IEEE SA
STANDARDS ASSOCIATION

INDUSTRY CONNECTIONS REPORT

IC21-013 SYNTHETIC DATA

TOWARD PRACTICAL ANONYMITY: A WHITE PAPER ON PRIVACY RISK, METRICS, AND GOVERNANCE IN SYNTHETIC DATA

Authored by:

Alexander Boudewijn
Maurice Coyle
Alexandra Ebert *(Chair)*
Abdullah Elbi
Mark Elliot
Matteo Giomi
Faris Haddad
Shannon Kroes
Aldo Maurizio Lamberti *(Vice Chair)*
Simone Mangiante
Rudolf Mayer
Grady Nye
Manal Slokom

TRADEMARKS AND DISCLAIMERS

IEEE believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

The ideas and proposals in this specification are the respective author's views and do not represent the views of the affiliated organizations.

ACKNOWLEDGMENTS

Special thanks are given to all members of the IEEE Synthetic Data IC who contributed their expertise to this white paper.

The Institute of Electrical and Electronics Engineers, Inc. 3 Park Avenue, New York, NY 10016-5997, USA

PDF: STDVA28346 979-8-8557-2728-9

Find IEEE standards and standards-related product listings at: http://standards.ieee.org.

NOTICE AND DISCLAIMER OF LIABILITY CONCERNING THE USE OF IEEE SA INDUSTRY CONNECTIONS DOCUMENTS

This IEEE Standards Association ("IEEE SA") Industry Connections publication ("Work") is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the IEEE SA Industry Connections activity that produced this Work. IEEE and the IEEE SA Industry Connections activity members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE SA Industry Connections activity members disclaim any and all conditions relating to: results; and workmanlike effort. This IEEE SA Industry Connections document is supplied "AS IS" and "WITH ALL FAULTS."

Although the IEEE SA Industry Connections activity members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE OR IEEE SA INDUSTRY CONNECTIONS ACTIVITY MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder. The policies and procedures under which this document was created can be viewed at http://standards.ieee.org/about/sasb/iccom/.

This Work is published with the understanding that IEEE and the ICCom members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

EXECUTIVE SUMMARY

Structured synthetic data generated for privacy-preserving purposes is a maturing technology that facilitates compliance with data protection regulations. One of the major challenges in the deployment of synthetic data is assessing the extent to which it mitigates the privacy risks of data subjects, as this requires deep knowledge of both the technology and the relevant regulatory frameworks. At the time of this writing, no universal formula or standard for defining when a particular synthetic dataset can be considered anonymous is available. However, as the technology matures and is adopted, a set of **best practices** is emerging that provides effective guidance for organizations deploying synthetic data.

Consideration of whether a synthetic dataset is anonymous typically involves an **examination of the requirements of applicable statutes and regulations** and a risk assessment of the potential disclosure risks arising from use of the dataset. It is generally accepted that anonymized datasets do always have some residual risk, and whether that risk is acceptably small enough is a judgment call.

The legal and privacy implications of any particular deployment will depend on the type of data and the use case. For this reason, and given the wide variety of synthetic data use cases and deployment scenarios, there is **not (yet) a universal threshold** on any risk metric that can be used to decide when synthetic datasets can be considered anonymous. However, **the practical anonymity of a synthetic dataset can be demonstrated on a case-by-case basis**.

Most recent work on privacy assessment for synthetic data has focused on adversarial-based evaluations. These methods simulate attacks using algorithms that could be employed by adversaries to disclose information about data subjects in the original dataset by using the synthetic dataset for leverage. Such methods make use of *threat modeling* to embody realistic privacy threats. The derived risks are thus easier to contextualize, especially when well-defined attack vectors such as *singling out*, *linkability*, *attribute inference,* and *membership inference* are modeled.

A thorough empirical analysis of these risks, modeling adversaries of various abilities, can also inform the application of **additional privacy-preserving techniques** (such as, for example, prior generalization or pseudonymization) to the synthetic dataset to mitigate the residual risks. Finally, **organizational and technical measures** (such as limiting access to the synthetic data) can also be employed to further reduce the tractability of potential threats.

The field of structured synthetic data is rapidly evolving, driven by advancements in AI, machine learning, and increased computational capacity. New methods are gaining attention, but further research is needed to better understand why different synthesis models work more effectively with specific data types; the reasons for this are not yet fully understood. Concomitantly, further research is needed to understand how to best select and tune synthesis models. The **future outlook** may see dynamic, context-aware approaches to privacy settings. **Ethical and governance concerns** are equally important in synthetic data generation, especially when addressing issues of transparency and fairness.

The importance of this white paper extends beyond its role as a comprehensive guide. It is envisioned as a foundational step toward establishing **agreed-upon definitions** and **meaningful standards** for privacy-preserving data synthesis. The paper therefore recommends the production of a standard project authorization request (PAR), which would be a significant milestone for this IEEE Industry Connection Report. This collective development is crucial for advancing the field, ensuring that the emerging standard is robust, comprehensive, and reflective of the full range of cognate expert insight.

TABLE OF CONTENTS

TABLE OF CONTENTS

TOWARD PRACTICAL ANONYMITY: A WHITE PAPER ON PRIVACY RISK, METRICS, AND GOVERNANCE IN SYNTHETIC DATA

ABSTRACT

Structured synthetic data is presented as a maturing technology for privacy preservation that facilitates compliance with data protection regulations. The difficulty of defining when such data can be considered anonymous under existing legal frameworks is highlighted, given the absence of a universal standard, even as emerging best practices begin to provide guidance for privacy compliance. A comprehensive analysis is provided of the relevant legal and regulatory context, empirical methods for privacy risk evaluation, and adversarial threat modeling approaches. Privacy risk metrics and technical mitigation techniques are examined, and governance considerations for enterprise data management and compliance are addressed. Finally, the need for industry standard-setting initiatives is underscored, and a recommendation is made to pursue formal standards for privacy-preserving data synthesis.

1. INTRODUCTION

1.1. MOTIVATION

In the rapidly evolving landscape of data-driven innovation, the need to balance the benefits of data utilization with the imperative of privacy protection has never been more pressing. As organizations across all sectors in modern information economies increasingly rely on data to drive decision-making and innovation, the risks associated with handling sensitive personal information have become a significant concern. Synthetic data has emerged as a promising solution to this challenge. By generating artificial datasets that replicate the statistical properties of real data without containing actual personal information, synthetic data offers the hope of mitigating privacy risks while still allowing the harnessing of the power of data.

Privacy is a contested and much-discussed term. Daniel Solove, one of the foremost privacy scholars, describes it as "a concept in disarray" (Solove [36]) and it is not within our remit to solve this issue.[1] To be clear, this paper refers to one particular type of privacy: *informational privacy*. The paper further constrains the consideration of privacy risks to the risk of disclosure of information about an individual. The disclosure itself is considered a harm without weighing the importance of that disclosure or the downstream harms that might befall the data subjects as a consequence of the disclosure. This is partly to constrain the discussion to what is tractable, but also to focus attention on the problem that synthetic data is trying to solve—the prevention of unwanted disclosures about data subjects.

Ensuring that synthetic data is truly privacy-preserving, yet still useful for its intended applications, requires a deep understanding of the trade-offs involved. While synthetic data is inherently less risky than real data, questions remain about the extent to which it can effectively protect privacy and the associated legal and ethical considerations. The growing adoption of synthetic data across industries underscores the need for a clear, standardized framework to evaluate its privacy and utility.

Mission statement: The mission of the IEEE Synthetic Data Industry Connection (IC) is to provide a thorough and accessible framework for understanding and evaluating synthetic data, particularly structured synthetic data, used for privacy protection. This white paper aims to equip business users, data professionals, and policymakers with the knowledge and tools needed to assess when a synthetic dataset

[1] Numbers in brackets correspond to the references listed in Section 7.

is sufficiently privacy-safe and anonymous for practical use, while also addressing the broader regulatory and technical challenges associated with synthetic data.

1.2. KEY QUESTIONS

This white paper addresses several critical questions that are central to the effective use of synthetic data: How effective is synthetic data in safeguarding individual privacy, and what metrics can be used to evaluate this effectiveness? What are the current legal and regulatory challenges related to synthetic data, and how can organizations navigate these complexities? How can synthetic data be effectively integrated into enterprise data ecosystems while maintaining compliance with privacy standards? Additionally, what specific privacy risks are associated with synthetic data, particularly in the context of informational privacy and the risk of data disclosure?

The primary objective of the paper is to address the practical question that many business users face: How can one determine when a synthetic dataset is sufficiently private to be used with confidence?

1.3. CONTRIBUTIONS

This white paper makes several contributions to the field of structured synthetic data. It provides a comprehensive overview of the current challenges and debates surrounding the privacy implications of synthetic data and highlights the need for standardized privacy metrics. The white paper outlines potential frameworks for evaluating the privacy of synthetic data, enabling business users and data professionals to make informed decisions. It also explores the role of synthetic data as a valuable asset in organizational data catalogs and industry-specific marketplaces, offering insights into its integration across various domains.

1.4. PAPER OUTLINE

The remainder of this paper is structured as follows. Section 2 provides an overview of what synthetic data is, its types, and its uses. Section 3 examines the complex relationship between privacy laws and the generation and use of synthetic data, emphasizing the importance of understanding legal definitions, conducting risk assessments, and ensuring compliance with data protection principles throughout the synthetic data lifecycle. Section 4 explores the various methods and metrics for assessing the privacy and

utility of synthetic data, focusing on how to evaluate the risk of reidentification and the effectiveness of privacy-preserving techniques. Section 5 provides a comprehensive guide on governing synthetic data within an enterprise, covering legal considerations, governance processes, and the implementation of technical and organizational measures to ensure compliance and effective management throughout the data lifecycle. Finally, Section 6 concludes the report with open areas of research, the broader adoption within organizations, as well as recommendations for future standardization efforts of synthetic data.

2. WHAT IS SYNTHETIC DATA AND WHY IS IT IMPORTANT?

2.1. WHAT IS SYNTHETIC DATA?

Synthetic data consists of artificial entities and/or events that mimic real-life patterns and trends. Although in principle the synthesized entities can represent anything—households, schools, businesses, administrative units, or accounts—typically, the goal of synthetic data is to create records of artificial individuals. Using statistics or machine learning, a *model* is created that describes associations in a real, privacy-sensitive dataset (often referred to as the *original data*). For example, in a medical dataset, the model may describe the relationship between demographic information and the likelihood of heart failure. Using this model, new, synthetic entities (e.g., patients) are generated that mimic the relationships found in the original data. These synthetic entities will not have a determinable one-to-one mapping to the original entities. Therefore, synthetic data is not the same as statistical disclosure control or pseudonymization, in which an original dataset is partly altered but the one-to-one mapping of the original and treated units remains.

Synthetic data can be generated "from scratch," either from a model of a specific *original dataset* or from some specific statistical outputs generated using the original dataset (both referred to as *fully synthetic data*). The alternative is where synthetic data are deliberately written over an existing data frame (referred to as *partially synthetic data*). Another key distinction is between *complete* and *incomplete,* which refers to whether all of the original dataset has been synthesized or just part of it. This white paper focuses on complete, fully synthetic data derived from a real, privacy-sensitive *original dataset* as this is the paradigmatic form. Most of the points made, however, do apply to other forms.

One desideratum of a synthetic dataset is that it should be similar enough to the original dataset to be usable for analysis or other applications and yield outcomes sufficiently similar to those obtained from the original data. The extent to which synthetic data meets this desideratum is referred to as the *utility* of the synthetic data. On the other hand, a synthetic dataset should not be so similar that it leaks information about the data subjects in the original data. For example, the synthesis model may have embodied the training data at a level of granularity that enables sensitive information about specific entities to be inferred from the synthetic records.

2.2. WHY IS SYNTHETIC DATA IMPORTANT?

With machine learning (ML) algorithms and artificial intelligence (AI) being developed at a rapid pace, the opportunity to learn from data has grown with it. As a result, data is more valuable than ever and there is widespread ambition to utilize that value. However, when this data is personal and/or includes sensitive information, privacy and security concerns arise. According to IBM's *Cost of a Data Breach Report 2023* [22], the most common type of data stolen or compromised was customer and employee personal identifying information (PII).

Privacy and data protection laws require organizations to protect the privacy of individuals and ensure that personal data is processed in a fair, lawful, and transparent manner. One mechanism for meeting these obligations is to use privacy-enhancing technologies (PETs); emerging techniques aim to unlock the value of data while preserving the privacy of individuals and limiting the security risks associated with providing access to data. One such privacy-enhancing technology is *synthetic data generation*.

2.3. WHY SYNTHETIC DATA IS NOT A SILVER BULLET

Synthetic data, while powerful in protecting privacy, is not a universal solution that fits every scenario seamlessly. First, it is crucial to dispel the notion that synthetic data can provide either perfect privacy or perfect utility. Perfect privacy would require random data and, therefore, no utility; perfect utility would, in effect, be the reproduction of the original data and therefore not privacy-safe. There is always an inherent trade-off between privacy and utility; enhancing privacy often comes at the cost of utility and vice versa. Yet, for synthetic data, this trade-off is substantially more favorable than for anonymization methods that merely mask or distort parts of real data in an effort to achieve anonymity, rather than

generating new data from scratch. Second, each use case requires tailored adjustments to the privacy–utility trade-off. Simply generating a synthetic dataset does not automatically make it suitable for all purposes; the settings need to be fine-tuned depending on the specific requirements and constraints of the intended application. Thus, rather than seeking an impossible ideal, the goal should be to achieve a well-understood balance between privacy and utility tailored to the actual application. This understanding allows for informed, risk-based decisions, where stakeholders can weigh the potential privacy risks against the benefits of data utility. By doing so, they can choose an appropriate level of privacy that aligns with their specific needs and regulatory requirements, ensuring that the synthetic data serves its intended purpose without compromising sensitive information.

2.4. HOW DOES SYNTHETIC DATA GENERATION DIFFER FROM OTHER PRIVACY-PRESERVING TECHNIQUES?

Synthetic data generation bears some resemblance to traditional anonymization techniques (particularly statistical disclosure control [SDC]) and to other PETs. This section discusses the relationship between synthetic data and these other technologies/methods.

2.4.1. STATISTICAL DISCLOSURE CONTROL (SDC)

Classic SDC consists of disclosure risk assessment and control steps. The risk assessment is invariably in the form of some sort of statistical model, and the control mechanism causes a set of adjustments to the data, such as adding noise to the data (perturbation) or aggregating information (e.g., representing age in ten-year bands instead of a single year). Different frameworks can be used to define and quantify the level of protection achieved; a well-known example is k-anonymity.

Like synthetic data, SDC aims to prevent successful confidentiality attacks; however, SDC techniques, such as generalization and aggregation, result in treated data that retains a determinable mapping back to the original data (albeit a less informative one). This can be an advantage as such adjustments are typically more tractable and explainable. On the other hand, the retention of the mapping leaves open a route for attackers to extract sensitive information from the treated data. To disrupt this, the required number of adjustments may be so large that the treated data is no longer informative. As Elliot, Mackey, and O'Hara say, the point of any anonymization process is to provide safe data, but it only makes sense if the data

provided is useful [13]. Where perturbative methods are used (such as adding noise), the structure of the data is not necessarily preserved.

2.4.2. OTHER PETs

Other well-known PETs include secure multi-party computation (SMPC), homomorphic encryption (HE), and federated learning (FL). These techniques are typically used in a setting where multiple parties perform a computation together, and the required data for these computations cannot be shared between parties. Using these techniques, privacy leakage can be mitigated during the computation process, allowing control over information access for different parties. They are, however, not intended to protect the *output* of the computation once a party has access. Synthetic data is in the same form as the original data. Parties do not necessarily need to agree on which computation or analysis will be carried out. Instead, the privacy enhancement takes place before the analyses, and users can retain more flexibility in deciding or tweaking their analyses (Elliot, Mackey, and O'Hara [13]). Note that it is possible to combine PETs, such as when the training data are distributed between parties, and other PETs may be used to enable parties to generate a synthetic data set *together*. For example, recent developments in *federated synthesis* are reviewed by Little et al., "Federated learning" [30] and Kairouz et al. [27].

Another technique that is sometimes combined with data synthesis is differential privacy (DP) (Ji, Lipton, and Elkan [25]; Solove [36]). DP is a provable guarantee that a mechanism has prevented leakage beyond that of a specified parameter; this involves perturbing the generation process. These other PETS can complement synthetic data, enhancing privacy protection, but they require careful consideration and technical expertise to implement effectively.

2.5. WHAT IS SYNTHETIC DATA USED FOR?

Synthetic data's range of uses is large and growing; the following list is not meant to be exhaustive but to give a flavor of the variety of uses:

- **Internal data sharing:** Data privacy regulations not only restrict data sharing between organizations but can also prevent the flow or re-use of data within an organization. Obtaining data access permissions can take weeks, which can hinder collaboration, or a lawful basis for processing real personal data may not be available. Organizations can speed up innovation and unlock the value of data with enhanced collaboration between teams by leveraging synthetic data. Even if personal data

can be legally shared internally, organizations may choose to use synthetic data instead to reduce the risks of leakage and manage their data safely and ethically.

- **External data sharing:** Innovation in many sectors relies on partnering with third-party organizations. Synthetic data enables enterprises to evaluate third-party vendors by sharing confidential data with them to gain business insights or build ML models, while reducing security and compliance risks. Third-party organizations might include researchers, marketing agencies, competitors, or other cognate organizations—depending on the use case. Such sharing will generally be subject to additional legal controls.
- **Augmentation:** Refers to the addition of data to an existing original dataset.
- **Reducing bias in models:** Synthetic data generation can be used to reduce the bias in datasets by representing data with appropriate balance, density, distribution, and other parameters to address data quality problems for downstream tasks.
- **Data supplementation for model accuracy:** Predicting rare events such as fraud or manufacturing defects is hard since small data volumes may lead to inaccuracies in ML models. Generating synthetic instances of such events could increase model accuracy.
- **Software testing:** A synthetic data twin (or an augmented data set) might be generated from existing data in a production system. That data could then be used by developers to test much earlier in the software development cycle. This also allows for data to be shipped to offshore development teams.
- **Cloud and data space migration:** Cloud services, as well as data spaces, offer a range of innovative products for many sectors. However, moving private data to cloud infrastructures involves security and compliance risks. Thus, migrating synthetic data may enable organizations to leverage the benefits of cloud services and data spaces.
- **Data retention:** A synthetic version of a privacy-sensitive data set can be retained when the original data needs to be deleted in accordance with legal or contractual data destruction requirements. This could then be used in long-term analyses, such as detecting the seasonality patterns over several years of data.

3. HOW DO PRIVACY LAWS AFFECT THE GENERATION AND USE OF SYNTHETIC DATA?

Privacy and data protection laws generally apply to the processing of "personal data," which is typically defined broadly to include any information that relates to an identified or identifiable natural person. While the specific requirements of privacy laws vary, privacy laws generally aim to ensure that personal data processing activities adhere to key data protection principles.

Synthetic data itself can be a tool to enable an organization to apply and comply with data protection principles and privacy laws. This section explains the basic relationship between privacy laws and the use of sensitive datasets to generate synthetic data and the use or disclosure of the synthetic data itself. It raises the right questions to ask when collaborating with attorneys and other privacy professionals to unlock the full privacy-enhancing benefits of synthetic data.

Remember that privacy and data protection laws vary across jurisdictions, and specific sectoral laws may also apply. Always consult with a qualified attorney, privacy expert, or your data protection officer (as applicable) to ensure that the data processing is in compliance with all applicable legal requirements.

3.1. UNDERSTANDING THE DIFFERENCE BETWEEN PERSONAL DATA AND ANONYMIZED DATA IN THE CONTEXT OF SYNTHETIC DATA

The definitions of personal data used in privacy laws typically cover any information that relates to an identified or identifiable natural person, even if the data does not include any directly identifiable fields and is only associated with pseudonymous identifiers. These "quasi-identifiers" are fields or combinations of fields that have high resolution for an individual or an event that can be associated with an individual. Anonymization, therefore, requires breaking the link between the individual and the data such that the data no longer relates to an identifiable natural person.

Anonymization thresholds vary under different laws. A leading example is the EU's General Data Protection Regulation (GDPR) Recital 26, which defines anonymous data as data that "does not relate to

an identified or identifiable natural person," or "personal data that has been rendered anonymous in such a manner that the data subject is not or no longer identifiable."[2] The answer to whether a dataset is "anonymous" typically involves an examination of the text of applicable statutes, regulatory guidance on the topic of anonymization, and a risk assessment of the potential for reidentification of individuals in the dataset. Evaluating the risk of reidentification may—depending on the applicable legal requirements—consider factors such as the context in which the data is being processed (including internal versus external use), the potential for reidentification within the dataset itself, the resources and technologies that would be required for reidentification, and how the administrative and technical controls applied to the dataset may affect any of these risks.

In other words, anonymization is a context-driven exercise that needs to be revisited over time, taking into account both the data itself and the context in which it is processed. Consider whether the objective is *absolute anonymization* (i.e., rendering data impossible for anyone to reidentify) or *relative anonymization* (i.e., rendering data sufficiently difficult to reidentify, taking into account the time, cost, effort, tools, and additional information that would be required to reidentify an individual in the dataset by those who obtain access to the data).[3] For example, the UK Information Commissioner's Office (ICO) guidance on privacy-enhancing technologies suggests that parties should evaluate the context and purposes of processing when determining which risk-mitigation measures are appropriate to ensure that synthetic data is anonymous, indicating that the risk of attack may be reduced when synthetic data is processed in a secure setting, such as a trusted research environment (ICO "Privacy enhancing" [24]).

When assessing the risk of reidentification *within* a particular dataset, the following three criteria, identified by the Article 29 Data Protection Working Part [4], are especially instructive:

- *Singling out*—The possibility of isolating some or all records that identify an individual in the dataset
- *Linkability*—The ability to link at least two records concerning the same data subject or a group of data subjects
- *Inference*—The possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes

[2] Recital 26 is on page 5 Of GDPR and available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
[3] See, for example, the UK Information Commissioner's "motivated intruder test" (ICO "Chapter 2" [23]).

These criteria and how they apply to the evaluation of synthetic data are discussed further in Section 4.

Processing personal data—or data that can be linked to identifiable individuals—without complying with the requirements of privacy and data protection laws can result in enforcement actions, violation of individuals' rights, complaints, fines, and orders to cease processing data. Once data has been anonymized, however, it is generally no longer subject to the requirements of privacy and data protection laws. This makes anonymization an attractive objective as a data practice. But it also serves as a reminder of the high standards to which anonymized data should be held. As the world of big data evolves, privacy regulators are likely to seek ways to extend the protections embedded in privacy and data protection laws to all datasets that retain a meaningful link to individuals and the exercise of their fundamental rights and freedoms.

What does this mean for synthetic data? In practice, it is almost certainly not accurate to say that synthetic data is anonymous under existing privacy and data protection laws simply because it has been constructed synthetically. Additional contextual factors that may affect the anonymity of synthetic data include the specific synthesis algorithm or technique used, the nature of the underlying data used to inform the data synthesis, and which privacy-enhancing measures or configuration options are applied to the synthetic data creation and its subsequent use. Organizations generating synthetic data should take steps to evaluate the disclosure risk associated with a synthetically generated dataset using the approaches described in Section 4, in addition to applying risk-mitigating controls to the use cases and settings in which the data will be generated and used.

The results of this risk assessment, and any mitigating controls applied, should be documented in a privacy impact assessment (as described in Section 5) and reevaluated periodically as the risk and/or use cases change over time, taking into account the latest guidance from regulators. While the specific standards prescribed by regulators may evolve, organizations that demonstrate accountability by undertaking, documenting, and periodically reevaluating an appropriate risk assessment will be better positioned to reduce the risk of non-compliance. In particular, certain elements of the risk assessment might be required every time a dataset is synthesized, given that the input data extract, size, synthesis configurations, intended use cases, and processing context may change on a case-by-case basis.

3.2. HOW DO PRIVACY LAWS AFFECT THE USE OF TRAINING DATA DURING THE DATA SYNTHESIS?

Privacy and data protection laws generally apply to all personal data processing activities. *Processing* is usually defined broadly to include any operation or set of operations performed on personal data, including collecting, recording, organizing, structuring, storing, adapting or altering, retrieving, using, consulting, disclosing, disseminating or making available, aligning or combining, anonymizing, restricting, erasing, or destroying such data.

What does this mean for synthetic data? This means that if real training data is used to build synthetic data, and that training data contains personal data, then personal data is being processed—and privacy and data protection laws apply to such training data and the generated synthetic data. This is true even if the ultimate objective in turning to synthetic data is to generate an anonymous dataset not subject to privacy laws. When the synthesis technique involves a model being trained over the source data to capture the correlations, distributions, and other relationships within the data, the model itself may potentially be personal data, especially if a "model inversion" attack can be used to reproduce personal data used as part of the training process.[4]

Any time personal data is processed—including for a privacy-preserving purpose, such as the implementation of a privacy-enhancing technology like synthetic data—a privacy impact assessment (PIA) or data protection impact assessment (DPIA) (as explained in Section 5) should be carried out, as applicable, to address the compliance requirements and assess and mitigate any risks associated with the data processing activity. Not only should this privacy impact assessment address the risk of reidentification in the synthetic dataset (as described previously), but it should also address the compliance requirements associated with the processing of personal data in the underlying training dataset.

Privacy impact assessments typically follow a form or template established by an organization's data protection officer or privacy professionals. Nevertheless, as training datasets are considered for the

[4] See Veale et al. [42]and ICO "Privacy enhancing" [24]. Note that model inversion attacks reproducing personal data may not be realistically feasible in the context of more complex models, and further research on this topic is needed to refine our understanding of the risks.

purpose of generating synthetic data, one threshold requirement merits further consideration here: the requirement to have a lawful basis for personal data processing activities.

3.3. WHY IS A LAWFUL BASIS NEEDED FOR DATA SYNTHESIS, AND WHAT LAWFUL BASES ARE AVAILABLE?

Privacy and data protection laws generally require that personal data processing activities be "lawful." Some privacy laws—sectoral laws in particular—specify that personal data may only be processed for certain enumerated purposes (for example, laws that prohibit the provision of credit reports except for purposes authorized by statute, or laws that prohibit the collection of children's data without explicit parental consent). Other privacy laws—especially general laws like the EU's GDPR—articulate a set of available lawful bases for processing personal data, and the data controller must establish that they have an appropriate lawful basis for each personal data processing activity that they undertake.

Typical lawful bases for processing personal data include consent of the data subject, performance of a contract to which the data subject is a party, compliance with a legal obligation, processing that is necessary for the protection of the vital interests of a person, processing that is necessary for a task carried out in the public interest, or processing that is necessary for purposes of legitimate interests pursued by the controller, except where overridden by the interests or fundamental rights and freedoms of the data subject. Not all of these lawful bases are available in every jurisdiction, and each is accompanied by limiting principles that require careful consultation with a qualified attorney, privacy expert, or data protection officer (as applicable) to ensure they are employed appropriately. In some cases, further processing of personal data for a new purpose–including anonymization–may be possible if it is considered "not incompatible" with the purpose(s) for which the data was collected.

What does this mean for synthetic data? Prior to using personal data to train a synthetic data generator (or: model), a privacy attorney/expert or data protection officer (as applicable) should establish whether the organization has an appropriate lawful basis for processing the personal data for this purpose. This analysis should be performed in addition to establishing whether the organization has other legal rights (e.g., commercial or contractual rights) to process the data in question.

In jurisdictions where it is available, “legitimate interest” may be a viable lawful basis for using personal data to train and generate synthetic data. To assess whether this is appropriate, conducting and documenting a balancing test may be necessary. The balancing test weighs the legitimate interests of the organization creating the synthetic data against the interests and fundamental rights and freedoms of the individual. If the privacy-preserving nature of the synthetic dataset can be established, this strengthens the argument that reliance on legitimate interest is an appropriate lawful basis for the activity. While transparency and notice requirements, as well as providing an opportunity for data subjects to opt out, may still apply to the use of legitimate interest grounds for personal data processing, synthetic data may be a more attractive option than the usually more onerous requirements of using consent as the lawful basis.

3.4. SUMMARY

Synthetic data has tremendous privacy-preserving potential and serves as a useful tool in demonstrating and applying privacy by design. Nevertheless, privacy and data protection laws still play a significant role in the lifecycle of synthetic data generation, and synthetic data cannot be considered anonymous by default or “safe because it is fake,” even if it does not contain any real, direct identifiers. The legal aspects of synthetic data are complex, and the summary given here should not be considered definitive. For readers interested in going further, a detailed legal analysis of the synthetic data and its implications on privacy, data protection, and competition is provided by Gal and Lynskey [16].

Nevertheless, there are some practical considerations that can be noted. Before using real data to train and generate synthetic data, a privacy impact assessment should be conducted to (1) determine whether there is an appropriate lawful basis for processing the underlying personal data, (2) identify and mitigate any privacy risks, and (3) assess whether the resulting synthetic data will meet the applicable anonymization thresholds for the relevant jurisdictions. Repeat this assessment at intervals, taking into account the developing legal knowledge and court cases in the relevant jurisdictions, as well as available techniques and technologies that might affect the risk of reidentification.

Section 4 turns to techniques for measuring and assessing privacy risks in synthetic data.

4. HOW TO MEASURE AND MITIGATE THE PRIVACY RISKS OF SYNTHETIC DATA?

The anonymity of data is not a binary characteristic, but rather one that exists on a spectrum: there are no unequivocal measures or thresholds for determining when a given dataset—synthetic or not—can be considered anonymous. Whenever an organization seeks to anonymize a dataset, it is crucial to assess the residual privacy risks in the context of the particular use case to be able to determine whether said dataset can be considered "anonymous" in the legal framework in which the organization operates.

This section will first outline how and why synthetic data can reveal personal information about the individuals whose data has been used to generate it (Section 4.1), and which types of disclosure risks are relevant to consider (Section 4.2).

The steps for following a risk assessment of the synthetic data are as follows:

- Define threat models that describe goals, incentives, and means of how adversaries attempt (or "might") infer personal data (Section 4.3)
- Select similarity and/or attack-based privacy metrics (Sections 4.4 and 4.5)
- Perform this evaluation for all defined threat models (Section 4.5)

Finally, additional mechanisms are discussed that can be used when generating synthetic data to further reduce privacy risks (Section 4.6).

4.1. ADDITIONAL SOURCES OF RESIDUAL PRIVACY RISKS

Various factors can potentially compromise the privacy protection that synthetic data might offer to data subjects. First, similar to the publication of summary statistics, synthetic data can leak general information about the original data. Particularly for categorical variables, adversaries could learn from the synthetic data that certain (rare) categories occurred in the original data. Rare category protection is effective in addressing this residual risk. Additionally, some synthesization methods may generate data that reflect

the minima and maxima of variables in the original data set. Here, protection of the value range is recommended to eliminate this risk.

Second, synthetic data generated without appropriate privacy mechanisms might mimic characteristics of specific records in the original data with sufficient accuracy and precision so that the data leaks information about individual subjects. However, even if appropriate privacy mechanisms are used during data synthesis, it is possible that synthetic records will match original records exactly or approximately without this automatically leading to this risk. This is more likely in a simpler dataset. For example, if a dataset only has attributes for ZIP code, sex, and job title, some synthetic records will be exact duplicates of records in the original data; this is expected and generally not a privacy risk. On the other hand, in more complex datasets (for example, a financial dataset with hundreds of attributes per customer and a year of financial transactions), direct matches are not to be expected and may be problematic. Lastly, as many synthetic data generators draw random samples across the learned statistical distribution of the dataset they were trained on, it can happen that a synthetic record matches a real-world record/ data subject, despite that specific real-world record not being part of the synthetic data generator's training sample (Imagine telephone numbers that follow a specific structure in each country. A synthetic data generator trained on telephone numbers of New York City citizens could, by chance, randomly generate a telephone number that belongs to a real person, despite that person not being in the training data of the generator.)

In such a situation, the direct "match" of one or some attributes can be considered as chance, rather than a result of the synthesis model learning and replicating the original records and is unlikely to constitute a privacy risk. Particularly in complex datasets with hundreds or thousands of columns per data subject, it is statistically highly unlikely that a synthetic record will match all of the attributes of a real-world record/data subject that was not part of the training sample.

Factors that can impact the level of privacy include the following:

- The method used to generate the synthetic data is as follows:
    - Fundamental characteristics of the synthetic data generator or model: some generators may store exact values of the original data in the model, which makes it more likely that these will occur in the synthetic data. On the other hand, other methods may "smooth over" the values, and the model thus does not "remember" the original values. Similarly, some synthetic data methods strictly adhere to the minima and maxima of attributes, whereas others can exceed these

boundaries. For (black-box) deep learning methods (such as generative adversarial networks, [GANs]), it may be more difficult to reason about the exact traits the model has saved. It is known that these methods are prone to memorization (Feng et al. [14] and overfitting. As a result, records from the original data may be duplicated in the synthetic dataset.

  - Whether additional privacy-enhancing techniques are used. Depending on the selected data synthesis approach, additional privacy-preserving techniques such as differentially private noise or disclosure controls can decrease the residual risks. Specific methods are detailed in Section 4.6.

- The characteristics of the original data:

  - Small datasets will often have a disadvantageous privacy–utility trade-off (i.e., they have to compromise more heavily between the levels of privacy protection and usefulness of the data for the analyst).

  - The homogeneity of the data (if there is little or almost no variation across some variables in the data set). The more homogeneous it is, the less risky it is.

  - The existence of outliers (since outliers are more distinctive than other data points, they may constitute elevated privacy risks if reproduced during synthesis, hence they should be considered and managed as part of a risk-based approach to synthetic data use).

