institutionalized decision-making with comparable and transparent standards. The result, lamentable therefore, would be uncertainty, inconsistency, and opacity.

The stakes are high here. Due to the scale of data collection and processing in the digital economy, a clear sense of which sets of data fall into which category is vital for data controllers and processors to effectively plan their compliance costs, and for law enforcement agencies to conduct their oversight in a consistent and efficient manner. The relevant legal rules currently in force have failed to provide a sufficiently predictable and objectively auditable benchmark for the various parties to arrange actions accordingly. The lamentable result is that both legal certainty and technological efficiency would suffer.

Differential privacy (DP) can help resolve this thorny problem. DP is not only a specific technique for database perturbation, as recognized by several official guidelines of various jurisdictions [1], [2]; it is, more importantly, an objective and mathematical standard of defining and measuring privacy. Although DP was initially proposed to

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed.

survey tabular data, after nearly 20 years of development, researchers have adapted it for various scenarios, covering much more complex human data, such as browsing histories [3], human trajectories [4], biometric identifiers [5], artificial intelligence models [6], and high-dimensional crowdsourced data [7], [8]. Additionally, researchers have even proposed two privacy definition frameworks, called Pufferfish Privacy [9] and Blowfish Privacy [10], which can help data processors easily tailor DP to their specific data processing scenarios. Nowadays, DP has become a de facto standard for data privacy protection in the computer science community and has been applied by giant companies such as Apple [11] and Google [12] in their real-world products and services.

The wide applicability of DP, plus its characteristics of robustness, compositionality, and auditability, has enabled such a privacy standard to become not only a specific technique of privacy protection and anonymization but also a criterion that can be incorporated into the institutional practices of information risk management, helping us delineate the boundaries among personal, anonymous, and pseudonymous data. This is the case because the key distinction that separates these three categories of data is the degree of identification risk, which is precisely what DP aims to address and calibrate. By appropriately selecting the privacy budget under DP, we can predefine the thresholds for anonymization and pseudonymization and manage our data practices accordingly. This approach synchronizes data categorization with processing, and incorporates them into the holistic practice of data management. It is an expansive and innovative example of data protection by design: not only can technology help us achieve legal and regulatory goals, but also it can facilitate the very definition and understanding of key concepts in law, such as personal, anonymous and pseudonymous data. The mechanism of defining data categories proposed by this article, which is objective, proactive, and cooperative, also offers insights into reforming the general framework of data protection and governance.

This article is structured as follows. Section II introduces the issue of defining data categories: why it is important and what the current approaches are for dealing with this issue. This part argues that existing approaches are undesirable because they are either too absolute to be practical or too vague to be manageable. Section III gives a brief introduction to the notion of differential privacy, as well as some of its advantages compared to other techniques. Section IV argues why this mathematical standard of privacy could help us define personal, anonymous, and pseudonymous data. Section V proposes a procedure consisting of six steps for using DP to define categories of data. Such an approach for quantifying privacy risks and classifying information, as illustrated in Section VI, is consistent with the normative goal of information privacy and can be used as an important mechanism for information risk management. Section VII lists some possible difficulties for adopting the new approach; in the meantime, some preliminary responses and directions for future research are explored. The final section offers concluding remarks.

II. BACKGROUND

Major data protection regimes in the world have focused their legal and regulatory frameworks on personal data or personal information while leaving non-personal or anonymous data outside the binding scope of the law.[1] The issue of whether data are personal or anonymous is thus vital for all parties: for data controllers and processors, it determines whether their data practices have to conform to the obligations prescribed by data protection laws; for law enforcement agencies, similarly, they are not required to check the compliance of anonymous data against legal rules; for data subjects, it influences their rights significantly, as they cannot claim the rights to data rectification, portability, and erasure as long as the data remain anonymous.[2] Because of the reduced risk brought by anonymization to both controllers and subjects, ''[a]nonymising data wherever possible is therefore encouraged'' by the Information Commissioner's Office (ICO) of the UK [13].

However, defining the boundary between personal and anonymous data, or the threshold of successful anonymization, is not easy. There are generally two approaches adopted by regulatory regimes. One is the absolutist approach. China, for example, has described anonymization as an ''irreversible process of making personal information unable to be used to identify specific natural persons''.[3] India has similarly employed stringent terms like ''irreversible'' and ''cannot'' to define anonymization.[4]

The problem with this absolutist criterion is that it is simply impossible to meet. The absolutist approach assumes that the reidentification risk can be reduced to zero. In practice, nonetheless, the risk cannot be eliminated completely because of the ''auxiliary information problem'': any piece of data can be linked, combined, or matched with other data to identify individuals; we can never know the content, quantity and location of such ''auxiliary information'' [14]. Several high-profile incidents, such as the AOL case, the case of the Massachusetts governor, and the case of Netflix prize, are examples of how the release of seemingly harmless data could be reidentified by linking them with some additional information [15]. The laws adopting the absolutist approach have failed to recognize the interconnectedness of data [16]. Such failure is not trivial, since ''a command that cannot be obeyed serves no end but confusion, fear, and chaos''. After all, data protection laws are not solely for the protection of individuals' data privacy, but aim to strike a balance between

[1] See Recital 26 of the EU's General Data Protection Regulation (GDPR), Article 4 of China's Personal Information Protection Law, Article 2(3) of India's Personal Data Protection Bill (2018).

[2] See Article 11(2) of EU's GDPR.

[3] See Article 73 of China's Personal Information Protection Law.

[4] See Article 3 of India's Personal Data Protection Bill (2018).

data privacy and data use or flow.[5] In this sense, if laws really mean to offer anonymization as a channel for data controllers to minimize the risk and relieve their burden of compliance, then the threshold should not be set at an absolute level. Doing so would just be too costly for controllers to implement anonymization in practice [17], [18].

Another approach of anonymization is more realistic, requiring not zero risk, but a risk small enough to be immune to any reasonable efforts of reidentification. We may call this the reasonableness approach. It has been endorsed by the EU, the UK, and the US. Recital 26 of the EU's GDPR, for example, explicitly states that to determine the boundary between personal and anonymous data, ''account should be taken of all the means reasonably likely to be used''.[6] The Article 29 Working Party's opinion also stressed that if the possibility of identification ''does not exist or is negligible'', then the data are anonymous [19]. The UK ICO's new guidance similarly states that ''a sufficiently remote level'' or identification risk is enough [20]; in other words, ''data protection law does not require anonymisation to be completely risk-free'' [20]. As for the US, even though there is no comprehensive data protection law at the federal level, several state laws have touched on this matter and adopted such a reasonableness approach.[7] This approach correctly avoids the trap of impracticability of the absolutist approach, but introduces the new problem of ambiguity. ''Reasonable'', a word that is so familiar to lawyers, is itself vague, amorphous, and easily manipulable.

