BIG DATA & SOCIETY

Commentary

Synthetic data protection: Towards a paradigm change in data regulation?

Big Data & Society
January–March: 1–5
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517241231277
journals.sagepub.com/home/bds
Sage

Ana Beduschi[1]

Abstract

Synthetic data generated through machine learning algorithms from original real-world data is gaining prominence across sectors due to their potential to provide privacy-preserving alternatives to traditional data sources. However, recent studies have raised concerns about the re-identification risks of synthetic data. This article examines the legal challenges surrounding synthetic data protection, with a focus on the European Union's General Data Protection Regulation (GDPR). After briefly explaining the methods of synthetic data generation and discussing their potential for privacy preservation, the article analyses the shortcomings of the personal/non-personal dualist approach under the GDPR. It then assesses the possibility of a paradigm change in data protection legislation, moving beyond this binary categorisation. The article argues in favour of establishing clear guidelines for the generation and processing of synthetic data, prioritising the principles of transparency, accountability and fairness.

Keywords

Artificial intelligence, data protection, GDPR, machine learning, non-personal data, synthetic data

Introduction

Synthetic data generated through machine learning algorithms from original real-world data (i.e. data relating to existing individuals or events) is gaining importance in various sectors. An example is the Global Victim-Perpetrator Synthetic Dataset, which is derived from actual case records on human trafficking victims and survivors (IOM, 2022). Synthetic data also holds great promise in healthcare and policymaking (Arora and Arora, 2022; Hrade et al., 2022).

Much of the interest in synthetic data is due to its potential to provide privacy-preserving alternatives to traditional data sources. Depending on how synthetic datasets are generated, they may be able to preserve the statistical properties of the original real-world data while minimising the personal information shared (Bellovin et al., 2019). Synthetic data can also be utilised to expand small datasets, potentially increasing their diversity and reducing biases (Jordon et al., 2022). These expanded datasets could then be used to train increasingly data-hungry artificial intelligence (AI) systems (Jacobsen, 2023).

However, not all synthetic datasets are fully artificial – some may contain personal information (Ruiz et al., 2018) or present a risk of re-identification (Stadler et al., 2022). From a legal standpoint, re-identification plays a crucial role in determining the applicability of data protection laws such as the European Union's General Data Protection Regulation (GDPR). That is because laws such as the GDPR only apply to the processing of personal data (Article 4-1 GDPR). Nonetheless, it remains unclear what level of re-identification risk would be sufficient to trigger their application in the context of synthetic data processing.

The article examines these issues, drawing primarily on the GDPR. This choice is justified by the GDPR's scope, reach and influence on legislative processes worldwide. The article also builds on academic and professional literature on legal, policy-oriented and technology-facing aspects to evaluate synthetic data's potential for privacy preservation and the significant challenges it presents.

The analysis proceeds in three steps. First, the article examines the context and methods employed for synthetic data generation, with particular attention to privacy protection measures. Second, it assesses the questions arising under the GDPR. Third, the article reflects on whether a paradigm change in data protection regulation is needed to align legal protection to technological developments.

[1] Law School, University of Exeter, Exeter, UK

Corresponding author:
Ana Beduschi, Law School, University of Exeter, Amory Building, Rennes Drive, Exeter EX4 4RJ, UK.
Email: a.beduschi@exeter.ac.uk

Finally, the article draws conclusions on the future of synthetic data protection.

Privacy-preserving potential and limitations

Synthetic data differs from real-world data as it is generated by algorithmic models known as synthetic data generators, such as Generative Adversarial Networks or Bayesian networks (Jacobsen, 2023; Jordon et al., 2022; Kaur et al., 2021; Stadler et al., 2022). The aim of synthetic data generation is to closely mimic real-world data, maintaining most of its statistical characteristics while also ensuring the confidentiality of the information and the privacy of the data subjects. In principle, that should allow synthetic data to retain enough utility to be useful for research and analysis without revealing much sensitive or identifiable information about real-world data subjects.