Besides the risks associated with memorization, synthetic data can also be affected by risks related to its context and use. Such risks are best avoided through stringent auditing and other technical and organizational measures (TOMs). Risk factors include the following:

- Information availability. In most scenarios, an adversary will leverage additional information sources. These can include information about the generative model used, or knowledge about some of the individuals in the real dataset. The information available to the attacker is part of the threat model (see Section 4.3).
- Improper use. A trusted party involved in the synthetic data generation process may (deliberately or otherwise) misuse their access to confidential information.
- Implementation. Software systems that provide synthetic data generation techniques may have weaknesses that can be exploited; for example, additional privacy-enhancing techniques that do not function as intended.

4.2. MEASURING DISCLOSURE RISKS USING PRIVACY ATTACKS

Section 3 introduced the risks: *singling out*, *linkability*, and *inference* that, according to the EU Article 29 Working Party, need to be appropriately mitigated to consider a dataset to be anonymous. These three risks map to specific types of attack that might be carried out against synthetic data. Simulating such attacks can then be used (by a data controller) to estimate the risk that the synthetic data could lead to a reidentification or other form of statistical disclosure. For inference, this paper considers two specific cases: attribute and membership inference.

It is also worth noting the subtle difference between disclosure and inference; *disclosure* is normally used when a particular piece of information is attributed to a known population unit (person) with confidence, while *inference* is normally a probabilistic outcome (possibly in the form of a posterior probability distribution).[5] The following uses the term *inference*.

There is a wider literature on privacy attacks against other types of privacy-preserving data publishing techniques; however, not all of the risks discussed there are relevant in the context of synthetic data deployment.

4.2.1. SINGLING OUT

Singling out occurs when an attacker isolates a single data record within the original dataset with a unique combination of attributes. While not implying reidentification, singling out enables the isolation of a data unit, providing control or facilitating other privacy attacks. The attacker uses the synthetic dataset to learn attribute combinations that single out data subjects in the original dataset. Careful consideration and protection are essential to mitigate the risks associated with singling out attacks. A thorough formalization of the singling out risk in an anonymized dataset is presented by Cohen and Nissim [9], who introduce mathematical definitions for the risk that are directly derived from an examination of the regulation.

[5] To illustrate the difference, imagine a person whose height is recorded in their deidentified electronic health record. If that record is identified as belonging to the person, then their height is *disclosed* (at the level of confidence in the accuracy of their health record). On the other hand, if all that is known about a person is that they belong to a particular population, then the best inference that can be made about their height is the mean height for that population. Now suppose that a person is in the dataset but it is unknown which record is theirs, but through statistical matching, there is a high probability that it is one of a small number of records all of which are above average height; now the accuracy of the *inference* about the person's height will have improved, even without specifically identifying which record is that of the person.

Based on this formalization, Giomi et al. [18] introduce practical attacks to measure the singling out risk in synthetic data.

4.2.2. LINKABILITY

Linkability refers to the possibility of connecting records belonging to the same individual or group using background knowledge. In a linkability attack, the adversary will use the synthetic data to link together records of the same individual that appear in multiple different datasets. The risk of linkability is one of the main privacy concerns that afflict more traditional privacy-enhancing technologies that do maintain a 1-to-1 connection between an original and a protected record, as demonstrated by many successful linkage attacks in the literature (e.g., Narayanan and Shmatikov [32], Sweeney, “Achieving k-anonymity” [39]). Synthetic data does not, by construction, preserve the link-to-original data; however, linkability risks may persist in synthetic data and need to be measured. Practical implementations of linkability measures based on attacks are provided by Giomi et al. [18] and Stadler, Oprisanu, and Troncoso [36].

4.2.3. ATTRIBUTE INFERENCE

This attack involves correctly *and* confidently inferring previously unknown attribute values of an individual in the original data. Successful attacks could lead to attribute inference from synthetic data, particularly if the generation method memorizes parts of the underlying dataset. Inference attacks targeting synthetic data have been implemented by Giomi et al. [18] and Stadler, Oprisanu, and Troncoso [36]. The disclosure risk generally depends on the number of known attributes, emphasizing the importance of contextual evaluation during synthetic data generation. This paper distinguishes between two types of attacks that lead to attribute inference.

First, an attribute inference attack formulated as a prediction problem leverages machine learning algorithms to predict sensitive attributes. These attacks fall into different categories based on the resources and knowledge available to the attacker, as delineated by Liu et al [31]. For instance, Jia et al. [26] and Slokom, Hanjalic, and Larson [35] investigated the inference of individual sensitive attributes in online social networks and recommender systems. Second, attribute inference based on statistical disclosure focuses on matching to estimate the value of the sensitive attributes. For instance, Elliot [11] and Hittmeir, Mayer, and Ekelhart “A baseline” [20] focus on the correct attribution probability (CAP). CAP measures the proportion of matches leading to correct attribution out of total matches.

4.2.4. MEMBERSHIP INFERENCE

In membership inference, the attacker aims to determine if a particular individual's data was used to train the generative model. Membership information can be privacy-sensitive in itself, especially when highly sensitive information, such as medical or judicial data, is involved. Membership disclosure occurs when the attacker establishes the presence of an individual's data in the training data, leading to an information leakage by obtaining contextual information from the membership of the individual (e.g., the individual is likely carrying a certain disease when participating in a medical dataset on that disease). Unintended memorization during synthetic data generation can replicate identifiable features, risking membership disclosure (see Section 4.3).

4.3. THREAT MODELING

To evaluate the risks associated with any data, one should first define the threat model and the scenarios through which an adversary might attempt to breach privacy (Elliot and Dale [12]). This also serves as a guiding framework for developing countermeasures. It includes the adversary's *goals and objectives,* their *knowledge* of the synthetic data and generation process, as well as their *capabilities* and *resources*.

- The *adversary's objective* defines what the adversary seeks to accomplish, with *potential goals* including reidentification, attribute inference, or membership inference attacks. A threat model might further specify at which step in the machine learning process the attack might happen.
- The adversary can be assumed to have different levels of access to the model used to generate synthetic data. In a *no box setting,* the adversary has access only to a fixed set of synthetic data records, with no information on the synthetic data generator. In a *black box setting,* the adversary can access the generator and create a (potentially large) number of synthetic records at will. In a *white box setting,* the adversary has access to both the generator used and the parameters of the trained generator. *Gray box settings* describe any setting that falls between, e.g., being able to generate synthetic data and also knowing the method used.
- The *capability of the attacker* might include being able to modify training data or the training process, or just the inference step; in the case of generative models, the inference stage could refer to any action happening during or after generating the synthetic samples.
- The *adversary's resources* place limits on what the adversary can do. The resources can be the

possession of auxiliary data (e.g., data scraped from social media). There are other possible resources available to the adversary. For example, the original data owner might have trained a predictive machine learning model (e.g., a classification model) on the original data, and the adversary may have some form (black, gray, or white box) access to that model. The adversary can then use that model to learn more about the data, and combine that information with information gleaned from attacking the synthetic data; for more details, see Liu et al. [31].

There are many ways an adversary can obtain resources and knowledge. For example, any adversary is free to make use of publicly available open data, for instance, census outputs, public registers, housing information from estate agents' listings, and even through personal contacts such as neighbors, friends, family members, colleagues, etc. Organizations—and their employees—have access to databases (e.g., customer databases). Data breaches from an attack on a computer system can provide needed attack data. Data subjects can unwittingly provide the adversary with opportunities to exploit, for instance, by posting information on the web (social media), and a data controller might make their synthesis model available without considering the consequences.

Detailed templates for two different threat models, membership inference attacks and attribute inference attacks, are provided in Appendix A, which details the types of knowledge, access, and capabilities that may need to be considered. The following specific example is a threat model for a membership attack based on learning a so-called “shadow model.” This example for a membership inference attack and the assumption it makes is based on a well-known technique in literature, called *shadow modeling,* which assumes that an adversary has access to *auxiliary data* (called the *reference dataset*) that has the same (or at least similar) underlying distribution as the original dataset used to generate synthetic data, but **no box access** to the model. In the attack, the adversary generates multiple shadow datasets by sampling from the reference dataset; in half of these datasets, a specific record is included, and in the other half, a random other record is added. Then, for each of the datasets, a synthetic data generator (SDG) is learned, and a meta-classifier is then trained on this set of datasets to recognize whether the record was included or not; the meta-classifier is then applied to the synthetic data and the target record to be attacked to decide if the record was a member of the original synthetic data generator, or not.

Not every potential opportunity for an attack will be exploited—attackers will likely consider the required costs and efforts for carrying out an attack in comparison to the expected benefit (i.e., the expected value of any new information learned about data subjects). This observation relates to GDPR Recital 26, which

states that one should account for “all the means reasonably likely to be used” and this should take into account “all objective factors, such as the costs of and the amount of time required for identification.”

Finally, the **vulnerability-countermeasure** component of the threat model suggests potential solutions to protect against specific attacks. Privacy-preserving techniques, such as obfuscation, data masking, perturbation, and differential privacy, can be combined with synthetic data to enhance protection. These are detailed in Section 4.6.

While implementing countermeasures is a critical step to address vulnerabilities and mitigate potential threats, it is also important to evaluate the effectiveness of the implemented mechanisms and assess the residual risks associated with the synthetic data. It is crucial to consider multiple threat models, covering all the disclosure risks outlined in Section 4.2, and covering adversaries with strong (but realistic) knowledge and resources.

4.4. OVERVIEW OF EMPIRICAL PRIVACY METRICS

One should not assume that synthesis alone eliminates all privacy risks. In some cases, additional mechanisms are needed to further reduce the risks. Regardless of which synthesizer method and which additional privacy mechanisms (discussed in Section 4.6) are used, it is always important to assess whether and to what extent the chosen constellation/deployment manages to achieve the necessary privacy protection: Are the privacy mechanisms implemented effective and work as intended? And what is the level of residual privacy risk?

To better answer these questions and to derive a risk assessment, carry out **empirical privacy assessments** of synthetic data. This step is necessary since the level of risk of synthetic data is not usually a parameter of the generation mechanism that can be controlled (and so is not comparable to, for example, the value of *k* in k-anonymized datasets). There are some exceptions to this (e.g., Chen, Taub, and Elliot [8]), however, these are experimental and this type of parameterization is not standard in delivery systems.

To empirically estimate the disclosure risk of synthetic data, one can distinguish two families of approaches: *statistical-based privacy evaluations* and *attack-based privacy evaluations*. A broad review of privacy metrics has been recently compiled by Boudewijn et al. [6]. Here is a brief sketch of the landscape of available methods, highlighting, where possible, established best practices.

Statistical-based privacy evaluations measure the privacy of the synthetic dataset by examining the similarities between the distributions of synthetic and original records (the similarities to the original training data or to any holdout data). Distance-based evaluations are the most common methods, where for each record in the synthetic dataset, the distance to the closest original data record is measured, and then the records or portions of records (often the assumed quasi-identifiers) that have been exactly reproduced are identified. A useful input to this process is the knowledge of where privacy risks lie within the data—that is, which attributes and/or records can be more vulnerable—so that a risk-based decision around data that has been replicated can be made. For example, in a financial transactions dataset the combination of (A) precise time, location, and amount may have a high likelihood of singling out a transaction, while (B) country, merchant name, and product may not have much likelihood of singling out a transaction. Knowing the privacy risks associated with (A) versus (B) means that when values are reproduced through synthesis, a risk-based decision may be made whether this is a problem or not.