This issue of ambiguity also hovers over the notion of pseudonymization. Pseudonymization is a process of replacing the identifiers (information that can identify an individual) with pseudonyms and then keeping the identifiers in a separate place with technical and organizational safeguards [21], the aim of which is to guarantee that the data cannot be used to identify individuals without the use of the separately kept identifiers.[8] Because pseudonymization can reduce data risk and help data controllers meet partial compliance duties under data protection laws, it has been encouraged and recommended by many data protection regimes. For example, Articles 25 and 32 of GDPR have recommended the use of pseudonymization to data controllers for the purpose of reducing data processing risks. Under the Personal Information Protection Law of China, deidentification (an equivalent of pseudonymization) is one of the recommended technical measures for data controllers to fulfill their compliance duty (Article 51). The UK ICO's guidance [20] is more detailed on this issue by listing several benefits of pseudonymization. However, for data controllers who consider pseudonymization as an option for compliance, puzzles abound: What are the identifiers? Can we predefine them accurately? What technical and organization measures are needed to keep the identifiers separate and secure? If pseudonymization is to partially reduce the risks of data instead of removing them entirely [13], [19], [21], then what level of risk reduction should be achieved?

Some recently issued opinions have tried to provide more detailed and operable guidance to the practice of anonymization and pseudonymization. The Article 29 Working Party, now replaced by the European Data Protection Board, specified three kinds of risks that anonymization should address: singling out, linkability, and inference [1]. The UK ICO has adopted a ''motivated intruder's test'' to delimit the scopes and types of potential attackers [22], an effort to make the anonymization practice more targeted. It also distinguishes limited access and public release of data and advises controllers to use different technical and organizational measures accordingly [23]. These guidelines, however, are far from enough. On the one hand, there exist tensions or contradictions among the regulatory texts, advisory opinions, and case laws regarding this matter. For example, some have argued that the Recital 26 of GDPR and one Article 29 Working Party's Opinion (WP 216) represent two contrasting approaches of anonymization [24]. And the Article 29 Working Party's interpretation of identification might be in tension with the CJEU's ruling in Breyer [25]. On the other hand, even though some tests or factors contained in the guidelines can reduce uncertainties to some extent, they do not offer measurable and reviewable, not to mention quantifiable, criteria for the practice. Without such measurable and quantifiable standards, ''frameworks would be just fancy documents sitting in cupboards with data scientists ungeared to implement the nice guidelines and best practices'' [26].

The academic scholarship has also noted the delicacies of different data categories and tried to offer more refined means of delineating them. For example, Polonetsky et al. [27] have borrowed concepts from chromatics to illustrate that reidentification risk of data is not simply black and white, but lies on a spectrum of shades of gray; in this spectrum, they named various categories of data: explicitly personal data, potentially identifiable and not readily identifiable data, pseudonymous and protected pseudonymous data, anonymous and aggregated anonymous data, among others. Khoury [28] has proposed to abandon the categorizations of data and viewing the status of data as comparable to the quantum superposition-constantly in flux and change. These proposals are undesirable because the more delicate categories require more interpretations, not less, and interpretations call for consistency and standards, which we do not have yet. Besides, dividing data into more categories without prescribing any measurable standards will further increase the stakes of subjective judgments, which are hard to predict and supervise.

Table 1 briefly summarizes the current approaches for defining data categories and their limits. The determination of which category data belongs to is a vital one, since it

[5]See Article 1 of the EU's GDPR; see also Article 1 of China's Personal Information Protection Law.

[6]See Recital 26 of the EU's GDPR.

[7]California Consumer Privacy Act (2020), 1798.140; Virginia Consumer Data Protection Act, § 59.1-575.

[8]See Article 4 of the EU's GDPR.

TABLE 1. Existing approaches for defining data categories.

Approach	Feature	Advantage	Disadvantage
Absolutist (China, India)	Requiring zero risk for anonymous data.	Easy to understand.	Impossible to implement (because reidentification risks cannot be eradicated).
Reasonableness (EU, UK, US)	Reasonableness-aware, not absolutely risk-free.	Possible to implement.	"Reasonable" is an ambiguous word.
Polonetsky et al. [27], Khoury [28]	More categories offered.	Delicate categorization.	More interpretations required, inviting greater subjectivity and inconsistency.

influences billions of compliance costs of data controllers as well as the rights and interests of countless data subjects. Such a determination is also very complex since it involves the management of massive amounts of data, often stored in databases of various kinds and formats. If we leave these decisions solely to the subjective judgments of data controllers, praying that they will exercise their best discretion with prudence, the result would be either that controllers will choose not to conduct anonymization or pseudonymization of data due to uncertainties, or that the industrial practices will be arbitrary, inconsistent, and chaotic. In the latter scenario, the formidable power exerted by data controllers and processors through the collection and computation of mass amount of data will be aggrandized by the unchecked discretion granted to them on defining data categories. A measurable, reviewable, and transparent criterion is, therefore, crucial to ensure that the categorization of data is efficient, consistent, and accountable.

So far, all the endeavors, no matter proposals from scholars or guidelines from government agencies, have paid inadequate attention to the role of technology. Even though Schwartz and Solove have proclaimed a decade ago that ''[t]he line between PII [personally identifiable information] and non-PII is not fixed, but depends upon technology'' [29], they and other researchers have not seriously considered the option of incorporating technology into the legal definitions of data. In an area that is dominated by rapid technological change, an exclusive focus on the legal without a glimpse at the technical tools would result in laws that ''frequently create substantial uncertainty for implementation, provide contradictory recommendations in important cases, disagree with current scientific technical understanding, and fail to scale to the rapid pace of technological development'' [30]. To sufficiently clarify and understand the concepts of anonymization and pseudonymization, we need not only legal norms but also technological standards. Differential privacy, as argued by the rest of this article, can serve as one such standard.

III. WHAT IS DIFFERENTIAL PRIVACY

Differential privacy (DP) [31] is a celebrated privacy concept that has garnered widespread interest in recent years. Unlike what some may assume, it is more of a mathematical definition and technical standard of privacy than a specific algorithm or technique of privacy protection [32], [33]. Formally, it can be defined as follows.

Definition 1 (ϵ-Differential Privacy [31]): Consider a data processing mechanism $\mathcal{M}$; for any two neighboring databases D and D' that differ only in one data record, if $\mathcal{M}$ yields outputs in accordance with the following condition, then $\mathcal{M}$ satisfies ϵ-differential privacy:

$$\frac{Pr[\mathcal{M}(D)]}{Pr[\mathcal{M}(D')]} \leq \exp(\epsilon)$$

DP requires that the presence or absence of any individual record in a database not significantly influence the output of the processing of the database. By perturbing the data through noise addition using algorithms such as the Laplace mechanism [31], DP prevents potential adversaries from determining whether an individual's data are in database D or D'. In other words, the goal of DP is to ensure that, when adversaries observe the output of a database, they cannot infer whether an individual is included in the database.

According to the definition, when processing the database, the impact of replacing or deleting one single record on the outcome is measured by $\frac{Pr[\mathcal{M}(D)]}{Pr[\mathcal{M}(D')]}$, which should be bounded by $\exp(\epsilon)$. In this case, the parameter ϵ measures every single data record's influence on the outcome of the processing mechanism. The smaller the ϵ is, the stronger the mechanism will be; this means that the mechanism sufficiently ''hides'' the presence of the individual record among other records in the database. By contrast, a larger value of ϵ means that the mechanism offers lesser protection. In the DP literature, ϵ is called privacy budget, or more precisely, the budget of privacy loss. It measures the maximum loss or risk of privacy that we could accept for a data processing mechanism.

As a technical tool for reducing and measuring data privacy risks, DP can used in a variety of occasions, such as inquiries on statistical databases, machine learning, data collection, and data synthesis [34]. In recent years, it has gained momentum in research and industry, triggering many new applications, such as combining it with other techniques like federated learning or multi-party secure computation [35]. Before elaborating on why DP could be used to define personal, anonymous, and pseudonymous data, four relevant characteristics of DP should be emphasized here: 1) Wide applicability: DP can be applied to various scenarios and problems. As a general privacy framework, DP does not build upon any assumptions about the types of personal data and data processing algorithms. Therefore, it has extremely broad applicability and can meet diverse demands in terms of personal data protection. 2) Robustness: DP also does not rely on any assumption of the background

knowledge of adversaries; actually, it resolves the auxiliary information problem by assuming that potential adversaries could have access to all the background information of the data subject. This guarantee against auxiliary information makes the protection of DP withstand future attacks [36]. 3) Compositionality: under DP, the privacy loss of any data processing mechanism is compositional; that means, if two processing mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$ cost privacy budgets $\epsilon(\mathcal{M}_1)$ and $\epsilon(\mathcal{M}_2)$, respectively, then a mechanism combining $\mathcal{M}_1$ and $\mathcal{M}_2$ will cost a budget of $\epsilon(\mathcal{M}_1 + \mathcal{M}_2)$ [37]. 4) Auditability: because the privacy loss of a differentially private mechanism is measurable by the parameter ϵ, and this parameter can be added, deducted, and distributed, we can then audit the privacy risk of a series of data processing practices in a systematic way [38], [39].

IV. WHY DIFFERENTIAL PRIVACY CAN BE USED TO DEFINE CATEGORIES OF DATA

This part argues that DP could be used as a standard for defining personal, anonymous, and pseudonymous data. By pre-selecting different privacy parameters ϵ and conducting noise addition to databases, we can group data into different categories on the basis of the corresponding risks. To understand why this approach works, let us take a closer look at how ϵ relates to the definitions of data.

First, the differences among personal, anonymous, and pseudonymous data are not ontological and fixed but are matters of contextual degree. What separates them is the degree of identifiability, or more precisely, the degree of identification risk [40]. The absolutist approach of anonymization wrongly assumes that such risk can be eliminated once and for all. Such a standard is theoretically possible, but impractical because it renders the data useless. The reasonableness approach is correct in acknowledging the relative and contextual nature of identification risk, but it relies merely on subjective judgments to measure the risk. Likewise, the dominant method for defining pseudonymization gauges the risk according to a predefined list of ''identifiers'', which is inevitably arbitrary. By contrast, differential privacy is an objective standard for measuring identification risks. DP hides the presence of an individual by reducing its impact on the output of the database, and such impact is precisely quantified by the parameter ϵ. This parameter measures exactly the identification risk because to identify, as the Article 29 Working Party correctly interprets, means that ''within a group of persons, he or she is 'distinguished' from all other members of the group'' [19]. Anonymity, as its antithesis, means to hide an individual in a crowd [41]. From the outside attacker's perspective, the higher the probability that the attacker could infer about the presence of an individual, the more visible the individual identity will be in the group (database). In other words, to identify is to unmask one's presence in a group-to make him or her visible, distinguishable, and special-the probability of which is quantified by the DP parameter.

Second, the identifiability protected by data protection laws is individuals' unique personal profile that has gradually formed and been elaborated upon in the digital world. Conversely, it is the epistemological leak that the law guards against. If we view privacy from the negative perspective, it can be described as non-identification or anti-identification. Identification, it should be emphasized, is a process rather than a result or state. The rationale lying behind data protection laws' encouragement of anonymization and pseudonymization is to protect the secrecy and integrity of individuals' identities [42]. Mass collection of personal data, automated processing and mining, and the greater power inequality generated by data economy all lead to the moral and legal concerns regarding the protection of personal identities from data collection and processing mechanisms. To protect identity in this process is to protect the individual from being discovered against his or her will. In this sense, identification is a process of increasing the knowledge about the individual, a process of his or her digital profile getting clearer and more detailed in others' eyes. Scholars [43] have suggested that privacy loss should be characterized as the change of belief of the adversary after data publishing. Identification without consent or other legal grounds is what legal regulations aim to prevent. The idea of differential privacy addresses exactly this concern because the very goal of this technique is to allow knowledge to be learnt from a community (population) without letting any knowledge to be inferred about every single individual [32]. This combination of group data utility and individual data privacy perfectly matches with the legislators' goals of protecting individual data subject's identity (by encouraging anonymization and pseudonymization) while at the same time accommodating the growing needs of data processing (and encouraging the data economy that builds upon it).

Third, because data protection laws aim to provide robust protection for personal data without compromising too much data utility in the meantime, a viable standard of defining data must be able to achieve and strike such balance. DP is such a technique. It recognizes that the tradeoff between data privacy and utility is unavoidable in practice [44]. According to the definition of DP, the balance is achieved by selecting an appropriate value for ϵ. The parameter ϵ not only measures this tradeoff, but also allows controllers and the public to choose the exact level of tradeoff before the data processing practice: note that ϵ is selected before data are computed or published. This feature makes the management and planning of data risks more convenient. Through institutional design, as will be discussed later in Section VI, the parameter could be made publicly available for comment to enhance transparency, accountability, and flexibility.

Therefore, differential privacy, with its parameter to measure identification risks and the tradeoff between data utility and privacy, can be used as an objective standard to define personal, anonymous, and pseudonymous data. Specifically, in the spectrum of identification risks, two thresholds delineate these three categories of data. Data

TABLE 2. An example that shows how DP can be used to define categories of data.

Value of ϵ	Iden. risk	Data category	Legal implication
$\epsilon \leq 1$	Too small, negligible	Anonymous data	Exempt from data protection laws.
$1 < \epsilon \leq 10$	Small	Pseudonymous data	Waived from some duties of data protection laws.
$\epsilon > 10$	Huge	(Explicitly) personal data	Fully covered by duties of data protection laws.

processors or publishers can select two values of ϵ and then use the standard of DP to make sure their data processing or publishing ϵ-differentially private. For example, as shown in Table 2, we can decide, by law or by consensus, that for a specific data processing or publishing context, if the privacy risk parameter ϵ is less than or equal to 1, the identification risk is negligible and acceptable, then the data can be defined as anonymous data, falling outside of the scope of data protection laws; if ϵ is between 1 and 10, the identification risk should be addressed but is relatively low, then the data are pseudonymous: on this occasion, it meets partial requirements of data protection laws, but is still personal data that should be protected by some technical and organizational measures; if ϵ exceeds 10, the identification risk is high, then the data is explicitly personal data, having to comply with all the compliance duties that the law assigns to the processing of personal data.

To be sure, the boundary values 1 and 10 are arbitrarily selected as examples for illustrating this point. How to set the values of ϵ is a difficult issue, as will be discussed in Section VII. In any case, through this approach, we can use DP as an adjustable, quantifiable, and objective standard to define personal, anonymous, and pseudonymous data, replacing the traditional definitional approach, which is binary, subjective, and ambiguous. What's more, this approach can be developed into a comprehensive mechanism of information risk governance that is more accountable, more transparent, and more inclusive. Before exploring this idea, a general framework of using DP to define data will be introduced in the next section.

