LUISS

Master of Science in Law, Digital Innovation and Sustainability

Chair of Data Protection Law

Synthetic Data: Between Anonymization and Parameterization.

A New Frontier for Data Protection?

Supervisor:

Prof. Filiberto Brozzetti

Co-Supervisor:

Prof. Sofia Hina Fernandes De Silva Ranchordas

Candidate:

Flaminia Tosto

631743

Academic Year 2024/2025

Table of Contents

List of figures

Introduction

The contemporary digital landscape is characterized by a seemingly contradictory dual nature: on the one hand, we witness the consolidation of a European culture built on the importance of personal data protection, as demonstrated by the adoption of rigorous regulations such as the General Data Protection Regulation (GDPR), which has now become an identity pillar of the Union; on the other hand, the emergence of artificial intelligence and advanced digital technologies requires increasingly broad access to high-quality datasets in order to fuel innovation and maintain Europe's competitiveness on a global scale. This fundamental tension between the protection of individual rights and the need for technological progress does not merely represent a technical or regulatory issue but rather touches the core of European identity and its ability to assert itself as an autonomous digital power. This challenge constitutes the central problem that the present research seeks to address and, if possible, resolve.

The emergence of synthetic data as an innovative technological paradigm offers a particularly promising perspective for reconciling these seemingly irreconcilable needs. However, their practical application raises legal and technical questions of considerable complexity, requiring a thorough multidisciplinary analysis to be adequately understood and managed.

The central question that runs through the entire research can be formulated as follows: to what extent can synthetic data be considered an effective solution to the paradox between personal data protection and technological innovation, and what regulatory and technical conditions are necessary to ensure that this solution is both legally compliant and practically usable?

This research question is articulated into several interconnected sub-questions. First, it becomes necessary to determine if and when synthetic data can be considered truly anonymous under a legal perspective, thereby excluded from the scope of the GDPR; this evaluation focuses not only on quantitative aspects but also on the qualitative dimensions that contribute to defining the degree of privacy protection offered by these technologies. Second, it is crucial to analyse the technical risks associated with synthetic data generation, with particular attention to the phenomenon of overparameterization and the associated re-identification risks. Ultimately, it becomes essential to explore complementary methodologies, such as machine unlearning and differential privacy, capable of strengthening privacy guarantees without compromising the utility of the generated data.

This thesis is structured into four interconnected chapters, each addressing specific aspects of the complex multidisciplinary relationship between synthetic data, personal data protection, and

technological innovation, with particular attention to the strategic role that the Italian company Aindo has managed to secure in this emerging sector.

Chapter 1, “Data as a Fundamental Asset for Europe”, provides the regulatory and strategic context necessary to understand the importance of data, especially personal data, in the European digital economy. The analysis focuses on the evolution that has transformed personal data from mere objects of legal protection into strategic resources for economic competitiveness and digital innovation. It examines how the European Union has sought to balance these objectives through the European Data Strategy, implemented through regulatory instruments such as the Data Governance Act and the Data Act. Particular attention is devoted to the emerging paradoxes between the fundamental principles of the GDPR, such as purpose limitation and data minimization, and the data reuse requirements promoted by the new regulations. The chapter concludes by identifying synthetic data as a potential solution to these regulatory tensions.

Chapter 2, “What is Synthetic Data? Understanding Its Role and Potential”, offers an in-depth technical analysis of synthetic data, examining generation methodologies from traditional techniques based on regression models and maximum likelihood estimation to modern architectures such as Generative Adversarial Networks (GANs). The theoretical analysis is integrated with practical examples in the financial and healthcare sectors, demonstrating how synthetic data can enable innovation in contexts characterized by high data sensitivity. The chapter also positions synthetic data within the Privacy Enhancing Technologies (PETs) framework, highlighting its potential as a privacy by design tool, capable of transforming the scarcity of accessible data into an abundance of resources for innovation.

Chapter 3, “The Overparameterization Problem”, tackles the crucial issue of synthetic data anonymization from a legal perspective. The analysis begins with the examination of the anonymization concept developed by European case law, with particular reference to Recital 26 of the GDPR and Opinion 05/2014 of the Article 29 Working Party on anonymization techniques. It also discusses the recent ruling of the Court of Justice of the European Union of 4 September 2025 (EDPS v SRB), which clarified the conditions under which pseudonymized data may be treated as anonymous by the recipient. Particular attention is given to EDPB Opinion 28/2024 on artificial intelligence models and its implications for synthetic data, analysing how this opinion, although not explicitly addressing synthetic data, delineates significant interpretative space for their legal qualification. The chapter then identifies three main categories of privacy risks: the perpetuation of algorithmic biases, the phenomenon of model collapse, and, above all, the overparameterization problem, which constitutes one of the most significant original contributions of this research.

Chapter 4, “Shaping the Future of AI and Data Protection: The Aindo Case”, proposes integrated technical solutions to address the risks identified in the previous chapter. The analysis focuses on two methodologies complementary to synthetic data: differential privacy, as a preventive mechanism to introduce mathematically rigorous privacy guarantees during the synthetic data generation phase, and machine unlearning, as a corrective tool to selectively remove specific information from already trained models. The chapter also includes a detailed case study of Aindo, an Italian start-up specializing in synthetic data that has turned privacy challenges into business opportunities, demonstrating the practical applicability of the proposed solutions.

The methodology adopted in this research is inherently interdisciplinary, combining the legal analysis of European data protection legislation with the technical examination of emerging technologies. On the legal front, the study is based on a comprehensive analysis of the European regulatory framework, including the GDPR, the Data Governance Act, the Data Act, and the AI Act, integrated with the opinions and guidelines provided by European supervisory authorities, in particular the EDPB and the Article 29 Working Party.

From a technical perspective, the research draws on a wide international scientific literature in machine learning, differential privacy, and synthetic data generation technologies. Particular attention has been given to the most recent contributions on overparameterization and re-identification risks, areas of rapid research evolution that require continuous source updating.

The empirical component of the research includes an in-depth interview with Daniele Panfilo, CEO of Aindo, and Alexander Boudewijn, Aindo’s Data Privacy Researcher, who provided practical insights into the implementation of synthetic data generation technologies, including how Aindo addresses intrinsic challenges such as bias management and outlier handling. This integration of theoretical analysis and practical experience lends empirical robustness that significantly enriches the scientific and applicative value of the results obtained.

The sources used range from academic peer-reviewed contributions to official regulatory documents, technical reports from international organizations, and industry publications. This diversification of sources reflects the intrinsically interdisciplinary nature of the field of study and the need to integrate legal, technical, and industry perspectives for a comprehensive understanding of the phenomenon.

It is important to underscore that the research field of this thesis is characterized by particularly rapid and dynamic evolution. Synthetic data, machine unlearning, and differential privacy are emerging technologies whose development is proceeding at an accelerated speed, with new scientific contributions and practical applications continuously emerging. This evolutionary nature implies that some of the sources used in this research may undergo changes, integrations, or supersessions in the

short term. However, this characteristic does not diminish the validity of the adopted methodological approach, but rather underscores the importance of establishing a solid and flexible conceptual framework capable of adapting to future developments while maintaining its theoretical and practical relevance.

The research therefore aims not only to provide an analysis of the current state of the examined technologies and regulations, but also to outline principles and methodologies that can remain valid even in the face of the rapid changes characterizing this field. In this sense, the objective is to contribute to the construction of a body of knowledge that can guide both policymakers and technology experts in navigating the future challenges of privacy-preserving innovation.

This research thus fits into an ongoing scientific and regulatory debate, with the ambition of making an original contribution that can positively influence the development of more privacy-respecting technologies while also supporting innovation and economic growth in the digital era. The proposal of an innovative legal framework for synthetic data, based on the integrated consideration of qualitative and quantitative elements, represents a concrete attempt to overcome current interpretative uncertainties and provide sector operators with legally robust tools for implementing these emerging technologies.

Chapter 1 – Data as a Fundamental Asset for Europe

1.1 The Data Economy and the rising value of personal data

Data, especially personal data, are gaining increasing importance, not only at the European level, but also on a global scale. From a broader perspective, it is evident that data have effectively become a source of wealth and power for nations.

With the introduction of the General Data Protection Regulation (GDPR)[1] in 2016, the primary objective was to empower data subjects to keep control over their personal data, since with the rapid rise of new technologies it was necessary to implement a more modern regulatory framework, one that reshaped data governance worldwide, reflecting the cross-border nature of digital information.

As a matter of fact, the GDPR stands as an emblematic example of the consolidation of the so-called "Brussel Effect"[2], this concept was coined and developed by Anu Bradford in 2012. It is a phenomenon through which the European Union (EU) successfully projects its regulatory influence globally, impacting many other countries and not limiting the culture of data protection only to Europe. By analysing in detail the strategic factors[3] that have contributed to the affirmation of this phenomenon in the area of data protection, it is possible to understand how the EU has affirmed itself at a global level.

One crucial factor to consider is the size of the European market. As a key indicator of economic influence, the EU's large consumer base, characterized by high purchasing power, makes access to this market essential for many global businesses. As a result, companies are strongly incentivized to comply with the strict data protection rules outlined in Regulation (EU) 2016/679, rather than risk losing access to such a lucrative market. This compliance incentive is further reinforced by the fact that the GDPR applies to inelastic objectives, meaning the personal data of EU citizens. Companies

[1] General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, *Official Journal of the European Union* L 119/1, May 4, 2016.

[2] "*The Brussels Effect refers to the EU's unilateral power to regulate global markets. Without the need to use international institutions or seek other nations' cooperation, the EU has the ability to promulgate regulations that shape the global business environment, leading to a notable 'Europeanization' of many important aspects of global commerce. Different from many other forms of global influence, the EU does not need to impose its standards coercively on anyone— market forces alone are often sufficient to convert the EU standard into the global standard as companies voluntarily extend the EU rule to govern their worldwide operations. Under specific conditions, the Brussels Effect leads to 'unilateral regulatory globalization,' where regulations originating from a single jurisdiction penetrate many aspects of economic life across the global marketplace*." Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (Oxford: Oxford University Press, 2020), p.xiv.

[3] *Ivi*, p.12.

wishing to offer goods or services to individuals within the EU cannot circumvent the regulation simply by relocating their operations or users to other jurisdictions. Since personal data are inherently tied to individuals residing within EU territory, they remain subject to European regulatory oversight regardless of geographical location[4].

It is evident that the key EU institutions have developed a significant regulatory capacity, allowing them to shape global standards. They possess the expertise and resources necessary to draft and enforce complex regulations such as the GDPR. Moreover, compliance with data protection laws is reinforced by substantial enforcement mechanisms, including the decision to impose fines of up to 4%[5] of a company's global turnover in cases of non-compliance. The risk of severe financial penalties, combined with potential reputational damage, serves as a strong deterrent, further incentivizing adherence to the GDPR.

Over the years, however, the perspective has broadened, and personal data have evolved from merely being assets to be protected into a resource with immense potential. No longer seen solely as a source of risk and vulnerability, data have become a key driver of profit and value for those capable of fully harnessing their potential.

Already in 2017, the European Commission wrote in the Communication "Building a European Data Economy"[6] that "*Data has become an essential resource for economic growth, job creation and societal progress. Data analysis facilitates the optimisation of processes and decisions, innovation and the prediction of future events. This global trend holds enormous potential in various fields, ranging from health, environment, food security, climate and resource efficiency to energy, intelligent transport systems and smart cities. The "data economy" is characterised by an ecosystem of different types of market players – such as manufacturers, researchers and infrastructure providers – collaborating to ensure that data is accessible and usable. This enables the market players to extract value from this data, by creating a variety of applications with a great potential to improve daily life (e.g. traffic management, optimisation of harvests or remote health care)*".[7] So, today, we can no longer speak of a *Data-Driven Economy* or a *Data-Driven Market*, we now live in a *Data Economy*, a *Data Market*, where data are not just the fuel powering the market but the very object of exchange.

---

[4] "*This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not*." *Ibid*., art. 3(1)

[5] *Ivi*, art. 83(5)

[6] European Commission, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Building a European Data Economy*, COM(2017) 9 final, Brussels, January 10, 2017. EUR-Lex - 52017DC0009 - EN - EUR-Lex

[7] *Ivi*, p. 2.

Unlike oil, which is finite and often mistakenly compared to data, data are limitless, constantly generated, and ever-changing. This shift represents an entirely new sociological paradigm.

The exponential growth of data has made this market increasingly flexible and dynamic, highlighting the need for regulatory frameworks to evolve accordingly.

1.1.1 The European Data Strategy

The European Commission has firmly focused its political agenda to the concept of the *twin transition*, a term that contains the parallel pursuit of digitalization and sustainability. This dual approach has become the cornerstone of European policy-making in recent years. Under the leadership of President Ursula Von der Leyen, the digital transformation of the European Union has been elevated to a top strategic priority. Digitalization, alongside environmental sustainability, has thus been placed at the core of the EU's long-term vision, building a shift toward an integrated model of governance capable of addressing both ecological and technological challenges in a coordinated manner.

What emerges is not merely a sectoral policy agenda, but a broader geopolitical strategy. In the current global landscape, marked by increasing technological competition and the consolidation of digital empires by major powers such as the United States and China, the European Union has recognized that asserting its sovereignty requires moving beyond traditional levers of power. Rather, digital sovereignty, understood as the capacity to develop, govern, and regulate key digital infrastructures, data flows, and emerging technologies independently, has become essential to preserving Europe's autonomy and relevance on the global stage.

In order to effectively put into action its European Data Strategy, the European Union has introduced two key legislative instruments: the Data Governance Act (DGA)[8] and the Data Act[9]. This strategy was implemented with the ambition to unlock the full potential of the vast amounts of data generated within Europe, with a particular focus on strengthening the internal market and increasing the value of European data. Central to this vision are objectives such as promoting data sharing and reuse across public and private sectors, encouraging data availability for research purposes, and leveraging data as a strategic asset for innovation, economic growth, and global competitiveness.

[8] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act), *Official Journal of the European Union*, L 151, June 3, 2022. Regulation - 2022/868 - IT - EUR-Lex

[9] Regulation (EU) 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, amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), *Official Journal of the European Union*, L 2854, December 22, 2023. Regolamento - UE - 2023/2854 - IT - EUR-Lex

First of all, the Data Governance Act, that is applicable since September 2023, serves as a comprehensive regulatory framework to facilitate and incentivize trust in voluntary data sharing and re-use of data across sectors. It introduces the concept of data intermediaries, as neutral entities designed to manage data sharing between parties, and promotes data altruism, enabling individuals and organizations to voluntarily make their data available for purposes of general interest. The DGA encompasses both personal and non-personal data, with the GDPR continuing to apply where personal data is involved. A key feature that emerges in this piece of legislation is that it incorporates additional safeguards to foster trust in data-sharing mechanisms, recognizing that building public confidence is essential to ensure a robust and thriving data economy.

Complementing the DGA, the Data Act, which entered into force on 11 January 2024, represents a cornerstone of the EU's data policy framework. Its overarching aim is to position Europe as a global leader in the data economy by capitalizing on the exponential growth of industrial data. The regulation seeks to foster a more equitable data environment by establishing clear rights and obligations concerning access to and use of data, thus enabling a more balanced distribution of data-derived value among businesses, consumers, and public bodies.

While these regulations mark a decisive step towards shaping a coherent data ecosystem within the Union, they also reveal certain limitations in the EU's approach. The strong regulatory focus, although beneficial in setting global standards, risks to become isolated, if it fails to engage with the broader geopolitical dynamics that influence data flows and technological development. By concentrating primarily on rules applicable within its own borders, the EU may overlook the inherently transnational nature of the digital environment. Moreover, the proliferation of regulatory frameworks, could create friction with innovation and undermine the capacity to adapt to fast-moving technological trends.

This raises an essential question for the next phase of the EU's digital agenda: how to strike a sustainable balance between fostering innovation and safeguarding fundamental rights, particularly the protection of personal data. As the EU continues to consolidate its leadership in the global regulation of technology, it must remain mindful of the need to combine normative ambition with practical flexibility, ensuring that the protection of privacy does not become a barrier to innovation, but rather a pillar of trustworthy technological advancement.

1.2 Balancing technological innovation and data protection

President Von der Leyen clearly expressed the ambition of Europe to have the digital transformation to power our economy and the aim is to find European solutions in the digital age.

Even the GDPR recognized, in Recital 6, that "*technology has transformed both the economy and social life, and should further facilitate the free flow of personal data within the Union and the transfer to third countries and international organisations, while ensuring a high level of the protection of personal data*."[10]

In order to realistically achieve the objectives of technological independence and development set by the European Union, it is essential to acknowledge that making Europe truly competitive, as it aspires to be, requires finding a delicate compromise between the protection of personal data and the promotion of technological innovation within the Union. This requires strategic solutions that foster competitiveness and economic growth without undermining the fundamental rights of individuals, particularly the right to data protection, which is recognized as a fundamental right in the EU Charter of Fundamental Rights[11].

As previously discussed, Europe's greatest asset lies in the quality and quantity of its data, and it's within this vast and valuable dataset that lies the potential to drive innovation and unlock a key competitive advantage. However, compared to other major global powers such as the United States and China, Europe lacks the same levels of investments, natural resources, and digital infrastructures necessary to assert itself as a global leader in innovation. As noted by Kai-Fu Lee, a prominent voice in the Chinese tech sector, the race for digital leadership allows no bronze medal, and Europe remains excluded from the leading group of nations driving the technological transformation[12]. In light of this, what remains for Europe is to capitalize on the vast potential of its data, a resource that is still considered as highly valuable and attractive. However, if this potential is not properly harnessed, or if access to such data remains overly burdensome, the uniqueness and desirability of European data may not last.

Finding a balance between technological innovation and data protection may appear simple in theory, but a closer, more critical examination reveals a significant paradox at the heart of this challenge.

[10] GDPR, Regulation (EU) 2016/679, Recital 6.
[11] Charter of Fundamental Rights of the European Union, 2000/C 364/01, December 18, 2000, art. 8
[12] Tyson Barker, "Europe Can't Win Its War for Technology Sovereignty. The European Union is running in circles in pursuit of 'digital sovereignty'." *Foreign Policy*, January 16, 2020, https://foreignpolicy.com/2020/01/16/europe-technology-sovereignty-von-der-leyen/.

It is now widely acknowledged that artificial intelligence (AI), particularly generative AI, is among the most disruptive tech innovations of recent years, bringing with it as many risks as opportunities. When examining this technology through the lens of data protection, numerous paradoxes inevitably emerge. As clearly remarked by President Von der Leyen: "*Artificial Intelligence is about big data, data, data and again data. And we all know that the more data we have, the smarter our algorithms. This is a very simple equation. And therefore, it is so important to have access to data that are out there. This is why we want to give our businesses, but also our researchers, and the public services better access to data. We all know that we are generating, by the day, data and data again […]*"[13]. From this statement, it emerges in a very clear way, that training an artificial intelligence system to achieve accurate and reliable results requires vast amounts of data, the more high-quality data is used, the better the outcome will be. While such resources are indeed available, the issue arises when, despite the efforts of the Data Governance Act and the Data Act to improve data accessibility, the use of personal data still involves critical elements that cannot be overlooked.

It is essential to identify and address the potential challenges in this field in order to develop adequate solutions, capable to find the right balance. In an era where falling behind in the digital transformation is no longer an option, it is equally a priority not to forget to protect the identity and the rights of the European citizens.

1.2.1 The paradox between data reuse and the principle of purpose limitation

There are different areas of tension between technological innovation and the protection of data, and it is sufficient to take a more critical look at the regulations issued on data-related matters to realize that problems may arise when these rules are applied to concrete cases. Specifically, conflicts and ambiguities may emerge due to substantial differences in the primary objectives pursued by each regulatory framework.

The GDPR outlines a set of principles that must always be considered when processing personal data. Among these, it is particularly relevant to mention the principle of data minimization[14], which states that it must be used the minimum amount of data possible, just the data strictly necessary to achieve the purpose of the processing and nothing more, only what is strictly proportionate, in substance, the less the better. According to this principle, a clear conflict emerges with the intrinsic nature of an AI model, where, on the contrary, the more data is used, the better the outcome. Of course, this must also

[13] European Commission, "Press remarks by President von der Leyen on the Commission's new strategy: Shaping Europe's Digital Future," Brussels, February 19, 2020, https://ec.europa.eu/commission/presscorner/detail/en/speech_20_294.
[14] GDPR, Regulation (EU) 2016/679, art. 5 (1)(c)

align with the principle of accuracy, as not only must the volume of data be significant, but its content and quality must also be appropriate and relevant to the intended purpose.

In addition to this conflict, the most significant paradox arises when comparing the principle of purpose limitation[15], as established by the GDPR, with the concept of data re-use, which represents one of the cornerstones of the Data Governance Act.

The purpose limitation principle states that, since the data subject must be informed about the reasons their data is being processed and for which purposes, it is the data controller who determines both the means and purposes of the processing activity. Therefore, once the purposes are clearly defined and made transparent, the data cannot be processed for other, unrelated purposes. Any further processing that is not previously declared, made transparent, or that is not compatible and connected to the original purpose is not lawful, except in cases where the data is used for statistical purposes.

This principle is challenged by recent data governance regulations, such as the Data Governance Act and the Data Act, which instead promote the re-use of data in order to unlock and enhance its value within the market.

The goal of this strategy is this one: European companies, public administrations and research centres as well as academia process personal information legitimately under the GDPR. At the end of their original processing activities, they are left with large amount of datasets which, under current rules, should be deleted once the processing activities are finished, since there is no more legitimate basis to retain or process them. But why should we give up on such a valuable resource? Could there be room to envision a legal and technical solution capable of reconciling these tensions?

1.2.2 Could synthetic data be the key to solve this dilemma?

In light of what has been analysed previously in this study, it is possible to affirm that unjustified restrictions on the free movement of data are likely to constrain the development of the EU data economy[16], and it is important to remember that privacy concerns are legitimate concerns but should restrict the free flow of data or limit the development of technological innovations.

To answer the question posed above, one possible solution that could balance the necessity to protect personal data while also ensure the advancement of technologies like artificial intelligence, lies in the synthetization of data.

[15] *Ivi*, art.5 (1)(b)
[16] European Commission, *Building a European Data Economy*, COM(2017) 9 final, p.3.

Synthetic data is accurately defined by Professor D'Acquisto as an artificially generated counterpart of real-world data. While real-world data are directly extracted from observed phenomena, synthetic data are computer-generated to emulate the key characteristics and properties of those phenomena that are relevant to a specific use case.[17]

This tool could represent a valuable solution for striking a balance between data protection and technological innovation, acting as a privacy-enhancing technology (PET). The generation of synthetic data can reduce or even eliminate the need to process personal data, achieving comparable outcomes to those obtained through real data. This is particularly useful in scenarios where access to real data is restricted due to legal constraints. Thanks to their application, organizations can validate their decisions under various simulated conditions without relying on direct observation, in a more controlled and less risky manner for individuals and resources.

Moreover, synthetic data is widely used across various domains, from the development and validation of machine learning algorithms to research in healthcare and social sciences, and even in strategic decision-making in political, economic, or military contexts. They can overcome data scarcity, allowing researchers and developers to augment and label their datasets at minimal cost.

Synthetic data emerges as a powerful accountability tool, offering the opportunity to strike a delicate balance between scientific progress and the rights of individuals. They are increasingly viewed as a way to implement a "data protection by design" approach in scenarios that involve the processing of personal data.

From this analysis, it emerges that synthetic data offers the possibility of using non-personal data for development, testing, analysis, and research purposes, thereby reducing the privacy risks associated with processing real personal data. At the same time, they help overcome limitations on access to real data and enable the generation of large amounts of data for technological innovation, representing a potential meeting point between the need to protect personal information and the urgency to advance science and technology. Nevertheless, it is essential to proceed with caution to ensure that data protection laws are not violated, as the line separating personal and anonymous data is often fine and context-dependent.

Therefore, the aim of this research will be to determine whether synthetic data can be truly considered anonymous, and therefore considered an effective tool for training AI models without exposing

[17] Giuseppe D'Acquisto, *"*Synthetic Data and Data Protection Laws*," Journal of Data Protection & Privacy 6*, no. 3 (2024): pp. 227–239.

personal data. Furthermore, it will explore whether anonymity should be assessed based solely on quantitative factors, or should we also take qualitative aspects into account.

Chapter 2 – What is Synthetic Data? Understanding Its Role and Potential

2.1 What is synthetic data?

As discussed in the concluding section of the previous chapter, one of the key challenges characterizing the current phase of the digital transformation lies in finding the right balance between the protection of personal data and the parallel need to develop advanced AI systems capable of solving complex real-world issues.

On the one hand, data protection regulations, such as the GDPR, impose stringent limitations on the ways personal data can be collected, processed and shared. On the other hand, the development of AI, especially those based on machine learning paradigms, are critically dependent on access to extensive, diverse, and high-quality datasets, without such data, the accuracy, reliability, and applicability of these models may be significantly compromised.

This intrinsic tension between the right to privacy and the imperative of technological progress necessitates the adoption of innovative methodologies that avoid framing the two objectives as mutually exclusive. Instead, it calls for approaches that enable a synergistic relationship between data protection and innovation.

In this regard, synthetic data has emerged as a particularly promising solution, offering the potential to reconcile these competing demands by providing a privacy-preserving alternative to real-world data.

Delving into the details, synthetic data refers to data that is generated artificially and is designed to replicate the statistical properties, structures, and relationships found in real datasets.[18] When generated properly, synthetic data offers a powerful solution for addressing limitations in traditional data availability. Through advanced generation techniques, this innovative kind of data can mirror the complexity and variability of real data, making it suitable for a wide range of applications, from training machine learning models to conducting simulations, testing software systems, or enabling secure data sharing between institutions.[19]

---

[18] Preeti Patel, "Synthetic Data," *Business Information Review* 41, no. 2 (2024): https://doi.org/10.1177/02663821241231101, pp. 48-52

[19] Centre for Information Policy Leadership (CIPL), *Privacy-Enhancing and Privacy-Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age*, December 2023, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-understanding-pets-and-ppts-dec2023.pdf, p. 41

In practical terms, synthetic data can support a wide variety of use cases, for instance, in machine learning, it can be used to augment training data when real examples are insufficient, allowing developers to work with values representative of the original data while mitigating privacy and legal concerns, this approach empowers organizations to build robust models without the risks associated with handling sensitive personal data. In software development, it can serve as a test environment that mimics production data without the associated privacy concerns. In data sharing between organizations, it enables collaboration without exposing confidential or regulated information.

Specifically, synthetic data enables the generation of large-scale, diverse datasets that can address gaps often present in real-world data collections. This includes rare events, edge cases, and underrepresented demographic or behavioural categories. As such, synthetic data plays a pivotal role in enhancing the representativeness and balance of training data, contributing to the reduction of algorithmic biases and increasing the fairness and robustness of AI systems.[20] In this way, synthetic data is not merely a workaround for regulatory challenges but a strategic asset in promoting more inclusive, ethical, and resilient technological development.

Even if this kind of data is occasionally identified with terms such as "fake" or "artificial", synthetic data should not be underestimated. Its true value resides not only in replicating real-world data on a one-to-one basis, but also in its capacity to support data-driven innovation in a privacy-conscious and resource-efficient manner.

As a matter of fact, the relevance of synthetic data lies in its potential to offer a privacy-preserving alternative to real data, since it does not correspond to any specific individual, it can be shared and used more freely, without exposing sensitive or personally identifiable information. Unlike traditional anonymization techniques, which often reduce data utility or fail to fully prevent re-identification, synthetic data can preserve the analytical value of the original dataset while significantly lowering the risk of privacy violations. This is particularly interesting in fields such as healthcare, finance, or social research, where the availability of rich datasets is crucial, but the ethical and legal risks of data misuse are high, and instead the use of synthetic datasets enables researchers and developers to simulate real-world scenarios without exposing individuals to the risks of identification or data misuse.[21]

The advantages highlighted, make synthetic data an increasingly strategic resource, not only for complying with data protection standards but also for driving ethical, inclusive, and effective

[20] Patel, "Synthetic Data"

[21] Jumai Adedoja Fabuyi, *"Leveraging Synthetic Data as a Tool to Combat Bias in Artificial Intelligence (AI) Model Training," Journal of Engineering Research and Reports* 26, no. 12 (2024): 24–46, https://doi.org/10.9734/jerr/2024/v26i121337.

technological development. Moreover, understanding what synthetic data is, how it works, and what its potential applications are is crucial for anyone navigating the intersection of data protection and innovation, since it leads to a more safe and beneficial way of sharing data.[22]

This research will also highlight, through a practical example, how synthetic data can become the core of a successful business model. In particular, it will examine the case of the Italian start-up Aindo, which assists its clients in the generation and use of synthetic data, thereby promoting the responsible adoption of artificial intelligence while enabling companies to fully harness the value of the data they possess.

2.1.1 Synthetic data as a Privacy Enhancing Technology

In light of the analysis conducted thus far, it is crucial to recognize that, within this context, synthetic data stands out as a privacy-enhancing technology, offering privacy-preserving alternatives to real world sensitive information. As such, it represents a set of technologies specifically designed to enhance the privacy and security of personal data.

The concept of PETs has been extensively examined in various reports published by the European Union Agency for Cybersecurity (ENISA), which defines them as "*a coherent system of ICT measures that protects privacy by eliminating or reducing personal data or by preventing unnecessary and/or undesired processing of personal data, all without losing the functionality of the information system*".[23][24] In this regard, these technologies can be seen as facilitating the collection, processing, analysis and sharing of personal data, without directly exposing them and safeguarding their confidentiality.

PETs constitute essential and highly effective tools for any entity seeking to preserve data privacy while still unlocking its full value. Organizations that decide to invest to develop a strategy that includes PETs, are able to demonstrate a tangible commitment to data protection by implementing a privacy by design approach. This result in an accountable use of data, ensuring that data protection becomes a core component of the technological development and the subsequent processing activities, rather than a threat to it. As such, PETs also serve as key business enablers, allowing companies and public sector organizations to access, share and use data that would otherwise be

[22] Patel, "Synthetic Data"
[23] European Union Agency for Cybersecurity (ENISA), *Data Protection Engineering: From Theory to Practice*, January 2022. Data Protection Engineering | ENISA
[24] European Union Agency for Network and Information Security (ENISA), *Readiness Analysis for the Adoption and Evolution of Privacy Enhancing Technologies*, version 1.0, December 2015, Readiness Analysis for the Adoption and Evolution of Privacy Enhancing Technologies

unavailable. In addition to safeguarding privacy, PETs also help protect confidential information, trade secrets, commercial interests, and ensure regulatory compliance.[25]

The importance of PETs, in particular of synthetic data, becomes crucial when it comes to the field of artificial intelligence, since they provide opportunities to protect privacy and cybersecurity in the development and deployment of AI.[26]

Among the different advantages of integrating PETs into AI systems, beyond protecting privacy or confidentiality, one significant benefit lies in the improvement of data quality, since PETs can help generate high-quality, diverse and representative datasets that can be used to train models. For instance, synthetic data can be used to address data scarcity, helping to reduce bias and leading to more fair and accurate outputs. Additional security measures provided by synthetic data enable more organizations to collaboratively develop AI models, share training data needed for development phase, or share outputs generated by AI models trained on synthetic datasets.[27]

Highlighting the relevance of this approach, the G7 Summit held in 2022 stated that: "*The use of PETs 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 we, as 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*."[28] This declaration emphasized how such technologies can constitute a prominent solution for reconciling innovation with data security, by adhering to the fundamental paradigms of protecting the confidentiality, integrity and availability of data.

For those reasons, PETs emerge not only as a technical safeguards, but also as a strategic measure of ethical innovation, allowing entities to pursue their objectives while maintaining high standards of privacy and trust.

[25] Centre for Information Policy Leadership (CIPL), *Privacy-Enhancing and Privacy-Preserving Technologies in AI: Enabling Data Use and Operationalizing Privacy by Design and Default*, March 2025. https://www.informationpolicycentre.com/uploads/5/
[26] *Ibid*.
[27] *Ivi*, p. 21.
[28] G7 Data Protection and Privacy Authorities, *Roundtable Communiqué*, 8 September 2022, 7, Microsoft Word - G7 DPA Roundtable Communiqué 8 sep 2022 final.docx

2.2 What there is in the background of synthetic data?

Understanding how synthetic data is generated is a fundamental step in contextualizing its legal and compliance-related implications. Before delving into the regulatory frameworks and assessing whether synthetic data can be truly considered anonymous, it is essential to examine the underlying processes through which such data is created. This analysis allows for a better understanding of the technical foundations that shape the privacy guarantees offered by synthetic data, as well as the limitations and risks that may arise.

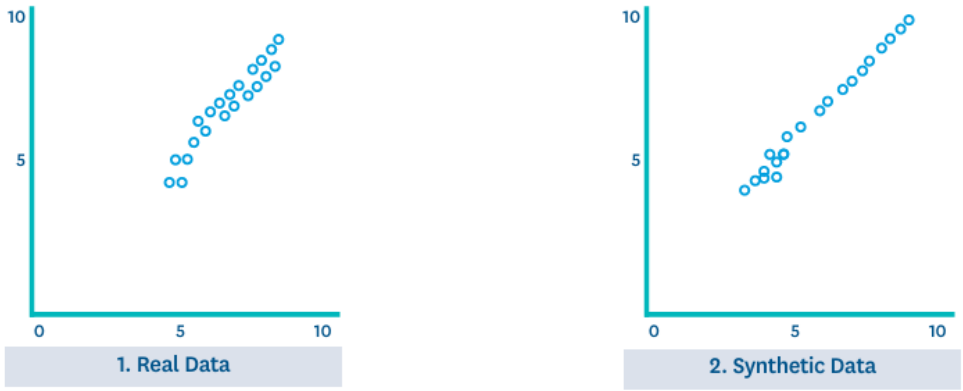

FIGURE 1: SYNTHETIC DATA

SOURCE: CENTRE FOR INFORMATION POLICY LEADERSHIP (CIPL), "PRIVACY-ENHANCING AND PRIVACY-PRESERVING TECHNOLOGIES: UNDERSTANDING THE ROLE OF PETS AND PPTS IN THE DIGITAL AGE", DECEMBER 2023.

In order to illustrate these key mechanisms that govern what is the background of synthetic data in a more tangible manner, it is possible to observe what is represented in the Figure 1 above. By looking to the comparison between a set of real data (graph 1) and its synthetic counterpart (graph 2), it is highlighted that the left side illustrates actual point drawn from an observed population, displaying a clear linear correlation between the two variables under consideration, while on the right side, the synthetic dataset replicates this statistical relationship without reproducing any of the original data points. Each observation in the synthetic dataset is newly generated, yet it maintains the same structural properties and correlation patterns evident in the real data. This demonstrates the capacity of synthetic data not merely to mimic the statistical distribution of the original dataset but also to preserve its analytical value, while ensuring that no individual from the original population is directly identifiable.

Therefore, in the context of this research, exploring which are the methods used to generate synthetic data is crucial to assess their practical benefits, particularly in terms of data protection, utility and regulatory compliance.

There are many ways to generate synthetic data, and without the aim to be exhaustive, this research will provide some methods.

One of them is through regression models. This method works by starting from a set of real data, called training set, which consists of a group of observations that collect multiple pieces of information together (for example age, gender, etc.), and are therefore multivariate data.

A regression function is applied with the aim of understanding how these pieces of information are connected to each other. This function can also be nonlinear, meaning it does not have to follow a straight line but can take on more complex curves to better fit the data. The goal is to find a function that accurately reconstructs the observed data, in other words, that interpolates it, meaning it connects the data in a realistic way and explains its behaviour.

It is important to understand that a regression works as an input-output model, where some variables are inputs, meaning the actual characteristics of the observed data, while the output variables are what one aims to predict. When performing a regression, a "family" of mathematical functions is selected, and these are functions that might fit the data well. Then, specific parameters are identified to make the function resemble the real data as closely as possible. The objective is to calculate, through the ordinary least squares approach, the parameters that produce the smallest possible overall error. To do this, the errors, meaning the differences between the real values and the values predicted by the function, are calculated, squared, and summed, and the function that results in the smallest total error is selected. Once the fitting is completed, the function can generate new, synthetic data, that follow the same structure as the original data.[29]

[29] D'Acquisto, *"Synthetic Data and Data Protection Laws",* p. 229.

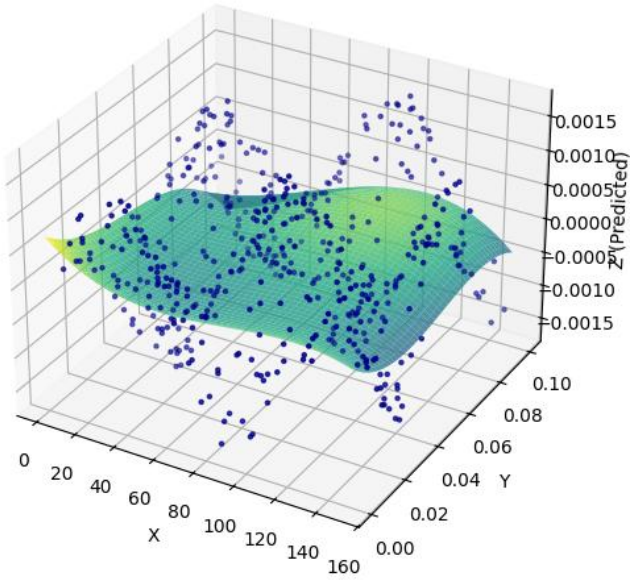

FIGURE 2: REGRESSION FUNCTION AS SYNTHESISER

SOURCE: AUTHOR ELABORATION, 2025

As a further solution for synthetic data generation, there is the maximum likelihood estimation (MLE) method. The purpose of this probabilistic approach is to determine the values of the parameters of a multivariate probability distribution. The application starts from the assumption that the real observed data come from a certain probability distribution, chosen from a known family of parametric distributions. The parameters of the distribution are selected in a way that they resemble the observed data as closely as possible. Following the similarity criterion, the objective is to identify those parameters that make it most likely that the observed real data were generated by that specific distribution.

In order to find the best parameters, from a technical point of view, a mathematical function called the likelihood function is applied. This function measures how likely it is that the observed data come from the assumed distribution, as a consequence, the greater the likelihood, the better the distribution fits the data. This function consists of the product of the probabilities calculated for each observed data point, and this product represents the value that express how "likely" the entire dataset is with respect to the given distribution. Once the distribution that best fits the data is found, it can be used to generate new synthetic data, which will follow the same proportions as the original data.[30]

[30] D'Acquisto, *"Synthetic Data and Data Protection Laws",* p.230.

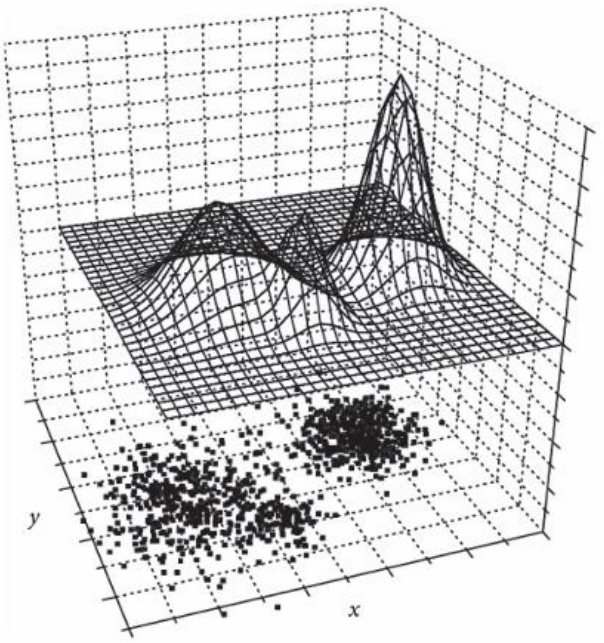

FIGURE 3: DATA SYNTHESIS AS A MAXIMUM LIKELIHOOD ESTIMATION PROBLEM
SOURCE: GIUSEPPE D'ACQUISTO, "SYNTHETIC DATA AND DATA PROTECTION LAWS," JOURNAL OF DATA PROTECTION & PRIVACY 6, NO. 3 (2024), P.230

These two methods described above fall within the category of traditional approaches to synthetic data generation; however, new methods based on artificial intelligence algorithms have recently emerged.

2.2.1 Generative Adversarial Networks

Synthetic data are widely employed to enable AI systems to have extensive access to data, without harming individuals and sensitive information, but at the same time, AI itself serves as a tool for the generation of synthetic data. As a matter of fact, one of the most prominent techniques for producing synthetic data is the use of generative adversarial networks (GANs). This approach has also been adopted by Aindo, a successful national startup recognized for its expertise in the field of synthetic data[31].

To better understand the dynamics of this algorithm, an insightful explanation was provided by Daniele Panfilo, CEO of Aindo, during the webinar "*Personal and non-personal data under the EU*

[31] While GANs constitute a fundamental paradigm for synthetic data generation, numerous alternative architectures exist, including Variational Autoencoders (VAE) (See Wikipedia contributors, "Variational Autoencoder," *Wikipedia*, last modified August 27, 2025, Variational autoencoder - Wikipedia), CART models (See Beata Nowok, Gillian M. Raab, and Chris Dibben, *synthpop: Bespoke Creation of Synthetic Data in R*, vignette del pacchetto synthpop, Comprehensive R Archive Network (CRAN), 2016, synthpop: Bespoke Creation of Synthetic Data in R). It is important to note that GANs present specific technical limitations, such as "model collapse", that will be mentioned in Chapter 3.

Regulation: exploring synthetic data" organized by Rödl & Partner. Panfilo illustrated the functioning of a GAN through an example involving two entities: a generator, that can be identified as a criminal, and a discriminator, that is a police officer.

Each entity is assigned specific tasks: the discriminator is responsible for understanding whether the money is real or counterfeit, while the generator is tasked with producing the fake money.

The two components, generator and discriminator, engage in a competitive training process, because as the criminal improves its ability to produce convincing counterfeit money, the police will be compelled to enhance its capacity to detect imitations. On the other hand, as the police becomes very good at distinguishing real from fake, the criminal will be forced to refine its counterfeiting techniques accordingly. Initially, the generator's output will not be of good quality, so for the discriminator will be very easy to identify real or fake money, however, in the long run, the generator will progressively improve its accuracy to the extent that the discriminator eventually becomes unable to distinguish between real and fake currency.[32]

These two entities in the example illustrate the functioning of a neural network, while the currency represents the data. At the beginning, the data generated are of a lower quality compared to the original ones, and the role of the neural network is to distinguish between authentic from synthetic observations. Over time these two entities are trained one against each other, in this process, the two neural networks engage in an adversarial interaction, each one striving to achieve its objective more effectively than the other. When the discriminator successfully identifies data as real or synthetic, the generator adjusts the parameters within its network to produce new data that is increasingly realistic, as a consequence, when the discriminator fails to classify correctly, it refines its capacity to distinguish synthetic data from real data in the subsequent attempts. After the training phase, the synthetic data generated can effectively serve as a proxy of the real data.[33]

[32] Daniele Panfilo, remarks during the webinar *"Personal and non-personal data under the EU Regulation: exploring synthetic data",* Rödl & Partner, April 16, 2025.
[33] D'Acquisto, *"Synthetic Data and Data Protection Laws",* p. 231.

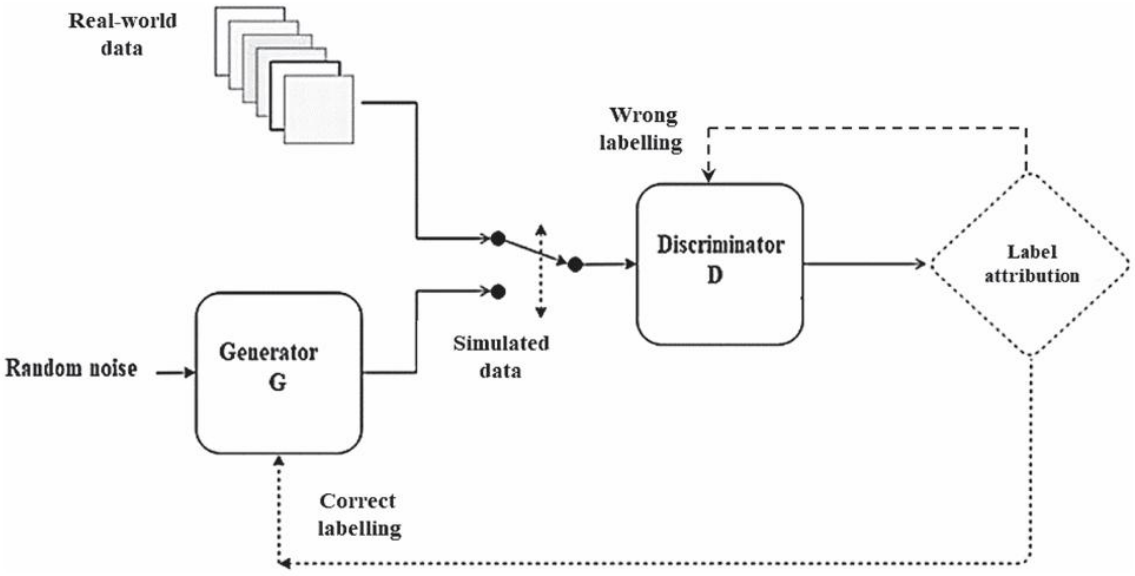

FIGURE 4: ARCHITECTURE OF GENERATIVE ADVERSARIAL NETWORKS

SOURCE: GIUSEPPE D'ACQUISTO, "SYNTHETIC DATA AND DATA PROTECTION LAWS," JOURNAL OF DATA PROTECTION & PRIVACY 6, NO. 3 (2024), P. 231

2.3 Exploring practical applications

As previously analysed, there are several reasons why organisations may consider adopting synthetic data. Frequently, entities face challenges related to data unavailability and furthermore, in today's highly regulatory landscape, organisations require to be legally compliant when it comes to data protection, and this may limit the use of original data. Another obstacle is that data flows, data sharing and data access within an organisation may be constrained by security concerns, for instance, certain information may be too sensitive to be migrated to a cloud infrastructure. In addition, for many organisations the financial costs associated with acquiring data from third parties or preparing and accessing it in-house may be prohibitive.

The synthesis process offers a way to overcome these challenges, enabling the generation of the necessary data, tailored to the specific needs of each sector, while also replicating types of information that are difficult to obtain from real-world dataset. As a matter of fact, the solution that this research aims to propose is to boost the adoption of synthetic data, given their applicability across multiple sectors and their capacity to deliver tangible benefits.

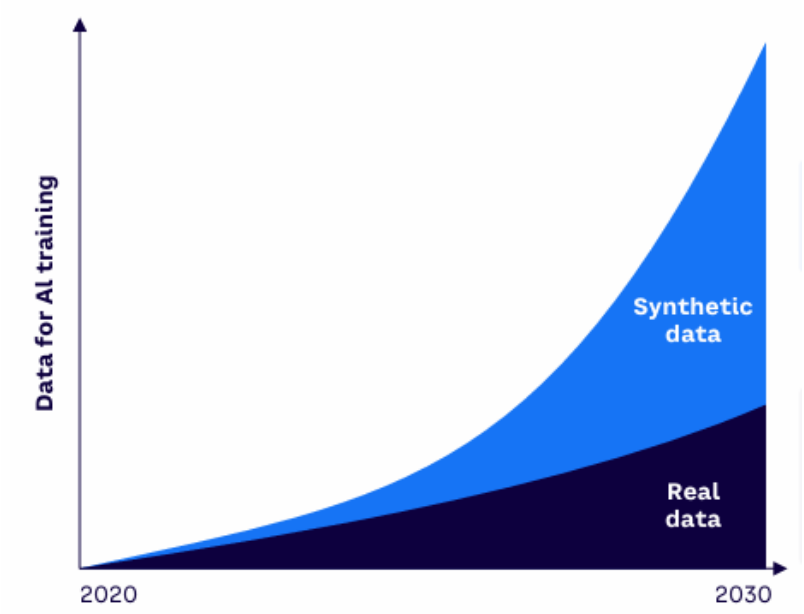

FIGURE 5: SYNTHETIC DATA EXPECTED TO OVERSHADOW REAL DATA IN AI MODELS BY 2030
SOURCE: SRI RAJAGOPAL, CORRINE BAI, AND MARK ROWLAND, *SYNTHETIC DATA: FACILITATING INNOVATIVE SOLUTIONS* (ARTHUR D. LITTLE, 2024), P. 4

Underscoring the rapid growth of this sector and encouraging organizations to reflect on the potential of this technology, the research conducted by Arthur D. Little and Gartner demonstrates that synthetic

data are not a temporary trend, on the contrary, it is projected that by 2030, synthetic data will completely overshadow real data in AI models.[34]

This statistic is not merely indicative of a technological shift, but rather a signal of a profound structural transformation in the way innovation will be conceived and operationalized. Indeed, from a strategic perspective, the integration of synthetic data implies that organizations capable of early adoption will not only migrate regulatory and ethical risks but also position themselves at the forefront of innovation, and instead organizations that fails to recognize this opportunity risk being locked out of future competitive advantages.

This new paradigm in data application is not merely a technical innovation but it represents a profound shift in how knowledge, risk and opportunity can be managed. Synthetic data can open the door to applications that were previously unthinkable due to ethical, legal or practical constraints, effectively transforming scarcity of accessible data into abundance. What makes this approach so powerful is not only its versatility across sectors, but also its ability to redefine the very boundaries of what is possible in industries where data sensitivity has historically limited progress.

Through my reflections on potential applications, enriched by the study of possible scenarios and real-world success cases, the following sections will illustrate how synthetic data can revolutionize contexts such as finance and healthcare, which are domains where the use of personal information is both critical and unavoidable. In these cases, data synthesis does more than reduce risks, since it enables organizations to compete more strategically, to unlock new levels of research and performance, and allows to find the right balance between data protection and innovation, instead of forcing to choose between compliance and progress, synthetic data suggests a future in which the two can evolve in harmony.

2.3.1 Financial sector

In highly regulated sectors such as banking, finance, and insurance, data represents both a strategic asset and a significant liability.

Financial institutions are increasingly digital and through customers interactions and engagement with products and services are generated and collected huge amounts of data, and therefore those entities are legally required to retain vast amounts of sensitive personal and transactional data, which can reveal much about the identity and behaviours of the data subjects, and they have to be kept over

[34] Sri Rajagopal, Corrine Bai, and Mark Rowland, *Synthetic Data: Facilitating Innovative Solutions* (Arthur D. Little, 2024), SYNTHETIC DATA: FACILITATING INNOVATIVE SOLUTIONS, p. 3

extended periods to comply with national and supernational regulations, such as anti-money laundering (AML) laws, tax compliance directives, and the GDPR, since it is vital that this data is protected and that consumers' right to privacy is safeguarded. Those legal obligations entail a substantial burden in terms of storage infrastructure, cybersecurity, and internal data governance, since any data breach or misuse could result in reputational damage, legal penalties, and loss of customer trust. As a result, the management and protection of data are not merely operational concerns, but strategic imperatives.

Simultaneously, this data can also drive valuable innovation, since today, an efficient usage of data has the potential to deliver societal benefits such as greater market efficiency and integrity, financial inclusion, and the prevention of financial crime. In this context, it is widely recognised that artificial intelligence holds significant potential in the financial services industry. However, the extent of its potential will depend on the wide availability and accessibility of data to innovators who will build the next generation of products and services. Whilst data privacy laws are critical to the protection of consumers' privacy rights, the challenges associated with access to financial data, especially for new market entrants, can inhibit the development of new products and services in the market. However, financial data, such as consumer transaction records, account payments, or trading data, is sensitive personal data subject to data protection obligations, as well as often being commercially sensitive. In instances where data is shared between or even within an organisation, it is vital that the appropriate protections and safeguards are in place to protect individuals' privacy. Consequently, the operation of data sharing can require several months of complex due diligence and onboarding processes.[35]

Within this framework, synthetic data offers a powerful solution to reduce these risks while still enabling institutions to innovate and derive value from data. By generating artificial data that preserves the statistical properties and structure of real datasets, but without being traceable to any actual individual, banks and insurers can develop and test AI-driven solutions such as fraud detection algorithms, credit scoring models, or customer profiling systems, without exposing sensitive information.

In order to better understand practical application of synthetic data in the financial sector, this paragraph has the aim to explore possible insightful solutions to real problems that could emerge in this filed when it comes to manage personal data.

[35] Financial Conduct Authority, *Synthetic Data to Support Financial Services Innovation,* March 2022. Available on: Call for input: Synthetic data to support financial services innovation

For example, one of the principal use cases regards the detection of suspicious activities within a bank. In practical terms, a bank might use synthetic data to train a machine learning model that detects unusual account activity patterns indicative of fraud, replicating realistic statistical patterns, like repeated small-value transfers or logins from unusual devices, without relying on actual customer data thereby avoiding the risks associated with processing real client transactions in the early development stages. A concrete example is offered by J.P. Morgan, which has publicly acknowledged its use of statistically accurate synthetic financial transaction data to enhance fraud detection and anti–money laundering systems.[36]

The same reasoning applies to the enhancement of credit scoring models, which could be widely employed by banks and financial institutions to assess a client's creditworthiness. Traditional models are often constrained by inherent biases or imbalances in the datasets on which they are trained, leading to frequent disadvantages for certain groups of individuals. In this context, synthetic data can be leveraged to mitigate these limitations through the generation of additional cases of underrepresented profiles, such as young individuals with limited credit histories or self-employed individuals with atypical cash flows. By enriching training datasets with realistic yet non-identifiable synthetic records, financial institutions can improve both the accuracy and the fairness of their credit scoring systems, thereby fostering greater inclusivity while maintaining compliance with the applicable regulatory frameworks, such as the principle of fairness established by the GDPR, which requires that the processing of personal data be conducted in a just and equitable manner.[37]

Similarly, insurance companies can use synthetic data to simulate customer profiles and claim scenarios for activities like actuarial analysis, underwriting processes, and policy pricing strategies that requires to analyse a huge amount of sensitive information. By using synthetic data, it is possible to reduce the exposure to real data thereby, diminishing the risks of violating the GDPR's data minimisation principle, which mandates that only personal data that is "*adequate, relevant and limited to what is necessary*" for the intended processing should be held.[38] This allows for improved accuracy and personalization of services, while remaining compliant with data protection regulations.

Cybersecurity constitutes another interesting and useful domain in which synthetic data can assume a pivotal role in safeguarding information assets. In data-intensive sectors like the financial one, where compliance and security are paramount, synthetic data not only serves as a tool for technical

[36] J.P. Morgan, "Synthetic Data for Real Insights," *J.P. Morgan Technology Blog*, accessed August 10, 2025. Available on: Synthetic Data for Real Insights
[37] GDPR, Regulation (EU) 2016/679, art. 5 (1)(a)
[38] GDPR, Regulation (EU) 2016/679, art. 5 (1)(c)

experimentation but also as a strategic enabler, transforming what is traditionally seen as a cost and a risk into a resource that supports responsible innovation. This large amount of retained personal data, in addition to representing a significant cost for IT infrastructure, exposes banks to considerable risk in the event of data breaches, unauthorized access, or cyber incidents. By converting part of this data into synthetic data to be used for testing, analysis, and simulations, banks can avoid the constant manipulation of real data, thereby reducing the overall risk surface.

Ultimately, another critical application, as has already been remarked, lies in complex challenge of sharing personal data across multiple actors. Although the Data Governance Act, as outlined in the first chapter, explicitly seeks to encourage and facilitate data sharing across different actors in the European Union, its practical implementation collides with significant complexities, most notably when personal data is involved in the processing activities. In the insurance and banking sectors, for instance, the exchange of information between different entities, such as partner companies, internal departments, or research institutes, remains indispensable for innovation and oversight, yet is simultaneously constrained by strict legal requirements and significant risks of data exposure. In this respect, synthetic data offers a valuable alternative, an insurance company may, for example, share synthetic datasets with an Insurtech startup to co-develop a new insurance product, thereby avoiding the lengthy negotiation of data processing agreements and mitigating the risk of GDPR violations.

2.3.2 Healthcare sector

Data is the lifeblood of modern healthcare bearing the potential to improve patient care by powering clinical research, and advancing public health initiatives. However, the promise of real-world data to provide personalized medical care, to guide policymaking, and to respond to rapidly changing conditions is tempered by the significant challenges inherent in accessing high-quality datasets.[39] Synthetic data generation is an active area of research, driven primarily by the need to share sensitive data for research and development. This is particularly relevant in the healthcare sector, where the application of synthetic data can bring revolutionary benefits, especially when it comes to do research. The main problem in this domain, when dealing with personal data, lies in the fact that health data are collected on the basis of consent, but their reuse, that in this case would be a secondary use, is prohibited. For instance, if a company collects data with consent to conduct a study on a rare disease, it cannot later reuse the same data for another type of investigation unless explicit and informed consent for that additional purpose was obtained at the time of collection. Therefore, unlocking the

[39] Mauro Giuffrè and Dennis L. Shung, "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy," *NPJ Digital Medicine* 6, no. 1 (2023): 186, https://doi.org/10.1038/s41746-023-00927-3, p. 1

potential of data in healthcare through the creation of synthetic datasets also means enabling their safe reuse, thereby significantly enhancing their potential to generate tangible benefits for research in this sector.

One of the most common applications of synthetic data in the medical field is described by Aindo, which actively collaborates with hospital institutions, particularly those that frequently face the challenge of being unable to provide clear data whenever pharmaceutical companies conduct Real World Evidence (RWE) studies, which are based on the analysis of data generated by patients during routine clinical practice. However, with Aindo's technology, hospitals are able to generate synthetic data directly from clinical records. In practice, the startup installs their technology within the hospital, enabling the institution to synthetize the data and subsequently provide it to the pharmaceutical company. Thanks to this method, synthetic datasets can facilitate cross-institutional collaboration between hospitals, research centres, and pharmaceutical companies, allowing them to share insights and conduct joint studies without the legal and ethical constraints normally associated with personal health data, in this way, synthetic data acts as a bridge between innovation and compliance, since this process not only safeguards patients privacy, but also optimizes the bureaucratic procedures that typically hinder the adoption of such research processes.[40]

Furthermore, synthetic data can enhance the development and validation of AI-driven diagnostic tools by providing large, balanced datasets that include rare conditions or underrepresented demographics.[41] Enhancing pharmaceutical and genetic research can help meet the growing demand for data, accelerating drug discovery, uncovering promising compounds, and predicting their efficacy, while reducing the risks linked to handling real patients information.

From a broader perspective, the healthcare sector offers multiple promising scenarios for the application of synthetic data, where the preservation of patient privacy is paramount. Beyond supporting research by enabling the secondary use of clinical data without infringing consent restrictions, synthetic data can be harnessed in several innovative ways.

For instance, it can be employed to build realistic simulation models that reproduce complex dynamics within healthcare systems, fluctuations in patient volumes, the availability of medical equipment or variations in staff training levels. These models allows hospitals and policymakers to

[40] Tiziana Tripepi, "Dati e intelligenza artificiale, Aindo è la startup del mese," *InnLifes*, April 29, 2025. Accessed August 18, 2025 https://www.innlifes.com/startup/aindo-dati-startup

[41] Barbara Draghi, Zhenchen Wang, Puja Myles, and Allan Tucker, "Identifying and Handling Data Bias within Primary Healthcare Data Using Synthetic Data Generators," Heliyon 10, e24164 (2024), https://doi.org/10.1016/j.heliyon.2024.e24164

stress-test healthcare infrastructures under different scenarios, such as pandemics or sudden increases in emergency admissions, without exposing sensitive personal information.

Another promising use case lies in the creation of digital twins, specifically of patients, where synthetic data enables the replication of individual clinical profiles in a way that safeguards identity. Such digital twins can then be used to simulate treatment responses, optimize therapeutic plans, and ultimately improve patient outcomes, particularly in precision medicine.[42] This approach is particularly relevant in light of the article 9 of the GDPR, which places strict limitations on the processing of health-related data, making synthetic data an attractive privacy by design and by default solution for advancing medical innovation while maintaining compliance.

To provide greater solidity to the analysis, several real-world success cases have been examined in which synthetic data has been applied effectively within different contexts of the healthcare sector, ranging from policy evaluation to diagnostics and emergency response.

One relevant adoption of synthetic data in the healthcare sector is its application in analysing the implications of demographic aging on the healthcare systems. By combining multiple national health surveys and generating enriched synthetic samples through microsimulation techniques, the researchers were able to construct scenarios assessing the future demand for healthcare services under different assumptions. This allowed them to evaluate, for instance, variations in the average number of medical visits, variations in morbidity and disability, community support, and doctor behaviour. The study demonstrates how synthetic data can serve as a valuable tool for policymakers, enabling the anticipation of resource allocation needs and the design of more resilient health policies without exposing sensitive patient information.[43]

The utility of synthetic data also can bring benefits in the field of mental health research, where through the use of synthetically generated clinical discharge reports is possible to train Natural Language Processing (NLP) models aimed at identifying diagnoses and phenotype of complex psychiatric conditions. Since mental health records are highly sensitive and often stored in unstructured textual formats, the generation of synthetic text provides a reliable means of protecting patient confidentiality while still allowing researchers to develop effective classification systems.[44]

---

[42] Patel, "Synthetic Data", p. 51

[43] Peter Davis, Roy Lay-Yee, and Janet Pearson, "Using Micro-Simulation to Create a Synthesised Data Set and Test Policy Options: The Case of Health Service Effects under Demographic Ageing," *Health Policy* 97, no. 2–3 (2010): 267–274, https://doi.org/10.1016/j.healthpol.2010.05.014

[44] Julia Ive et al., "Generation and Evaluation of Artificial Mental Health Records for Natural Language Processing," *NPJ Digital Medicine* 3, no. 1 (2020): https://doi.org/10.1038/s41746-020-0267-x. p. 69.

A further area of application emerged during the COVID-19 pandemic, where the scarcity of clinical images represented a significant barrier to the development of accurate diagnostic tools. Synthetic chest Computed Tomography (CT) scans were created to augment the limited real-world datasets available. These artificial images were used to train machine learning models, with the goal of improving their accuracy in distinguishing COVID-19 patients from individuals with pneumonia or normal patients. This case highlights how synthetic data can play a decisive role in emergency contexts, where the rapid expansion of available datasets is crucial to enhance disease detection and support timely clinical responses[45][46].

[45] Yifan Jiang, Han Chen, Murray Loew, and Hanseok Ko, "COVID-19 CT Image Synthesis with a Conditional Generative Adversarial Network," *arXiv* (preprint), July 29, 2020, https://doi.org/10.48550/arXiv.2007.14638.
[46] Hari Prasanna Das, Ryan Tran, Japjot Singh, Xiangyu Yue, Geoffrey Tison, Alberto Sangiovanni-Vincentelli, and Costas J. Spanos, "Conditional Synthetic Data Generation for Robust Machine Learning Applications with Limited Pandemic Data," *Proceedings of the AAAI Conference on Artificial Intelligence* 36, no. 11 (2022): 11792-11800, https://doi.org/10.1609/aaai.v36i11.21435.

Chapter 3 – The Overparameterization Problem

3.1 About the concept of anonymization

After having examined, from a technical standpoint, the mechanisms underlying the generation of synthetic data and their potential practical applications, it is now necessary, for the purposes of this research, to address a fundamental legal question, whether synthetic data may legitimately be regarded as anonymous.

This chapter will therefore analyse the concept of anonymization as developed under European data protection laws, with particular reference to the GDPR and the interpretative guidance provided by the European Data Protection Board (EDPB). The inquiry will further examine the criteria that must be fulfilled for data to be considered truly anonymous and thus fall outside the scope of data protection law. At the same time, particular attention will be devoted to the specific challenges posed by synthetic data, including the risk of re-identification, the potential presence of hidden personal information within training datasets, and the ambiguities surrounding their qualification under the current legal framework.

Above all, it is essential to begin with the concept of anonymization, bearing in mind that non-personal or anonymous data are not subject to the principles and safeguards established under the GDPR.

As provided by Article 4(1) GDPR, personal data is defined as "*any information relating to an identified or identifiable natural person*". From this definition, it can be deduced that, on the contrary, data qualify as "non-personal", and thus as anonymous and for this reason excluded from the scope of the Regulation, when they cannot be linked, even indirectly, to an identified or identifiable natural person.

In contrast to pseudonymization, data anonymization is aimed at ensuring that the re-identification of individuals within the dataset is no longer possible, instead the definition of pseudonymization, which is provided in the Article 4(5) of the GDPR states that "*the 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 information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person*". In light of what has emerged, the distinction between anonymization and pseudonymization is therefore both clear and fundamental, while pseudonymization constitutes a processing activity that alters the representation of data, it does not

remove their nature as personal data, rather, it is offered as a security measure designed to minimize risks arising from data processing, so even if the actual identities are kept separately from the pseudonymized dataset, the original identity can still be re-traced by using the appropriate tools. Anonymization, on the other hand, means that those data cannot identify any more under any circumstances a natural person, because once anonymized the data cease to be considered personal data, and thus fall outside the material scope of the GDPR.

To enrich this debate, the recent decision of the Court of Justice of the European Union (CJEU) dated 4 September 2025 place a fundamental role[47]. The decision was about the judgement T-557/20 (EDPS v SRB)[48], commonly referred to as the "Deloitte judgment". The Court clarified that pseudonymised data does not automatically remain personal data in the hands of the recipient if two conditions are met:

- the recipient lacks access to the means reasonably likely to re-identify the data subject, taking into account existing technologies and safeguards; and
- robust technical and organisational measures effectively eliminate any real possibility of re-identification.

In such circumstances, pseudonymised data may be treated as anonymous by the recipient, while still remaining personal data for the original controller, who retains the key or means to re-identify, thus preserving their obligations under the GDPR.

This judgment is, in essence, is very helpful for the strengthening of the innovation ecosystem. The development of AI models that are both effective and unbiased necessitates access to large and diverse datasets. By adopting a risk-based approach, under which the recipient's inability to re-identify individuals is fact-specific and demonstrable, the CJEU has provided a legally robust pathway to share pseudonymised data for research and development, while maintaining the highest privacy standards through technical and organisational safeguards. Such a balanced framework is fundamental to preserving European competitiveness in the digital era.

[47] Court of Justice of the European Union, *Judgment of the Court (First Chamber) of 4 September 2025, EDPS v SRB, Case C-413/23 P*, ECLI:EU:C:2025:645, EDPS v SRB (Notion de données à caractère personnel)
[48] General Court of the European Union, *Judgment of the General Court (Eighth Chamber, Extended Composition) of 26 April 2023, SRB v EDPS, Case T-557/20*, ECLI:EU:T:2023:219, SRB v EDPS

3.1.1 Insights from the Recital 26 of the GDPR

This reflection is further reinforced by Recital 26 of the GDPR, which states that "*Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person*."

At the same time, Recital 26 endorses a substantive approach, requiring consideration of the actual circumstances in which data are processed, and in particular whether there is a realistic possibility of identifying the data subject, such assessment must take into account factors as the costs and the amount of time required, the technologies available at the relevant point in time, and the broader context of the processing activity.

Moreover, Recital 26 remains of the most significant interpretative references concerning anonymization, clarifying that "*The principles of data protection should therefore not apply to 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. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes*."

Although these statements appear categorical, several caveats must be highlighted. First of all, Recitals, unlike Articles, do not have binding legal force, but serve as interpretative tools. Secondly, through its substantive approach, Recital 26 emphasizes that "*account should be taken of all the means reasonably likely to be used*" to identify a natural person, so this formulation introduces a degree of ambiguity, particularly regarding the scope of the concepts of "reasonability" and "probability", as well as the permissible margin of residual risk of a potential re-identification.[49]

3.1.2 Working Party 29 Opinion 05/2014 on anonymisation techniques

Further contributions to the debate on anonymization can be found in other EU writings, most notably the Opinion 05/2025 of 10 April 2014 by the Article 29 Working Party, known as WP216.

In this document, anonymization is defined as "*a technique applied to personal data in order to achieve irreversible de identification*". The explicit reference to "irreversibility"[50] seems to exclude the admissibility of any residual risk of re-identification, this entails the need for a carefully designed

[49] Carmine Andrea Trovato and Chiara Rauccio, "L'anonimizzazione è morta? Un'analisi dei dati sintetici come proposta per superare la dicotomia "dato personale-non personale"." *Ciberspazio e Diritto* 23, no. 71 (2-2022): p. 240

[50] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Techniques*, WP216 (Brussels: European Commission, April 10, 2014), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, p. 7

process aimed at eliminating any possibility that data, even indirectly, could be traced back to the original individual.

A critical conceptual understanding that is described by the WP216 is that even when a data controller processes personal data by removing or masking direct identifiers and subsequently shares a portion of this dataset, the resulting dataset retains its classification as personal data if the original (identifiable) event-level data are not deleted. This determination hinges on the principle that as long as the data controller, or indeed any other party, maintains access to the original raw data, the potential for re-identification persists. Consequently, a dataset can only be unequivocally deemed anonymous if the data controller proceeds to aggregate the information to a level where individual events become non-identifiable, thereby irreversibly preventing identification.[51]

As indicated by the Opinion, there are three key factors that may lead to the identification of an individual and can also be useful to assess the degree of robustness of certain anonymization practices and techniques, which are[52]:

- *Singling out:* which corresponds to the possibility to isolate some or all records which identify an individual in the dataset;
- *Linkability:* which is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases). If anybody can establish (e.g. by means of correlation analysis) that two records are assigned to a same group of individuals but cannot single out individuals in this group, the technique provides resistance against "singling out" but not against linkability;
- *Inference:* which is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

The document further emphasizes the importance of adopting a dynamic and continuously updated approach to data security, in response to technological advancements and emerging attack methods, since safeguarding data privacy thus requires an ongoing commitment to revisiting and adjusting protective strategies. Effective anonymization techniques are therefore designed to mitigate the risks associated with these conditions of identifiability, relying on a risk-based assessment that takes into account a variety of factors, including the state of the art of available technologies. Given the rapid pace of technological evolution, such an assessment must be conducted periodically to determine

[51] *Ivi*, p. 9
[52] *Ivi*, p. 11

whether the measures adopted at an initial stage remain valid over time, or whether new or different safeguards are required in order to preserve the anonymity of the data.

The application of the anonymization techniques provided by the WP216 must therefore be context-specific, carefully considering the nature of the data, the potential threats to privacy, and the existing technological landscape.

Taking into account what has been analysed until this point, data synthesization as an advanced methodology for the generation of synthetic data offers new insights into the debate on anonymization. Synthetic data produced through advanced mechanisms, such as those examined in Chapter II of this thesis, are capable of reducing the risk of re-identification to a sufficiently remote level. In line with the reasoning set out by the WP216, synthetic data generated through such advanced methodologies could, under certain conditions, qualify as anonymous data.

3.1.3 Legal framework of synthetic data

As illustrated by Carmine Andrea Trovato and Chiara Rauccio, experts in the field of data protection, from a regulatory perspective it is important to underline that anonymization constitutes a further processing of personal data, since it changes the very nature of such data by rendering them anonymous, in line with the principles relating to processing of personal data laid down in Articles 5 and 6(4) of the GDPR. As an additional processing of personal data, anonymization must therefore satisfy the requirement of compatibility with regard to the legal basis or the circumstances of the further processing, as also recalled in Recital 50 GDPR, moreover, the data subject must be informed about the anonymization process pursuant to Articles 13 and 14 of the GDPR. If synthetic data, from a regulatory point of view, are regarded as a technique of anonymization, these conditions would equally apply to the procedures for their generation, which would also have to be considered a further processing activity. Indeed, if synthesis is employed with the purpose of achieving irreversible de-identification, it is essential that the original personal data used for this process have been collected and processed in compliance with the applicable legal framework. Subsequently, the generation of synthetic data qualifies as an anonymization process and, therefore, as a "further processing," which must either pass the compatibility test under Article 6(4) and Recital 50 of the GDPR, or rely on a specific and adequate lawful basis.[53]

Furthermore, synthetic data are increasingly gaining prominence as a regulatory tool, as they are explicitly mentioned in recent European regulations, such as the Data Governance Act. As outlined

[53] Trovato and Rauccio, "L'anonimizzazione è morta? Un'analisi dei dati sintetici come proposta per superare la dicotomia "dato personale-non personale"." pp. 248-249

in the first chapter of this thesis, the DGA aims to create a European digital space in which data, including non-personal data, may be used independently of their physical location of storage within the Union, thereby enabling sharing among public and private entities and enhancing the availability of information, especially in strategic sectors. Among the techniques considered suitable by the DGA to safeguard personal data and reduce the risk of re-identification of data subjects, the use of synthetic data is expressly referred to in Recital 7. This explicit recognition highlights the European legislator's openness to synthetic data as a valid privacy-preserving methodology, whose application is placed on the same level as anonymization or differential privacy. In other words, data synthesis can be used, within the areas identified by the DGA, as an effective anonymization technique for accessing, sharing and reusing, data without disclosing personal information. For this reason, data synthesis is presented both as a tool that complies with data protection requirements and as a driver of technological innovation.[54]

The theme of synthetic data is also present in the Artificial Intelligence Act (AI Act)[55], which pursues the primary objective of fostering the development and adoption, by both public and private actors, of safe and trustworthy artificial intelligence systems subject to strict rules under a risk-based approach. In particular, Articles 10 "Data and data governance" and 59 "Further processing of personal data for developing certain AI systems in the public interest in the AI regulatory sandbox" refer to synthetic and anonymous data in terms of equivalence. This emerges, for instance, from Article 59, which regulates the conditions under which, within the so-called "AI regulatory sandboxes," previously collected personal data may be reused exclusively for the development, training, and testing of AI systems. A literal interpretation of paragraph 1(b) suggests that the use of synthetic data, similarly to anonymous or non-personal data, should be considered the prevailing and preferable option.

Moreover, Article 10(5), authorizes providers of high-risk AI systems, only to the extent strictly necessary for detecting and correcting negative biases or distortions, to exceptionally process personal data belonging to special categories. However, such processing must be accompanied by appropriate safeguards for the rights and freedoms of individuals, including technical limitations on reuse and the

[54] Giulia Finocchiaro, Antonio Landi, Gianluca Polifrone, Davide Ruffo, and Francesco Torlontano, *The Regulatory Future of Synthetic Data: Data Synthesis as a Resource for Scientific Research, Innovation, and Public Policy in the European Legal Landscape* (Rome: Istituto Italiano per la Privacy e la Valorizzazione dei Dati and Data Intermediaries Alliance, July 15, 2024), Il_futuro_regolatorio_dei_dati_sintetici__1721197747 (1).pdf, p. 38

[55] European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 of 13 June 2024 Establishing Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, OJ L 1689, July 12, 2024, Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (regolamento sull'intelligenza artificiale)Testo rilevante ai fini del SEE.

adoption of state-of-the-art security and privacy measures. The provision also sets out strict cumulative conditions for the processing of special categories of personal data for the purposes indicated. The first condition establishes that processing under Article 10 is permissible only if the intended objective, namely, the identification and correction of biases, cannot be effectively achieved through the use of synthetic or anonymized data. Conversely, if providers are able to detect and mitigate biases by relying on synthetic or anonymized data, they are obliged to do so. The second condition concerns the application of technical restrictions on the reuse of special categories of personal data, along with advanced security and privacy safeguards, including pseudonymization.[56]

In this context, the European legislator appears to identify the use of synthetic data, together with anonymized data, as the preferred method for detecting and addressing biases in the datasets used for the development of high-risk AI systems. Finally, the explicit inclusion of synthetic data within the normative framework of the AI Act constitutes an acknowledgment of their legal significance and a confirmation of the growing awareness of the potential of this technology.

[56] Finocchiaro et al., *The Regulatory Future of Synthetic Data: Data Synthesis as a Resource for Scientific Research, Innovation, and Public Policy in the European Legal Landscape*, 2024, pp. 43-46.

3.2 The EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models

The European Data Protection Board (EDPB) Opinion 28/2024 of 17 December 2024[57] addresses specific aspects of data protection concerning the processing of personal data protection in the context of AI models. The Irish Supervisory Authority The Irish Supervisory Authority requested the EDPB to issue an opinion on questions of general application pursuant to Article 64(2) GDPR, as these technologies raise important concerns about data protection. In particular, the request focused on the processing of personal data during the development and deployment phases of AI models. In summary, the key issues raised by the Irish SA were:

- When and how an AI model can be considered "anonymous";
- How controllers can demonstrate the appropriateness of legitimate interest as a legal basis in the development and deployment phases;
- Which are the consequences of unlawful processing of personal data during the development phase of an AI model on the lawfulness of subsequent processing or operation of that model.

The following analysis aims to identify potential areas in which synthetic data can complement the reflections developed by the EDPB, offering possible solutions to enable the safe dissemination of AI models.

3.2.1 On the circumstances under which AI models could be considered anonymous and the related demonstration

Regarding the circumstances under which AI models may be considered anonymous and how such anonymity can be demonstrated, the EDPB emphasizes that the concept of "personal data" as indicated in the Article 4 (1) of the GDPR, is broad, covering any information relating to an identified or identifiable natural person. Data protection principles, as previously mentioned, do not apply to anonymous data, i.e. information which does not relate to an identified or identifiable person, taking into account "all the means reasonably likely to be used" by the controller or by others.

For the assessment of anonymity, the EDPB considers that AI models trained with personal data cannot in all cases be regarded as anonymous. The determination of anonymity must be assessed on

[57] European Data Protection Board, *Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models*, adopted December 17, 2024, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models | European Data Protection Board

a case-by-case basis by the competent Supervisory Authorities. In particular, for an AI model to be considered anonymous, the likelihood of direct, including probabilistic, extraction of personal data about individuals whose data was used to develop the model, as well as the likelihood of retrieving such data through queries, intentionally or unintentionally, must be insignificant. This evaluation must take into account all the means reasonably likely to be used by the controller or by others.

A key issue arises even when an AI model is not deliberately designed to produce personal information from training data, as personal data may nonetheless remain "absorbed" into its parameters and be extractable or obtainable, directly or indirectly, through reasonably available means. This is why Supervisory Authorities must scrutinize the documentation provided by the controller to demonstrate anonymity. Such assessment should consider that, with reasonable means: (i) training data-related personal data cannot be extracted from the model, and (ii) any output generated by querying the model does not relate to data subjects whose personal data was used for training. The EDPB refers also to the criteria mentioned in the WP29 Opinion 05/2014, which states that if it is not possible to single out, link, or infer information from a supposedly anonymous dataset, then such data may be considered anonymous.

Moreover, Supervisory Authorities must further take into account: (i) characteristics of the training data, the AI model, and the training procedure; (ii) the context in which the model is released/processed; (iii) additional information available to a person; (iv) costs and time required to obtain additional information; and (v) available technology and future developments.

Another risk that must be assessed is the possibility of identification by the controller or by different categories of "other persons," including unintended third parties who gain access to the AI model, in this context, the levels of testing and resilience against attacks may vary significantly depending on whether the model is public or internal.

Among the various factors available to assess the residual probability of identification, it is possible look at the AI model's design, in particular: the selection of data sources, the preparation and minimization of data, the methodological choices including privacy-preserving techniques and output controls.

Another fundamental aspect is model testing and resistance to attacks, analysing the breadth, frequency, quantity, and quality of tests conducted, especially against known and state-of-the-art attacks.

Additionally, it is crucial that controllers properly document their processing operations, including information relating to DPIAs, or the reasons for their absence, advice or feedback from the DPO,

technical and organizational measures adopted during the model's design to reduce the probability of identification, including threat models and risk assessments, measures applied throughout the model's lifecycle to contribute to or verify the absence of personal data, documentation demonstrating theoretical resistance to re-identification techniques, and controls limiting the success of attacks (e.g. training data/parameter ratio, metrics on re-identification probability, testing reports).

At this point, the role of synthetic data becomes particularly relevant since, if properly generated, synthetic data aims to mimic the statistical properties of real datasets without containing information that can be traced back to identifiable individuals. In this sense, the EDPB's strict interpretation of anonymity highlights both the potential and the limitations of synthetic data, while synthetic datasets are often presented as an "anonymized" solution, they may still carry re-identification risks if the generation process is inaccurate or if traces of the original personal data are retained in the model parameters. Therefore, according to what has emerged through the Opinion 28/2024, synthetic data should not automatically be equated with anonymity; rather, its use must be rigorously tested against the same criteria outlined by the EDPB.

3.2.2 On the appropriateness of legitimate interest as a legal basis for processing of personal data in the context of the development and deployment of AI Models

In its analysis of the legitimate interest as a legal basis under Article 6(1)(f) of the GDPR, the EDPB reiterates that three cumulative conditions must be met: the pursuit of a legitimate interest by the controller or a third party; the necessity of the processing for such interest; and a balancing test ensuring that the interests or fundamental rights and freedoms of the data subjects are not overridden. Supervisory Authorities must therefore verify that controllers have carefully assessed and documented these conditions.

A key element concerns the necessity test, which requires that the processing of personal data to be essential to the stated purpose and that no less intrusive alternative is available. From this perspective, the EDPB provides an implicit opening to the use of synthetic data, since if controllers can achieve the same objectives, such as training an AI model, by using synthetic datasets that replicate statistical properties without processing actual personal data, then the processing of real personal data may not satisfy the necessity condition. Consequently, synthetic data can serve as a decisive tool in demonstrating compliance with the principle of data minimization and in strengthening the legitimacy of AI development processes.

Furthermore, when assessing whether the legitimate interest pursued by the controller or a third party prevails over the fundamental rights and freedoms of data subjects, a careful balancing test is required.

This evaluation must take into account not only the nature of the interests involved, but also the degree of potential interference with the data subjects' privacy and autonomy. Where the interests of the individuals concerned appear to prevail over the controller's objectives, mitigation measures may be introduced to reduce the impact of the processing. Such measures must be tailored to the specific context of the processing, examples include technical solutions that do not amount to full anonymization, such as robust access controls, pseudonymization techniques, or the introduction of output filters and post-training "unlearning" methods designed to prevent the storage, regurgitation, or unintended generation of personal data.

Within this framework, synthetic data assumes particular importance, since this technology can be seen as a powerful risk-mitigation tool in the balancing test, by mitigating significantly the risks to the rights and freedoms of data subjects. By enabling controllers to pursue their legitimate interests, such as innovation, testing, or model improvement, while substantially lowering the risks of infringing upon data subjects' rights, synthetic data contributes to aligning the necessity and proportionality of the processing with the fundamental principles of the GDPR. In this sense, synthetic data not only reduces the likelihood of harm but also reinforces the argument that less intrusive alternatives to personal data exist, thereby strengthening the lawfulness of processing under Article 6(1)(f) of the GDPR.

3.2.3 On the possible impact of an unlawful processing in the development of an AI model on the lawfulness of the subsequent processing or operation of the AI model

Lastly, the EDPB examines the implications of unlawful processing of personal data during the development of an AI model for its subsequent use or operation. In order to answer in a complete manner, the Opinion 28/2024 identifies different scenarios.

In cases where personal data are unlawfully processed by the controller during the development phase and the resulting model continues to process personal data always by the same controller, the illegality of the initial processing will directly affect the lawfulness of subsequent processing, potentially leading to corrective measures such as data deletion that would prevent further use of the model.

Similarly, if the unlawfully developed model is later deployed by another controller, each controller remains independently accountable for ensuring that the model's operation complies with the GDPR. This requires careful due diligence, particularly with respect to the provenance of the data and the findings of Supervisory Authorities.

In the final scenario considered by the EDPB, a controller unlawfully processes personal data during the development of an AI model, but subsequently ensures that the model is anonymized before either the same or another controller engages in further processing at the implementation stage. In such circumstances, Supervisory Authorities remain competent to intervene both with respect to the initial unlawful processing and the subsequent anonymization of the model. Once it can be convincingly demonstrated that the functioning of the AI model no longer involves the processing of personal data, the GDPR ceases to apply. Consequently, the unlawfulness of the initial processing should not affect the lawfulness of the subsequent operation of the anonymized model. Nevertheless, the EDPB underlines that a mere assertion of anonymization is insufficient. Claims of anonymization must be rigorously assessed by Supervisory Authorities against the criteria set out in the Answer 1, particularly with regard to the risk of re-identification through reasonably available means. This requirement reflects the principle that anonymization is not a formal declaration, but the outcome of a demonstrable and verifiable process.

At this stage, synthetic data can play a particularly valuable role, since if a model initially trained on unlawfully processed data is later subjected to a rigorous anonymization process, the subsequent replacement or supplementation of sensitive elements with synthetic datasets could further mitigate the risk of residual re-identification. In this sense, synthetic data are not portrayed as a definitive solution, but rather as a corrective instrument that can reinforce the credibility of the model's anonymization, particularly in cases where Supervisory Authorities must assess whether the unlawfulness of the initial processing continues to affect subsequent stages.

In conclusion, although the EDPB does not explicitly address synthetic data within the analysed Opinion, its reasoning leaves a significant interpretative opening that this thesis seeks to exploit. By stressing the centrality of anonymization as a decisive threshold for the applicability of the GDPR, and by requiring supervisory authorities to rigorously assess claims of anonymization, the EDPB implicitly delineates a conceptual space where synthetic data may be situated. Indeed, synthetic data can be understood as a technical and legal response to the concerns raised by the EDPB. On the one hand, they represent a potential path toward effective anonymization, ensuring that the subsequent use of datasets and AI models falls outside the scope of the GDPR, on the other hand, they simultaneously expose the structural limits of the GDPR, which was designed without specific reference to synthetic data and therefore lacks explicit criteria for determining when such data can be truly anonymous. This regulatory gap forces reliance on broad principles, such as the substantive approach of Recital 26, and on the interpretative vigilance of Supervisory Authorities, thereby

creating uncertainty for developers and controllers who seek legal certainty when deploying synthetic datasets.

This interpretative space, though not formally codified, is crucial since it highlights how synthetic data could be regarded not merely as a technological instrument but as a possible legal mechanism for reconciling innovation with data protection. At the same time, it underscores the responsibility of Supervisory Authorities in scrutinizing whether synthetic datasets truly achieve the standard of anonymity demanded by EU law.

Furthermore, the debate on personal data protection in the context of AI models built through the use of personal data, even when synthetically generated, must be examined across three critical phases: the training of the model, the generation of outputs, and the subsequent inspection of the model. It is particularly within this last phase that risks such as overparameterization may arise, which can inadvertently enable the reconstruction or disclosure of personal information, thereby challenging the very premise of anonymity.

3.3 Privacy risks of synthetic data

Although synthetic data presents itself as a promising solution to address the stringent challenges of privacy and data access in the age of artificial intelligence, their adoption is intrinsically connected to a complex series of risks and limitations, that require a careful analysis and management before their operational deployment.

The impact of such technologies is not univocal and can manifest both in positive externalities, such as increased competition and innovation through the reduction of barriers to data access, and negative ones, amplifying pre-existing challenges such as the intensification of power asymmetries and data externalities or introducing entirely new problems, requiring a balance between utility and data protection[58].

The following subsections will examine three critical privacy and utility risks that emerge from synthetic data deployment: the perpetuation and amplification of algorithmic bias, the phenomenon of model collapse resulting from recursive training on synthetic content, and the complex challenge of re-identification risks, with particular attention to the problem of overparameterization in generative models. Understanding these risks is essential for developing appropriate safeguards and ensuring that synthetic data successfully achieves its promise as a privacy-enhancing technology while avoiding the creation of new vulnerabilities that could undermine both individual privacy and the integrity of AI systems.

3.3.1 The challenge of bias inheritance in synthetic datasets

A significant risk lies in the potential perpetuation or even introduction of bias and inaccuracies[59]. These risks require careful management and analysis, as they can compromise the fairness, accuracy, and reliability of artificial intelligence systems trained on such data. This risk emerges from the fact that models employed for synthetic data generation learn patterns, distributions, and correlations from real data. Consequently, they inherently replicate the prejudices and inaccuracies present in real-world datasets, unless these are proactively identified and corrected during the design and generation phases[60]. Inaccurate or incomplete original data can have a harmful impact on the quality of synthetic

[58] Michal S. Gal and Orla Lynskey, *Synthetic Data: Legal Implications of the Data-Generation Revolution*, LSE Law, Society and Economy Working Papers 6/2023 (London: London School of Economics and Political Science, 2023), https://ssrn.com/abstract=4414385.

[59] See CIPL, *Privacy-Enhancing and Privacy-Preserving Technologies in AI: Enabling Data Use and Operationalizing Privacy by Design and Default*, 2025; Gal and Lynskey, *Synthetic Data: Legal Implications of the Data-Generation Revolution*, 2023.

[60] CIPL, *Privacy-Enhancing and Privacy-Preserving Technologies in AI: Enabling Data Use and Operationalizing Privacy by Design and Default*, 2025, p.8

data, as the models themselves may become imprecise due to errors in input data, misleading baseline information, or incorrect assumptions about distributions and correlations between variables, leading to a significant reduction in the overall quality of generated data[61].

Furthermore, although synthetic data can be used to mitigate bias, for example by adding examples of underrepresented categories[62], this process requires deep awareness and expertise; inaccurate social engineering could inadvertently perpetuate or introduce other types of prejudices that programmers might be less aware of.

A well-known example, although the case did not use synthetic data for bias perpetuation, clearly illustrates how biases in real data are transferred to AI systems. Amazon attempted to train an algorithm to evaluate candidates for technical positions using resumes submitted in previous years. Since the industry was historically male-dominated, the algorithm learned to judge male candidates as superior and penalized references indicating that the candidate was female. If synthetic data had been generated from this original dataset without targeted intervention, they would have perpetuated the same bias. In this context, conversely, a proper application of synthetic data could have been used to counter such bias by adding synthetic examples of successful women's resumes[63].

In another example, an algorithm trained to distinguish images of Husky dogs from wild wolves showed bias. Although it was often accurate, the separating principle it had adopted was insufficiently representative: it focused on the background, having learned that a white background (snow) meant a wolf. To teach the algorithm not to focus solely on the background, a data perturbation technique was applied, mixing parts of different datasets. This generated synthetic data in which wolves appeared on a variety of backgrounds, allowing the algorithm to learn to distinguish animals based on more relevant parameters[64].

3.3.2 The reality distortion of model collapse

Another risk arises from a phenomenon of growing concern, especially in the field of generative artificial intelligence models, known as "model collapse."

[61] Gal and Lynskey, *Synthetic Data: Legal Implications of the Data-Generation Revolution*, 2023, p. 18
[62] *Ivi*, pp. 10, 17, 47.
[63] Xinyu Chang, "Gender Bias in Hiring: An Analysis of the Impact of Amazon's Recruiting Algorithm," in *Proceedings of the 2023 International Conference on Management Research and Economic Development* (Seattle: University of Washington, 2023), 1–9, https://doi.org/10.54254/2754-1169/23/20230367.
[64] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, arXiv preprint arXiv:1602.04938v3 [cs.LG], August 9, 2016, 1602.04938

This term describes a situation in which a model's performance progressively degrades following repeated training cycles on synthetic data. This condition leads to a "reality distortion", meaning a statistical drift in which the model gradually loses the fundamental statistical properties of the original real data. Scientific research has identified the effects of this distortion as a progressive loss of diversity in the synthetic world and a blending of patterns, culminating in predictions that tend to be uniform and poorly aligned with phenomena observed in reality when real data are absent[65].

This problem is particularly relevant in the context of Large Language Models (LLMs), where there is a risk that an increasing number of models will be trained on synthetic data produced by other publicly available LLMs, replacing real data obtained via web scraping. The phenomenon is not confined to text, it also occurs in models trained on successive cycles of synthetic images, which produce visual glitches and distortions. Although model collapse cannot be completely avoided when training relies exclusively on synthetic data, it can be mitigated by integrating real data into training, continuously evaluating the accuracy of synthetic data, and implementing traceability and watermarking techniques to identify and monitor synthetic data, thereby reducing the risk of over-reliance on them[66].

3.3.3 Re-identification threats and outlier exposure

The risk of re-identification and the issue of outliers represent significant challenges for synthetic data, undermining its promise of privacy protection and utility. Although synthetic data are artificially generated and presumed not to have a direct link to real records, research shows they can still expose sensitive information.

The re-identification risk manifests through various types of attacks, like singling-out attacks, where synthetic instances too closely resemble real individuals or linkability attacks, where synthetic data are linked to real identities by matching attributes across datasets and inference attacks, which deduce sensitive information from statistical patterns[67], precisely the three key identifiability factors identified by Working Party in its Opinion 05/2014 on anonymization techniques, as described in paragraph 3.1.2 of this thesis. These privacy breaches occur when an attacker can learn new

[65] Giuseppe D'Acquisto, "Dati sintetici: cosa sono, le applicazioni e i rischi da gestire", *Agenda Digitale*, May 6, 2024, Dati sintetici: cosa sono, le applicazioni e i rischi da gestire - Agenda Digitale
[66] CIPL, *Privacy-Enhancing and Privacy-Preserving Technologies in AI: Enabling Data Use and Operationalizing Privacy by Design and Default*, 2025, p.9.
[67] *Ivi*, pp. 8-9.

information about a specific individual, such as age or income, because strong correlations between attributes present in real data are replicated in synthetic data[68].

A crucial aspect of this risk relates to outliers, meaning data points significantly different from the rest of the dataset[69]. If synthetic models capture outliers from the original data, these records, having rare attributes or extreme numerical values, remain highly vulnerable to disclosure, enabling malicious actors to infer real-world information. Specifically, if an outlier is observed in the real world, it is unlikely to be reproduced by synthetic data; conversely, if an outlier appears in a synthetic dataset, it is unlikely to correspond to any real individual[70].

To mitigate these risks, it is essential to remove direct identifiers and identified outliers and to apply differential privacy[71], as will be discussed in detail in the next chapter. Indeed, privacy protection inherently requires "hiding" vulnerable data points such as outliers, although this may come at the cost of data utility, since extreme cases, often drivers of scientific progress, are easily observable in the real world but rarely captured in synthetic data. These considerations are critical for applying data protection laws because they may cause false positives and negatives, potentially resulting in discriminatory inclusion or exclusion, as well as true positives and negatives, leading to undesirable individual exposure[72].

3.3.4 Overparameterization in synthetic data: definition, risks, and privacy implications

Closely linked to the previously discussed risks of re-identification and outliers, overparameterization represents one of the most significant and technically complex threats to individual privacy in the context of synthetic data generation. This phenomenon constitutes an amplifying factor of the vulnerabilities already identified, transforming what might be considered theoretical weaknesses into actual mechanisms of privacy compromise.

The concept of overparameterization refers to the situation where a machine learning model is designed with a number of parameters significantly greater than the available data samples for training, such that the model's capacity to adapt to data becomes excessive relative to its ability to generalize to new cases[73]. Under normal conditions, an excess of parameters may facilitate the

[68] Emiliano De Cristofaro, *Synthetic Data: Methods, Use Cases, and Risks*, arXiv preprint arXiv:2303.01230v3 [cs.CR], February 27, 2024, 2303.01230, p.4.
[69] CIPL, *Privacy-Enhancing and Privacy- Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age*, 2023, p. 47.
[70] D'Acquisto, "Dati sintetici: cosa sono, le applicazioni e i rischi da gestire" 2024.
[71] *Ivi*, p. 43
[72] D'Acquisto, *"Synthetic Data and Data Protection Laws",* p. 234.
[73] Ahmet Cagri Duzgun, Samy Jelassi, and Yuanzhi Li, *How Does Overparameterization Affect Features?*, arXiv preprint arXiv:2407.00968v1 [cs.LG], July 1, 2024, https://arxiv.org/pdf/2407.00968v1

learning of complex patterns, but it also entails the possibility that the model memorizes specific details of the training data rather than extracting only the general statistical trends.

In the context of synthetic data, overparameterization can translate into a significant risk for personal data protection. Indeed, when a generative model exhibits excessive parametric capacity, it tends not only to capture general correlations between variables, but also to replicate almost identically individual records from the original dataset, a phenomenon known as overfitting[74]. As a matter of fact, while overparameterization refers to designing a model with more parameters than are necessary for the available data, overfitting represents the empirical outcome when such a model memorizes training data instead of generalizing patterns.

Analysing the domain of synthetic data, this excess complexity implies that the generator may produce observations so similar to real ones as to enable, through simple proximity comparisons or similarity metrics, the reconstruction or identification of individual subjects present in the training dataset[75].

Such risk manifests particularly in the form of membership inference attacks, where an attacker, possessing a target record and appropriate analytical methods, can determine with high probability whether that record was part of the training set and thus deduce sensitive information about the corresponding individual[76]. Indeed, it has been demonstrated that as generative models become increasingly overparameterized, their vulnerability to membership inference attacks proportionally increases, thereby elevating the risk of privacy breach through unintended memorization of training data[77].

Beyond membership inference, overparameterization of synthetic models entails other privacy risks. As analysed in the previous paragraph, it facilitates the re-identification of unique individuals or those belonging to restricted groups, for example patients with rare pathologies, since the generator, by memorizing peculiar characteristics, may return exactly such data trajectories. In this context, an oversized model makes evident the inclusion of details relating to specific individuals, compromising the protection mechanisms designed to guarantee the impossibility of extracting personalized information.

---

[74] Daniel Susser et al., "Synthetic Health Data: Real Ethical Promise and Peril," *The Hastings Center Report* 54, no. 5 (2024): 8–13, https://doi.org/10.1002/hast.4911, p. 10.
[75] Amy Steier, Lipika Ramaswamy, Andre Manoel, and Alexa Haushalter, *Synthetic Data Privacy Metrics*, arXiv preprint arXiv:2501.03941v1 [cs.LG], January 7, 2025, 2501.03941v1.
[76] Jasper Tan, Blake Mason, Hamid Javadi, and Richard G. Baraniuk, *Parameters or Privacy: A Provable Tradeoff Between Overparameterization and Membership Inference*, arXiv preprint arXiv:2202.01243v2 [stat.ML], November 30, 2022, 2202.01243.
[77] *Ibid*.

From a legal perspective, this eventuality constitutes a violation of the data minimization principle established by the GDPR, as the synthetic dataset does not remain a sufficiently distant derivation from real data but, through its punctual reproductions, continues to contain personal information attributable to the original subjects, effectively configuring a processing of personal data that would require additional normative safeguards. Furthermore, the principle of accountability is also involved, as the controller's ability to demonstrate having adopted adequate measures to prevent leakage is called into question. Finally, if a model generates "synthetic" data that nevertheless contains personal data in practice, the controller risks basing subsequent processing on a false premise (considering them anonymous), thereby compromising the entire chain of lawfulness. For those reasons, the availability of data that do not adequately respect anonymity exposes the data controller to the risk of sanctions for non-compliance with data protection requirements and to potential claims from data subjects, with consequent reputational and financial damages.

To mitigate such risks, it is therefore essential to adopt regularization strategies for the generative model and apply privacy-by-design methodologies in the synthetic data generation process. Techniques such as the introduction of statistical noise, for example, through differential privacy algorithms or by introducing innovative machine unlearning techniques, can enable significant reduction of the tendency toward overfitting and, consequently, the danger of personal information leakage. In any case, the design of a synthetic dataset must always be accompanied by a preventive assessment of re-identification risk, through membership inference tests and similarity analyses, to ensure that the synthetic output leaves no room for individual identities. Only through an integrated approach that contemplates both technical aspects and normative constraints is it possible to exploit the benefits of synthetic generation while minimizing the impact on data subjects' privacy.

Chapter 4 – Shaping the Future of AI and Data Protection: The Aindo Case

4.1 Addressing privacy threats

Despite the significant privacy risks and technical challenges outlined in the preceding chapter, synthetic data generation can be transformed into a fundamentally secure and highly effective privacy enhancing technology, through the strategic implementation of complementary methodologies and robust safeguards.

Rather than accepting the false dichotomy that has historically forced organizations to choose between privacy protection and data utility, emerging approaches demonstrate that sophisticated technical solutions can achieve an optimal balance across all critical dimensions, meaning privacy preservation, data fidelity, and practical utility, thereby positioning synthetic data as a cornerstone technology for the secure development and deployment of artificial intelligence systems.

This transformative potential is particularly significant when considered within the framework of the European Union's Ethics Guidelines for Trustworthy AI, which emphasize seven fundamental requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability.[78] The integration of advanced privacy-enhancing techniques with synthetic data generation not only addresses these ethical imperatives but creates a synergistic ecosystem where each component reinforces the others, ultimately delivering AI systems that are both technically superior and ethically sound.

Central to this comprehensive approach is the recognition that transparency and respect for data subjects' rights must be embedded throughout the entire AI development lifecycle, from initial data collection through model deployment and ongoing operation. As highlighted by Professor D'Acquisto, two promising strategies to effectively operationalize the right to opt-out in such computation-intensive contexts are the adoption of randomisation techniques, including differential privacy, and machine unlearning, each embodying a distinct philosophy of privacy protection[79].

[78] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (Brussels: European Commission, April 8, 2019), ai_hleg_ethics_guidelines_for_trustworthy_ai-en_87F84A41-A6E8-F38C-BFF661481B40077B_60419.pdf

[79] D'Acquisto, *"Synthetic Data and Data Protection Laws",* p. 236.

By incorporating differential privacy techniques during the pre-training phase of synthetic data generation, organizations can introduce mathematically rigorous privacy guarantees that provide formal protection against re-identification attacks while preserving the statistical properties essential for effective model training[80]. This approach addresses one of the most critical vulnerabilities identified in Chapter 3: the risk that overparameterized models may inadvertently memorize and subsequently expose sensitive information from their training data. Complementing this preventive approach, post-training machine unlearning techniques offer a powerful mechanism for selectively removing specific data points or patterns from trained models, providing organizations with the flexibility to respond to evolving privacy requirements, regulatory mandates, or individual data subject requests without necessitating complete model retraining. This capability is particularly valuable in addressing the dynamic nature of privacy compliance, where changing regulations or individual preferences may require retroactive adjustments to model behaviour[81].

The convergence of these technologies, synthetic data generation, differential privacy, and machine unlearning, represents more than a mere technical advancement since it constitutes a paradigm shift toward the integration of the privacy-by-design approach in the development of artificial intelligence. This integrated approach not only strengthens each component of the trustworthy AI framework but provides essential protection against the overfitting phenomenon that Chapter 3 identified as one of the most significant risks in synthetic data deployment. By implementing regularization strategies and continuous monitoring throughout the model lifecycle, organizations can ensure that their synthetic data generation processes remain both technically robust and privacy-preserving.

Furthermore, this comprehensive methodology is particularly crucial when deploying AI systems in critical sectors such as healthcare and finance, where the sensitivity of the underlying data demands the highest standards of privacy protection without compromising the analytical insights necessary for innovation and scientific advancement. The successful implementation of these integrated privacy-preserving techniques not only ensures compliance with existing regulatory frameworks but positions organizations to adapt to future privacy requirements while maintaining their competitive edge in an increasingly data economy.

---

[80] Alex Bie and Umar Syed, "Generating Synthetic Data with Differentially Private LLM Inference," *Google Research Blog*, March 18, 2025. Accessed September 14, 2025 https://research.google/blog/generating-synthetic-data-with-differentially-private-llm-inference

[81] Saskia Keskpaik, "Machine Unlearning," *European Data Protection Supervisor – TechSonar*, Accessed September 14, 2025, https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/machine-unlearning_en

The following sections will examine in detail how these complementary technologies can be systematically integrated to create a new generation of AI systems that support individual privacy rights while advancing the frontiers of artificial intelligence research and application, ultimately demonstrating that the apparent tension between privacy and innovation is not an insurmountable challenge but an opportunity for creating more trustworthy, ethical, and effective technological solutions.

4.1.1 Differential Privacy

As already introduced within this analysis, differential privacy, which fits perfectly within the framework of Privacy Enhancing Technologies as defined by the CIPL[82], represents an innovative and crucial methodology in the field of data protection, particularly in machine learning and advanced analytics contexts, and as will also be explored in the context of synthetic data generation. It can be defined as a robust mathematical framework that ensures that the inclusion or exclusion of a single data point within a dataset used for model training does not significantly alter the final output of the model itself [83]. This is formalized through the concept of ε-indistinguishability, which ensures that for every model output (called a "transcript"), the probability of obtaining it from two databases that differ by only one record is very similar[84]. This mechanism is based on the idea that two different scenarios, one in which a specific data point is included in the training dataset and another in which the same point is excluded, must produce statistically indistinguishable output distributions, making it impossible to infer whether a particular element was included in the training dataset simply by analysing the output[85].

[82] CIPL, *Privacy-Enhancing and Privacy- Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age*, 2023, p. 38.
[83] Aindo, "Enhancing Model Training with Differential Privacy," *Aindo Blog*, March 5, 2025, accessed September 14, 2025, https://www.aindo.com/blog/differential-privacy
[84] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith, *Calibrating Noise to Sensitivity in Private Data Analysis* (Microsoft Research Technical Report, 2006), https://people.csail.mit.edu/asmith/PS/sensitivity-tcc-final.pdf, p.2
[85] Aindo, "Differential Privacy," 2025.

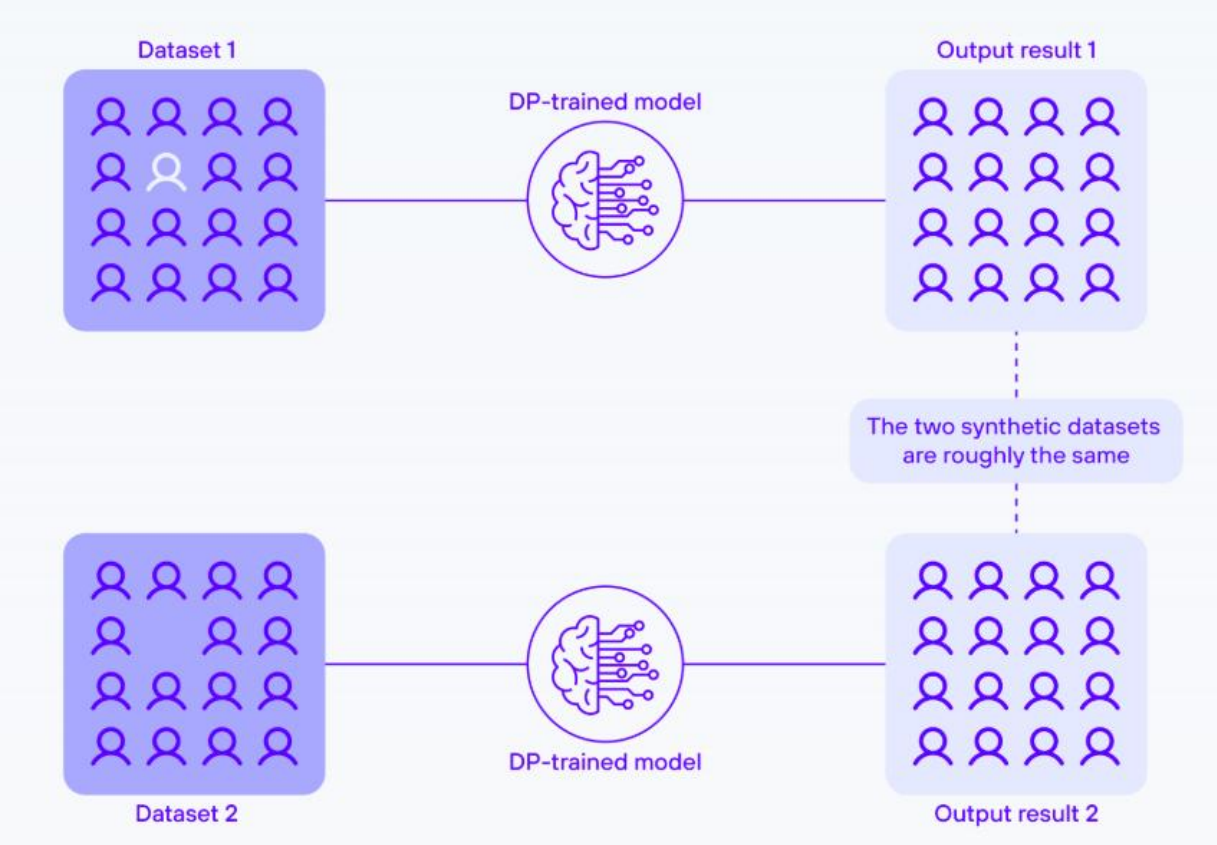

FIGURE 6 : HOW DIFFERENTIAL PRIVACY WORKS

SOURCE: AINDO, "ENHANCING MODEL TRAINING WITH DIFFERENTIAL PRIVACY," AINDO BLOG, MARCH 5, 2025, ACCESSED SEPTEMBER 14, 2025, HTTPS://WWW.AINDO.COM/BLOG/DIFFERENTIAL-PRIVACY

In other words, this technology introduces an appropriate amount of random "noise" into the responses or calculations extracted from a dataset, making it practically impossible to infer, by analysing the output, whether a particular individual or their specific data was included in the training dataset. The primary objective is therefore to make data usable for statistical analysis and predictive model construction while maintaining a strong guarantee that individual information will not be compromised.

The utility of differential privacy is significantly amplified when coupled with the use of synthetic data. This type of data, being artificially generated to emulate the statistical properties of real data without containing direct information about individuals, already constitutes a technique for disclosure limitation. However, the application of differential privacy during the process of generating this synthetic data offers an additional and more robust level of protection, particularly to mitigate the risks they may pose. When synthetic data is created through an algorithm that preserves differential privacy, it guarantees that even if an attacker were to access both the generative model and the

synthetic data produced, they would not be able to determine whether a specific record from the original real data was used for training[86].

A fundamental property of differential privacy, known as the "post-processing property"[87], further ensures that any subsequent operation on differentially private synthetic data cannot cancel the privacy protection effect, a characteristic of great importance also in relation to regulations such as the GDPR. Precisely in light of what has been analysed, this synergy between differential privacy and synthetic data allows for unlocking new possibilities for research and innovation, enabling the use of sensitive information in an ethical manner that complies with regulations.

Another relevant characteristic of this technology consists in the fact that differential privacy offers an effective solution to the problem of overparameterization and overfitting in machine learning models, as well as re-identification of rare cases such as outliers. Overfitting occurs when a model learns the training data too thoroughly, including noisy details or specificities of individual data points, losing the ability to generalize to new data. By introducing calibrated noise during training, the contributions of individual data points are "hidden", forcing the model to identify and learn only general trends and significant relationships within the dataset. This prevents the model from excessively adapting to a single entity or to specific and potentially identifiable details, improving its generalization capacity and reducing the risk of membership inference attacks.

Furthermore, some implementations of differential privacy, as highlighted by Aindo [88], incorporate "early stopping" procedures. This is a mechanism that interrupts model training when the reduction in validation loss, that is, the error made by the model on a dataset not used for learning, which becomes marginal and falls below a predetermined threshold. This prevents overfitting, a phenomenon whereby the model tends to "memorize" the training data instead of learning general patterns, compromising both generalization capacity and privacy protection. The adoption of early stopping thus allows for limiting the risk that sensitive information becomes incorporated into the model parameters, while ensuring the maintenance of the formal protections that differential privacy intends to guarantee. In summary, this technology not only protects individual privacy but also contributes to building more robust, reliable machine learning models that are less inclined to memorize sensitive information.

[86] *Ibid*.
[87] *Ibid*.
[88] *Ibid.*

4.1.2 Machine Unlearning

Machine unlearning emerges as a fundamental paradigm in modern artificial intelligence, offering machine learning models the ability to selectively "forget" specific information or data points used during their training phase, without the need for costly and resource-intensive complete retraining[89].

This need arises from the awareness that contemporary computer systems hold enormous quantities of personal data, whose persistence in AI models can constitute a threat to privacy and weaken trust between users and technology. Unlike traditional databases, where data removal is straightforward, machine learning models often "remember" the learned information, making simple deletion from the back-end insufficient. Machine unlearning directly responds to regulations such as the "right to be forgotten", also known as right to erasure, present in the Article 17 of the GDPR, allowing organizations to comply with requests for personal data removal and safeguard individuals' privacy.

Beyond privacy, machine unlearning proves valuable for improving model security, for example by removing adversarial data introduced by adversarial attacks that are very similar to the original data to the extent that a human cannot distinguish between the real and fake data. This adversarial data is designed to force the deep learning models into outputting wrong predictions, which frequently results in serious problems. Another scenario in which the potential of machine unlearning can be exploited is also to mitigate undesired bias that may emerge from training data, thus promoting fairness and reducing the risk of discriminatory outcomes. Its utility also extends to model adaptability, allowing them to remain relevant in dynamic data contexts, eliminating obsolete or no longer representative information[90].

In a context where synthetic data generation is increasingly encouraged to overcome the scarcity of real data or restrictions on their accessibility, and to develop machine learning algorithms compliant with data protection regulations, machine unlearning assumes a crucial complementary role. Although synthetic data can be placed on the same level as artificial data, the training phase of "synthesizers", that are the models that generate synthetic data, involves real personal data, therefore, machine unlearning offers the possibility to remove the influence of specific individual data points from the synthesizer parameters, when required, without having to retrain the entire synthetic data

[89] See D'Acquisto, *"Synthetic Data and Data Protection Laws",* p. 237; Pecan AI, "The Rise of Machine Unlearning," *Pecan Blog*, June 25, 2024, accessed September 15, 2025 https://www.pecan.ai/blog/the-rise-of-machine-unlearning; Keltin Grimes, Collin Abidi, Cole Frank, and Shannon Gallagher, "3 Recommendations for Machine Unlearning Evaluation Challenges," *Software Engineering Institute Insights (blog)*, Carnegie Mellon University, August 26, 2024, accessed September 15, 2025, https://www.sei.cmu.edu/blog/3-recommendations-for-machine-unlearning-evaluation-challenges

[90] Thanh Tam Nguyen et al., *A Survey of Machine Unlearning*, arXiv preprint arXiv:2209.02299 [cs.LG], 2024, https://arxiv.org/pdf/2209.02299

generation process.[91] This reinforces the “privacy by design” approach associated with synthetic data, enabling more granular and reactive information management.

Furthermore, machine unlearning offers a direct solution to the problem of overparameterization and overfitting, phenomena by which machine learning models, especially deep neural networks, tend to excessively memorize training data instead of learning general patterns. Such "memorization" as already specified, can lead to poor generalization capability on new data and a decline in performance. Machine unlearning can “repair” over-trained deep neural networks by actively removing useless, obsolete or redundant data samples that contribute to this excessive memorization. By selectively eliminating this information, the model can be guided to focus on more significant patterns, preventing overfitting and improving its overall accuracy, acting similarly to selective compression of useful data[92].

Therefore, machine unlearning is not only a compliance tool, but also a mechanism to optimize the architecture and intrinsic performance of artificial intelligence models, promoting more robust, fair and reliable systems.

However, it is correct to highlight that, being a very innovative method, especially when associated with the use of synthetic data, there are challenges that must be overcome with more in-depth studies and experimentation. If on one hand this technique, when applied correctly, can allow working with more precise data compared to when differential privacy is adopted, it is extremely complex to understand how an individual’s data has influenced model optimization. This is extremely complex, because during training the data is “fused” into weights and parameters, and it is not at all simple to isolate the specific contribution of a record. For this reason, machine unlearning is still an ongoing research objective since theoretically this technology is promising, but in practice it presents difficulties that must be taken into consideration and analysed accurately[93].

[91] D’Acquisto, *“Synthetic Data and Data Protection Laws”,* p. 237
[92] Nguyen et al., A *Survey of Machine Unlearning*, 2024, p.20
[93] D’Acquisto, *“Synthetic Data and Data Protection Laws”,* p. 237

4.2 The Aindo experience: a case study in applied synthetic data innovation

Aindo, a company previously referenced throughout this thesis as a source of inspiration and as an example of success in the applied synthetic data domain, provides a compelling case study for examining the practical implementation of synthetic data technologies in industry settings. Founded in 2018 as an artificial intelligence start-up emerging from the International School of Advanced Studies (SISSA) in Italy, Aindo initially positioned itself as a consulting firm specializing in artificial intelligence algorithm development.

However, Aindo's strategic shift toward synthetic data generation emerged from a key insight about the challenges facing AI innovation. The company discovered that the main obstacle to AI advancement was not developing sophisticated algorithms, which have increasingly become a commodity, but rather gaining access to high-quality data, especially in sensitive fields like healthcare. This insight revealed a fundamental paradox: while organizations possess valuable databases with enormous analytical potential, privacy regulations and compliance requirements severely restrict their use. In response to this challenge, Aindo developed a comprehensive data valorisation platform based on synthetic data generation, effectively transforming data scarcity from a barrier into an opportunity for innovation.

Aindo's approach to synthetic data quality assessment is multidimensional, and aims to address the question about the quality of the data that they generate. This evaluation framework encompasses three critical dimensions that collectively determine the viability of synthetic data for practical applications: privacy, fidelity, and utility[94]. These dimensions work together to ensure that synthetic data not only protects the privacy of original data subjects but also maintains the essential characteristics of the original datasets while remaining effective for AI model training and other analytical purposes.

Fidelity and utility represent two distinct but interconnected aspects of synthetic data quality that are often confused but serve fundamentally different evaluative purposes. Fidelity measures the extent to which synthetic data accurately reproduces the statistical properties and distributional patterns of the original dataset. This is assessed through statistical metrics, high fidelity indicates that the synthetic data preserves the overall shape, variance, and statistical relationships present in the real data at a

[94] Daniele Panfilo et al., "A Deep Learning-Based Pipeline for the Generation of Synthetic Tabular Data," *IEEE Access*, vol. 11, 2023, IEEE Xplore Full-Text PDF:

macro level. Utility, conversely, evaluates how effectively the synthetic data performs in downstream applications compared to the original data. A synthetic dataset may achieve high fidelity by accurately reproducing statistical distributions yet demonstrate low utility if it fails to capture the nuanced patterns necessary for specific analytical tasks. For instance, in medical research, synthetic data might perfectly replicate population-level health statistics (high fidelity) but prove inadequate for training diagnostic models for rare diseases (low utility) due to insufficient representation of critical edge cases. This distinction becomes particularly relevant when managing outliers: while their removal might minimally impact overall fidelity metrics, it could significantly compromise utility for applications specifically requiring the modelling of rare but scientifically important phenomena.

Aindo's privacy framework incorporates both technical and organizational measures[95]. Technical privacy measures include: control privacy mechanisms implemented a priori (such as differential privacy), measurement privacy approaches that quantify potential risks a posteriori through statistical metrics, and adversarial testing to verify attack resistance. Organizational privacy measures encompass local deployment ensuring only data owners access real data, restricted generator access limiting black-box attack vectors, and formal certification processes, including Europrivacy certification obtained following inspection by the Italian Supervisory Authority, which assess that their activities are compliant with the GDPR[96].

4.2.1 Aindo's Perspective on Biases and Outliers

To strengthen the empirical foundation of this research and provide comprehensive analysis of Aindo's approach, I conducted an in-depth interview with the CEO Daniele Panfilo and Alexander Boudewijn, Aindo's Data Privacy Researcher. Specifically, two main issues that constitute a potential obstacle to the adoption of these technologies were explored in depth:

- since synthetic data reproduces the statistical patterns of real data, there is a risk that biases present in the original datasets may be replicated or even amplified. How do you address this issue within your models? In particular, which techniques do you use to identify hidden biases in large real-world datasets, and how do you manage to correct or mitigate them through synthetic data generation, while ensuring representativeness and reliability?

[95] Daniele Panfilo, remarks during the webinar *"Personal and non-personal data under the EU Regulation: exploring synthetic data",* Rödl & Partner, April 16, 2025.
[96] Aindo. "What Is Europrivacy and Why Does It Matter?" Aindo Blog, June 23, 2025. Accessed September 19, 2025. What is Europrivacy and why does it matter? - Aindo AI

- Outliers, by definition, represent rare but potentially significant cases, as they may pose a re-identification risk and are therefore often attenuated or masked through techniques such as differential privacy. However, these extreme cases can be crucial drivers of scientific progress, especially in fields such as medical research. How do you balance the need to protect privacy by reducing re-identification risk with the need to preserve the accuracy and informational value of synthetic data, without completely excluding outliers?

The first question concerned the risk that synthetic data, in reproducing the statistical patterns of real data, might replicate or amplify biases present in the original datasets.

It is essential to note that Aindo clearly distinguishes between different categories of bias that require differentiated approaches. Fairness biases are primarily linked to demographic characteristics such as ethnicity and gender and have direct ethical implications insofar as they can influence access to resources and opportunities allocated by artificial intelligence systems. These contrast with statistical biases, such as selection bias (when only certain groups are represented in the data) and responder bias (when only people particularly interested in the topic respond to a survey), which emerge during the original data collection process and are characterized by the absence of certain groups or information in the dataset.

To address fairness biases, Aindo has developed a systematic approach structured in three phases: amplification prevention, identification, and mitigation.

Regarding amplification prevention, the company emphasizes how this phenomenon is relatively easy to detect through standard fidelity metrics, since bias amplification would result in a statistically different distribution compared to the original data, compromising the overall quality of synthetic data.

For bias identification, in addition to traditional metrics such as Statistical Parity, which measures whether a favourable outcome occurs with equal probability for all demographic groups, Aindo prefers an approach based on sources external to the original dataset, cross-referencing the generated data with scientific literature, aggregate population statistics, and domain expert opinions. This method allows for the identification not only of biases present in the data, but also those external to the dataset that could compromise representativeness.

For mitigation, they explain that the technique of “conditional generation” is used as the primary method for mitigating fairness-related biases. Specifically, when a demographic group is under-represented (for example, women in prestigious professions), separate generators are trained for each

group, one for males and one for females, and during generation they can oversample from the female generator to generate more females with prestigious jobs in order to rebalance the dataset so that there are proportionally more females with prestigious jobs. At the end, the aim is to merge the male and female synthetic datasets and get a more fair overall synthetic set. This methodology, while introducing a trade-off between fairness and fidelity, as in the example, the average income of females will be larger than in the real-world dataset, allows for systematic correction of identified biases while maintaining the overall utility of the dataset.

The second issue addressed the delicate balance between privacy protection and preservation of scientific information represented by outliers. As explained by Alexander Boudewijn, outliers constitute a particularly insidious privacy risk because they can be easy prey for singling-out attacks, and it is emphasized that this was shown in several studies in which outliers were the most, and nearly only, sensitive to membership inference attacks[97]. Even the absence of outliers in synthetic data can reveal information about the original data, for example, the absence of ultra-centenarians in synthetic data could still allow inference of their presence in the original dataset through analysis of the overall age distribution. To address this challenge, Aindo proposes three solutions.

The first is a hybrid approach that combines deep-learning-based synthetic data generation with simulation techniques[98], in which outliers are removed from the real data before training the generator, and subsequently artificial outliers are created manually through simulation, this process is guided by external knowledge (scientific literature, expert opinions, specialized ontologies) rather than through learning from real data.

The second solution involves pooling multiple datasets, since what constitutes an outlier in one dataset might not be one in another, making privacy protection possible through source diversification.

Finally, it is recognized that in some cases synthetic data might simply not represent the most appropriate privacy-enhancing technology for the specific use case, suggesting the need for case-by-case evaluation.

[97] See Chapter 3, par.3.3 “Privacy Risks of Synthetic Data”
[98] Aindo. “Unlocking the Future with Simulation.” *Aindo Blog*, September 17, 2025. Accessed September 19, 2025. Unlocking the future with simulation - Aindo AI

4.2.1 Final Remarks

This dialogue with Aindo has confirmed that, despite the technical challenges identified, practical and methodologically rigorous solutions exist to address the main obstacles to synthetic data implementation. The company's experience, demonstrates that it is possible to develop industrial approaches capable of effectively balancing privacy protection, scientific accuracy, and operational utility.

It is essential to emphasize that, based on the analysis conducted throughout this research, synthetic data generation constitutes a processing activity, which stands on the same legal terms as the anonymization process. This legal classification does not merely represent a technical matter, but defines the regulatory framework within which synthetic data can be legitimately used as a privacy protection tool.

The institutional validation of this approach finds concrete confirmation in Aindo's experience, which represents the only European company specializing in synthetic data to have obtained Europrivacy certification for synthetic data generation in healthcare. This certification, issued following an in-depth inspection by the Italian Supervisory Authority, constitutes formal recognition that the synthetic data generation processes implemented by the company meet the highest European data protection standards. Europrivacy certification does not represent a mere formal recognition, but attests to the compliance of Aindo's processes with GDPR provisions and European data protection authorities' guidelines[99]. This regulatory precedent establishes a replicable model for the industry, demonstrating that synthetic data can effectively qualify as anonymous data when generated through technically robust methodologies and subjected to rigorous evaluations of resistance to re-identification. Aindo's position as a certified pioneer in this sector not only confirms the technical validity of the synthetic data approach, but also provides a procedural roadmap for other organizations intending to implement these technologies in compliance with European regulations.

The empirical evidence gathered through this research, combined with the institutional recognition represented by Europrivacy certification, definitively supports the thesis that synthetic data can represent a reliable and legally sound solution to the paradox between technological innovation and personal data protection. In this context, synthetic data do not simply constitute a technical expedient to circumvent regulatory restrictions, but represent a genuine privacy by design tool that transforms

[99] Aindo, "What is Europrivacy and why does it matter?", 2025

data protection from a constraint into an enabling factor for European innovation in the artificial intelligence sector.

Conclusions

In light of the analysis carried out throughout this research, it is now appropriate to underline some fundamental considerations that not only recapitulate the key findings, but also provide a reasoned response to the research question and outline the broader implications of this work.

There is a fundamental problem that must be solved: making data more accessible in order to harness its potential virtuously for society. Nowadays, we are surrounded by an infinite quantity of personal data belonging to natural persons which, although it must be protected according to existing laws, is often viewed as a risk and cost rather than the strategic resource it could represent.

The objective of this research has focused on a fundamental paradigm shift: transforming the traditional perception of personal data from regulatory constraint to strategic resource for European innovation. The research aimed to demonstrate how personal data can be reconceptualised as an endogenous informational asset, capable of enhancing advanced research processes and technological development without the need to depend on external sources or compromise the environmental sustainability of innovation processes. This approach is based on the recognition that Europe already possesses a high-quality data ecosystem that, if managed through rigorous privacy-preserving methodologies, can generate significant economic and social value while maintaining the Union's full digital autonomy.

The implementation of advanced anonymisation techniques, such as synthetic data, enables the transformation of this informational wealth into a strategic competitiveness tool that respects the fundamental data protection principles that characterise European normative identity. The practical applications analysed in the healthcare and financial sectors have concretely demonstrated how synthetic data can enable innovations that would otherwise be impossible or excessively risky. Furthermore, these data can be the subject of intelligent and profitable exchange for both parties, since those who possess the resources can both disseminate them following the principle of "data altruism" promoted by the Data Governance Act, while also requesting appropriate economic compensation, given the value that lies behind such resources.

After establishing the strategic importance of data, especially personal data, the time has come to address the crucial question of how to "handle with care" these latter, and synthetic data responds to this necessity. These tools, as demonstrated by the research conducted, if developed accurately through technically robust methodologies, can enable the exchange and dissemination of those data we so desire without compromising individual privacy protection.

After conducting a comprehensive analysis of how these data can be applied concretely, how they function at a technical level, and how they can be interpreted at a legal level, it is possible to respond firmly to the research question posed at the beginning, with reference to determining whether synthetic data can be truly considered anonymous, and therefore considered an effective tool for training AI models without exposing personal data. Furthermore, it will explore whether anonymity should be assessed based solely on quantitative factors, or should we also take qualitative aspects into account. The response is categorical: synthetic data, if developed correctly without compromising individual privacy and using the correct metrics to prevent personal data from re-emerging during the various phases of preparation, development, testing, and use of AI models based on synthetic data, can be considered anonymous data.

From the perspective of legal analysis, the European regulatory framework provides solid foundations for this conclusion. The framing of synthetic data within today's legislative landscape is essential to enable everyone to use them correctly in a safe and proper manner: being considered an anonymisation technique, they can be treated from a legal standpoint just like anonymous data. Recital 26 of the GDPR, which adopts a substantive approach requiring consideration of the concrete circumstances in which data are processed, and in particular whether there is a realistic possibility of identifying the data subject, constitutes the theoretical foundation of this position. Such assessment must take into consideration factors such as the costs and time required, the technologies available at the relevant time, and the broader context of the processing activity. Opinion 05/2014 of the Working Party 29, which identifies the three key identifiability factors, singling out, linkability, and inference, provides concrete operational criteria against which synthetic data can be evaluated. The recent judgment of the Court of Justice of the European Union of 4 September 2025 (EDPS v SRB) further clarified that the substantive approach to anonymisation must consider the actual capacity for re-identification, not merely the theoretical possibility. Even Opinion 28/2024 of the EDPB on artificial intelligence models, although not directly citing synthetic data, has outlined a significant interpretive space that this research has been able to exploit, defining rigorous conditions to determine when an AI model can be considered anonymous.

From a technical perspective, the research has demonstrated that the overparameterization problem, one of the most significant risks in using synthetic data, can be effectively mitigated through the implementation of complementary technical solutions. The integration of differential privacy during the synthetic data generation phase introduces mathematically rigorous guarantees that provide formal protection against re-identification attacks while preserving the essential statistical properties for effective model training. Machine unlearning techniques offer a powerful mechanism to

selectively remove specific data points or patterns from trained models, providing organisations with the flexibility to respond to evolving privacy requirements without necessitating complete model re-training. The convergence of these technologies, synthetic data generation, differential privacy, and machine unlearning, represents more than mere technical progress, constituting a paradigm shift towards integrating the privacy-by-design approach in artificial intelligence development.

The Aindo case proves illuminating and essential for empirically validating these theoretical conclusions. The Italian company, unique in Europe in having obtained Europrivacy certification for synthetic data generation in the healthcare sector, has demonstrated that it is possible to develop industrial approaches capable of effectively balancing privacy protection, scientific accuracy, and operational utility. The Europrivacy certification, issued following an in-depth inspection by the Italian Supervisory Authority, constitutes formal recognition that the synthetic data generation processes implemented by the company respect the highest European data protection standards. This regulatory precedent establishes a replicable model for the industry, confirming that synthetic data can effectively qualify as anonymous data when generated through technically robust methodologies and subjected to rigorous evaluations of resistance to re-identification. Aindo's multidimensional approach to evaluating synthetic data quality - considering privacy, fidelity, and utility - demonstrates how it is possible to systematically address identified technical challenges, from bias management to outlier management.

This assumption must be made considering a broader framework that goes beyond only the statistical methods that are in the background of their generation: qualitative factors must certainly be considered, which are based on a more substantial approach that takes into consideration various cases. Conditions that can be defined as indispensable must exist to consider synthetic data anonymous: implementation of complementary techniques, adoption of anti-overparameterization strategies, demonstrated resistance to singling out, linkability, and inference attacks, substantial analysis from a legal perspective (available technologies, costs, time, processing context), dynamic and continuous evaluation of re-identification risks, and submission to certification processes. At the same time, this field is not black or white, but there are a thousand nuances that must be taken into consideration through case-by-case evaluations.

The explicit inclusion of synthetic data within the Data Governance Act and the AI Act, where they are placed on the same level as anonymous data, represents significant regulatory recognition that testifies to the maturity achieved by these technologies. However, official documents have not yet been produced that definitively attest to this interpretation by competent bodies such as the EDPB. It would indeed be desirable for specific guidelines to be developed that outline necessary conditions

to indicate the correct path to pursue when dealing with synthetic data in the European legal landscape. This would represent a crucial step in providing legal certainty to sector operators and promoting responsible adoption of these emerging technologies.

Looking ahead to future developments, Europe faces an historic opportunity. Synthetic data can represent the key to achieve the ambitious objectives of the European data strategy, transforming privacy protection from a regulatory constraint into a driver for innovation in the era of artificial intelligence. The transformation of personal data from mere objects of legal protection into strategic resources for economic competitiveness and digital innovation finds in synthetic data a privileged tool for reconciling the apparently conflicting needs of individual protection and collective progress. In a geopolitical context where digital sovereignty has become essential to preserve European autonomy, synthetic data can allow Europe to capitalise on its most precious asset, which lies in the quality and richness of its data, without compromising the fundamental principles that characterise the European identity.

Future research should focus on deepening the mechanisms for dynamic evaluation of re-identification risks, developing standardised metrics for measuring the utility of synthetic data in different application domains, and exploring the ethical and social implications of large-scale adoption of these technologies. It would be particularly interesting to investigate how synthetic data can contribute in achieving the data altruism objectives promoted by the Data Governance Act, creating sharing mechanisms that maximise social benefit while maintaining the highest privacy protection standards.

Ultimately, under the conditions specified in this research – implementation of complementary techniques during generation, adoption of anti-overparameterization strategies, demonstrated resistance to singling out, linkability, and inference attacks, substantial analysis from a legal perspective (available technologies, costs, time, processing context), dynamic and continuous evaluation of re-identification risks, and submission to certification processes – synthetic data can legitimately be considered anonymous data. This does not represent simply a technical solution, but an innovative paradigm that transforms privacy protection from obstacle into an engine for the European innovation.

Synthetic data reflects exactly what society and Europe needs now: a tool that allows unlocking the immense potential of European data while respecting the fundamental privacy protection values that characterise the Union’s identity. This technology represents concrete demonstration that it is possible to build a digital future that does not force a choice between innovation and data protection, but

achieves both objectives synergistically, positioning Europe as a world leader in responsible innovation in the era of artificial intelligence.

Bibliography

Books:

Bradford, Anu. *The Brussels Effect: How the European Union Rules the World*. Oxford: Oxford University Press, 2020.

Izenman, Alan Julian. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* New York: Springer, 2013.

Academic Articles and Reports:

Brozzetti, Filiberto E. "EU Digital Sovereignty: How Long Will the 'Brussels Effect' Last?" *Rivista Internazionale di Filosofia del Diritto*, Serie V, no. 2 (2024): 343–371.

Centre for Information Policy Leadership (CIPL). *Privacy-Enhancing and Privacy-Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age*. December 2023. cipl-understanding-pets-and-ppts-dec2023.pdf

Centre for Information Policy Leadership (CIPL). *Privacy-Enhancing and Privacy-Preserving Technologies in AI: Enabling Data Use and Operationalizing Privacy by Design and Default.* March 2025. cipl_pets_and_ppts_in_ai_mar25.pdf

Chang, Xinyu. "Gender Bias in Hiring: An Analysis of the Impact of Amazon's Recruiting Algorithm." In *Proceedings of the 2023 International Conference on Management Research and Economic Development*, 1–9. Seattle: University of Washington, 2023. https://doi.org/10.54254/2754-1169/23/20230367.

D'Acquisto, Giuseppe. *"*Synthetic Data and Data Protection Laws*." Journal of Data Protection & Privacy 6*, no. 3 (2024): 227–239.

Das, Hari Prasanna, Ryan Tran, Japjot Singh, Xiangyu Yue, Geoffrey Tison, Alberto Sangiovanni-Vincentelli, and Costas J. Spanos. "Conditional Synthetic Data Generation for Robust Machine Learning Applications with Limited Pandemic Data." *Proceedings of the AAAI Conference on Artificial Intelligence* 36, no. 11 (2022): 11792-11800. https://doi.org/10.1609/aaai.v36i11.21435.

Davis, Peter, Roy Lay-Yee, and Janet Pearson. "Using Micro-Simulation to Create a Synthesised Data Set and Test Policy Options: The Case of Health Service Effects under Demographic Ageing." *Health Policy* 97, no. 2–3 (2010): 267–74. https://doi.org/10.1016/j.healthpol.2010.05.014

De Cristofaro, Emiliano. *Synthetic Data: Methods, Use Cases, and Risks*. arXiv preprint arXiv:2303.01230v3 [cs.CR], February 27, 2024. 2303.01230

Draghi, Barbara, Zhenchen Wang, Puja Myles, and Allan Tucker. "Identifying and Handling Data Bias within Primary Healthcare Data Using Synthetic Data Generators." *Heliyon* 10, e24164 (2024). https://doi.org/10.1016/j.heliyon.2024.e24164

Duzgun, Ahmet Cagri, Samy Jelassi, and Yuanzhi Li. *How Does Overparameterization Affect Features?* arXiv preprint arXiv:2407.00968v1 [cs.LG], July 1, 2024. https://arxiv.org/pdf/2407.00968v1

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. *Calibrating Noise to Sensitivity in Private Data Analysis*. Microsoft Research Technical Report, 2006. https://people.csail.mit.edu/asmith/PS/sensitivity-tcc-final.pdf

Fabuyi, Jumai Adedoja. *"Leveraging Synthetic Data as a Tool to Combat Bias in Artificial Intelligence (AI) Model Training." Journal of Engineering Research and Reports* 26, no. 12 (2024): 24–46. https://doi.org/10.9734/jerr/2024/v26i121337.

Financial Conduct Authority. *Synthetic Data to Support Financial Services Innovation*. March 2022. Call for input: Synthetic data to support financial services innovation

Finocchiaro, Giulia, Antonio Landi, Gianluca Polifrone, Davide Ruffo, and Francesco Torlontano. *The Regulatory Future of Synthetic Data: Data Synthesis as a Resource for Scientific Research, Innovation, and Public Policy in the European Legal Landscape*. Rome: Istituto Italiano per la Privacy e la Valorizzazione dei Dati and Data Intermediaries Alliance, July 15, 2024. Il_futuro_regolatorio_dei_dati_sintetici__1721197747 (1).pdf

Fontanillo López, César Augusto, and Abdullah Elbi. *On the Legal Nature of Synthetic Data*. NeurIPS 2022 Workshop on SyntheticData4ML. Leuven: Center for IT and IP Law, KU Leuven, 2022. pdf

Gal, Michal S., and Orla Lynskey. *Synthetic Data: Legal Implications of the Data-Generation Revolution*. LSE Law, Society and Economy Working Papers 6/2023. London: London School of Economics and Political Science, 2023. https://ssrn.com/abstract=4414385.

Giuffrè, Mauro, and Dennis L. Shung. "Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy." *NPJ Digital Medicine* 6, no. 1 (2023): Article 186. https://doi.org/10.1038/s41746-023-00927-3.

Ive, Julia, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. "Generation and Evaluation of Artificial

Mental Health Records for Natural Language Processing." *NPJ Digital Medicine* 3, no. 1 (2020): 69. https://doi.org/10.1038/s41746-020-0267-x.

Jiang, Yifan, Han Chen, Murray Loew, and Hanseok Ko. "COVID-19 CT Image Synthesis with a Conditional Generative Adversarial Network." *arXiv* (preprint), July 29, 2020. https://doi.org/10.48550/arXiv.2007.14638.

Lucini, Fernando "The Real Deal About Synthetic Data." *O'Reilly*. October 20, 2021. The Real Deal About Synthetic Data

Nguyen, Thanh Tam, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. *A Survey of Machine Unlearning*. arXiv preprint arXiv:2209.02299 [cs.LG], 2024. https://arxiv.org/pdf/2209.02299

Nowok, Beata, Gillian M. Raab, and Chris Dibben. *synthpop: Bespoke Creation of Synthetic Data in R*. