Synthetic data: A convergence between Innovation and GDPR

Shalini Kurapati[a], Luca Gilli[a]
[a] *Clearbox AI*

Abstract. This paper investigates the role that synthetic data could potentially play in generating a convergence between the protection of the fundamental right to personal data protection on the one hand and innovation and data sharing on the other. Synthetic data as an expression of privacy-enhancing technologies could be a useful means to foster data sharing and reutilisation, a crucial aspect of the Open Science approach. Despite the multiple applications, of which the paper offers an overview, there are still two major challenges that the analysis underlines: (i) difficulties in communication between legal experts and tech practitioners; (ii) legal uncertainty, due to the fact that European authorities and policymakers have not yet clearly expressed themselves on synthetic data. The intent of the paper is to propose an introductory analysis of the state of the art on synthetic data and its use, which enables one to envisage future developments. Methodologically, by adopting the perspective of tech practitioners, the paper intends to contribute to the legal debate on technology and the protection of data protection and privacy, with particular reference to the relationship with innovation and data sharing.

Keywords: Synthetic Data, Artificial Intelligence, Innovation, open science, GDPR, data protection by design, scientific research, Privacy Enhancing technologies

1. Introduction: Data is key for innovation

> *"Data is the lifeblood of economic development."- European Commission, 2020*[1]

Data is shaping and will continue to reshape how we produce, consume, and live, influencing every aspect of our lives. Data are the key driver of innovation and currently also play a crucial role in scientific research (Leonelli 2020). Data is essential to train Artificial Intelligence (AI, hereinafter) systems resulting in innovative products and services based on decision support, pattern recognition, forecasting, and insights for enhanced decision-making. We are at a critical juncture of human technological advancement, where massive volumes of data are matched by the unprecedented power of AI algorithms to harness this data for economic and societal benefit (Durante 2021).

Artificial Intelligence, which is largely data-driven, could contribute up to $15.7 trillion to the global economy in 2030.

[1] See: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX\%3A52020DC0066`

There is a fierce global competition to claim these economic benefits from businesses and governments worldwide. The issue also inevitably involves the scientific research sector, both public and private, which can benefit considerably from the adoption of AI techniques, both as a result of the research project and as a means of implementing research projects that only a few years ago were inconceivable without such tools (Paseri, 2022a). The success of AI outcomes largely depends on having access to data, and most cases require personal data. Against this backdrop, the need to foster data sharing as much as possible clearly emerges (Brambilla, Taddeo 2021). In the field of scientific research, openness of research data, represented in terms of sharing and reuse of research data, is a fundamental pillar of the Open Science approach (Tanlongo, et al. 2020). The Open Science approach is "the new way of conducting science, which aims to foster the openness of every phase of the scientific research process from data collection to the dissemination of scientific results, within the scientific community, and externally, towards society" (Paseri, 2022b). According to some scholars, however, there is a substantial divergence between Open Science intentions and the legal framework (Erb, et al. 2021, Dennis, et al. 2019).

Processing personal data is non-trivial since it is governed by data protection laws such as the European "General Data Protection Regluation" (GDPR, hereinafter)[2].

The common opinion among data practitioners, and professionals applying innovation in organizations on GDPR is that it only restricts data processing, thereby stifling innovation. While the private sector has commercial and business strategy constraints on top of legal compliance issues, accessing and sharing personal data is complex even in research settings where the awareness and resources for doing it efficiently are lacking. . In both contexts there is a considerable knowledge gap on the functions of pseudonymization, and the related privacy/residual risks. Second, when it comes to GDPR, the common practice is to completely avoid using personal data or fully anonymize it without considering technologies on preserving both privacy and utility. This could lead, among other consequences, to an under-use of the potential of personal data (Pagallo 2022a), which in certain sectors can generate a considerable economic impact (in particular, on the health sector, see: Pagallo 2022b). In this paper, we argue that GDPR implementation has

[2] 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), `ELI:http://data.europa.eu/eli/reg/2016/679/oj`.

yet reached its potential to achieve the goal of enabling personable data flows while respecting individual rights.

To close this gap, we introduce the importance of using new privacy-enhancing technologies, focusing on synthetic data, in order to implement the GDPR principles and enable responsible innovation. We explain the opportunities and challenges presented by synthetic data for GDPR particularly with regard to scientific research. Finally, we conclude with thoughts on future developments.

2. The GDPR Gap between Lawyers and Tech Practitioners

The European legal framework on data protection, i.e., the GDPR, was adopted in 2016 and became enforceable as of 25 May 2018, is arguably the world's most comprehensive and influential data protection law. The GDPR effectively harmonized the data protection laws of 27 EU countries and three EEA states, Iceland, Liechtenstein, and Norway, with a single standard for data protection. GDPR sets a clear legal basis for personal data processing, empowers data subjects with comprehensive rights and obligations to data controllers, and sets heavy penalties for non-compliance. Any entity required to process data from anywhere in the EU and the three EEA countries, regardless of location, must comply with GDPR (Kuner, Docskey, Bygrave 2020).

In recent years, the GDPR has inspired similar data protection laws, including Brazil's General Data Protection Law (LGPD), Japan's Act on the Protection of Personal Information (APPI), and California's Consumer Privacy Act (CCPA). While the GDPR is often hailed as a global victory for data protection, it is often vilified as a roadblock to innovation[3]. However, this is a misplaced view on both GDPR, since its essence, like any data protection law, is balancing rights while promoting economic activity and innovation (Durante 2021).

Let us briefly revisit the history of data protection laws to understand this misplaced view.

First and foremost, related to individual rights, the European Convention on Human Rights (ECHR) in 1950 codified the right to privacy, as well as the right to freedom of expression. However, the critical point is that the ECHR explicitly strikes a balance (Fuster, Van Brakel, De Hert 2022). With the liberalization of trade policies worldwide in the 1960s, the need to provide guidelines to strike a balance between protecting privacy rights and enabling trade flows gave rise to data protection laws.

[3] See: `https://cepr.org/voxeu/columns/gdpr-effect-how-data-privacy-regulation-shaped-firm-performance-globally`.

All the ensuing global and European data protection initiatives, including the OECD guidelines on data protection in 1980[4], the Council of Europe Convention 108 in 1981[5], the Data Protection Directive of 1995[6], up until GDPR in 2016, aim at ensuring responsible data flows by providing clear guidelines to data controllers while protecting individual rights.

However, in terms of implementation, primarily related to the digital age and GDPR, there has been a disproportionate focus only on the restriction of data use rather than on the utility and the responsible use of personal data for innovation. A prime example of this phenomenon is anonymization. The following section will discuss the data protection versus innovation trade-off related to anonymization.

3. Privacy and Utility:Beyond Anonymization

"Data can be either useful or perfectly anonymous but never both." - Paul Ohm

Anonymization is the process of removing all personal identifiers, both direct and indirect, that may lead to an individual being identified. Direct identifiers include name, address, postcode, telephone number, and photograph. At the same time, indirect identifiers refer to personal data that gives away identity when with other sources of information, including place of work, job title, salary, postcode, or health condition. In order to foster sharing and reuse of data (also research data), it is necessary to develop automated or semi-automated processing processes, and techniques of anonymisation or pseudonymisation can represent a useful instrument (Podda, Palmirani 2022).

According to the GDPR, anonymization is a process that transforms personal data into anonymous data "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."

According to Art. 4 (5) "pseudonymisation is processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional

[4] See: https://www.oecd-ilibrary.org/science-and-technology/oecd-guidelines-on-the-protection-of-privacy-and-transborder-flows-of-personal-data_9789264196391-en.

[5] See: https://www.coe.int/en/web/data-protection/convention108-and-protocol.

[6] See:https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX\%3A31995L0046.

information". The Recital 26 of the GDPR states that anonymous data are not considered as personal data.