V. PROPOSED STEPS FOR USING DP TO DEFINE DATA

In this section, we propose six steps that use DP to define personal, anonymous, or pseudonymous data, as summarized in Figure 1.

A. STEP 1: CHECKING THE APPLICABILITY OF DP

The first step of using DP to categorize personal, anonymous, and pseudonymous data is to check whether DP is suitable for the context. Generally, there are three conditions for applying DP: the object of processing or publishing is a database or dataset [45], the data records in the database cannot be too few (otherwise, it will need too much noise to mask the presence of individual records in the group) [46], and accurately targeting or identifying individuals

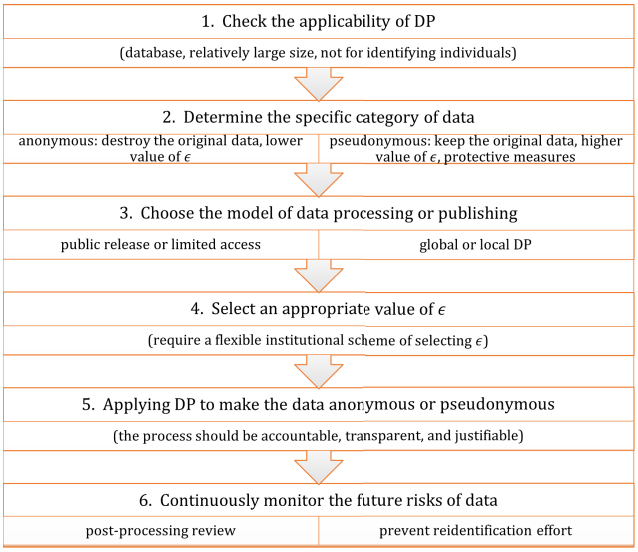

FIGURE 1. Six steps for using DP to define data categories.

is not necessarily required (counterexamples may include scenarios like crime forecasting [47] and terrorist financing tracking [48]).

DP is most suitable for circumstances where the goal is to seek statistical distributions or overall trajectories of the database, or to construct group profiles in order to describe features of a group rather than any single individuals. UK ICO's guidance echoes this point by arguing that differential privacy is "useful in the context of statistical analysis and broad trends, rather than for detecting anomalies or detailed patterns within data" [2]. Even though the three conditions may seem to pose a high threshold for applying DP, there are many contexts that fit with DP in practice, such as the government administrative database, scientific research statistics database, and corporate business database-as long as these databases are large and are not collected for the very purpose of identifying individuals.

B. STEP 2: DETERMINING THE SPECIFIC CATEGORY OF DATA

The second step is to determine whether the goal of applying DP is to make the data anonymous or pseudonymous. This determination makes two differences. First, if the aim is to anonymize data, then data controllers and processors shall not keep the original data after applying the standard of DP to process the database. That is because if the original data has not been destroyed or deleted permanently, then there will still be substantial risk for the DP-perturbated data to be reidentified-a risk that will be actualized if the original data is breached or disclosed in the future. If the aim is to pseudonymize data, by contrast, then the original data can be kept as long as proper technical and organizational safeguards are provided-as required by the very definition of pseudonymization. Second, anonymous data requires stronger protection and a lower value of ϵ than

do pseudonymous data. The cost accompanied by the stronger protection, though, is reduced data utility.

C. STEP 3: CHOOSING THE MODEL OF DATA PROCESSING OR PUBLISHING

Different models of data processing or publishing have different implications for privacy risk. Here, at least two choices have to be made. First, whether the data will be released to the general public or to an enclosed group with limited access: if the former, the data will need stronger protection by applying a smaller ϵ on its processing; if the latter, while we can choose a higher value of ϵ to achieve more data utility, some technical or organizational measures should be taken to enforce the access limitation [21], [49]. Second, whether there exists a trusted central processor of data who is responsible for conducting the noise-addition process under DP: if yes, the classic or ''global'' DP will be used; otherwise, we may consider the local DP model [50] in which all data records are perturbed under DP in each subject's local side before sending to the processor—in this circumstance, even though the risk from the untrusted centralized processor can be avoided, the noise that needs to be added to the data is greater than that in the centralized DP, resulting in more loss of data utility [51].

D. STEP 4: SELECTING AN APPROPRIATE VALUE OF ϵ

Before processing data under the standard of DP, a proper value of the privacy budget ϵ should be selected. As argued before, the value directly determines the identification risk and the tradeoff between data utility and privacy. It is not an easy decision though, as admitted by the Article 29 Working Party [1]. Researchers have proposed various algorithms or methods that help controllers choose the value [52], [53], [54]. It should be noted that as the value of ϵ must be adjusted to specific data contexts horizontally and the future development of technologies vertically, the mechanism of selecting ϵ should be flexible enough to accommodate these changes [55]. The Article 29 Working Party of the EU has listed some contextual factors to be considered in picking the value of ϵ, including the security measures of the database, the size of the database, the availability of public information, the scope of data release, the sensitivity of the data, and the knowledge/motive/capability of the potential attackers [1]. However, as will be argued in Section VII, institutional design is indispensable, requiring joint efforts by technical and legal experts.

E. STEP 5: APPLYING DP TO MAKE THE DATA ANONYMOUS OR PSEUDONYMOUS

The fifth step is to process the data according to the selected value of ϵ to make the data anonymous or pseudonymous. The reason for choosing ϵ, the algorithm for noise addition, as well as other parameters and management procedures in the process should be made public, or at least, disclosed to other data processors in the industry and the reviewing agencies to ensure accountability. This is vital not only for public oversight, but also for mutual learning and knowledge transfer among different parties [56]. The processors should rest assured because algorithmic disclosure under DP will not compromise the level of privacy protection [57].

F. STEP 6: CONTINUOUSLY MONITORING THE FUTURE RISKS OF DATA

As new technologies may emerge, new information may be generated, and new risks will appear, we should give up the old paradigm of ''release and forget'' which leaves the data unmanaged after publishing, even though the data are pseudonymous or anonymous at that moment. Rather, controllers and processors must continuously monitor the subsequent use and flow of the data and offer protection through appropriate technical or legal measures [58]. As the nature of anonymity is a contextual issue determined by the data environment [59], the environment must be regularly checked to assess the status of data on an ongoing basis. That's why Ohm warned us that ''[t]echnology cannot save the day, and regulation must play a role'' [15]. On the one hand, data processors or publishers should periodically review the risk of the data they have processed or published, to make sure that the privacy budget is not used up and that risks will not substantially increase in the future [22]; if the privacy budget is exhausted or new risks emerge, there should be a mechanism that alerts the processor or publisher [1]. On the other hand, the duties of relevant parties should be confirmed and supervised to make sure that they will not reidentify individuals based on the released anonymous or pseudonymous data—such duties can be specified on a legal, contractual, or fiduciary basis [60].

VI. TOWARD A NEW MODE OF DATA PROTECTION AND MANAGEMENT

Using differential privacy to define personal, anonymous and pseudonymous data is not merely a new application of technical tools in the field of legal regulation. Indeed, it has far broader repercussions and implications. Such an application implies the possibility of ushering in a new era of information privacy protection characterized by an objective definition of privacy rather than subjective descriptions, a cooperative and open mode of information governance rather than a segmented and closed one, and a more profound understanding of data protection by design (DPbD) as well as the mutual engagement between law and technology.