Due to their relative simplicity and robustness, distance-based approaches were among the first risk metrics applied to synthetic data. However, for most of these metrics, it is unclear how, in practice, they map onto privacy implications and what concrete privacy risks exist for individual data records (Ganev and De Cristofaro [17]). To overcome these limitations, more recent work uses the success rate of concrete **privacy attacks** to quantify privacy **(attack-based privacy evaluations)**. These were presented in Section 4.2. By modeling a very specific attack, such attack-based evaluations make it easier to interpret the privacy risks. The main drawback of these evaluations is that, in general, they do not provide an upper bound to the risk, since they cannot exclude the existence of other, more powerful attacks. Also, using incorrect assumptions for the threat model (e.g., assuming an attacker has less knowledge than is actually available) can lead to incorrect estimates.

4.5. USING PRIVACY METRICS IN PRACTICE

Section 4.2 described how inferences are often not absolute (i.e., they cannot be 100% certain that they are correct), but still, harm can befall a data subject. For example, in 2021, a U.S. priest resigned after being affected by a data leak from the location-based hookup app Grindr.[6] He was identified because his mobile device was emitting signals at both his home and office addresses and was also active while

[6] https://www.pillarcatholic.com/p/pillar-investigates-usccb-gen-sec

traveling for a bishop's conference. In this example, one could have a high level of confidence that the information on the priest was correctly inferred, even without *absolute* certainty.

In many other examples of attribute or membership inference, the conclusion is often not that clear-cut, and the practical relevance is sometimes difficult to judge. Staying with the example above, if for some other individuals, participation in the database can be inferred with a 60% correctness, is that sufficient to draw the conclusion with the "significant probability" that the Article 29 working party opinion document sets out as a quasi-standard? In addition, the uncertainty of the inference needs to be compared to what might be assumed from prior knowledge. If prior knowledge allows an adversary to infer a certain attribute at 10% certainty and that increases to 60%, that is likely a more important inference than if the prior knowledge was already at 50%. This is the basis for metrics such as differential correct attribution probability (DCAP) (Taub [39]) that compares the level of inference allowed by synthetic data with that of a prior knowledge baseline (usually the univariate population distribution of the attribute in question).[7]

Similar issues also arise in other privacy-enhancing technologies, such as k-anonymity or differential privacy. While these measures come with in-built quantification (*k* and epsilon are built onto the models), there is no fixed threshold that defines when disclosure risk is ruled out; in other words, there is no specified minimum value of *k* or a maximum value of epsilon defined, nor indeed a principled methodology for determining what the level should be on a case-by-case basis. Therefore, it is not easy to assess the practical relevance of the score on a specific metric, and one must therefore adopt a pragmatic approach. Some general, practical guidelines follow.

4.5.1. SELECTING AN APPROPRIATE MEASURE FOR THE PRIVACY RISK

The academic literature abounds with proposed frameworks that offer empirical privacy measures. While each is a valid contribution to this evolving research field, some of these methods are more useful in practice than others, depending on the use case. When deciding on which evaluation framework to use, several factors must be kept in mind: Is the framework applicable to your data? Do you need to evaluate the privacy of specific individuals or subpopulations? Does the metric provided by the framework model

[7] The level of inference that prior knowledge allows is highly related to the sensitivity of the information. Take for instance the binary variable "has AIDS." Common sense tells us that that would be sensitive if true (rare) and not sensitive if false (common).

a relevant privacy risk? Are the assumptions that went into the calculation of the metric realistic and understandable?

The first question is technical in nature, and its answer depends on the type and volume of data that need to be examined. Some proposed frameworks, such as those implementing shadow training techniques (Stadler, Oprisanu, and Troncoso [36]) do not scale well to large volumes of data, resulting in limited applicability (see Section 4.3). While it might be tempting to address this limitation by evaluating only a subsample of synthetic records, this practice is discouraged. Due to the nature of synthetic data, information on original records is spread over the entire synthetic data. Only by examining it in its entirety can one be certain to have assessed the risk thoroughly.

Many of today's datasets are characterized by the presence of different subpopulations that are imbalanced. In these cases, the privacy protection offered by synthetic data is often not uniformly distributed across classes (Strobel and Shokri [37]). This factor has to be kept in mind when performing privacy evaluations. Sometimes, a single measure of privacy, computed for the entire dataset, might hide concerning privacy leaks affecting a specific subpopulation. The ease of customization of the analysis frameworks needs to be taken into consideration, as some may allow a finer grade of control over the target population. Currently, the method outlined by Stadler, Oprisanu, and Troncoso [36] is the primary option for evaluating the privacy of a *given individual record* within training sets, albeit at a significant computational cost.

Another important factor when measuring the privacy of a synthetic dataset is understanding what a given evaluation framework measures and the assumptions that went into the calculation, in order to determine if that particular method is applicable to the use case. Of particular importance is the understanding of the hypothesis and threat model (see Section 4.3) underpinning the privacy metric. In fact, the measured privacy risk varies greatly depending on what exactly the attacker can have access to (Nasr [32]). A relevant question in this case is how realistic the threat model is. Some proposed attacks, for example, might require adversarial access to the synthetic data generator or knowledge of its architecture and configuration. These assumptions might or might not be applicable to the particular synthetic data deployment under consideration.

Besides checking the metric assumptions, it is also important to understand the results of the metric and their implications for the privacy of the individuals whose data is used to generate the synthetic dataset. Due to the complex technicalities involved in synthetic data generation and evaluation and the diverse

regulatory landscapes, a gap exists between the technical and academic work on synthetic data privacy evaluations and the requirements and obligations dictated by privacy and data protection laws. As a consequence, the output of some privacy metrics might be difficult to understand for a non-technical audience. At the time of this writing, the synthetic data evaluation framework that comes closest to filling this gap is Anonymeter (Giomi et al. [18])] due to its GDPR-aligned attack models measuring singling out, linkability, and inference risks, and due to the favorable opinion of the Commission Nationale de l'Informatique et des Libertés (CNIL) [10], the French national data protection authority. The correct attribution probability (CAP) family of Metrics (Elliot [11]; Little et al. "Comparing the utility" [29]; Taub et al [39]) also incorporates all three components by driving attribution disclosure risk assessment through the mechanisms of singling out and linkage and also allows inference from non-singled out synthetic records. So, it has the advantages of allowing for an adversary who has response knowledge (i.e., already knows that an individual is represented in the training data) and providing a single, interpretable metric that allows for comparison of what the adversary might be assumed to know based on prior knowledge. Finally, as in other areas of computer security, opting for open source and peer-reviewed metrics is advantageous, as they ensure transparency, reliability, and validation by the wider research community, fostering trust in their effectiveness and applicability.

4.5.2. CHOOSING A THRESHOLD ON THE MEASURE

Once a suitable privacy risk measure has been chosen, it is necessary to select an appropriate threshold on the measured risk that the synthetic data needs to meet to be considered anonymous. It is impossible to establish a universal risk threshold that is always valid, independent of the dataset and nature of the use case. Available measures do not take account of the nature of the deployment nor the use case and so a universal risk threshold will inevitably be too loose for some synthetic data deployments, endangering privacy, and too strict in others, unnecessarily sacrificing utility.

When deciding on the risk threshold that is suitable for a specific use case, several factors need to be taken into account. Contextual analysis is essential, considering factors like the sensitivity of the data. For instance, healthcare data, or data that falls into the **special categories of personal data** outlined in Article 9 of the GDPR, may require stricter risk thresholds due to its sensitive nature. Moreover, the data governance and characteristics of the data deployment, such as the **presence and scope of technical and organizational measures** that secure the data, affect the required threshold. For example, a healthcare organization deploying synthetic data for research might implement strict access control measures that

can allow for a higher threshold for privacy risks, enabling greater data utility. Finally, it must be remembered that privacy risk evaluation is and must always be an ongoing process (Article 29 Data Protection Working Party [4]) that needs to be carried out regularly to help ensure that the synthetic data remains anonymous in the face of newly arising data protection issues originating from technological progress and the compounding of data releases. As a consequence of this process, the threshold itself might also need to be adjusted.

4.5.3. BEST PRACTICES WHEN EVALUATING THE PRIVACY RISKS

Once the privacy evaluation metrics and the threshold are established, it is crucial to comprehensively assess privacy risks for the previously defined threat models. One good practice is to consider various adversaries with different abilities and background knowledge levels through comprehensive scenario analysis (Elliot, and Dale [12]). For instance, an adversary with background knowledge about individuals in the dataset may pose a more significant threat than one without such information. Additionally, exploring different attack methods—such as singling out, linkability, attribute inference, or membership inference—allows for a more comprehensive evaluation of privacy protections.

It is also important to vary parameters within threat models, such as the size and type of the adversary's auxiliary dataset or the level of precision in their knowledge. By exploring this range of threat models, one can identify vulnerabilities and weaknesses in the synthetic data and develop more robust solutions to effectively mitigate these risks. As a concrete example, the analysis of the inference risk might reveal that some attributes in the dataset are at a high risk of being inferred. Mitigations, such as masking or generalization, can be employed on these values to reduce their exposure.

In order for the evaluation to be relevant and reliable, several practical aspects should be considered:

- Often the results of privacy evaluations, especially if of an adversarial nature, have an intrinsic variability. For example, the privacy risks can vary depending on the target records that have been considered. In other cases, the attack in itself is non-deterministic, as is the case when training machine learning algorithms is involved. In these cases, it is good practice to repeat the privacy evaluation multiple times and the results aggregated together, to obtain a statistically more reliable estimate of the privacy risks, and to avoid over-optimistic risk estimates stemming from a dataset that by chance was more private than other generations would be.

- Avoid concluding that privacy risks for a subsample of the dataset will generalize to the whole population. For reliable results, the synthetic dataset used in the evaluation should be the one that will actually be used.
- Avoid concluding that if a synthesizer was evaluated to be fulfilling the privacy requirements on a specific dataset, it will also hold true for a different dataset. As with any machine learning evaluation, more evidence collected from multiple repeated runs (repeated holdout, cross-validation, etc.) will arrive at a better estimate for an unseen dataset, but peculiarities of the dataset in question might still render such assumptions wrong.

Finally, it is important, both in the context of auditing as well as for future iterations of the evaluation, to be able to document the process and findings. Document the entire process, including assumptions (e.g., the adversary's resources and knowledge) and parameter settings (e.g., for the trained synthesizer, generation, performed attack, etc.) via methods such as model cards,[8] and use the best coding practices, such as version control. If it is feasible to retain the required data, the privacy evaluation should be conducted in a controlled environment with sufficient statistical power to ensure reproducible results.

4.6. HOW TO MITIGATE PRIVACY RISKS?

Privacy-preserving synthetic data aims to be useful and privacy-safe at the same time. Arguably, it would be hard to surpass the accuracy of a synthetic dataset where its generator simply memorized or overfitted to the real training data. Yet, this would be a disaster from a privacy standpoint. Hence, whenever a synthesis process aims to create a privacy-safe end result, a powerful synthetic data generator alone will usually not be sufficient, but additional privacy protection mechanisms before, during, or after synthesis may need to be additionally implemented.

This section does not aim to provide an exhaustive list of all privacy protection mechanisms for synthetic data generators that are relevant today as well as in the future, but instead gives an overview of some of the most common mechanisms used to mitigate different types of privacy risks during synthesis.

[8] See https://huggingface.co/docs/hub/en/model-cards.

4.6.1. PRIVACY-PRESERVING MACHINE LEARNING

Many synthetic data generators make use of deep learning models (such as generative adversarial networks [GANs] or variational autoencoders [VAEs], for example) to capture the statistical properties of the original data. As seen in Section 4.1, some synthetic data generation models, especially, but not only, those based on deep learning models, are also prone to memorizing specific characteristics of some data samples, opening up the potential for privacy violations. One approach to further increase the privacy of these models is to modify the learned model to be less specific to the trained data, with several different techniques available.

4.6.1.1 DIFFERENTIALLY PRIVATE SYNTHETIC DATA

Differential privacy (DP) is a technique to enforce privacy-preserving learning of the generator. It protects the specifics about the examples in training data, while still allowing the model to learn to perform its tasks well. This is achieved by a modification of the algorithm that is used by the model to "learn." Regardless of the particular architecture or application of a model, the "learning" can be thought of as the search for a set of model parameters that yields good (if not the bestl) performance. Differentially private synthetic datasets can be generated by using a DP version of the algorithm used to search for these model parameters (Abadi et al. [1]) during the training phase, in which the generative model is exposed to the (sensitive) original data and learns from it. In this case, the DP guarantee assures the extent to which learning would have resulted in the same generative model, regardless of whether any specific data subject had contributed to the training dataset.

Training the synthetic data generator with differential privacy generally increases the privacy protection offered by the synthetic data (DPSYN, DPGAN, PATEGAN). However, deploying DP synthetic data suffers from some practical limitations that can negatively impact the data utility as well as privacy. These originate ultimately from the difficulty of obtaining a strong enough theoretical privacy guarantee from differential privacy while maintaining some utility in the synthetic data. As Little et al. demonstrate, differential privacy is outperformed by sampling from the original data in terms of the level of attribute disclosure risk for a given level of utility (Little et al. "Comparing the utility" [29]) and the relationship between the privacy budget (epsilon) and the level of risk is non-linear with the unfortunate property that, at high levels of epsilon, utility is damaged without any appreciable impact on risk. This leaves users uncertain as to what "differentially private" actually means. Other recent research has shown that the

actual privacy offered by training machine learning models with a differentially private optimizer is, under practical attack scenarios, greater than what can be demonstrated theoretically (Nasr et al. [32]).

Another important factor about generating synthetic data in a differentially private manner is that to achieve a fully differentially private system, all channels of information connecting the original data to the synthetic one, for example, pre-processing numerical attributes or encoding categorical ones, must use DP algorithms.

To conclude, differential privacy is often not sufficient by itself to fully ensure the privacy of the synthetic data, and its use does not exempt the data controller from performing a thorough empirical evaluation of the privacy risk. However, it has been shown to improve the privacy of the synthetic data generators and their robustness to privacy attacks, and can be a useful tool to mitigate privacy risks.

4.6.1.2 REGULARIZATION TO REDUCE OVERFITTING

While differential privacy helps ensure that the output of a data processing algorithm does not reveal information on whether a specific individual is included in the data, other approaches aim at reducing the overfitting of the synthetic data generation model. Overfitting can, on the one hand, lead to poor quality data, and on the other hand, also contribute to memorization and privacy problems as described in Section 4.1. Regularization generally modifies the objective function within the optimization within a machine learning algorithm and tries to find a trade-off between a model learning too much from the training data and a model that generalizes better to unseen data. It is a technique frequently used for models, such as neural networks. While studied extensively for the purpose of preventing overfitting, the side effect of increasing privacy by reducing memorization is explored to a much lesser extent. However, recent work has shown that information leakage from generative models can be reduced via regularization, for example, in GANs using Lipschitz regularization (Wu et al. [41]) or maximum entropy loss (Shateri et al. [33]). The potential gain in privacy is not a property of the algorithm but needs to be evaluated empirically, as described in Section 4.5.

4.6.1.3 DIRECT MODIFICATION OF SYNTHESIZER MODELS

Other approaches might directly interfere with how the generative model is built, which is especially feasible for interpretable models like graph-based ones. For example, Hittmeir, Mayer, and Ekelhart “Efficient Bayesian” [21] introduce an approach to modify the structure of a Bayesian network used to

generate synthetic data, with the goal of protecting certain selected, sensitive attributes. This is achieved by deliberately distorting certain connections and thus correlations between attributes in the dataset of the Bayesian network.

In addition to adjusting the model (training process) to increase privacy, it is possible to include disclosure control techniques in the model. For example, by incorporating generalization into the model, it is less likely that exact values (such as the exact transaction amount) recur in the synthetic data. Furthermore, when variable relations are partially removed by the model, this reduces opportunities for attackers to combine different pieces of background information. Both of these techniques have been combined with synthetic data generation (Kroes et al. [28]).

The potential gain in privacy is not a property of the algorithm and therefore needs to be evaluated empirically, as described in Section 4.4.

4.6.2. COMBINING DATA SYNTHESIS WITH OTHER PRIVACY MITIGATIONS

Synthetic data generators learn the probability distribution of the data and draw samples from it. If the domain of these distributions is also inferred by the training data, as is often the case in automatic systems, the synthetic data becomes liable to expose secret values from the original data. Synthetic data can, however, be combined with other approaches used in statistical disclosure control as a pre- or post-processing step.

For example, extreme values in the input space, such as categories that appear with very low frequency, outlying numerical values, or values that are close to the extreme of the range, are susceptible to being reproduced in the output synthetic data. To mitigate these problems, techniques such as redaction and masking (deletion of whole attributes or parts thereof) or generalization (replacing attribute values with a more general value) can be employed.

Another source of problems originates from the fact that values of categorical variables are sampled from those in the original data. If these variables contain privacy-sensitive information—direct identifiers such as addresses, bank account details, or other identifiers—the synthetic data might reproduce them in the output, potentially leading to very serious privacy violations. In these cases, it is important that the generation algorithm is equipped with techniques to detect and protect these values, for example, via masking or pseudonymization.

5. HOW TO GOVERN SYNTHETIC DATA WITHIN AN ENTERPRISE ORGANIZATION?

Organizations that collect, process, and share data in support of their activities should have proportionate and effective data governance processes and policies in place. Given that synthetic data is typically generated using real data as input, these data governance policies and procedures should generally be followed when generating synthetic data as well.

Synthetic data may bring specific challenges in its deployment (Arthur et al. [3]), but does not necessarily require an entirely separate governance process. Overly burdensome or unnecessarily duplicative governance requirements may serve only to disincentivize organizations from using privacy-protective synthetic data, especially if less costly alternatives using real data (with fewer privacy protections) remain available. Accordingly, data governance processes should be attuned to the inherent advantages, from a risk management perspective, of using synthetic data without losing sight of any residual risks. It may be appropriate for organizations to consider creating a "fast track" version of existing governance processes to counterbalance the additional cost or effort required to generate and use synthetic data, thereby promoting privacy-protective data innovation. It is equally beneficial, where appropriate, for regulators to provide space for the creation of industry-led guidance, frameworks, and standards that will foster responsible adoption of synthetic data (as part of the toolkit of privacy-enhancing technologies).

Nevertheless, the generation and use of synthetic data introduce unique issues that go beyond traditional data governance. Synthetic data has its own lifecycle, beginning with the underlying source data and following an often-recurring pattern of data generation and use, refreshing the synthesis process periodically, or creating entirely new datasets designed to meet the needs of one or more use cases. An end-to-end synthetic data process generally goes through three phases: pre-synthesis, synthesis, and post-synthesis. The producers of synthetic data (those that participate in the synthetic data generation process) may not always be the same as the consumers of that data (those that leverage the data for purposes of a given use case). An effective governance process will ensure that necessary controls and/or standards are applied throughout the lifecycle to both types of participants in the process. Therefore, when introducing synthetic data into an organization, a critical examination of the ability of existing governance processes to account for these unique factors is warranted.

Additionally, it is important to note that companies that sell synthetic data generation platforms/products ("vendors") often play an essential role in the creation of synthetic data. More often than not, synthetic data generation capabilities are *bought* rather than *developed* in-house, further necessitating a governance framework that can tie together the synthetic data generation methodology and risk identification/mitigation framework—that might be heavily reliant on the services of a vendor—with the deployment of the synthetic data against a particular use case driven by the needs of the organization. Where appropriate, third-party risk management processes should evaluate the capabilities of the vendor to ensure that any potential risks are identified and mitigated. The contract with the vendor may also need to establish the appropriate data processing roles and responsibilities of the parties in the form of a data protection agreement.

While specific needs will vary depending on the size and nature of the organization, consideration should be given to at least the following factors when evaluating how synthetic data fits into overall data governance practices:

- **Do not neglect the proper use of underlying source data in the synthetic data lifecycle.** While the synthetic data generation process itself does not entail the "collection" of data in the traditional sense, it may involve the use of data that was originally collected for other purposes. Proper data governance processes should ensure that the data used to train synthetic data generation models is being used appropriately, complying with privacy and data protection laws (as noted in Section 3), operating within the constraints of any applicable licenses or other commercial rights to the data, and adhering to any other internal or contractually imposed limitations on the use, storage, or retention of the source data. Due consideration should be given to the fact that synthetic data may contain confidential or proprietary information that needs to be appropriately handled and protected, even if the synthetic data has been generated in such a way that privacy risks have been appropriately mitigated.
- **Consider how to manage and track synthetic data within a broader data environment or ecosystem.** To ensure that synthetic data is appropriately used and differentiated from the underlying "real" data, an organization may need to develop a structure of metadata (e.g., apply tags) to synthetically generated data and/or ensure that it remains contained in a segregated environment where controls can be applied and persisted for the duration of a particular use case, or for the entire retention period of the synthetic dataset. Ideally, an inventory should contain information regarding the source data that was synthesized, the technique used for synthesis, all

parameters or configurations to the synthesis process, especially those relating to privacy and utility, the intended use case for the synthetic data, and a period within which the synthetic data may be used or is considered useful. This metadata should be retained and made available to all users of a synthetic dataset to enable informed, risk-based decisions when using the data for any purpose. In some cases, proper management of the synthetic dataset within a broader environment may include finding ways to ensure that access to the synthetic data is limited to a specific group of individuals. Additionally, depending on the size and scope of the organization, the ecosystem may need to manage multiple "producers" and "consumers" of synthetic data—and it may be necessary to establish a central mechanism to help ensure consistent standards are applied whenever synthetic data is used.

- **Ensure that the governance process includes a means for evaluating the suitability of synthetic data for particular use cases.** Suitability can mean the privacy-protective nature of synthetic data (which is the primary focus of this paper), but it also means evaluating other key attributes of the data, such as data quality, utility, accuracy, reliability, and/or potential bias. The process for evaluating suitability needs to be able to adapt to the specific use case for which the synthetic data is being generated. Different use cases may lead to different legal assessments or data utility needs, thereby calling for different parameters for the synthetic data, even when derived from the same source data. Given that no generic claims can be made on how private the synthetic data is (as has been generally discussed in Section 3 and Section 4), enterprises will likely find themselves needing to define and tune an acceptable level of risk and utility for every use case and synthesized data set.
- **Education and training are essential ingredients to successfully unlocking the value of synthetic data.** Just like other emerging privacy-enhancing technologies, synthetic data can be technically complex, requiring niche knowledge or expertise to implement it effectively. This will be new territory for many stakeholders within an organization. A good data governance process will educate all relevant stakeholders about the benefits, capabilities, and limitations of synthetic data—as well as the role they play in the synthetic data lifecycle—to better identify meaningful use cases and to encourage investment.

6. OUTLOOK AND NEXT STEPS

The future of structured synthetic data is poised to be a pivotal force in reshaping data-driven industries. As organizations grapple with the immense challenges of data privacy and data scarcity, synthetic data emerges as a powerful solution.

With an increasingly regulated landscape, including the EU's GDPR, the EU's AI Act, the California Consumer Privacy Act, ISO 27001, and others, synthetic data is becoming a vital tool for organizations seeking to comply with privacy standards while still extracting value from data. Moreover, advancements in machine learning are driving significant improvements in the utility and fidelity of synthetic data, making it increasingly indistinguishable from real data. This will enable a broader range of applications and further solidify synthetic data's role as a strategic asset.

6.1. OPEN AREAS OF RESEARCH

6.1.1. TECHNICAL DEVELOPMENTS

6.1.1.1 NEW SYNTHESIS METHODS AND APPLICATIONS

Data synthesis is a hot research area, fueled by increasing computational capacity and an eruption of AI and machine learning–based methods. Models that start from low fidelity (possibly even random data) and move "toward" the original data (including genetic algorithms and diffusion models) are one new area of interest. There is no systematic theory of synthesis; different methods appear to work better with different types of data. The reasons for this are not fully understood, and more research is needed to understand how to best select and tune synthesis models.

Federated synthesis, where the inputs come from multiple datasets—effectively fusing the technologies of linkage and synthesis—is also a growth area. High utility/low fidelity (HULF) approaches, where the data is deliberately overfitted to particular high-level features of the original data to provide specific functionality, look promising for providing bespoke negligible-risk datasets.

6.1.1.2 DIFFERENTIAL PRIVACY

The question of which epsilon value in differential privacy ensures that synthetic data is privacy-safe or anonymous is a complex and ongoing area of research. There's no one-size-fits-all answer, as the

appropriate epsilon value depends on the specific context and the data's sensitivity. The future of this field is likely to involve a more dynamic and context-aware approach to setting epsilon values. One might see the development of adaptive algorithms that can adjust epsilon based on the specific dataset and the intended use of the synthetic data. These algorithms would consider factors like the data's sensitivity, the potential risks of data exposure, and the required data utility for specific applications. There may also be advancements in creating standardized guidelines or frameworks for different industries or types of data.

6.1.2. ETHICS, LEGAL, AND SOCIAL ASPECTS (ELSA)

6.1.4.1 ETHICS

The ethical concerns surrounding the use of synthetic datasets, particularly those deemed "provably private," extend beyond the realm of privacy assurance. When a decision is made to release such a dataset publicly, control over who accesses it and how they utilize it is relinquished. This opens the door to a wide range of uses, not all of which may be ethical. The critical point here is that the privacy-preserving features of synthetic data do not release data owners from other governance responsibilities. Responsibilities such as access control and project approval continue to play a crucial role in the ethical management of these datasets. This highlights the necessity for a comprehensive governance framework (as outlined in Section 5) that encompasses various aspects of data management, ensuring responsible and ethical usage even if data has been made public.

6.1.4.2 TRANSPARENCY

Transparency is key in the field of structured synthetic data, as it builds trust and ensures the data generation process is effective. A transparent approach helps users understand the methods and assumptions used in creating synthetic data, allowing them to recognize its strengths and limitations. Transparency also facilitates compliance with regulatory standards and audit processes, while encouraging collaboration and innovation by inviting feedback to enhance synthetic data quality. However, achieving such transparency poses challenges. It involves balancing the need to provide enough information to gain trust, but not so much that privacy protection is undermined. This may require simplifying complex algorithms for easier understanding, ensuring thorough documentation, and maintaining the privacy of original data sources.

6.1.4.3 FAIRNESS

Ensuring fairness in synthetic data generation is a critical issue, as standard methods do not automatically eliminate biases present in the original data. The challenge lies in developing techniques that can either identify and mitigate these biases for fairness or, alternatively, maintain them when necessary. The choice between removing or retaining biases depends on the intended use of the synthetic data. If the goal is to produce unbiased data, the original dataset should be adjusted for biases before the synthetic data generation. However, if the aim is to accurately mirror real-world data, including its inherent biases, then these should be retained in the synthetic version.

6.2. NEXT STEPS FOR THIS IEEE IC

6.2.1. CALL TO ACTION AND PARTICIPATION

This white paper is anticipated to be a catalyst for fostering more in-depth discussions on the topic of structured synthetic data, and it invites feedback from stakeholders worldwide. This inclusive approach ensures that the insights and concerns of a broad range of participants, including those from different geographic, cultural, and professional backgrounds, are considered. Such comprehensive feedback is vital for developing a well-rounded understanding and effective strategies in the realm of synthetic data.

To capitalize on this promising future, the IEEE IC will intensify its engagement with existing ecosystems, such as the Organisation for Economic Co-operation and Development (OECD), and explore new collaborations to foster the development of best practices and standards for synthetic data. By working closely with these organizations, the aim is to accelerate the adoption of synthetic data and ensure its ethical and responsible use.

To increase its reach and impact, this white paper will be presented at webinars and events. This approach not only aims to gain more feedback but also to create additional awareness about the IC's work. The intention is to attract more members, which will in turn enable the IC to address a broader range of topics beyond just synthetic data privacy.

Looking ahead, there are plans to expand its scope of work. This could include developing a white paper on the role of synthetic data in AI governance and responsible AI, with a focus on aspects such as fairness, explainability, privacy, and external AI assurance ecosystems, which are vital for areas like AI bias audits.

Additionally, there could be a discussion and subsequent white paper focusing on recommendations for governance processes for synthetic data and its generative models within enterprise organizations.

6.2.2. STANDARD SETTING

The importance of this white paper extends beyond its role as a comprehensive guide. It is envisioned as a foundational step toward establishing unified definitions as well as a meaningful standard in the field of structured synthetic data privacy. By providing a detailed exploration of the subject, the document lays the groundwork for a future standard. This involves issuing a recommendation for a standard project authorization request, which will be a significant milestone for this IEEE IC. This collective effort is crucial for advancing the field, ensuring that the emerging standard is robust, comprehensive, and reflective of a wide range of expert insights.

7. REFERENCES

The following sources have either been referenced within this paper or may be useful for additional reading:

[1] Abadi, Martín, Chu, Andy, Goodfellow, Ian, McMahan, H. Brendan, Mironov, Ilya, Talwar, Kunal, and Zhang, Li, "Deep learning with differential privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security,* 24 Oct. 2016, pp. 308–318.

[2] Abay, Nazmiye Ceren, Zhou, Yan, Kantarcioglu, Murat, Thuraisingham, Bhavani, and Sweeney, Latanya, "Privacy preserving synthetic data release using deep learning," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD),* Dublin, Ireland, 10–14 Sep. 2018, pp. 510–526.

[3] Arthur, Lauren, Costello, Jason, Hardy, Jonathan, O'Brien, Will, Rea, James, Rees, Gareth, and Ganev, Georgi, "On the challenges of deploying privacy-preserving synthetic data in the enterprise," *arXiv,* 9 Jul. 2023, arXiv:2307.04208.

[4] Article 29 Data Protection Working Party, "Opinion 05/2014 on anonymisation techniques," 10 Apr. 2014.[9]

[5] Beigi, Mandis, Shafquat, Afrah, Mezey, Jason, and Aptekar, Jacob, "Simulants: Synthetic clinical trial data via subject-level privacy-preserving synthesis," *AMIA Annual Symposium Proceedings,* Washington, DC, 29 Apr. 2022, pp. 231–240.

[6] Boudewijn, Alexander, Ferraris, Andrea Filippo, Panfilo, Daniele, Cocca, Vanessa, Zinutti, Sabrina, De Schepper, Karel, and Chauvenet, Carlo Rossi, "Privacy measurement in tabular synthetic data: state of the art and future research directions," *arXiv,* 29 Nov. 2023, arXiv:2311.17453.

[7] van Breugel, Boris, Kyono, Trent, Berrevoets, Jeroen, and van der Schaar, Mihaela, "DECAF: Generating fair synthetic data using causally-aware generative networks," *arXiv,* 25 Oct. 2021, arXiv:2110.12884.

[9] Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[8] Chen, Yingrui, Taub, Jen, and Elliot, Mark, "The trade-off between information utility and disclosure risk in a GA synthetic data generator," *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality,* The Hague, The Netherlands, 29–31 Oct. 2019.

[9] Cohen, Aloni, and Nissim, Kobbi, "Towards formalizing the GDPR's notion of singling out," *Proceedings of the National Academy of Sciences,* vol. 117, no. 15, 31 Mar. 2020, pp. 8344–8352.

[10] Commission Nationale de l'Informatique et des Libertés (CNIL), "Opinion on the Anonymeter privacy evaluation framework," 13 Feb 2023.[10]

[11] Elliot, Mark, "Final report on the disclosure risk associated with the synthetic data produced by the SYLLS Team," Technical report, University of Manchester, Manchester, United Kingdom, 2014.

[12] Elliot, Mark, and Dale, Angela, "Scenarios of attack: The data intruder's perspective on statistical disclosure risk," *Netherlands Official Statistics* 14, Spring 1999, pp. 6–10.

[13] Elliot, Mark, Mackey, Elaine, and O'Hara, Kieron, *The Anonymisation Decision-Making Framework 2nd Edition: European Practitioners' Guide,* Manchester, United Kingdom: UKAN Publications, 2020.

[14] Feng, Qianli, Guo, Chenqi, Benitez-Quiroz, Fabian, and Martinez, Aleix, "When do GANs replicate? On the choice of dataset size," *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* Montreal, Canada, 2021, pp. 6681–6690.

[15] Gainetdinov, Ainur, "GAN mode collapse explanation," *Towards AI*, 7 Mar. 2023.

[16] Gal, Michal, and Lynskey, Orla, "Synthetic data: Legal implications of the data-generation revolution," *109 Iowa Law Review,* 20 Apr. 2023, Forthcoming LSE Legal Studies Working Paper No. 6/2023.

[17] Ganev, Georgi, and De Cristofaro, Emiliano, "On the inadequacy of similarity-based privacy metrics: reconstruction attacks against 'truly anonymous synthetic data'," *arXiv,* 8 Dec. 2023, arXiv:2312.05114.

[18] Giomi, Matteo, Boenisch, Franziska, Wehmeyer, Christoph, and Tasnádi, Borbála, "A Unified Framework for Quantifying Privacy Risk in Synthetic Data," *arXiv,* 18 Nov. 2022, arXiv:2211.10459.

[10] Available at https://github.com/statice/anonymeter/blob/main/cnil/CNIL_opinion_anonymeter_courtesy_translation.pdf

[19] Guillaudeux, Morgan, Rousseau, Olivia, Petot, Julien, Bennis, Zineb, Dein, Charles-Axel, Goronflot, Thomas, Vince, Nicolas, Limou, Sophie, Karakachoff, Matilde, Wargny, Matthieu, and Gourraud, Pierre-Antoine, "Patient-centric synthetic data generation, no reason to risk reidentification in biomedical data analysis," *npj Digital Medicine*, vol. 6, no. 37, 10 Mar. 2023, pp. 1–10.

[20] Hittmeir, Markus, Mayer, Rudolf, and Ekelhart, Andreas, "A baseline for attribute disclosure risk in synthetic data," *ACM Conference on Data and Application Security and Privacy, CODASPY,*" New Orleans, LA, 16 Mar. 2020, pp. 133–143.

[21] Hittmeir, Markus, Mayer, Rudolf, and Ekelhart, Andreas, "Efficient Bayesian network construction for increased privacy on synthetic data," *2022 IEEE International Conference on Big Data (Big Data),* Osaka, Japan, 17–20 Dec. 2022, pp. 5721–5730.

[22] IBM, "Half of Breached Organizations Unwilling to Increase Security Spend Despite Soaring Breach Costs," 24 Jul. 2023.[11]

[23] Information Commissioner's Office (ICO), "Chapter 2: How do we ensure anonymisation is effective? Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance," Oct. 2021.[12]

[24] Information Commissioner's Office (ICO), "Privacy enhancing technologies (PETs),"(no date).

[25] Ji, Zhanglong, Lipton, Zachary C., and Elkan, Charles, "Differential privacy and machine learning: A survey and review," *arXiv,* 25 Dec. 2014, arXiv:1412.7584.

[26] Jia, Jinyuan, Wang, Binghui, Zhang, Le, and Gong, Neil Zhenqiang, "AttriInfer: Inferring user attributes in online social networks using Markov random fields," *Proceedings of the 26th International Conference on the World Wide Web,* Perth, Australia, 3–7 Apr. 2017, pp. 1561–1569.

[27] Kairouz, Peter, McMahan, H. Brendan, Avent, Brendan, Bellet, Aurélien, Bennis, Mehdi, Nitin Bhagoji, Arjun, Bonawitz, Kallista, Charles, Zachary, et al., "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning,* vol. 14, no. 1–2, 23 Jun. 2021, pp. 1–210.

[11] Available: https://newsroom.ibm.com/2023-07-24-IBM-Report-Half-of-Breached-Organizations-Unwilling-to-Increase-Security-Spend-Despite-Soaring-Breach-Costs

[12] Available at https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/how-do-we-ensure-anonymisation-is-effective/

[28] Kroes, Shannon K. S., Van Leeuwen, Matthijs, Groenwold, Rolf H. H., and Janssen, Mart P., "Generating synthetic mixed discrete-continuous health records with mixed sum-product networks," *Journal of the American Medical Informatics Association,* vol. 30, no. 1, Jan. 2022, pp. 16–25.

[29] Little, Claire, Elliot, Mark, and Allmendinger, Richard, "Comparing the utility and disclosure risk of synthetic data with samples of microdata," *Privacy in Statistical Databases, Lecture Notes in Computer Science,* vol. 13463, 14 Sep. 2022, pp. 234–249.

[30] Little, Claire, Elliot, Mark, and Allmendinger, Richard, "Federated learning for generating synthetic data: A scoping review," *International Journal of Population Data Science,* no. 8, vol. 1, 16 Jan. 2023.

[31] Liu, Bo, Ding, Ming, Shaham, Sina, Rahayu, Wenny, Farokhi, Farhad, and Lin, Zihuai, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys,* vol. 54, no. 2, art. 31, 5 Mar. 2022.

[32] Narayanan, Arvind, and Shmatikov, Vitaly, "Robust de-anonymization of large sparse datasets," *2008 IEEE Symposium on Security and Privacy (sp 2008),* Oakland, CA, 2008, pp. 111–125.

[33] Nasr, Milad, Song, Shuang, Thakurta, Abhradeep, Papernot, Nicolas, and Carlini, Nicholas, "Adversary instantiation: Lower bounds for differentially private machine learning," *arXiv,* 11 Jan. 2021, arXiv:2101.04535.

[34] Shateri, Mohammadhadi, Messina, Francisco, Labeau, Fabrice, and Piantanida, Pablo, "Preserving privacy in GANs against membership inference attack," *IEEE Transactions on Information Forensics and Security,* vol. 19, 13 Dec. 2023, pp. 1728–1743.

[35] Slokom, Manel, Hanjalic, Alan, and Larson, Martha, "Towards user-oriented privacy for recommender system data: A personalization-based approach to gender obfuscation for user profiles," *Information Processing & Management,* vol. 58, no. 6, Nov. 2021, p. 102722.

[36] Solove, Daniel J., *Understanding Privacy,* First Harvard University Press paperback edition, Cambridge, Massachusetts/London, England: Harvard University Press, 2009.

[37] Stadler, Theresa, Oprisanu, Bristena, and Troncoso, Carmela, "Synthetic data—Anonymisation Groundhog Day," *arXiv,* 24. Jan 2022, arXiv:2011.07018v6.

[38] Strobel, Martin, and Shokri, Reza, "Data privacy and trustworthy machine learning," *IEEE Security & Privacy,* vol. 20, no. 5, Sep.-Oct. 2022, pp. 44–49.

[39] Sweeney, Latanya, “Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 10, no. 5, 2002, pp. 571–588.

[40] Sweeney, Latanya, “Simple demographics often identify people uniquely,” (working paper), 2000, pp. 1–34.

[41] Taub, Jennifer, Elliot, Mark, Pampaka, Maria, and Smith, Duncan, “Differential correct attribution probability for synthetic data: An exploration,” *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018,* Valencia, Spain, 26–28 Sep. 2018, pp. 122–137.

[42] Veale, Michael, Binns, Reuben, and Edwards, Lilian, “Algorithms that remember: Model inversion attacks and data protection law,” *Philosophical Transactions of the Royal Society A,* 15 Oct. 2018, 376: 20180083.

[43] Wu, Bingzhe, Zhao, Shiwan, Chen, ChaoChao, Xu, Haoyang, Wang, Li, Zhang, Xiaolu, Sun, Guangyu, and Zhou, Jun, “Generalization in generative adversarial networks: a novel perspective from privacy protection,” *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS),* Red Hook, NY, 8 Dec. 2019, pp. 307–317.

[44] Xie, Liyang, Lin, Kaixiang, Wang, Shu, Wang, Fei, and Zhou, Jiayu, “Differentially private generative adversarial network,” *arXiv,* 19 Feb. 2018, arXiv:1802.06739.

[45] Yoon, Jinsung, Jordon, James, and Schaar, Mihaela van der, “PATE-GAN: Generating synthetic data with differential privacy guarantees,” (poster), *International Conference on Learning Representations (ICLR).* New Orleans, LA, 2019.

APPENDIX A

THREAT MODEL OVERVIEW

This appendix contains details on the threat models introduced in Section 4 and provides examples of these threat models for specific attacks.

THREAT MODEL FOR MEMBERSHIP INFERENCE ATTACKS

Adversary's goal and objective: An attacker wants to know if they can tell whether the record(s) of individuals they possess were used to train the synthetic data model. The goal is to infer additional information about the individual that is associated with membership, such as a medical condition.

Defender's goal: The defender's aim is to safeguard patient privacy within the dataset used for training the synthetic data model while maintaining data utility for further research.

Adversary's knowledge:

- The adversary is in possession of one or multiple auxiliary records of individuals.
- **Adversary's access to the model:** various levels of access to the model that was used to generate the synthetic data:
    - S1. **Full black box setting**, also known as **no box setting**: the adversary only has access to a given amount of synthetic data records and no information on the synthetic data generator.
    - S2. **Black box setting**: the adversary has access to the generator and can create a (potentially large) number of synthetic records.
    - S3. **White box setting**: the adversary has access to the generator used, as well as the parameters of the trained generator.
- Adversary's knowledge of the data records to disclose:
    - K1. No auxiliary knowledge.

- K2. Limited auxiliary knowledge: The adversary could leverage limited additional side information about the training set. For example, the adversary has incomplete knowledge of members of the test dataset, the training dataset, or both.
- K3. Auxiliary knowledge: For instance, the adversary has access to a reference dataset that follows the same underlying distribution of the original training data, and the adversary has access to the size of the original training data.

Option 2: Table X: Adversary's knowledge

Knowledge of records/model	No auxiliary knowledge	Limited auxiliary knowledge	Auxiliary knowledge
No box			
Black box			
White box			

Option 3: Table Y: Adversary's access to the model

No box	Black box	White box
Access to a given amount of synthetic data records; no information on the synthetic data generator	Access to the generator; can create a (potentially large) number of synthetic records	Access to the generator and the parameters of the trained generator

Option 3: Table Z: Adversary's auxiliary knowledge

None	Limited	Auxiliary knowledge
	An adversary could leverage limited additional side information about the training set, for example, incomplete knowledge of members of the test dataset, the training dataset, or both	Adversary has, for example, access to a reference dataset that follows the same underlying distribution of the original training data, and the adversary has access to the size of the original training data

THREAT MODEL FOR ATTRIBUTE INFERENCE ATTACKS

Adversary's goal: this attack differs in terms of the goal, as the adversary wants to infer information on one or more attributes of an incomplete record

Adversary's knowledge:

- The adversary possesses one or multiple incomplete records of individuals.
- Adversary's access to the model: the same as for membership inference above.
- **Adversary's knowledge of the data records to disclose:**
    - K1. No background knowledge of the data. Basically, the adversary can only infer information from the whole population, for example, by computing mean/median values.
    - K2. The adversary has access to a set of "quasi-identifiers" of the records from the original dataset, except for the sensitive attribute to be inferred. The task is very similar to the one of data imputation, in other words, using the synthetic data to estimate the values of an incomplete record.
    - K3. The adversary has access to a set of "quasi-identifiers" of the records from the original dataset, except for the sensitive attribute to be inferred, and has additional background knowledge of the original data. For instance, the adversary knows the true correlation of some attributes, the distribution of the sensitive attribute, or has access to a similar reference dataset, and can use this data in addition to the synthetic data to improve the estimation of the unknown values.

APPENDIX B

SYNTHETIC DATA GOVERNANCE

Basic questions to ask when evaluating potential synthetic data use cases include the following:

- Where does the real data come from?
- Is the synthetic data for internal or external sharing?
- Where is the synthetic data (or “artifact”) stored, and who has access to it?
- How much control will there be over who has access to the synthetic data and/or the generator model, and for which purposes it can be used?
- Is the downstream application of the synthetic data known in advance?
- How accurate do analytical results generated from the synthetic data need to be?
- Will business or other decisions be made as a result of analysis performed using the synthetic data?
- Will users of the synthetic data know that the data is synthetic, from what source the data was generated, and with which privacy and accuracy configurations?
- Is external expertise needed to help validate the approach?
- How many times have synthetic data been generated and by whom?

BASIC DATA GOVERNANCE

A common data governance framework in enterprises includes the following:

- Data scoping and ownership: to identify high-risk/value data, define owners
- Data model: to harmonize data in a standard format
- Data interoperability: to exchange data securely and effectively
- Data catalog and lineage: to document and track assets

- Data quality: to define rules and thresholds for the data to be fit-for-purpose
- Data sharing: to ensure the appropriate use of data in line with risk assessments
- Data retention and archiving: to define correct periods to keep the data for its purpose and implement archiving/deleting strategies

RAISING THE WORLD'S STANDARDS

3 Park Avenue, New York, NY 10016-5997 USA http://standards.ieee.org

Tel.+1732-981-0060 Fax+1732-562-1571