However, these two concepts are often used synonymously by many practitioners, even according to ENISA (ENISA 2022).

Since completely anonymous data falls out of the GDPR scope, many organizations strive to make the personal data in their environment anonymous using various methods.

Some of the more common anonymization methods are as follows:

- Aggregation: Data is displayed as totals, so no data relating to or identifying any individual is shown. Small numbers in totals are often suppressed through 'blurring' or omitted altogether.

- Data masking: This involves stripping out obvious personal identifiers, such as names, from a piece of information to create a data set in which no personal identifiers are present.

- Data Perturbation: the values from the original dataset are modified to be slightly different.

- Data Permutation/Shuffling: The purpose of swapping is to rearrange data in the dataset such that the individual attribute values are still represented in the dataset but, generally, do not correspond to the original records. This technique is also referred to as shuffling and permutation.

- Suppression: Record suppression refers to the removal of an entire record in a dataset.

Although these techniques are considered anonymization techniques, they do not necessarily produce anonymous data, according to the GDPR.

For each of these techniques, there are documented re-identification risks such as identity disclosure/singling out, linkage, and inference (Farzanehfar, Houssiau, de Montjoye 2021). Not to mention that, often, in scientific research projects, anonymisation of data may risk compromising the results of the project itself (Shabani, Borry 2018).

The risk of re-identification of an individual from that data is zero. The trouble with this is that it is practically impossible to achieve and assure 100% anonymity if one wants to derive an iota of value from the resulting anonymous dataset.

This is called the privacy-utility conundrum. The higher the privacy, the lower the utility, and vice-versa. This conundrum has brought a renewed focus on Privacy Enhancing Technologies (PETs) to solve the

gap, that can be extremely beneficial in the private sector, in business, and in scientific research.

Privacy-enhancing technologies (PETs) are digital solutions that allow the collection, analysis, and sharing of personal data while protecting confidentiality and privacy (Tavani, Moor 2001; Ghioni, Taddeo, Floridi 2023).

They refer to privacy-preserving data sharing and analytics technologies that enable data sharing and analysis among participating parties while maintaining dissociability and confidentiality. Examples include, but are not limited to, secure multiparty computation, homomorphic encryption, zero-knowledge proofs, federated learning, secure enclaves, differential privacy, and synthetic data generation.

The discussion on each of the PETs is beyond the scope of this paper. However, our focus is on synthetic data, which among the various privacy-enhancing technologies, has been demonstrated to provide the highest privacy vs. utility scores, especially when dealing with large-scale personal data (Hradec et al. 2022; Digital Dubai 2022).

4. Synthetic Data and Its Application in a Nutshell

Synthetic Data is artificially generated data, often using AI algorithms upon real-world 'seed' data. It has the same statistical properties and predictive power as the real data on which it was generated. Using synthetic data may represent a safe proxy for real data since it contains no real personal information for several AI and analytics use cases, such as data science/AI projects, test automation, and, most importantly, privacy preservation.

The synthetic data generation for privacy preservation usually consists of four elements: (i) The starting data set that needs to be synthesized, (ii) a generation method, (iii) the resultant synthetic data and insights into the privacy and (iv) utility metrics of the synthetic data. There are two main types of synthetic data (a) partially synthetic data, which has only some synthetic variables compared to the original data and (b) fully synthetic data where all the variables are synthetic.

There are many generation techniques for synthetic data generation. The two most common methods used are based on Generative Adversarial Networks (GANs) and Variational AutoEncoders (VAEs). GANs are deep neural networks that can generate new data samples based on an adversarial training process consisting in training a neural network, i.e., the generator, to produce new data points, and another one, i.e., the discriminator, to predict whether the points produced by the generator are real or fake (Goodfellow et al. 2020).

Variational Autoencoders (VAE) are probabilistic models that can learn to compress (encode) data into a meaningful and tractable representation that can be used to sample original probability distributions using a reconstruction function (decoder) (Kingma, Welling 2013).