First, the approach of defining categories of data using DP replaces the subjective notion of privacy with an objective and quantifiable one. In the legal literature, privacy has been defined and described as subjective feeling or psychological perception, in the forms of unwanted breach or observation [61], or loss of control over one's own information [62]. However, these subjective descriptions are hard to gauge in practice, due to the fact that privacy harms are often psychological rather than material, invisible rather than

visible, and intuitive rather than calculable. To be sure, what law protects are values and interests, which are not always convertible into numerical values. However, lawyers should also refrain from leaving every key decision to their discretion and from being blind to technological tools in other areas [63]. In the sphere of information privacy and personal data protection, this is especially the case because technical affordances and constraints have been the major forces driving the evolution of privacy and data protection in history [64], [65], [66], [67]. In this sense, legal regulators should be open to the possibility of incorporating technical and mathematical tools into legal definitions.

DP, as a mathematical standard, can help us measure privacy risk quantifiably and planned its budget in advance: this new approach transforms the subjective, reactive, and backward-looking definition of privacy into an objective, proactive, and future-oriented standard. Moreover, as a technical tool, ''differential privacy is unique in enabling data subjects and other parties to assess the relative quality of . . . a firm's privacy practices prior to purchase or participation, permitting an informed decision'' [56]. By adopting the standard of DP, data controllers can decide how and to what extent privacy should be protected before processing or publishing the data: on the one hand, they can plan, adjust, and distribute the privacy budget according to their specific needs or the regulatory mandates; on the other hand, they can also consider which datasets need to be anonymized or pseudonymized, what the corresponding risks are, and how much computational cost should be invested. All these conveniences and efficiencies are brought about by the objectivity and quantifiability of DP. The planning of privacy budget and the definition of data categories can thus be incorporated into a general framework of data management, facilitating greater adaptability and flexibility in governance.

Second, a unified definition based on DP, which quantifiably delineates the boundaries among personal, anonymous, and pseudonymous data, would promote consistency across different regulatory contexts, legal systems, and jurisdictions, thereby facilitating data flow. Different laws and regulations may use different phrases, methods, or standards to define the content and scope of personal data. Schwartz et al. [29] have famously summarized that there exist three modes of defining personal data in the legal systems of the world: the tautological approach that defines personal data as personally identifiable data, the non-public approach that excludes public data from the category of personal data, and the specific-types approach that defines personal data by listing typical examples. In reality, the definitions of data are more amorphous and they sometimes incorporate different approaches in one legal system, causing confusion and ambiguity. For example, China's the Personal Information Protection Law (in article 4) defines personal information as information relating to any identified or identifiable natural persons, excluding anonymous information. By contrast, the Civil Code (article 1034) and Cybersecurity Law of China (article 76) do not explicitly exclude anonymous data from their application. In addition, China has issued a national standard called ''Guide for De-identifying Personal Information'', in which the definition of personal information is broader than the laws mentioned above, including not only information that relates to individuals' identity but also information that reflects individuals' activities. In the meantime, while it is common for national data protection laws to explicitly exclude anonymous data from their binding scope, the standards of anonymization they endorse are divergent and uncertain [68]. These inconsistencies and uncertainties are disastrous for data flow across systems and borders, since data controllers have to design different compliance schemes to meet the different requirements of anonymization or pseudonymization, as well as invest tremendous costs to reprocess the data before transferring them to another system or jurisdiction.

A DP-based definition of data facilitates the smooth flow of data by managing risks through a uniform parameter: controllers will only need to adjust their privacy budget ϵ before data transfer instead of delving into the esoteric legal texts, translating them into technical measures, and transforming one technical parameter into another accordingly. In addition to compliance, the uniform definition based on DP can also make the data risk governance more cooperative and transparent. This is because the quantifiable standard of privacy budget allows comparisons across platforms, systems, and even jurisdictions [56]. By using a common parameter ϵ, different controllers and platforms can compare their privacy budget on the one hand, and regulators can get a clearer view of the level of data protection on the other [56]. Additionally, as the algorithm of DP is relatively independent of its computing data, controllers can share the DP algorithm with others without compromising the security and privacy of the data [33]. These features and applications facilitate collaborative governance and mutual learning among regulators, controllers, and data subjects: all parties can collaboratively participate in the management of data privacy, share their rationales of choosing specific models, parameters, and algorithms, and learn from one another in the process. A more participatory and transparent governance framework will then emerge from adopting the DP-based approach.

Third, the DP-based approach of defining categories of data stimulates us to rethink the relationship between law and technology and to explore more thoroughly the meaning of data protection by design (DPbD). In fact, the approach is not only about data definition, but also about data management: it does not define data in the traditional legal manner by labeling what already exists in the world with certain concepts or ''tags''; rather, it processes data according to the pre-selected category and standard. In other words, personal, anonymous, and pseudonymous data are not out there, but what we choose them to be. Unlike old definitional approaches, which only reactively delineate what technical practices happen

to produce, this new approach proactively prescribes the technical practices by combining definition of data with processing and management of data. Definition here is part of the holistic management mechanism: to define data means to process the data according to the practical and regulatory requirements. Remember that anonymization and pseudonymization are themselves data processing practices. In this way, the new approach combines the defining, processing, and managing of data according to our desirable levels of risks and utilities in certain contexts.

Such a holistic framework implicates a direction for developing the currently trending Data Protection by Design (DPbD).[9] DPbD expands the traditional relationship between law and technology by reminding us that not only the law can guide the development of technology, but technology can also facilitate the implementation of law by incorporating legal rules and values into the technical design of products. This mode of thinking is still compliance-based, though, since what technical experts do under this paradigm is to comply with what legal regulators mandate [36]. The DP-based approach goes further: it tells us that technology can reshape and reformulate the legal landscape as well. By quantifiably measuring the definitional thresholds of legal terms, technical standards can revolutionize lawyers' thinking about data, privacy, and risk. What technology can do is not only comply with the law but critically reform the legal definition, regulation, and governance. Legal definitions then evolve into techno-legal frameworks that encompass a broader scenario of managing data by both legal and technical means. Such incorporation provides insights for the DPdD enterprise by opening doors for greater cross-fertilization between law and technology.

VII. CHALLENGES AND FUTURE RESEARCH

To use the DP as a definitional standard of data and a general framework for data management, several challenges remain to be addressed. This section sketches three typical difficulties and addresses them in a preliminary way. The analyses and tentative answers provided in this section aim not to conclude the debate but to invite further research on these issues in the future.

A. TOO ROBUST?

For the DP-based approach to be widely applicable in practice, it must be accepted as an industry standard for balancing data privacy and utility. One reason some controllers may hesitate to adopt DP is that even though DP provides robust protection to data, it is just too robust. DP resolves the auxiliary information problem by making a strong assumption about the auxiliary information attackers could have access to. It assumes that if the dataset contains data records about N individuals, then the attacker knows the background knowledge of $N-1$ individuals and tried to find out the last one. DP requires that even under this stringent condition, the attacker could not infer with confidence whether the last individual is in the dataset. The problem with this assumption is that it is too strong in reality: attackers in most occasions do not have as much knowledge as what DP assumes. To meet this assumption, too much noise should be added. The result is that DP may sacrifice too much data utility in return for the strong protection it promises [69].

