Synthetic Data in AI Development: Ensuring Data Protection and Ethics

Maria Catarina Batista

NOVA School of Law, Lisbon, Portugal

Abstract

The generation and use of synthetic data have transformed AI system development, enabling a shift from reliance on real-world data to artificial data that preserves the statistical properties of real data while mitigating privacy concerns. As a Privacy Enhancing Technology, data synthesis strikes a balance between data protection mandates and data utility. However, synthetic data introduces ethical challenges, such as bias, misinformation, and public distrust, which this study addresses. This paper emphasizes the necessity of urgent measures to uphold public trust in AI systems and ensure the responsible use of synthetic data in research, especially in sensitive areas like healthcare. It evaluates the British perspective on synthetic data use in research, presenting it as an initial approach to these challenges.

Keywords

Synthetic Data, AI, GDPR, Data Protection, Ethics, Data Governance

1. Introduction

Back in the day, data used for research were mostly collected from sources in the physical world, encompassing a wide range of information. However, with the generation of synthetic data, this scenario has suffered radical changes. This paper examines the impact of using synthetic data to train AI systems on privacy and ethics in our society. In the first section, key concepts around data synthesis are delineated. In section two, we explore issues such as bias, loss of public trust, and the principle of data accuracy, with a practical scenario involving health data accuracy. The third section assesses the British perspective on the use of synthetic data for research. Finally, the fourth section draws conclusions and outlines future approaches to ethical standards for the use of synthetic data .

For the purpose of this paper, data obtained from real-world sources to generate synthetic data are referred to as ”real data”. When this data is related to an identified or identifiable natural person, they are categorized as ”personal data”, as per Article 4 (1) of the General Data Protection Regulation (“GDPR”) [1].

According to Dr Khaled El Emam, a leading figure on data synthesis and anonymization, at a conceptual level, synthetic data can be defined as “data that has been generated from real data and that has the same statistical properties as the real data” [2]. This definition recognizes the artificial nature of synthetic data while retaining the statistical characteristics of the real data. It is crucial to understand that synthetic data refers to data that is artificially created to mimic the patterns and insights found in a real dataset without directly copying information about the individuals represented in that dataset [2]. This type of data can be produced either by using an actual dataset or through deductions and rules established by the coder.[3] [4] Such inferences can be drawn from AI systems, or via human analysis, contingent on the variables present within the source dataset [4].

Data synthesis is a Privacy Enhancing Technology (“PET”) that has been developed as a promising solution to address Data Protection concerns while enabling valuable insights to be extracted from real data [5]. The imperatives of privacy dictate that synthetic data should not solely repeat the statistical patterns and correlations of the real data used for the data synthesis procedure. The GDPR and other

AIEB 2024: Workshop on Implementing AI Ethics through a Behavioural Lens | co-located with ECAI 2024, Santiago de Compostela, Spain

catabatista1999@gmail.com (M. C. Batista)

0009-0002-9817-8642 (M. C. Batista)

Data Protection frameworks demand an inherent trade-off between the safeguarding of data subjects and the practical utility of such data [2]. This trade-off is quantified by measuring the accuracy of the synthetic data in relation to the real data [2]. The higher the degree of privacy preservation incorporated, the more likely the synthetic data is to diverge from the statistical relationships present in the real data, thus having lower utility [2]. This balancing test is crucial in scenarios where the preservation of certain attributes from the real data, for example for analytical accuracy and reliability, is necessary to achieve the purpose for which the synthetic data was generated [6]. For instance, if the purpose of the data synthesis is to generate synthetic data to train AI models for consumer prediction, the demand for high utility is superior [2]. In opposition, when data synthesis' purpose is to assess a software's capability to manage an extensive volume of transactions, the interest in the utility of such data would be significantly reduced [2].

In short, by examining the balance between data protection and utility, especially within the framework of GDPR, we underscore the importance of maintaining data accuracy while safeguarding individual privacy. The next section of this paper aims to provide a comprehensive understanding of the responsible use of synthetic data in AI systems. packages.

