Alexander Boudewijn*
Andrea F. Ferraris**

Legal and Regulatory Perspectives on Synthetic Data as an Anonymization Strategy

In an increasingly digital world, the protection of personal data is paramount for individuals, organizations, and regulators. As data collection technologies evolve, so must methods for ensuring data privacy. This paper explores synthetic data as a promising privacy-enhancing technology (PET) for anonymization, focusing on legal, theoretical, and practical perspectives. Synthetic data, generated algorithmically, do not pertain to real individuals, making them valuable for data science and AI development while preserving privacy. We examine the regulatory context, particularly under the GDPR, and identify privacy risks and attacks that anonymization must defend against. We argue that synthetic data, when properly generated, can meet anonymization standards and provide deployment recommendations to mitigate privacy risks. Our findings contribute to a standardized framework for synthetic data privacy assurance, aligning with current and future data protection regulations.

Keywords: *Synthetic data, anonymization, privacy-enhancing technology, GDPR.*

1. Introduction

In an increasingly digital world, the protection of personal data has become a critical concern for individuals, organizations, and regulators. As data collection and processing technologies evolve, so too must the methods for ensuring that personal information remains secure. One such method is anonymization, which aims to transform data in such a way that individuals are no longer identifiable. This process is essential for complying with privacy regulations like the General Data Protection Regulation (GDPR) and for protecting individuals' privacy against misuse. This paper explores synthetic data as an emerging privacy enhancing technology (PET)[1] and more specifically, anonymization technology from a legal and theoretical viewpoint.

Synthetic data are not collected empirically, but generated through algorithms. As such, they do not pertain to real individuals, but can be useful in data science activities and artificial

* PhD, Data Privacy Researcher at Aindo SpA, Trieste, Italy.
** Research Specialist / PhD Candidate in Law, Science and Technology, Data Valley Consulting srl, Milan, Italy & University of Bologna.
[1] OECD, "Emerging Privacy Enhancing Technologies: Current Regulatory and Policy Approaches," OECD Digital Economy Papers (OECD, 2023).

intelligence (AI) development. The introduced perspectives and interpretations on the technology and the concept of privacy contribute to a standardized and sound framework for synthetic data privacy assurance.

The remainder of this paper is structured as follows: in section 2, we provide an overview of the regulatory definition of anonymization, focusing on the GDPR. In particular, we provide an analysis of personal data, anonymous data, and the requirements on anonymization processes, including possible types of attacks they should protect against. In section 3, we use the developed legal theory to show that synthetic data generation from real data through generative artificial intelligence can serve as an anonymization technology if carried out correctly. In particular, we consider both scientific and regulatory evidence. In Section 4, we use the developed concepts to provide recommendations for proper deployment of synthetic data based on potential privacy hazards and open research topics.

2. The Concept of Anonymization

In this section, we provide an introduction to the concept of anonymization as also outlined in previous work.[2] Anonymization is the process of transforming personal data to ensure that individuals are no longer identifiable, either directly or indirectly, within a dataset. This technique is essential for safeguarding privacy in our increasingly data-centric world, aligning with regulations like the GDPR. By effectively removing or altering identifiable elements, anonymization protects against unauthorized access and misuse of personal information while still allowing for meaningful data analysis and utilization. Before exploring the various approaches to achieve anonymization, it is crucial to first understand the elements of personal data as defined by the GDPR.

2.1. Personal Data

The General Data Protection Regulation (GDPR)[3] defines personal data as "***any information relating to an identified or identifiable natural person*** *('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*".

This definition includes four key elements, as analyzed by López & Elbi[4]:

1. "**any information**": This term is comprehensive, encompassing all types of information about a person, regardless of its nature (objective or subjective) or the context in which the person acts (e.g., as a consumer, patient, employee). It covers a wide spectrum, from sensitive data to general information about private, family, or professional life.

[2] *Panfilo D., Boudewijn A. T., Ferraris A. F., Cocca V., Zinutti S., De Schepper K., & Chauvenet C. R.,* Measuring Privacy Protection in Structured Synthetic Datasets: A Survey, in: *Hideyuki Matsumi, Paul De Hert et al. (ed)*, Privacy and Data Protection: Ideas that Drive Our Digital World, 2024.

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), (GDPR) art. 4.1.

[4] *López C.A.F., Elbi A*., On The Legal Nature of Synthetic Data, NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, 2022.

2. "**relating to**": Information relates to a person in three ways: through its content (specific data about an individual, like medical records), purpose (if it's used to evaluate or affect an individual's status or behavior), and result (if its use impacts the individual's rights and interests). These forms of relation contribute to an understanding at a population level rather than providing precise individualized information.
3. "**an identified or identifiable**": This focuses on the ability to distinguish a person within a group through identifiers, which may be direct (like a name) or indirect (like a unique combination of information). The protection extends to personal data regardless of the identification method.
4. "**natural person**": The protections cover all human beings, emphasizing the universal aspect of a "natural person" in alignment with human rights principles. It applies to living individuals who are identifiable or can be identified.

These components collectively ensure a consistent and comprehensive approach to data protection, balancing the broad scope of personal data with the need for identifiable linkage to an individual.

2.2. Anonymization and Anonymized Data

Anonymized data refers to information devoid of personally identifiable markers, ensuring no individual can be discerned either directly or through auxiliary data accessible to third parties. Properly anonymized data is exempt from data protection regulations like the GDPR since it ceases to be personal data. Such anonymization safeguards individual privacy against unlawful or unethical use.[5] Recital 26 of the GDPR delves deeper into personal data, further elucidating the concept of anonymization. It posits that anonymized data is "*information that does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a way that the data subject is not or no longer identifiable*."

The possibility of re-identification, or the process of transforming anonymized data back into personal data, is assessed based on its likelihood within a given dataset[6]. This might occur via data matching techniques or other similar methods. Recital 26 elucidates that to determine whether a person can be identified, one should take into account all the methods likely to be utilized, such as singling out, either by the controller or by another person.

"To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly."

The notion of "*reasonableness*" is crucial in this context, referring to the means used to identify the person, which may involve significant costs, time, and operations. The assessment should take into account the state of existing and available technologies at the time of data processing, as well as possible future developments.[6]

[5] *Foglia C.*, Il Dilemma (ancora aperto) dell'Anonimizzazione e il Ruolo della Pseudonimizzazione nel GDPR, in Circolazione e Protezione dei Dati Personali, tra Libertà e Regole del Mercato. Commentario al Regolamento UE n. 2016/679 (GDPR) e al Novellato d.lgs. n. 196/2003 (Codice Privacy), 2019. *Stalla-Bourdillon S., Knight A.,* Anonymous Data v. Personal Data-False Debate: an EU Perspective on Anonymization, Pseudonymization and Personal Data, Wisconsin International Law Journal, 2017.

[6] *Tempestini L., D'Acquisto G.,* Il dato personale oggi tra le sfide dell'anonimizzazione e le tutele rafforzate dei dati sensibili,in Le nuove frontiere della privacy nelle tecnologie digitali, Aracne, 2016.

The principle of data protection does not apply to anonymized data if it is only possible to identify a person through unreasonably extensive means. Hence, the challenge in identifying an individual from anonymized data lies in the minimal presence of potentially identifiable elements.[7]

The Article 29 Working Party (WP29), in Opinion 4/2007, delved into the concept of personal data, concentrating on the "identified or identifiable" aspect of the definition[8]. The robustness of the anonymization process is measured by the ``means reasonably to be used'' test.

"To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."

The need for feasibility is particularly relevant in the context of statistical data. If even aggregated data, due to a small sample size or availability of other identifying information, potentially leads to identification, it underscores the inadequacy of the anonymization process.[9]

An effective anonymization strategy prevents any entity from isolating an individual in a dataset or linking two records within a dataset. Simply removing direct identifiers is insufficient to ensure anonymity.[11] It often necessitates additional measures, the nature of which depends on the context and purpose of the data processing. It is crucial to weigh the effort and cost of anonymization against the increasing technical capacity to identify individuals, the availability of public datasets, and instances of incomplete anonymization.[10][11]

The term "identifiable" extends to any entity in control of the data, not just the original data controller. If the controller retains the identifiable raw data and shares a modified version, it remains classified as personal data. Conversely, if the controller renders the data unidentifiable at an aggregate level and only shares these aggregate statistics, it is deemed anonymous data. When applying anonymization techniques, data controllers need to evaluate the level of guarantee provided by the selected method in relation to the current technological state.

2.3. Types of Privacy Leaks and Attacks

Three keycategories of attacksidentified by WP29 act as a benchmarkfor a proper anonymization, namely: Singling Out, Linkability, and Inference, as defined below.[12]

- **Singling Out**: Singling Out refers to the capability to isolate some or all records which identify an individual in a dataset. It suggests the ability to distinguish an individual in a group, even without precisely knowing who the individual is. It pertains to the distinctiveness of an individual's record, which might lead to identification.

[7] *D'Acquisto G., Naldi M.*, Big Data e Privacy by Design, Anonimizzazione, Pseudonimizzazione, Sicurezza, Giappichelli, 2017.
[8] *Article 29 Data Protection Working Party (WP29)*, Opinion 4/2007 on the Concept of Personal Data, 2007.
[9] *Information Commissioner's Office (ICO)*, Anonymisation: Managing Data Protection Risk Code of Practice, 2012.
[10] *Agencia Española de Protección de Datos (AEDP) and European Data Protection Supervisor (EDPS)*.
[11] Misunderstandingrelated to Anonymization, 2021, <https://www.edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en>[31.07.2024].
[12] Article 29 Data Protection Working Party (WP29). "Opinion 05/2014 on Anonymisation Techniques", 2014.

The singling out phenomenon was strikingly highlighted in the 1990s when an MIT student, Latanya Sweeney, successfully re-identified the supposedly anonymized health records of Massachusetts Governor William Weld.[13] She cross-referenced these records with public voter registration data, underscoring that ZIP code, birth date, and gender alone could uniquely identify a vast majority of the US population.

- **Linkability**: Linkability pertains to the ability to link at least two records concerning the same data subject or a group of data subjects from the same data set or two different data sets. This means that if different pieces of information, collected separately, can be linked together to possibly identify an individual, they are not effectively anonymized.

  Linkability surfaced prominently during the 2006 Netflix data breach. Anonymized customer data was intricately linked with public data from the Internet Movie Database (IMDb) by researchers Narayanan and Shmatikov. They managed to reverse the anonymity for users who had posted movie ratings under their names on IMDb, bringing to light how seemingly unrelated pieces of information could jeopardize data anonymization.
- **Attribute inference**: attribute inference refers to the ability to deduce, with significant probability, the value of an attribute from the values of a set of other attributes. In a legal context, it relates to the possibility of deducing unknown information about an individual from known information. Inference attacks can take advantage of statistical dependencies in the data to deduce sensitive information from non-sensitive information.

  It is interesting in this sense a 2014 case concerning the New York City Taxi and Limousine Commission's dataset as it illuminates the issue of inference.[14] The anonymized dataset contained extensive details of every taxi ride in the city. However, researchers could infer not only specific drivers' activities and earnings but also pinpoint individuals, including celebrities, demonstrating that even supposedly anonymized data could be vulnerable to inference attacks.

A robust anonymization procedure, thus, safeguards against the possibility of isolating individuals (singling out), associating records within or amongst datasets (linkability), and deducing information on specific individuals contained in the dataset (inference).[15] Simply eliminating overtly identifiable elements is inadequate; often, the processing context and purpose necessitate additional steps. Data protection laws apply as long as the identification or attribution remains possible, irrespective of the data controller or recipient's intentions.

Recent views of privacy classify attacks into two categories: identity disclosure and attribute inference.[16] These correspond roughly to singling out attacks and attribute inference attacks under the WP29 definition, respectively. The underlying intuition here is that upon successfully conducting an attack, the attacker must either have correctly identified an individual in a dataset (identity disclosure), or must have gained some information about a specific individual (parameter inference). In this light, linkability is a means to achieve one of these outcomes, but not a distinct attack category. Due to the abundance of information

[13] *Barth-Jones D.*, The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now, 2012. *Narayanan A., Shmatikov V.*, How to Break Anonymity of the Netflix Prize Dataset (arXiv preprint cs/0610105), 2006.
[14] *Hern A.,* New York Taxi Details Can Be Extracted from Anonymised Data, Researchers Say, The Guardian, 2014.
[15] Article 29 Data Protection Working Party (WP29), "Opinion 05/2014 on Anonymisation Techniques", 2014.
[16] *Hu J., Bowen C.M.*, Advancing Microdata Privacy Protection: A Review of Synthetic Data, 2023.

available in the current digital world and the limited means needed to relate it to individuals[17] are at times raised as arguments supporting this perspective.[18]

3. Synthetic Data Generation as an Anonymization Technology

Following Jordon et al., we define the concept of synthetic data as follows:[19] *"Synthetic data (SD) is data that has been generated using a purpose-built mathematical model or algorithm (generator), with the aim of solving a (set of) data science task(s)."*

In the remainder, we restrict our analysis to *synthetic data generated through machine learning algorithms from original real-world data* (hereafter "synthetic data" or "SD"). This form of synthetic data has become the most discussed in application domains and the legal community.

In the remainder of this section, we outline why SD generation classifies as an anonymization technique as defined in Section 2. To this end, we first discuss computer scientific perspectives on the matter. We then discuss the dynamic and evolving legal landscape surrounding the technology. Finally, we discuss proper deployment conditions by considering potential risk factors and providing concrete deployment recommendations. Combined, these viewpoints not only show that SD can serve as an anonymization tool.

3.1. Scientific Perspectives

Anonymization and privacy protection are widely considered a core application of SD.[20] SD technology lends itself particularly well to this use case because it breaks the direct correspondence between data and real individuals. Legacy anonymization technologies rely on obscuring real data, for instance through randomization and generalization.[21] By contrast, SD generation proceeds by inferring a stochastic model of a given real dataset. This model captures the patterns of the real data in a probabilistic manner. Subsequently, this model can be sampled to generate new, entirely artificial records. Combined, a set of such artificial records exhibits the same patterns as the real dataset, since the underlying distribution is the same.

This process is best illustrated by a thought experiment.[22] Suppose one wants to create an artificial dataset that accurately represents the properties of some real population. Through research, the individual has knowledge of some properties of this population. For example, they know that the female to male ratio of the population is 1:1. Furthermore, they know that roughly one in six people have blue eyes, while the others have brown eyes. To create the data of an artificial person, they now do the following experiment: first, they flip a coin. If it lands on heads

(a chance of 1/2), they mark down "female". If it lands on tails, they mark down "male". Next, they roll a die. If it lands on the face with six eyes (a chance of 1/6), they mark down

[17] See, e.g., *De Montjoye Y. A., Hidalgo C. A., Verleysen M., & Blondel V. D.*, Unique in the Crowd: The Privacy Bounds of Human Mobility, Scientific Reports, 3(1), 2013, 1-5.
[18] *Beduschi A.*, Synthetic Data Protection: Towards a Paradigm Change in Data Regulation? *Big Data & Society*, 11(1), 2024.
[19] *Jordon J., Szpruch L., Houssiau F., Bottarelli M., Cherubin G., Maple C., Weller A.*, Synthetic Data-What, Why and How? 2022.
[20] Ibid.
[21] Article 29 Data Protection Working Party (WP29), "Opinion 05/2014 on Anonymisation Techniques", 2014.
[22] Thisthoughtexperiment was previously described in a blogpost, <https://aindo.com/blog/is-synthetic-good/> [30.07.2024].

"blue eyes". If it lands on any other face (1, 2, 3, 4 or 5, a chance of 5/6), they mark down "brown eyes". By repeating this experiment many times, the individual obtains an entirely synthetic population that has all the statistical and mathematical properties of the real population.[23] Yet, the records are generated by a series of probabilistic experiments, not by reference to real individuals.

Generative AI-based SD generation works in an analogous, but automated and more sophisticated manner. The thought experiment requires prior knowledge about the population (the distribution of male to female and of eye colors). When using generative AI, this is not a requirement: all relevant patterns are extracted directly from the data. Furthermore, generative AI-based generators accurately replicate correlations between attributes. In the thought experiment, for example, the overall proportion of blue-eyed individuals may be 1/6, but perhaps blue eyes are more common among females than males. A generative model could automatically infer and replicate this correlation. We refer the reader to Finocchiaro et al.[24] for an accessible overview of machine learning-based SD generation techniques.

The degrees of both privacy protection and realism of properly generated SD are widely recognized. The United States of America's National Institute for Standards and Technology (NIST) compared a multitude of de-identification algorithms.[25] Their results table[26] shows that SD generators excel at combining data utility with privacy protection. This empirical evidence for SD as a reliable anonymization tool is corroborated by the NeurIPS 2020 Hide-and-Seek Privacy challenge.[27][28] Submissions to the challenge demonstrated a privacy protecting SD algorithm (hiders), or a re-identification algorithm based on membership inference attacks (seekers). The challenge showed that only for a single (likely improperly synthesized) hider could any method re-identify significantly more than random guesses would.

3.2. Policy and Regulatory Perspectives

Like in the scientific community, SD generation as an anonymization process is gaining recognition in the legal scholarly community. In a seminal paper, Bellovin et al. state that:[29] *"Synthetic data offers progress. Though not a silver bullet, the method allows us to put an end to the de-identification–re-identification arms race and focus on what matters: useful, private data. To this extent, we recommend the privacy community accept synthetic data as a valid, next step to the database privacy problem."*

[23] By a theorem called "The Law of Large Numbers", see any introductory text on probability theory.

[24] *Finocchiaro G., Landi A., Polifronei G., Ruffo D., Torlontano F.*, Il Futuro Regolatorio Dei Dati Sintetici. La Sintetizzazione dei Dati Come Risorsa per Ricerca Scientifica, Innovazione e Politiche Pubbliche nel Panorama Giuridico Europeo, 2024.

[25] See: *Task C., Bhagat K., Howarth G.,* SDNist v2: DeidentifiedDataReport Tool, 2023, <https://data.nist.gov/od/id/mds2-2943> [30.07.2024].

[26] See NIST Collective Research Cycle (CRC), <https://pages.nist.gov/privacy_collaborative_research_cycle/pages/archive.html> [30.07.2024].

[27] *Jordon J., Jarrett D., Saveliev E., Yoon J., Elbers P., Thoral P., van der Schaar M.*, Hide-and-Seek Privacy Challenge: Synthetic Data Generation vs. Patient Re-Identification.

[28] Competition and Demonstration Track, 2021.

[29] *Bellovin Steven M., Dutta, Preetam K., Reitinger N.*, Privacy and Synthetic Datasets. *S*tanford Technology Law Review, Vol. 22, 2018.

This view is echoed in a more recent legal analysis by Finocchiaro et al., concluding that SD, if properly generated, constitute "*an advanced anonymisation technique that complies with current and future data protection regulatory requirements.*"[30]

Outside of the scholarly community, SD generation is gaining recognition as an anonymization process among policy-makers and even independent authorities. The Spanish Data Protection Agency (AEDP) explicitly recognizes that SD can be, at least under certain conditions, equated to anonymized data, suitable for personal data de-identification.[31] The Data Governance Act[32]and the Artificial Intelligence Act (AIA)[33] recognize SD generation as a privacy-preserving data processing method, equating it with anonymous or non-personal data. The European Union's *Joint Research Centre* (JRC) extends this acknowledgment, perceiving properly designed synthetic data not only as free of privacy-related risks but also as a strategic enabler in the domain of Artificial Intelligence.[34]

Lastly, interesting in France a law called the "Loi pour une République numérique" empowers the CNIL (Data Protection Authority in France) to validate and certify anonymization techniques. In 2020, the CNIL did use such power and certified a methodology for generating SD, affirming its anonymous nature based on criteria like individualization, linkability, and inference, thus acknowledging SD as (at least potentially equiparable to) anonymous data[35]. The CNIL even made a step further expressing interest in the development of a framework for empirically evaluating the degree to which specific synthetic datasets protect privacy[36] so as to be able to rely on quantitative and more objective analysis in validating different SD generation techniques.

4. Recommendations for Proper Deployment of Synthetic Data

In Section 3, we provided scientific and regulatory perspectives on SD generation as an anonymization technology. Throughout, we noted that SD generation serves as a strong tool in achieving anonymity, *provided that it is generated properly*. In this section, we analyze the notion of "proper generation" by studying sources of potential privacy risks, as well as some theoretical implications of SD generation. Throughout, we provide concrete recommendations to minimize risks. This fosters proper use of the technology, allowing resulting SD methods to achieve anonymization status.

[30] *Finocchiaro G., Landi A., Polifronei G., Ruffo D., Torlontano F.*, Il Futuro Regolatorio Dei Dati Sintetici. La Sintetizzazione dei Dati Come Risorsa per Ricerca Scientifica, Innovazione e Politiche Pubbliche nel Panorama Giuridico Europeo, 2024.

[31] Agencia Española de Protección de Datos (AEDP), "Approach To Data Spaces From GDPR Perspective", 2023.

[32] See Recital 7 of the Regulation (EU) 2022/868.

[33] See art. 10.5.a and art. 59.1.b of the Regulation (EU) 2024/1689.

[34] See *Hradec J., Craglia M., Di Leo M., De Nigris S., Ostlaender N., Nicholson N.,* Multipurpose Synthetic Population for Policy Applications, JRC Technical Report, 2022, 58-60.

[35] See Octopize's communication on the CNIL evaluation in the following presentation: <https://documentation-snds.health-data-hub.fr/files/presentations/meetup-snds7/20210318_Octopize_Deck-OctopizeHdhCom_MLP-2.0.pdf> [30.07.2024].

[36] SeeCNIL's Letterto Statice GmbH on Anonymeter, <https://www.anonos.com/hubfs/Documents/Reports/CNIL_Anonymeter.pdf> [30.07.2024].

4.1. Synthetic Data Risk Factors

Core risk factors are identified in our previous work,[37] namely: 1) the quality of the generator; 2) the approach to synthesis; 3) properties of the real dataset (presence of outliers, sparsity, etc.); 4) the information available to the attacker (threat model). These factors align closely to the caveats for SD technology identified by the European Data Protection Supervisor (EDPS)[38]. These risks can be evaluated and typically mitigated in practice by adopting a proactive and quantitative approach to privacy assessment. Worryingly, a recent literature study in the medical field found that while 85% (78/92) of papers include synthetic data for privacy use cases, only 42% of papers (39/92) use a privacy quantification method.[39] This shows that synthetic data privacy protection is not always quantified. Instead, it is presumed to offer sufficient protection by default, providing a false sense of security.

Recommendation: practitioners working with synthetic data should be aware of potential privacy risks. Moreover, they should use quantitative and objective methods to assess, evaluate and mitigate privacy risks. Our recent survey on privacy quantification[40] provides an in-depth discussion of such methods and their merits. In the remainder, we also present additional practical recommendations for the optimal use of these tools. We further recommend the use of fidelity and utility metrics to make sure that a good balance between privacy and utility is achieved.

4.2. SD-based and Generator-based Privacy

A fundamental question in SD technology is whether privacy protection is a property of SD (here after: "SD-based") or of the generator that produces the SD (hereafter: "generator-based"). These viewpoints affect how privacy is quantified in practice. To illustrate the different perspectives, consider the so-called "infinite monkey theorem".[41] This theorem roughly states that, given an infinite amount of time, a monkey randomly hitting keys on a typewriter will type the works of Shakespeare at some point. Similarly, it is easy to imagine a generator that, due to its stochasticity, produces a record (nearly) identical to a real individual's, without that real individual having been in its training set.[42] The question then arises: is this individual's privacy breached? Under SD-based privacy, the answer is "yes", as the synthetic dataset exposes the individual. Under generator-based privacy, the answer is

[37] *Panfilo D., Boudewijn A. T., Ferraris A. F., Cocca V., Zinutti S., De Schepper K., Chauvenet C.R.*, Measuring Privacy Protection in Structured Synthetic Datasets: A Survey, in: Hideyuki Matsumi, Paul De Hert et al. (ed) Privacy and Data Protection: Ideas that Drive Our Digital World, 2024.

[38] *Wiewiórowski W.*, Synthetic Data: What Use Cases as Privacy Enhancing Technology? IPEN Webinar on Synthetic Data, European Data Protection Supervisor.

[39] *Kaabachi B., Despraz J., Meurers T., Otte K., Halilovic M., Prasser F., Raisaro J. L.*, Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics, 2023.

[40] *Boudewijn A. T. P., Ferraris A. F., Panfilo D., Cocca V., Zinutti S., De Schepper K., Chauvenet C. R.*, Privacy Measurements in Tabular Synthetic Data: State of the Qrt and Future Research Directions, NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI, 2023.

[41] See most introductory texts on probability theory.

[42] Unlike the monkeys and Shakespeare, the generator will not *almost surely* generate the record of any given individual. This is because a generator models a specific distribution very accurately, unlike the uniform distribution over keys hit in the theorem. Thus, a generator does not necessarily generate all combinations of all possible attribute values given an infinite amount of time. Still, the scenario where a generator, due to stochasticity, produces a record overly similar to a real individual whose data was not in the training set is well within the realm of possibilities.

"no", since the production of the individual's data was not a consequence of the model's lack of protection of records in its training set.

The SD-based privacy viewpoint is more common in practice: a recent literature review on quantification of both utility and privacy in SD in medical contexts found that 81.2% of papers (18 out of 22) on SD using a privacy quantification approach are SD-based; only 9.1% (2 out of 22) are generator-based. The remaining 9.1% use a combination of both.[43] This popularity of SD-based privacy is likely due to two reasons. Firstly, the comparison with the infinite monkey theorem exposes a shortcoming of model-based privacy. Secondly, data-based privacy can be measured through metrics with clear real-world interpretations. In particular, these metrics typically quantify the likelihood that produced records are memorized from real ones; hazardous similarities between synthetic and real records; or the success rates of deliberately conducted attacks.[44]

Like SD-based privacy, generator-based privacy has several merits. Firstly, it is typically added to the generation process in an a priori, user-controlled manner. The most widespread examples are generalization control through early stopping and differential privacy,[45] with the user specifying a so-called *privacy budget*. Secondly, a number of *composition theorems* apply to differentially private systems, roughly stating that no further mathematical manipulation of a differentially private generator can remove its differentially private status. Recall that possible future developments should be taken into consideration when determining whether anonymization was successful. This makes the composition property particularly appealing. Shortcomings include that privacy budgets have no clear real-world interpretation and are hard to choose in practice.[46][47] Furthermore, the composition theorems apply to mathematical manipulation, not the use of compounding auxiliary information.

Recommendation: In light of the above discussion, we recommended employing a combination of data-based and generator-based methods. Data-based methods focus on the synthetic datasets produced, utilizing empirical privacy evaluations and similarity assessments to ensure that no identifiable information remains. Conversely, model-based techniques provide strong theoretical guarantees by embedding privacy mechanisms directly into the data generation process. These methods ensure that the generative models themselves do not expose individual data points. A combined approach maximizes the robustness of privacy protection, addressing both the model and the data perspectives.

4.3. "Relating to in Content" and Synthetic Data

López & Elbi argue that only access to information relating to an individual "in content" should be considered a breach.[48] This theoretical viewpoint seems evident in an applied setting, as aggregate statistics, such as those released by public institutions, can already relate

[43] *Kaabachi B., Despraz J., Meurers T., Otte K., Halilovic M., Prasser F., Raisaro J. L.*, Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics, 2023.

[44] For a thorough survey on both SD-based and generator-based quantification of privacy, see: *Boudewijn A. T. P., Ferraris A. F., Panfilo D., Cocca V., Zinutti S., De Schepper K., Chauvenet C. R.*, Privacy Measurements in Tabular Synthetic Data: State of the Qrt and Future Research Directions, NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI, 2023.

[45] *Dwork C., Roth A.,* The Algorithmic Foundations of Differential Privacy, Foundations and Trends in Theoretical Computer Science, 9(3-4), 2014, 211-407.

[46] *Lee J., Clifton C.*, How Much is Enough? Choosing for Differential Privacy.

[47] International Conference, ISC 2011, Proceedings 14, Springer Berlin Heidelberg, 2011, 325-340.

[48] *López C.A.F., Elbi A.*, On The Legal Nature of Synthetic Data, NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, 2022.

to specific individuals "in purpose" and "in result". For instance, an overview of average incomes per region can lead to prejudice about individuals residing in a particular region.

Applying this viewpoint to SD, the question becomes whether access to SD allows attackers to infer knowledge pertaining to a specific individual. SD (and/or its generator) poses a privacy risk if its access leads to *specific information about* (i.e. relates to in content) a specific individual that could not have been inferred from mere general information. This is particularly relevant in the context of attribute inference attacks, as they deal with information attributable to an individual.

The Anonymeter framework by Giomi et al.[49] includes an implementation of an attribute inference attack that models the described notion of specificity and genericity. In particular, they compare the success rates of attribute inference attacks informed by SD to those of attribute inference attacks informed by a control group. The control group consists of real data that was not used for training the generator. If these success rates are comparable, then the information gained by SD access was generic in nature. If, on the other hand, significantly more can be deduced about a specific individual through SD than through the control set, this information relates to a specific individual in content and constitutes a privacy risk. A comparable approach to the use of control groups is also advocated in using distance-based privacy indicators.[50]

Recommendation: To refine the assessment of privacy (and risk of attribute inference attacks in particular), we recommend systematically including control groups of real data not used in the training of the generator. By doing so, researchers can more accurately determine whether the information derived from synthetic data is genuinely generic or if it improperly reveals specific individual details, posing a privacy risk.

4.4. Generator Training as a Form of Data Processing

While properly generated synthetic data (SD) is considered anonymous, the process of obtaining it involves training a generator using real data, classified under data protection laws as personal data processing. The United Kingdom's Information Commissioner's Office (ICO)[51]explicitly characterizes this training as such, emphasizing the need for stringent adherence to the legal framework.

In particular, SD generator needs to bear in mind from a regulatory perspective the following aspects:

- Legal Basis and Purpose: Data controllers must establish a clear and lawful basis for processing personal data, particularly when it involves sensitive categories such as health data (GDPR Article 6 and Article 9). The purpose of this processing should be explicitly defined and documented, aligning with the principle of purpose limitation.
- Data Protection Impact Assessment (DPIA): A DPIA is crucial when processing is likely to result in high risks to the rights and freedoms of natural persons (GDPR Article 35). This assessment should detail the data flows, evaluate the risks

[49] *Giomi M., Boenisch F., Wehmeyer C., Tasnádi B.*, A Unified Framework for Quantifying Privacy Risk in Synthetic Data, 2022.
[50] *Platzer M., Reutterer T.*, Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data, Frontiers in Big Data, 2021.
[51] *Information Commissioner's Office (ICO)*, Anonymisation: Managing Data Protection Risk Code of Practice, 2012.

associated with the synthesis process, and describe measures to mitigate these risks. It should be noted that the DPIA is not a one-time requirement but should be revisited and updated to reflect changes in the processing activities or to address any data breaches or security issues that may have arisen.