One possible solution is to introduce some relaxations to DP to make it more context-adaptive. The challenge with this solution is that the introduction of those parameters will increase the complexity of DP and compromise its advantage of simplicity. It is not easy for data controllers, subjects, and government officials to fully understand the meaning of parameter ϵ. More parameters will make the understanding more difficult. Another solution is to keep the strong assumption of protection intact while adjust the level of protection by changing the single parameter ϵ. But this general relaxation of protection by increasing the value of ϵ is "divorced from context, and . . . runs the risk of exposing a few data subjects to unnecessary risks" [70]. The underlying dilemma here is that ϵ itself does not directly measure the knowledge of the potential attacker. Therefore, for this solution to be feasible, the relationship between the two must be further explored to enable DP flexible enough to adjust to different contexts with heterogeneous types of attackers.

B. HOW TO CHOOSE THE PARAMETER?

Under the new approach proposed in this article, the importance of ϵ needs no more emphasis: it is the key for categorizing data. But how to select its value? This issue has spurred heated discussions within the technical community: researchers have offered various algorithms or mechanisms of selecting ϵ. For example, technicians can adjust the value of ϵ according to either a fixed level of risk or a fixed level of utility [71]. Some visualization tools are also developed for technicians to see and compare the different outcomes gendered by different values of ϵ in specific contexts [72]. However, what's missing in these discussions are the voices from legal regulators, data subjects, and the general public at large. Their engagement matters because the determination of the boundaries of personal, anonymous, and pseudonymous data is a value judgment that influences people's rights and duties. Data subjects' right to privacy and right to data protection, data-related companies' right to conduct businesses as market entities, and scientific researchers' freedom of academic research are all protected by the law. The balancing of these rights through the selection of ϵ is only legitimate upon broader participation and input. As Dwork has admitted, "[t]he parameter ϵ is public, and its selection is a social question" [73].

There are at least two models of realizing the collective selection of ϵ. One is the deliberation mechanism, in which representatives from the regulatory agencies, data controllers, and data subjects convene to deliberate on the standard and mechanism of selecting ϵ. If we choose this mode, we have to work on the institutional design of selecting representatives

[9] See Article 25 of EU's GDPR.

and the procedure of decision-making in those conventions. Another mode is the voting mechanism [74], in which data subjects of a database vote for the value of ϵ. The problem with this mode is how to protect the privacy of the voters since if an attacker gets to know who has voted, then this very fact will disclose the voter's presence in the database, and the goal of DP will flounder. Another more formidable problem is to teach the general public about the notion of DP and the meaning of ϵ, especially the relationship between the value of ϵ and the risks of their data [75]. For the general public, these may seem too technical and cryptic. Legal regulators and technical experts should then develop lucid or visual guidelines for educating the public. Whatever mode we adopt, the value of ϵ is not selected once and for all; rather, a mechanism of adjusting the value according to changing conditions must be put into place. This is also a collective endeavor rather than the sole duty of data controllers. In any case, future implementation of DP and the DP-based approach to defining data requires deeper cooperation among different professional fields as well as the general public.

C. HOW TO ACCOMMODATE PERSONAL PREFERENCES OF PRIVACY?

DP hides the presence of an individual data record in a group of records or a database. According to the classic model, DP guarantees the same level of protection to all the data records in the database by adding noises to them on the basis of a uniform parameter ϵ. However, there may be huge gaps between different individuals' privacy preferences: some may care anxiously about their data privacy, and some may be more willing to share their data [76]. How to account for these divergences between different individuals when they are contained in the same database? A uniform value of ϵ seems undesirable because it offers either over-protection or under-protection to individual participants in the database.

Researchers in the technical field have proposed several algorithms (e.g., [76], [77], [78], [79]) to realize the idea of the so-called personalized or heterogeneous DP. They tried to incorporate divergent preferences into the computation by using multiple values of ϵ as inputs. These approaches, however, need to address two thorny issues: First, enabling personalized levels of protection to meet users' diverse privacy preferences may introduce bias into data processing results [80], potentially leading to unfair decisions such as those related to loans and hiring [81]. Second, as privacy preference is itself information that needs to be protected, the mechanism of soliciting and computing individual choices of ϵ should be carefully designed to make sure that information of these choices is duly protected [78]. To protect the confidentiality of privacy preferences, other technical measures like encryption can be used.

VIII. CONCLUDING REMARK

This article outlines the rationale, steps, and implications of using DP to define different categories of data. The introduction and adoption of a technical standard do not mean abandoning the legal ones. In essence, the determination of whether a piece of data is personal, anonymous, or pseudonymous is a value judgment about whether, how, and to what extent legal regulations should intervene to protect some treasured liberties or interests of human beings. It hinges upon what we want data protection law to be. In this sense, legislators' and regulators' participation is necessary for defining the landscape of data categorization. Technical standards, though, are indispensable because data protection is itself technology-oriented and will be continuously shaped by technical developments in the future. Learning and borrowing technical tools and standards, as this article demonstrates, are beneficial for lawyers significantly since they could make legal enterprises more predictable, consistent, and transparent. Many unsolved issues remain to be explored in the future. Those issues, along with most issues in the data protection field, could not be properly solved unless legal and technical experts engage with each other in a more profound way.

REFERENCES

[1] Article 29 Data Protection Working Party. (2014). *Opinion 05/2014 on Anonymisation Techniques*. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[2] ICO. (2022). *Anonymisation, Pseudonymisation and Privacy Enhancing Technologies Guidance: Privacy-Enhancing Technologies (PETs)*. [Online]. Available: https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf

[3] G. Kellaris, S. Papadopoulos, X. Xiao, and D. Papadias, "Differentially private event sequences over infinite streams," *Proc. VLDB Endowment*, vol. 7, no. 12, pp. 1155–1166, Aug. 2014.

[4] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2013, pp. 901–914.

[5] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[6] A. Triastcyn and B. Faltings, "Bayesian differential privacy for machine learning," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 9583–9592.

[7] X. Ren, C.-M. Yu, W. Yu, S. Yang, X. Yang, J. A. McCann, and S. Y. Philip, "LoPub: High-dimensional crowdsourced data publication with local differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2151–2166, Sep. 2018.

[8] R. Liu, Y. Cao, M. Yoshikawa, and H. Chen, "FedSel: Federated SGD under local differential privacy with top-k dimension selection," in *Proc. 25th Int. Conf. Database Syst. Adv. Appl.*, 2020, pp. 485–501.

[9] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, pp. 1–36, Jan. 2014.

[10] X. He, A. Machanavajjhala, and B. Ding, "Blowfish privacy: Tuning privacy-utility trade-offs using policies," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, Jun. 2014, pp. 1447–1458.

[11] Apple. *Differential Privacy Overview*. Accessed: Aug. 10, 2023. [Online]. Available: https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

[12] Google. (2021). *How We're Helping Developers With Differential Privacy*. Accessed: Aug. 10, 2023. [Online]. Available: https://developers.googleblog.com/2021/01/how-were-helping-developers-with-differential-privacy.html

[13] ICO. (2022). *Guide to the General Data Protection Regulation (GDPR)*. [Online]. Available: https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr-1-1.pdf

[14] I. S. Rubinstein and W. Hartzog, "Anonymization and risk," *Washington Law Rev.*, vol. 91, no. 2, pp. 703–760, 2016.

[15] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA Law Rev.*, vol. 57, no. 6, pp. 1701–1778, 2009.

[16] A. Narayanan and V. Shmatikov, "Myths and fallacies of 'personally identifiable information,'" *Commun. ACM*, vol. 53, no. 6, pp. 24–26, 2010.

[17] K. El Emam and C. Alvarez, "A critical appraisal of the article 29 working party opinion 05/2014 on data anonymization techniques," *Int. Data Privacy Law*, vol. 5, no. 1, pp. 73–87, Feb. 2015.

[18] M. Finck and F. Pallas, "They who must not be identified—Distinguishing personal from non-personal data under the GDPR," *Int. Data Privacy Law*, vol. 10, no. 1, pp. 11–36, 2020.

[19] Article 29 Data Protection Working Party. (2007). *Opinion 4/2007 on the Concept of Personal Data*. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf

[20] ICO. (2022). *Anonymisation, Pseudonymisation and Privacy Enhancing Technologies Guidance: Introduction to Anonymisation*. [Online]. Available: https://ico.org.uk/media/about-the-ico/consultations/2619862/anonymisation-intro-and-first-chapter.pdf

[21] ICO. (2022). *Anonymisation, Pseudonymisation and Privacy Enhancing Technologies Guidance: Pseudonymisation*. [Online]. Available: https://ico.org.uk/media/about-the-ico/consultations/4019579/chapter-3-anonymisation-guidance.pdf

[22] ICO. (2022). *Anonymisation, Pseudonymisation and Privacy Enhancing Technologies Guidance: How do We Ensure Anonymisation is Effective?* [Online]. Available: https://ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf

[23] ICO. (2022). *Anonymisation, Pseudonymisation and Privacy Enhancing Technologies Guidance: Accountability and Governance*. [Online]. Available: https://ico.org.uk/media/about-the-ico/consultations/4019713/chapter-4-anonymisation-guidance-accountability-and-governance.pdf

[24] E. M. Weitzenboeck, P. Lison, M. Cyndecka, and M. Langford, "The GDPR and unstructured data: Is anonymization possible?" *Int. Data Privacy Law*, vol. 12, no. 3, pp. 184–206, Sep. 2022.

[25] P. Alexander and E. Davis, "Facial detection and smart billboards: Analysing the 'identified' criterion of personal data in the GDPR," *Eur. Data Protection Law Rev.*, vol. 6, no. 3, pp. 365–377, 2020.

[26] R. Bhaskar, D. Kaafar, and H. Asghar. *From Probably Private to Provable Privacy: On the Need for Rigorous Privacy Treatment for Data-Driven Organisations*. Accessed: Jul. 20, 2023. [Online]. Available: https://www.mq.edu.au/partner/access-business-opportunities/innovation-entrepreneurship-and-it/cyber-security-hub/news/news/from-probable-to-provable-privacy

[27] J. Polonetsky, O. Tene, and K. Finch, "Shades of gray: Seeing the full spectrum of practical data de-intentification," *Santa Clara Law Rev.*, vol. 56, no. 3, pp. 593–630, 2016.

[28] A. El Khoury, "Personal data, algorithms and profiling in the EU: Overcoming the binary notion of personal data through quantum mechanics," *Erasmus Law Rev.*, vol. 11, no. 3, pp. 165–177, Dec. 2018.

[29] P. M. Schwartz and D. J. Solove, "The PII problem: Privacy and a new concept of personally identifiable information," *NYU Law Rev.*, vol. 86, no. 6, pp. 1814–1894, 2011.

[30] M. Altman, A. Cohen, K. Nissim, and A. Wood, "What a hybrid legal-technical analysis teaches u s about privacy regulation: The case of singling out," *Boston Univ. J. Sci. Technol. Law*, vol. 27, no. 1, pp. 1–63, 2021.

[31] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd Conf. Theory Cryptography*, 2006, pp. 265–284.

[32] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[33] A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. O'Brien, T. Steinke, and S. Vadhan, "Differential privacy: A primer for a non-technical audience," *Vanderbilt J. Entertainment Technol. Law*, vol. 21, no. 1, pp. 209–276, 2018.

[34] C. Dwork and A. Smith, "Differential privacy for statistics: What we know and what we want to learn," *J. Privacy Confidentiality*, vol. 1, no. 2, pp. 135–154, Apr. 2010.

[35] A. E. Ouadrhiri and A. Abdelhadi, "Differential privacy for deep and federated learning: A survey," *IEEE Access*, vol. 10, pp. 22359–22380, 2022.

[36] K. Nissim, A. Bembenek, A. Wood, M. Bun, M. Gaboardi, U. Gasser, D. R. O'Brien, T. Steinke, and S. Vadhan, "Bridging the gap between computer science and legal approaches to privacy," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 687–780, 2017.

[37] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, Jun. 2009, pp. 19–30.

[38] V. Feldman and T. Zrnic, "Individual privacy accounting via a Rényi filter," in *Proc. 35th Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 28080–28091.

[39] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[40] W. K. Hon, C. Millard, and I. Walden, "The problem of 'personal data' in cloud computing: What information is regulated?—The cloud of unknowing," *Int. Data Privacy Law*, vol. 1, no. 4, pp. 211–228, Nov. 2011.

[41] K. Stokes, "On computational anonymity," in *Proc. Int. Conf. Privacy Stat. Databases*, 2012, pp. 336–347.

[42] A. F. J. W. Westin and L. L. Review, "Privacy and freedom," *Washington Lee Law Rev.*, vol. 25, no. 1, p. 166, 1968.

[43] M. H. Afifi, K. Zhou, and J. Ren, "Privacy characterization and quantification in data publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1756–1769, Sep. 2018.

[44] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, Jun. 2011, pp. 193–204.

[45] A. Chin and A. Klinefelter, "Differential privacy as a response to the reidentification threat: The Facebook advertiser case study," *North Carolina Law Rev.*, vol. 90, no. 5, pp. 1417–1456, 2011.

[46] K. M. P. Shrivastva, M. A. Rizvi, and S. Singh, "Big data privacy based on differential privacy a hope for big data," in *Proc. Int. Conf. Comput. Intell. Commun. Netw.*, Nov. 2014, pp. 776–781.

[47] W. L. Perry, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Santa Monica, CA, USA: Rand Corporation, 2013.

[48] P. M. Connorton, "Tracking terrorist financing through SWIFT: When US subpoenas and foreign privacy law collide," *Fordham Law Rev.*, vol. 76, no. 1, pp. 283–322, 2007.

[49] Y. Lagos and J. Polonetsky, "Public v. nonpublic data: The benefits of administrative controls," *Stanford Law Rev. Online*, vol. 66, pp. 103–110, Jan. 2013.

[50] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, Jun. 2003, pp. 211–222.