2. Unveiling the Challenges and Risks in AI Systems Using Synthetic Data

Over recent years, data synthesis has developed into a refined tool that effectively tackles both privacy and accuracy issues in settings that rely heavily on data [5]. In 2020, Gartner acknowledged the significance of synthetic data, advising organizations to incorporate it into their overall data strategies [7]. They pointed out its scalable nature and compliance with privacy standards, underscoring its broad applicational potential [7]. By 2024, the use of synthetic data has expanded significantly, with both commercial enterprises and governmental institutions leveraging it to advance research, enhance services, and improve decision-making processes [6]. Nevertheless, it also holds significant accountability issues and ethical challenges, as it will be demonstrated in the following subsections.

2.1. Bias and Loss of Public Trust

Data synthesis is a technique that has the potential of enhancing the reproducibility and diversity of a dataset, thus, it can be used as a tool to reduce biases in datasets [8]. In the context of AI development, synthetic data generation enables the creation of edge cases and fills in missing data. This approach helps to address potential biases and inaccuracies in the input datasets, which are crucial for training models. By incorporating these diverse scenarios, synthetic data ensures that the models are more robust and less likely to produce harmful biased outputs.[8].

While having the potential to protect disadvantaged groups from harmful bias present in datasets, the use of data synthesis brings to the table many ethical challenges, such as synthetic media and deepfakes, enhancing the risks of misinformation and societal distrust [6]. It is essential to understand that the absence of information about the source and quality of synthetic data introduces a major challenge: discerning which information within the dataset is valid and which is not [9].

Synthetic media, which is a subset of synthetic data, focusing specifically on media content created using AI techniques, is a great example of the ethical concerns previously mentioned, since its main function is to replicate real-world content, such as images, videos, or audio [10]. Increasingly recognized for its problematic aspects in society, "deepfakes" involve manipulated media where images and videos are altered to falsely depict individuals saying or doing things they never actually did [10]. For instance, deepfakes involving fake sexual photos represent a severe violation of privacy and consent, intensifying ethical issues within synthetic media [10, 11]. Therefore, in this case, synthetic data's capacity for misrepresentation damages reputations, leads to misleading perceptions, and can cause significant emotional distress [12].

The widespread creation and distribution of synthetic media contribute to societal distrust in media,

further eroding social cohesion and heightening public scepticism towards legitimate information. This, in turn, poses challenges for maintaining trust in digital communications and media integrity [10]. Furthermore, synthetic data can lead to cases of mistaken identity. For instance, when creating a synthetic persona, it is possible that this fake person could be mistaken for a real person from the dataset used to generate the synthetic data [10].

While there appears to be no straightforward solution for synthetic data generated with malicious intent, it is possible to manage some ethical issues like bias and misrepresentation, through the implementation of risk mitigation measures previously and during the data synthesis procedure. Thus, in the next session we go through practical scenarios to evaluate the legal and ethical dimensions in synthetic data use cases and we provide our input to improve the compliance of such processing activities.

2.2. Principle of Data Accuracy

The principle of data accuracy, enshrined in Article 5(1)(d) of the GDPR, embodies the trustworthiness and reliability of the Data Subjects in the processing of Personal Data [13]. According to the GDPR, controllers and processors should maintain the precision of datasets and must immediately rectify any inaccuracies when they arise. However, the processing of synthetic data introduces a complex layer to this issue. Synthetic data, being an artificial construct, does not directly represent real individuals. From a Data Protection compliance perspective, it raised an important question: How can accuracy be ensured in synthetic data, which lacks a direct link to the individuals?

It is necessary to point out that once the synthetic data has been generated, the next step of the data synthesis procedure involves calculating its metrics [2, 14]. These metrics are then compared with those of the real data using a tool known as a discriminator [2]. This discriminator evaluates the utility of the synthetic data by examining whether its statistical properties closely mirror those of the real data [2]. During this metrics calculation phase, synthetic data developers are responsible for ensuring that the statistical patterns and correlations present in the real data are accurately replicated in the artificial data [2]. This ensures that the synthetic data maintains fidelity to the real data it represents [2]. Thus, when the comparison reveals that the synthetic data diverges from the real data, adjustments should be made to the generation parameters and a new and accurate dataset should be produced [2]. This process must be repeated until achieving accurate synthetic data [2].