However, synthetic datasets are not always completely artificial and free from personal data. Much depends on the type of synthetic data generated, which can be categorised into fully synthetic, partially synthetic and hybrid (Ruiz et al., 2018). In fully synthetic datasets, all attributes across all records are artificial. Partially synthetic datasets only synthesise sensitive attributes. Hybrid datasets combine both real-world and fully synthetic data (Ruiz et al., 2018: 61–62). The approaches taken for the generative process, which can be classified as 'vanilla' and 'differentially private', are also relevant (Bellovin et al., 2019). 'Vanilla' refers to a type of synthetic data that is created using a simple algorithm which does not incorporate any complex statistical models or advanced sanitisation techniques – in other words, it is 'data in, data out' (Bellovin et al., 2019: 41). Differentially private synthetic data is generated from sanitised data using a differential privacy technique (Arnold and Neunhoeffer, 2020; Bellovin et al., 2019). This technique aims to improve privacy by adding random information ('noise') to the original data before generating the synthetic data (Li et al., 2017). Preferably, the amount and type of random noise are carefully calibrated to maintain the overall statistical properties of the original data while increasing privacy protection, thus providing a fair balance between utility and privacy.

Accordingly, certain types of synthetic data could prove particularly useful in circumstances such as when AI systems necessitate larger training datasets, but the real-world data is too sensitive to share, too scarce, or of too low quality. That is, for instance, the case of AI systems trained on health data for medical research, which are often very sensitive (Chen et al., 2021) and humanitarian data for natural disaster response, which may be scarce or of low quality (IOM, 2022; Kuglitsch et al., 2022).

Although synthetic data is a promising technology (Jordon et al., 2022), it still has its limitations. During the generative process, information loss may occur due to sanitisation techniques such as suppressing sensitive information or adding noise to the data, resulting in a decrease in the utility of synthetic data. For instance, a study on synthetic geospatial and temporal epidemiologic data for COVID-19 showed mixed results in terms of dataset utility, with a decrease in utility for small sample sizes (Thomas et al., 2022).

Moreover, researchers have demonstrated that synthetic data poses varying degrees of information disclosure risk depending on the generation method used (Bellovin et al., 2019; Jordon et al., 2022; Ruiz et al., 2018; Stadler et al., 2022). That could potentially lead to the re-identification of the original real-world individual records, even in the case of differentially private synthetic data (Stadler et al., 2022) – thus transforming non-personal data into personal information. Such re-identification risks are important in determining whether existing data protection regulations apply to synthetic data, as elaborated further below.

Current issues under the GDPR

Data protection laws such as the GDPR only apply to the processing of personal data. The GDPR's definition of personal data is remarkably broad, encompassing 'any information relating to an identified or identifiable natural person' (Article 4-1 GDPR). That brings about the questions of whether synthetic data qualifies as personal data under the GDPR, and when compliance will be required. To answer these questions, three main points should be underscored.

First, it is crucial to decouple the generation of synthetic datasets from their processing. On the one hand, generative models that rely on personal data from a real-world dataset and utilise machine learning techniques to produce synthetic data are subject to GDPR regulations. On the other hand, the extent to which the GDPR applies to the processing of the resulting synthetic data varies depending on the type of synthetic data generated.

Second, the GDPR applies to the processing of partially synthetic and hybrid datasets, as they contain information about identified or identifiable individuals. In practical terms, the processing of these two types of synthetic data will have to comply with the principles outlined in Article 5 of the GDPR. These include lawfulness, fairness and transparency, as well as purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality (security) and accountability. Compliance also requires respect for the rights provided by the GDPR, such as the right to access information (Article 15 GDPR), the right to seek rectification of any inaccuracies (Article 16 GDPR), and the right to restrict the processing of personal data (Article 18-1-a GDPR). Furthermore, individuals also have the right to object to the processing of their personal data, even if that is contained in synthetic datasets, as outlined in Article 21 of the GDPR.

Third, by the same logic, the GDPR would not apply to fully synthetic datasets – those that have only artificial attributes across all records. In principle, fully synthetic data would qualify as anonymous data within the meaning of the GDPR. According to its explanatory text, 'anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable' is exempt from compliance with data protection principles and rights under the GDPR (Recital 26 GDPR).