[51] D. Feldman and E. Haber, "Measuring and protecting privacy in the always-on era," *Berkeley Technol. Law J.*, vol. 35, no. 1, pp. 197–250, 2020.

[52] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth, "Differential privacy: An economic method for choosing epsilon," in *Proc. IEEE 27th Comput. Secur. Found. Symp.*, Jul. 2014, pp. 398–410.

[53] M. Naldi and G. D'Acquisto, "Differential privacy: An estimation theory-based method for choosing epsilon," 2015, *arXiv:1510.00917*.

[54] M. F. St. John, G. Denker, P. Laud, K. Martiny, A. Pankova, and D. Pavlovic, "Decision support for sharing data using differential privacy," in *Proc. IEEE Symp. Visualizat. Cyber Secur. (VizSec)*, Oct. 2021, pp. 26–35.

[55] J. Lee and C. Clifton, "How much is enough? Choosing ε for differential privacy," in *Proc. 14th Int. Conf. Inf. Secur.*, 2011, pp. 325–340.

[56] C. Dwork, N. Kohli, and D. Mulligan, "Differential privacy in practice: Expose your epsilons!" *J. Privacy Confidentiality*, vol. 9, no. 2, Oct. 2019. [Online]. Available: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/689

[57] R. Gong, "Transparent privacy is principled privacy," *Harvard Data Sci. Rev.*, Special no. 2, Jun. 2022. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/ld4smnnf/release/5

[58] S. Stalla-Bourdillon and A. Knight, "Anonymous data v. personal data—A false debate: An EU perspective on anonymization, pseudonymization and personal data," *Wisconsin Int. Law J.*, vol. 34, no. 2, pp. 284–322, 2016.

[59] M. Elliot, K. O'Hara, C. Raab, C. M. O'Keefe, E. Mackey, C. Dibben, H. Gowans, K. Purdam, and K. McCullagh, "Functional anonymisation: Personal data and the data environment," *Comput. Law Secur. Rev.*, vol. 34, no. 2, pp. 204–221, Apr. 2018.
[60] R. Gellman, "The deidentification dilemma: A legislative and contractual proposal," *Fordham Intellectual Property Media Entertainment Law J.*, vol. 21, no. 1, pp. 33–62, 2010.
[61] R. Calo, "The boundaries of privacy harm," *Indiana Law J.*, vol. 86, no. 3, pp. 1131–1162, 2011.
[62] F. Z. Borgesius, J. Gray, and M. Van Eechoud, "Open data, privacy, and fair information principles: Towards a balancing framework," *Berkeley Technol. Law J.*, vol. 30, no. 3, pp. 2073–2131, 2015.
[63] A. Sokolovska and L. Kocarev, "Integrating technical and legal concepts of privacy," *IEEE Access*, vol. 6, pp. 26543–26557, 2018.
[64] D. J. Solove, "Conceptualizing privacy," *California Law Rev.*, vol. 90, pp. 1087–1155, Jul. 2002.
[65] J. Zittrain, "Privacy 2.0," *Univ. Chicago Legal Forum*, vol. 2008, no. 1, pp. 65–119, 2008.
[66] D. J. Solove, "Privacy and power: Computer databases and metaphors for information privacy," *Stanford Law Rev.*, vol. 53, no. 6, pp. 1393–1462, 2000.
[67] G. G. Fuster, *The Emergence of Personal Data Protection as a Fundamental Right of the EU*, vol. 16. Berlin, Germany: Springer, 2014.
[68] J. Scheibner, M. Ienca, S. Kechagia, J. R. Troncoso-Pastoriza, J. L. Raisaro, J.-P. Hubaux, J. Fellay, and E. Vayena, "Data protection and ethics requirements for multisite research with health data: A comparative examination of legislative governance frameworks and the role of data protection technologies," *J. Law Biosci.*, vol. 7, no. 1, Jul. 2020, Art. no. lsaa010.
[69] J. Yakowitz, "Tragedy of the data commons," *Harvard J. Law Technol.*, vol. 25, no. 1, pp. 1–68, 2011.
[70] J. Bambauer, K. Muralidhar, and R. Sarathy, "Fool's gold: An illustrated critique of differential privacy," *Vanderbilt J. Entertainment Technol. Law*, vol. 16, p. 701, Jan. 2013.
[71] K. Ligett, S. Neel, A. Roth, B. Waggoner, and S. Z. Wu, "Accuracy first: Selecting a differential privacy level for accuracy constrained ERM," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 2566–2576.
[72] P. Nanayakkara, J. Bater, X. He, J. Hullman, and J. Rogers, "Visualizing privacy-utility trade-offs in differentially private data releases," *Proc. Privacy Enhancing Technol.*, vol. 2022, no. 2, pp. 601–618, Apr. 2022.
[73] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, Jan. 2011.
[74] N. Kohli and P. Laskowski, "Epsilon voting: Mechanism design for parameter selection in differential privacy," in *Proc. IEEE Symp. Privacy-Aware Comput. (PAC)*, Sep. 2018, pp. 19–30.
[75] F. Liu, "A statistical overview on data privacy," *Notre Dame J. Law, Ethics Public Policy*, vol. 34, no. 2, pp. 477–500, 2020.
[76] Z. Jorgensen, T. Yu, and G. Cormode, "Conservative or liberal? Personalized differential privacy," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 1023–1034.
[77] H. Ebadi, D. Sands, and G. Schneider, "Differential privacy: Now it's getting personal," in *Proc. 42nd Annu. ACM SIGPLAN-SIGACT Symp. Principles Program. Languages*, 2015, pp. 69–81.
[78] M. Alaggan, S. Gambs, and A.-M. Kermarrec, "Heterogeneous differential privacy," *J. Privacy Confidentiality*, vol. 7, no. 2, pp. 127–158, Jan. 2017.
[79] B. Niu, Y. Chen, B. Wang, Z. Wang, F. Li, and J. Cao, "AdaPDP: Adaptive personalized differential privacy," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2021, pp. 1–10.
[80] S. Wang, L. Huang, M. Tian, W. Yang, H. Xu, and H. Guo, "Personalized privacy-preserving data aggregation for histogram estimation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.
[81] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2022.

TAO HUANG received the B.Eng. degree in computer science from Chaohu University, Anhui, China, in 2010, the J.M. degree in law from Peking University, Beijing, China, in 2013, the L.L.M. degree in law from the Harvard Law School, Cambridge, MA, USA, in 2016, and the S.J.D. degree in law from the Duke University School of Law, Durham, NC, USA, in 2021. From 2013 to 2015, he was a Staff with the Judicial Bureau, Municipal Government, Beijing. He is currently an Assistant Professor with the City University of Hong Kong, Hong Kong, China. His works have appeared or forthcoming in prestigious law journals, including *Harvard Human Rights Journal*, *Columbia Human Rights Law Review*, and *University of Cincinnati Law Review*. His major research interests include constitutional law, cyberlaw, and law and technology.

SHUYUAN ZHENG (Member, IEEE) received the B.E. degree in software engineering from Peking University, in 2018, and the Ph.D. degree in informatics from Kyoto University, in 2023. He is currently a specially-appointed Assistant Professor with the Graduate School of Information Science and Technology, Osaka University. His research interests include computer science, law, and economics, with a particular focus on data economy and data privacy.