2.2.1. Practical Case: Health Data Accuracy

This subsection highlights the vital role of synthetic data in enhancing healthcare through a practical perspective. AI systems, which demand extensive and accurate training data, are increasingly being employed in healthcare for various purposes, such as medical imaging, patient data analytics, and drug discovery [6, 15, 16]. A common issue in clinical trials is the inaccurate gender distribution among participants. For instance, when there is the predominance of male participants in a drug trial, it hinders the understanding of the medication's effects on females [17]. To face this issue, synthetic data, generated specially to replicate the health profiles typical of female participants, can be integrated into the analysis to create a more inclusive and balanced study [18]. Thus, synthetic data can be used as a strategic feature to improve the performance and reliability of AI systems to generate better informed results [6].

However, the use of not well-produced synthetic data might diminish societal trust in research, leading to doubts about the authenticity and integrity of a trial's findings [15]. When using poor generated synthetic data to represent a demographic insufficiently represented in the real trial, such data might lead to potential inaccuracies in understanding how the medication affects that specific group, leading to erroneous medical decisions, with dangerous consequences for real patients [6].

From an ethical stance, the use of synthetic data not only helps achieve an unbiased dataset but also supports broader demographic research, in any field of study [6]. Nevertheless, regardless of the precautionary measures taken by developers, researchers must be aware that there is always the risk that errors in the synthetic data generation may occur [6] Therefore, when processing synthetic data,

analysts and researchers must always proceed with caution, acknowledging that there is the possibility that not every pattern or correlation observed might be accurate.

A prime example of synthetic data's limitation is the partial synthesis of survey data collected by the Cancer Care Outcomes Research and Surveillance ("CanCORS") project [19]. In this instance, after evaluating the synthetic data created using the project's model, researchers determined that the dataset was suitable only for preliminary data analysis due to problems with data correlations [6, 19]. Consequently, it is essential for developers of synthetic data to maintain transparency about the dataset's quality and clearly communicate its limitations to end-users.

Although some synthetic datasets as CanCORs can only be used for preliminary data analysis, they can still offer significant value at this early stage of research. For example, synthetic data provides a safe, efficient, and flexible alternative to using real data during software testing [2]. Furthermore, incorporating synthetic data in the development phase can expedite the software refinement process and reduce computational demands, due to its high-quality labelling [2].

By incorporating synthetic data in the early stages of model development, the use of real data is deferred until the software's security has been verified. This strategy effectively reduces the risks associated with data processing, such as data breaches, thus enhancing the protection of the confidentiality, integrity, and availability of personal data [6]. Moreover, this method highlights the role of synthetic data in facilitating research advancements while simultaneously bolstering data security. This approach is particularly relevant for special categories of personal data, such as health data, where the fundamental right to data protection demands special attention.

3. From Theory to Practice: How the UK Implements Synthetic Data Strategies

In this section, we assess how synthetic data is being approached from a policy-making perspective. As mentioned in the previous sections, the use of synthetic data raises critical implications for transparency and communication to individuals [20]. When presenting research findings based on synthetic data, it becomes crucial to ensure that audiences are made highly aware of this combination of data [9]. Recognizing the gravity of these concerns, bodies like the UK Statistics Authority and the Office for National Statistics ("ONS") have taken proactive steps, formulating comprehensive guidelines on synthetic data [21].

The ONS Synthetic Data Policy stresses key legal and ethical issues, such as confidentiality and data disclosure risks, offering an essential framework for responsible synthetic data processing in statistical research [21]. This Policy determines the ethical handling of synthetic data in research and analysis, ensuring compliance with legal standards and reducing potential liabilities. This Policy is particularly significant as it marks the first documented guideline for managing synthetic data, thus providing orientation for researchers and analysts across all jurisdictions [21].