- Technical and Organizational Measures: To ensure the integrity and confidentiality of personal data, GDPR Article 32 requires controllers to implement appropriate technical and organizational measures. This includes secure data handling practices, data minimization, and pseudonymization techniques during the training phase of the SD generator.
- Documentation and Compliance: Documentation should be kept to demonstrate compliance with GDPR requirements[52]. This includes maintaining records of processing activities under GDPR Article 30, detailing the purpose of processing, categories of data subjects and personal data, and the recipients of the data.
- Certifications and Standards Compliance: Adhering to recognized standards such as ISO/IEC 27001 and obtaining certifications like Europrivacy™/® can further evidence compliance with GDPR and enhance the trustworthiness of the anonymization processes. Europrivacy™/®[53] assesses compliance with GDPR and is managed by the European Centre for Certification and Privacy (ECCP), aligned with ISO/IEC 17065 and Article 42 of the GDPR.
- Ethical Considerations: Beyond legal compliance, ethical considerations should guide the synthesis of personal data. This involves ensuring that the synthetic data generation does not reproduce or exacerbate biases present in the original data sets, thereby upholding ethical standards and promoting fairness in data usage.
- Transparency and Accountability: GDPR emphasizes transparency and accountability. Data controllers should be transparent with data subjects about the use of their data for synthesizing SD and the measures in place to protect their privacy. This includes clear communication through privacy notices and public disclosures of DPIA summaries where appropriate.

Recommendation: Practitioners involved in synthetic data (SD) generation should establish a legal basis and define a clear purpose for anonymization, adhering to GDPR requirements. It is advisable to perform regular audits and continuous monitoring, along with updates to the Data Protection Impact Assessment (DPIA) to keep pace with technological and regulatory changes. The DPIA, mandatory under GDPR Article 35, should assess risks and outline mitigation strategies.

Furthermore, aligning SD practices with international standards such as ISO/IEC 27001 and obtaining certifications like Europrivacy™/® can enhance compliance credibility. Europrivacy™/®, recognized across EU and EEA Member States, evaluates GDPR compliance and is managed by the European Centre for Certification and Privacy (ECCP), adhering to ISO/IEC 17065 and Article 42 of the GDPR.

This approach promotes legal and ethical synthetic data use, supporting ongoing research and development while maintaining public trust.

[52] As provided by the Accountability principle ex art. 5.2 GDPR.

[53] See Aindo's entry in the Europrivacy/registry: <https://repository.europrivacy.org/en/certifications/edit/d9064da7-603a-4377-b596-b654824e365f> [30.07.2024].

5. Conclusion

As we navigate an era marked by exponential growth in data and escalating privacy concerns, synthetic data generation presents itself as a promising anonymization strategy. This paper has thoroughly evaluated synthetic data from multiple dimensions—legal, theoretical, and practical—highlighting its effectiveness as a privacy-enhancing technology (PET). Our analysis confirms that when produced correctly, synthetic data meets the strict anonymization criteria set forth by the General Data Protection Regulation (GDPR) and other related frameworks.

The need for transformation of data to a state where individuals are indistinguishable is critical, as traditional anonymization techniques increasingly fail to withstand sophisticated re-identification techniques. Synthetic data offers a contemporary solution by creating artificial datasets that reflect the statistical patterns of real data without any direct links to individual identities. This not only maintains privacy but also ensures the usability of data, proving essential for advancements in data science and artificial intelligence.

Regulatory recognition of synthetic data as a viable anonymization method is gaining momentum. Esteemed bodies such as the Spanish Data Protection Agency, the European Union's Joint Research Centre, and France's CNIL have endorsed the capability of synthetic data to fulfill anonymization standards. The endorsement is further strengthened by certifications like Europrivacy™/®, which attest to the adherence of synthetic data processes to GDPR and other regulatory stipulations.

However, the implementation of synthetic data generation demands careful consideration. It is crucial to rigorously assess factors like the quality of the generator, the methodology of synthesis, the characteristics of the source data, and the prevailing threat model. Employing quantitative methods to evaluate privacy and adopting a comprehensive approach that melds data-based and generator-based privacy strategies are vital to mitigate potential risks effectively. Additionally, the inclusion of control groups in empirical privacy assessments and adherence to data protection laws during the training of generators are imperative to uphold stringent compliance standards.

Looking forward, the field of data anonymization is set to evolve with advanced technological methods, rigorous legal regulations, and detailed empirical risk evaluations. This paper emphasizes the significant potential of synthetic data to adeptly address the complexities of data privacy in our digitally evolving landscape, contingent upon its strategic and cautious deployment.

Bibliography:

1. European Commission, "Regulation of the European Parliament and of the Council: on European Data Governance (Data Governance Act)".
2. European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (ARTIFICIAL INTELLIGENCE ACT) and Amending Certain Union Legislative Acts".
3. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of such Data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).
4. *Agencia Española de Protección de Datos (AEDP),* "Approach to Data Spaces from GDPR Perspective", 2023.
5. *Agencia Española de Protección de Datos (AEDP) and European Data Protection Supervisor (EDPS)*, 10 Misunderstanding related to Anonymization, 2021.
6. *Barth-Jones D.*, The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now, 2012.
7. *Beduschi A.*, Synthetic Data Protection: Towards a Paradigm Change in Data Regulation? *Big Data & Society*, 11(1), 2024.
8. *Bellovin Steven M., Dutta, Preetam K., Reitinger N.*, Privacy and Synthetic Datasets, Stanford Technology Law Review, Vol. 22, 2018.
9. *D'Acquisto G., Naldi M.*, Big Data e Privacy by Design, Anonimizzazione, Pseudonimizzazione, Sicurezza, Giappichelli, 2017.
10. *De Montjoye Y. A., Hidalgo C. A., Verleysen M., Blondel V. D.*, Unique in the Crowd: The Privacy Bounds of Human Mobility, Scientific Reports, 3(1), 2013, 1-5.
11. *Dwork C., Roth A.,* The Algorithmic Foundations of Differential Privacy, Foundations and Trends in Theoretical Computer Science, 9(3–4), 2014, 211-407.
12. *Finocchiaro G., Landi A., Polifronei G., Ruffo D., Torlontano F.*, Il Futuro Regolatorio Dei Dati Sintetici. La Sintetizzazione dei Dati Come Risorsa per Ricerca Scientifica, Innovazione e Politiche Pubbliche nel Panorama Giuridico Europeo, 2024.
13. *Foglia C.*, Il Dilemma (ancora aperto) dell'Anonimizzazione e il Ruolo della Pseudonimizzazione nel GDPR, in Circolazione e Protezione dei Dati Personali, tra Libertà e Regole del Mercato. Commentario al Regolamento UE n. 2016/679 (GDPR) e al Novellato d.lgs. n. 196/2003 (Codice Privacy), 2019.
14. *Giomi M., Boenisch F., Wehmeyer C., Tasnádi B.*, A Unified Framework for Quantifying Privacy Risk in Synthetic Data, 2022.
15. *Hern A.,* New York Taxi Details Can Be Extracted from Anonymised Data, Researchers Say, The Guardian, 2014.
16. *Hradec J., Craglia M., Di Leo M., De Nigris S., Ostlaender N., Nicholson N.*, Multipurpose Synthetic Population for Policy Applications, JRC Technical Report, 2022.
17. *Hu J., Bowen C.M.*, Advancing Microdata Privacy Protection: A Review of Synthetic Data, 2023.
18. *Information Commissioner's Office (ICO),* "Anonymisation: Managing Data Protection Risk Code of Practice", 2012.

19. *Jordon J., Jarrett D., Saveliev E., Yoon J., Elbers P., Thoral P., van der Schaar M.*, Hide-and-Seek Privacy Challenge: Synthetic Data Generation vs. Patient Re-Identification, In NeurIPS 2020 Competition and Demonstration Track, 2021.
20. *Jordon J., Szpruch L., Houssiau F., Bottarelli M., Cherubin G., Maple C., Weller A.*, Synthetic Data-What, Why and How? 2022.
21. *Kaabachi B., Despraz J., Meurers T., Otte K., Halilovic M., Prasser F., Raisaro J. L.*, Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics, 2023.
22. *Lee J., Clifton C.*, How Much is Enough? Choosing ε for Differential Privacy, In Information Security: 14th International Conference, ISC 2011, Proceedings 14, Springer Berlin Heidelberg, 2011, 325-340.
23. *López C.A.F., Elbi A.*, On The Legal Nature of Synthetic Data, NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research, 2022.
24. *Narayanan A., Shmatikov V.*, How to Break Anonymity of the Netflix Prize Dataset, 2006.
25. *OECD*, "Emerging Privacy Enhancing Technologies: Current Regulatory and Policy Approaches," OECD Digital Economy Papers (OECD, 2023).
26. *Panfilo D., Boudewijn A. T., Ferraris A. F., Cocca V., Zinutti S., De Schepper K., Chauvenet C. R.*, Measuring Privacy Protection in Structured Synthetic Datasets: A Survey, in: Hideyuki Matsumi, Paul De Hert et al. (ed) Privacy and Data Protection: Ideas that Drive Our Digital World, 2024.
27. *Platzer M., Reutterer T.*, Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data, Frontiers in Big Data, 4, 2021.
28. *Stalla-Bourdillon S., Knight A.,* Anonymous Data v. Personal Data-False Debate: an EU Perspective on Anonymization, Pseudonymization and Personal Data, Wisconsin International Law Journal, 2017.
29. *Task C., Bhagat K., Howarth G.,* SDNist v2: Deidentified Data Report Tool, 2023, <https://data.nist.gov/od/id/mds2-2943> [30.07.2024].
30. *Tempestini L., D'Acquisto G.,* Il dato personale oggi tra le sfide dell'anonimizzazione e le tutele rafforzate dei dati sensibili, in Le nuove frontiere della privacy nelle tecnologie digitali, Aracne, 2016.
31. *Wiewiórowski W.*, Synthetic Data: What Use Cases as Privacy Enhancing Technology? IPEN Webinar on synthetic data, European Data Protection Supervisor.