However, that is not so straightforward. Some experts contend that there is always a remaining risk of re-identification in all types of synthetic data, including fully synthetic data and even when differentially private methods are used (Stadler et al., 2022). If there is a residual risk of re-identification in a fully synthetic dataset, the GDPR does apply and compliance is required. This is because some information within that dataset could qualify as personal data according to Article 4-1 of the GDPR and thus no longer be anonymous if re-identification occurs. Still, such a risk might not materialise, and the information contained in such a dataset could remain anonymous, thus creating considerable legal uncertainty for those processing synthetic datasets. That raises the question of whether it would be possible to quantify the risk of re-identification and establish an acceptable threshold for triggering GDPR compliance, which is further discussed below.

Guidelines for synthetic data regulation

It has been proposed that, in order to overcome the issues discussed above, data protection laws should integrate a *de minimis* rule for re-identification – an acceptable level of risk below which data protection laws such as the GDPR would not apply (Bellovin et al., 2019: 50). In this case, fully synthetic datasets would be considered anonymous data if the re-identification risk fell below that minimum threshold.

However, even if determining a specific amount of risk was feasible, such a strategy remains problematic. Particularly, its implementation could bring practical difficulties for synthetic data processors. They would need clear methods to establish precise risk levels and compliance with GDPR obligations or risk hefty fines. Moreover, if sensitive information is to retain its privileged status, processing should be prohibited in principle (Article 9 GDPR) unless any of the exceptions outlined in the GDPR apply (Article 9-1 GDPR). That would require synthetic data processors to filter for re-identification risks concerning sensitive and non-sensitive data, adding further obstacles to the process.

A potential solution to address the issue could be to move beyond the binary categorisation of personal and non-personal data, as is the case regarding the EU Data Act (Regulation 2023/2854). Personal and non-personal data are increasingly interwoven in large and complex datasets, making the distinction more difficult or even redundant. Security and confidentiality breaches of both types of data could also undermine trust in digital services.

Furthermore, such binary categorisation of data may become progressively more blurred as contemporary societies gradually rely more on data-driven AI. This is particularly noticeable with the emergence of generative AI and advanced language models such as DALL-E 3 and GPT-4, which can be trained on and have the ability to generate synthetic data (Goldstein et al., 2023). In turn, the potential large-scale generation of synthetic data, including images of non-existent individuals via DALL-E 3 (thus non-personal data), could facilitate the dissemination of misleading information and have detrimental effects on society (Goldstein et al., 2023).

In this sense, should this type of non-personal synthetic data be regulated somehow? If so, a variety of questions arise. In particular, will the increase in generative AI synthetic data transform the binary approach to data protection laws? Will the concepts of anonymous and personal data need to be revised? For instance, what would privacy and data protection mean not only for fully synthetic but also for hybrid and partially synthetic datasets that have a certain fidelity to real-world data in terms of overall statistical characteristics? Finally, would moving beyond the personal and non-personal data paradigm result in an improved or, conversely, an impoverished standard of protection overall?

While more research is needed to address these questions, some criticism has already been levied at the idea of surpassing the current data dualism orthodoxy in general, as it may result in confusion and diminish the effectiveness of the GDPR (Da Rosa Lazarotto and Malgieri, 2023). It is certainly crucial not to weaken the existing legal frameworks. However, this article suggests that there should be clearer guidelines for all types of synthetic data, including those that would qualify as non-personal data.

As a starting point for further discussion, it is suggested that these guidelines should reflect the principles of transparency, accountability and fairness. Transparency would require, for example, that synthetic data is clearly labelled as such and that information about its generation is provided to users. Accountability would entail establishing clear procedures for calling to account those responsible for the generation and processing of synthetic data. Fairness should include guarantees that synthetic data is not generated and used in ways that bring adverse effects on individuals and society, such as perpetuating existing biases or creating new ones. This initial framework should be developed further to allow for synthetic data production and use to be carried out in ways that minimise potential societal harm.

Conclusion

Synthetic data has emerged in recent years as an alternative to traditional datasets. Although it is not a panacea, synthetic data can be particularly useful in situations where the actual data is too sensitive to share, too scarce, or of too low quality.

While synthetic data has advantages, it also has drawbacks. One limitation is that during the generative process, information loss may occur due to techniques used to sanitise the data, such as suppressing sensitive information or adding noise to the dataset. This can result in a decrease in the utility of the synthetic dataset. Additionally, synthetic data may be vulnerable to information disclosure risks, which differ depending on the generation method. This could potentially result in the re-identification of the original real-world records, even when using differentially private synthetic data methods.

Existing data protection laws that only apply to personal data are not well-equipped to regulate the processing of all types of synthetic data. For instance, fully synthetic datasets are, in principle, exempt from GDPR compliance, except when there is a possibility of re-identification. That creates legal uncertainty and practical difficulties for the processing of such datasets. In this sense, it is important to question whether the increase in synthetic data generation will contribute to changing the current orthodoxy in data protection laws. And whether moving beyond the personal and non-personal data paradigm would result in an improved or, conversely, an impoverished standard of protection overall.

While more research is certainly needed to address these questions, this article has proposed, as a way forward, that clear guidelines for all types of synthetic data should be established. They should prioritise transparency, accountability and fairness. Having such guidelines is especially important as generative AI and advanced language models such as DALL-E 3 and GPT-4 (which can both be trained on and generate synthetic data) may facilitate the dissemination of misleading information and have detrimental effects on society. Adhering to these principles could thus help mitigate potential harm and encourage responsible innovation.

Acknowledgements

The author would like to thank Prof Kubo Mačák for his constructive feedback on an earlier version of this article, colleagues from the Centre for Science, Culture and Law at Exeter Law School for their feedback during informal discussions, and the Journal's editors and anonymous reviewers for their valuable insights and instructive comments.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Ana Beduschi https://orcid.org/0000-0002-8037-5384

References

Regulation 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) OJ L 2023/2854.

Arnold C and Neunhoeffer M (2020) Really useful synthetic data – a framework to evaluate the quality of differentially private synthetic data. In: 37th international conference on machine learning, Vienna. DOI: 10.48550/arXiv.2004.07740.

Arora A and Arora A (2022) Synthetic patient data in health care: a widening legal loophole. *The Lancet* 399(10335): 1601–1602.

Bellovin SM, Dutta PK and Reitinger N (2019) Privacy and synthetic datasets. *Stanford Technology Law Review* 22(1): 1–51.

Chen RJ, Lu MY, Chen TY, et al. (2021) Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* 5: 493–497.

Da Rosa Lazarotto B and Malgieri G (2023) The Data Act: a (slippery) third way beyond personal/non-personal data dualism? European Law Blog, 4 May 2023.

Goldstein JA, Sastry G, Musser M, et al. (2023) *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*. San Francisco: Open AI.

Hrade J, Craglia M, Leo MD, et al. (2022) *Multipurpose Synthetic Population for Policy Applications*. Luxembourg: Publications Office of the European Union.

IOM (2022, December 22) IOM-Microsoft release the first public dataset on victims and perpetrators of trafficking. Available at: https://www.iom.int/news/iom-microsoft-release-first-public-dataset-victims-and-perpetrators-trafficking (accessed 24 January 2023).

Jacobsen BN (2023) Machine learning and the politics of synthetic data. *Big Data & Society* 10(1): 205395172211453.

Jordon J, Szpruch L, Houssiau F, et al. (2022) *Synthetic Data-What, Why and How?* London: The Royal Society.

Kaur D, Sobiesk M, Patil S, et al. (2021) Application of Bayesian networks to generate synthetic health data. *Journal of the American Medical Informatics Association* 28(4): 801–811.

Kuglitsch MM, Pelivan I, Ceola S, et al. (2022) Facilitating adoption of AI in natural disaster management through collaboration. *Nature Communications* 13: 1–3.

Li N, Lyu M, Su D, et al. (2017) *Differential Privacy. From Theory to Practice*. Cham: Springer Nature.

Ruiz N, Muralidhar K and Domingo-Ferrer J (2018) On the privacy guarantees of synthetic data: a reassessment of the maximum-knowledge attacker perspective. In: Domingo-Ferrer J and

Montes F (eds) *Privacy in Statistical Databases*. Cham: Springer Nature, 59–74.

Stadler T, Oprisanu B and Troncoso C (2022) Synthetic data – anonymisation Groundhog Day. In: Proceedings of the 31st USENIX security symposium, pp.1451–1468. Boston: USENIX Association.

Thomas JA, Foraker RE, Zamstein N, et al. (2022) Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: results from analyzing >1.8 million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C). *Journal of the American Medical Informatics Association* 29(8): 1350–1365.