The UK Statistics Authority also established comprehensive guidance on synthetic data, including an overview of ethical considerations and mitigation strategies and an ethics checklist [22, 23]. This Authority has also developed ethical principles and an ethics self-assessment tool to guide researchers and statisticians in addressing ethical issues in various projects, including those involving synthetic data [23, 24]. Such principles emphasize the public good, data confidentiality, risk assessment, legal compliance, public perception, and transparency in data collection and usage [23, 24]. Therefore, by consistently incorporating a thoughtful ethical framework into each project, it is possible to address these concerns, ensuring both the integrity of the research and the continued trust of individuals in data synthesis [6].

Moreover, the Authority demonstrates the prominent need to balance utility, which is the data's practical usefulness, and fidelity, its authenticity [22]. Such a balance is a parameter that demonstrates the efficiency of synthetic data to serve its intended purpose while accurately representing the real data [22]. Essentially, utility represents if synthetic data satisfies specific research or analytical purposes [22]. Conversely, synthetic data retaining substantial fidelity accurately reflects the attributes of real data,

consequently serving as an accurate alternative for the real data [22]. A mirror reflecting a complex scene can exemplify fidelity; the clearer the reflection, the higher the fidelity [22]. High-fidelity datasets are very detailed and closely mimic real-world data, thus they are very useful for complex tasks like developing new medical treatments or training advanced AI models to predict patient outcomes. On the one hand, if synthetic data mirrors too closely the real data, it could inadvertently reveal personal data through inference, thus violating Data Protection norms and ethical considerations [22]. On the other hand, if synthetic data deviates too much from the real dataset, its utility for research might be compromised due to a lack of authenticity [22]. In opposition, low fidelity datasets are less detailed and more generalized, thus having a lower risk of revealing personal data, making it safer to use for research [22]. Low-fidelity datasets are also particularly useful for gaining a broad understanding of trends and patterns in research without delving into sensitive details.

Finally, the importance of these British guidelines lies in their function as a standard for ideal data management practices in the world while the statistical research industry is in harmony with wider legal norms like the European and UK GDPR. Therefore, by adhering to these guidelines, organizations and researchers will comply with legal mandates related to Data Protection, thereby reducing legal risks associated with the use and management of Synthetic Data, while also upholding societal ethical principles.

4. Conclusion

The exploration of synthetic data within the field of artificial intelligence has illuminated both its vast potentials and its ethical challenges. As this paper has discussed, synthetic data offers a crucial advantage by reducing reliance on real data, thereby enhancing privacy and reducing the risks associated with personal data breaches. However, the complexities of ensuring data accuracy, maintaining public trust, and managing potential errors cannot be overlooked.

The British perspective on synthetic data utilization in research advocates for a balanced approach, emphasizing the necessity of stringent Data Protection measures alongside the benefits of synthetic data. The UK's regulatory framework and ethical guidelines serve as a beacon for other nations, promoting a synthesis of utility and fidelity that respects both individual rights and the demands of technological advancement.

For synthetic data to truly benefit society, particularly in sensitive applications like healthcare, developers and regulators must work in concert to forge policies that not only enhance data utility but also prioritize transparency and accountability. Ensuring that synthetic data maintains its integrity without compromising on ethical standards is essential for its acceptance and success.

In conclusion, as synthetic data generation techniques continues to evolve, so too must our strategies for its regulation and use. Only through a concerted effort to address these legal and ethical challenges can we harness the full potential of synthetic data to propel AI development while safeguarding the fundamental rights of individuals. Moving forward, the lessons learned from the British model should inspire global standards that advocate for responsible and ethical synthetic data practices across all sectors.

References

[1] General data protection regulation (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), https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016.

[2] K. E. Emam, et al., Practical Synthetic Data Generation, O'Reilly Media, Inc, 1005 Gravenstein Highway North, Sebastopol, CA 95472, USA, 2020.

[3] R. D, Statistical disclosure limitation, J OFF STAT 9 (1993) 461–462. URL: https://www.scb.se/

contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf.
[4] G. M, L. O, Synthetic data: Legal implications of the data-generation revolution, 2023. URL: https://papers.ssrn.com/abstract=4414385, accessed: 21 September 2023.
[5] O. of the Privacy Commissioner of Canada, Privacy tech-know blog: When what is old is new again – the reality of synthetic data, 2022. URL: https://www.priv.gc.ca/en/blog/20221012/?id=7777-6-493564, accessed: 11 June 2023.
[6] M. Batista, Synthetic Data and GDPR Compliance: Navigating the Legal and Ethical Landscape, Master's thesis, Nova School of Law, University of Lisbon, 2024. URL: https://run.unl.pt/bitstream/10362/166398/1/Batista_2024.pdf, accessed: 15 May 2024.
[7] Judah, et al., Predicts 2021: Data and analytics strategies to govern, scale and transform digital business, 2020. URL: https://www.gartner.com/en/documents/3993855, accessed: 2 December 2023.
[8] R. L, H. B, In defense of synthetic data, 2019. URL: http://arxiv.org/abs/1905.01351, accessed: 1 December 2023.
[9] H. R, Ai image generator: This is someone thinking about data ethics · dataetisk tænkehandletank, 2022. URL: https://dataethics.eu/ai-image-generator-this-is-someone-thinking-about-data-ethics/, accessed: 7 October 2023.
[10] L. University, Lancashire cyber foundry an introduction to deepfakes, https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/cyber-foundry/lcf-articles/LCFArticle-Josh-Deepfakes_WEB.pdf, 2023. Accessed: 3 December 2023.
[11] D. B, et al., The deepfake detection challenge (dfdc) preview dataset, 2019. URL: http://arxiv.org/abs/1910.08854, accessed: 3 December 2023.
[12] S. Team, How to detect ai generated images with sensity in 2023, 2023. URL: https://sensity.ai/blog/deepfake-detection/how-to-detect-ai-generated-im/, accessed: 3 December 2023.
[13] K. C, et al., The EU General Data Protection Regulation (GDPR): A Commentary, Oxford University Press, 2020.
[14] K. E. Emam, Accelerating AI with Synthetic Data, O'Reilly Media, Inc, 2020.
[15] D. C, Synthetic data for healthcare: Benefits case studies in 2023, 2022. URL: https://research.aimultiple.com/synthetic-data-healthcare/, accessed: 29 November 2023.
[16] B. S, Synthetic data and privacy - experiences implementing data synthesis in a global life sciences company, 2021. URL: https://edps.europa.eu/system/files/2021-06/01_stephen_bamford_en_0.pdf, accessed: 3 December 2023.
[17] G. A, et al., Generation and evaluation of synthetic patient data, BMC Medical Research Methodology 20 (2020) 108. URL: https://doi.org/10.1186/s12874-020-00977-1, accessed: 4 November 2023.
[18] Gonzales, et al., Synthetic data in health care: A narrative review, 2023. URL: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082, accessed: 4 November 2023.
[19] L. B, et al., Disclosure control using partially synthetic data for large-scale health surveys, with applications to cancors, Statistics in Medicine 32 (2013) 4139. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.5841, accessed: 29 November 2023.
[20] W. W, R. J, Rti u.s. synthetic household population, https://www.rti.org/brochures/rti-us-synthetic-household-populationtm-database, 2020. Accessed: 4 November 2023.
[21] Office for National Statistics, Synthetic data policy - office for national statistics, https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/syntheticdatapolicy/, 2023. Accessed: 10 December 2023.
[22] U. S. Authority, Ethical considerations relating to the creation and use of synthetic data, 2023. URL: https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/, accessed: 27 September 2023.
[23] U. S. Authority, Ethical principles, 2023. URL: https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethical-principles/, accessed: 27 September

2023.

[24] U. S. Authority, Ethics self-assessment tool, 2023. URL: https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/, accessed: 27 September 2023.