However, the mere generation of synthetic data alone does not guarantee privacy and can be susceptible to the re-identification risks of regular personal data (Stadler, Oprisanu, Troncoso 2022). Therefore, all the generated synthetic data should have quantifiable privacy risks (differential privacy, for example). In addition to the relatively qualitative Data Protection Impact Assessment (DPIA), a data controllers should perform a computational privacy assurance assessment to ensure that the resulting synthetic data is not personal data. A differentially private synthetic dataset looks like the original dataset – with the same properties (e.g., correlations between attributes) – but it provides a provable privacy guarantee for individuals in the original dataset, thereby providing both the features of privacy and utility (Nears, Darais 2021).

Synthetic data has applications in many sectors. In the past years, many use cases in both the private and public sectors have demonstrated the value of synthetic data (ex multis, specifically in the sector of scientific research, see Azizi et al. 2021).

The demand for synthetic data is especially prevalent in regulated sectors such as finance, insurance, and healthcare and sectors like energy, mobility, and transport, where data is scarce, or data collection is expensive and unsafe.

Financial giants such as JP Morgan have dedicated research teams on synthetic data since financial services generate highly complex and varied data with customers' most sensitive and personally identifiable attributes (Assefa et al. 2020). These datasets are often stored in silos within organizations to meet regulatory requirements and business needs. Therefore accessing and sharing these datasets for innovations is severely limited, so they are exploring Synthetic Data as a viable alternative to working on improving their fraud detection algorithms and customer satisfaction rates.

The automotive industry giants, including Ford and BMW, use synthetic data to train their autonomous driving systems. In healthcare, medical imaging startups like Cure AI use synthetic data (in this case, 400000 synthetic patient records) to train AI models while protecting patient privacy (Andrews 2021).

Gartner predicts that 60% of AI models will use synthetic data in some form or another by 2025 (White 2021). Its recent market trends report also reiterated the importance of generative AI that fuels data generation in its recent hype cycle. While many advantages of synthetic

data drive this trend, data augmentation, cost-effective and safe data procurement, the main focus these days is its role in privacy preservation and enhancement. There have already been a few impactful examples of synthetic data use cases for privacy preservation. The US Census Bureau employs it in their public datasets and online tools to "balance the competing requirements of releasing statistics and protecting privacy" (US Census Bureau 2021).

A research project mandated by the city of Dubai showed that differentially private-synthetic data outperforms traditional data anonymisation techniques (such as removal, substitution, masking, and aggregation) in terms of protecting the privacy of individuals and boosting data utility (or usefulness) (Digital Dubai 2022). For a dataset containing traffic accidents, they almost completely protect individuals' privacy while preserving 90% of the utility, as measured against the original dataset.

Moreover, in light of these applications, it becomes very evident how scientific research can benefit from the use of synthetic data sets. From a technical standpoint, this could further facilitate openness, meaning the sharing and reuse of personal data processed for scientific research purposes without prejudicing the individuals involved in the research projects.

5. Opportunities and Legal Challenges of Synthetic Data

Synthetic data offers many advantages for data protection, thereby, can be considered an important instrument to comply with regulations such as GDPR, in particular in the scientific research sector in which the sharing is a crucial aspect of the scientific community's modus operandi. The key advantages are summarized below.

- Synthetic data can be an effective protection mechanism against direct re-identification, especially in the case of fully synthetic datasets.

- Synthetic data can be seen as a measure in line with the concept of data protection by design, since it allows for additional privacy safeguards like privacy risk assessment metrics.

- Synthetic data can capture the statistical characteristics of high-dimensional datasets, i.e., datasets that have a high number of dimensions. By obfuscating individual data in the statistical properties of the data, synthetic data provides a more precise depiction of complex datasets while safeguarding individuals' identities.

– Synthetic data can help improve imbalance or biased datasets, thereby improving the representativeness of the dataset, which also adheres to the ethical aspects of profiling in GDPR. Consider that the principle of non-discrimination, in addition to being a founding principle of the EU law, with regard to the field of scientific research is also one of the principles underlying the Open Science approach.

Some of the shortcomings of synthetic data regarding privacy preservation and data protection are related to the fact that synthetic data generation is an active research area and has yet to reach peak application maturity in all sectors. Other areas for improvement include the issues of replicating outliers in the original dataset and the dependence on the synthetic data quality based on the original dataset.

Beyond these privacy-enhancing features, synthetic data has also been considered a data minimization technique. Some regulators have explicitly recommended using it to comply with the data minimization principle of GDPR. Examples include the recommendation of the Italian Data Protection Authority's guidelines on using synthetic data for building AI models and the opinion of the Norwegian Data Protection Authority that fined a company that did not have a legal basis for performing software testing with this personal data. It opined that the testing could have been achieved by processing synthetic data (EDPB 2021).

The early applications of using synthetic data for data sharing are nevertheless emerging. The Netherlands Comprehensive Cancer Organisation (IKNL), has piloted the use of synthetic data to share breast cancer data. Another impactful example is Ireland's Central Statistics Office's (CSO) Synthetic House Price Dataset created as a statistical training tool for educators in universities and elsewhere.

6. Conclusions: Synthetic Data Applications' and Future Developments

The generation of synthetic data may represent one of the game-changing technologies able to shape the future of privacy, data protection and AI. Although supervisory authorities and regulators are closely following the application of PETs, such as synthetic data, still a gap in legal certainty persists. In other words, there is no assurance to companies that such technologies are GDPR compliant, even though they are more effective than "approved" methods such as anonymization. There has not been an updated guidance similar to the extensive guidance by the previous Article 29 Working Party on anonymisation techniques (WP29 2014).

However, there is slow but steady progress in how regulators embrace such emerging technologies.

Regulators, including the European Data Protection Supervisor and Office of the Privacy Commissioner of Canada, have set up dedicated task forces to study the potential of PETs, including synthetic data within the data protection regulatory framework.

The Communiqué of G7 Data Protection and Privacy Authorities rightly pointed out that PETs like synthetic data "can facilitate safe, lawful and economically valuable data sharing that may otherwise not be possible, unlocking significant benefits to innovators, governments and the wider public. In recognition of these benefits (. . .) the G7 data protection and privacy authorities (. . .) will seek to promote the responsible and innovative use of PETs to facilitate data sharing, supported by appropriate technical and organizational measures" (G7 Germany 2022).

Beyond statements, regulators should give comprehensive, updated, and unambiguous guidance for companies, organizations, universities and research centers on using PETs synthetic data to harness the power of new technologies responsibly and prove that data protection is not a deterrent but an enabler to innovation and societal prosperity and well-being and stay competitive.

References

Andrews, G. (2021) What is Synthetic Data?, Ndiva Blog, https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data/.

Assefa, S. A., et al. (2020) Generating synthetic data in finance: opportunities, challenges and pitfalls, Proceedings of the First ACM International Conference on AI in Finance, pp. 1-8.

Azizi, Z., et al. (2021) Can synthetic data be a proxy for real clinical trial data? A validation study, BMJ open 11.4, e043497.

Brambilla Pisoni, G., Taddeo, M. (2022) Apropos Data Sharing: Abandon the Distrust and Embrace the Opportunity, DNA and Cell Biology 41.1, pp. 11-15.

Dennis, S., et al. (2019) Privacy versus open science, Behavior research methods 51, pp.1839-1848.

Digital Dubai (2022), Digital Dubai Launches Framework on Synthetic data and Its Role in Enhancing Artificial Intelligence, https://www.digitaldubai.ae/newsroom/news/digital-dubai-launches-framework-on-synthetic-data-and-its-role-in-enhancing-artificial-intelligence.

Durante, M. (2021), Computational Power: The impact of ICT on law, society and knowledge, London, Routledge.

EDPB (2021) Norwegian DPA: Norwegian Confederation of Sport fined for inadequate testing, https://edpb.europa.eu/news/national-news/2021/norwegian-dpa-norwegian-confederation-sport-fined-inadequate-testing_en.

ENISA (2022) Data Protection Engineering. From Theory to Practice, https://www.enisa.europa.eu/publications/data-protection-engineering, pp. 1-42.

Erb, B., et al. (2021) Emerging Privacy Issues in Times of Open Science, Psyarxiv, pp. 1-5.

Farzanehfar, A., Houssiau, F., de Montjoye, Y-A (2021) The risk of re-identification remains high even in country-scale location datasets, Patterns 2.3, p. 100204.

Fuster, G. G., Van Brakel, R., De Hert, P. (2022) Introduction to Research Handbook on Privacy and Data Protection Law, Research Handbook on Privacy and Data Protection Law, London, Edward Elgar Publishing, pp. 1-8.

Ghioni, R., Taddeo, M., Floridi, L. (2023) Open source intelligence and AI: a systematic review of the GELSI literature, AI & society, pp. 1-16.

Goodfellow, I., et al. (2020) Generative adversarial networks, Communications of the ACM 63.11, pp. 139-144.

G7 Germany (2022) Roundtable of G7 Data Protection and Privacy Authorities Promoting Data Free Flow with Trust and knowledge sharing about the prospects for International Data Spaces, Communiqué, 8 September 2022.

Kuner, C., Docksey, C., Bygrave, L. (2020), The EU General Data Protection Regulation: A Commentary, Oxford, Oxford University Press. Hradec, J., et al. (2022) Multipurpose synthetic population for policy applications, Luxemburg, Publications Office of the European Union, Luxembourg, pp. 1-89.

Kingma, Di. P., Welling, M. (2013) Auto-encoding variational bayes, arXiv preprint arXiv, 1312.6114.

Leonelli, S. (2020) Scientific Research and Big Data, The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), `https://plato.stanford.edu/archives/sum2020/entries/science-big-data/`.

Near, J., Darais, D. (2021) Differentially Private Synthetic Data, NIST Blog Post, `https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data`.

Pagallo, U. (2022a), The Politics of Data in EU Law: Will It Succeed?, Digital Society 1.3, pp. 1-20.

Pagallo, U. (2022b), Il dovere alla salute: Sul rischio di sottoutilizzo dell'intelligenza artificiale in ambito sanitario, Milano-Udine, Mimesis.

Paseri, L. (2022a), *Il ruolo delle istituzioni in design, sviluppo e applicazione dell'IA per il settore della ricerca scientifica*, Queste istituzioni, 2022, pp. 213-226.

Paseri, L. (2022b), *From the Right to Science to the Right to Open Science. The European Approach to Scientific Research*, European Yearbook on Human Rights, Intersentia, 2022, pp. 515-541.

Podda, E., Palmirani, M. (2022) Anonimizzazione e Pseudonimizzazione di Sentenze Giudiziarie, La trasformazione digitale della giustizia nel dialogo tra discipline, Milano, Giuffrè, pp. 37-64.

US Census Bureau (2021) What Are Synthetic Data? `https://www.census.gov/about/what/synthetic-data.html`.

Shabani, M., Borry, P. (2018) Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation, European Journal of Human Genetics 26.2, pp.149-156.

Stadler, T., Oprisanu, B., Troncoso, C. (2022) Synthetic data–anonymisation groundhog day, 31st USENIX Security Symposium (USENIX Security 22).

Tanlongo, F., et al. (2020), I want to be an Open Scientist! Research evaluation and incentives to boost Open Science and research careers, Zenodo.

Tavani, H. T., Moor J. H. (2001) Privacy protection, control of information, and privacy-enhancing technologies, ACM Sigcas Computers and Society 31.1, pp. 6-11.White, A. (2021) By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated, Gartner Blog, `https:`

//blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/?__s=mo8bhbyt1tfaqoqncsa7.

WP29 (2014) Opinion 05/2014 on Anonymisation Techniques, 10 April 2014, pp. 1-37.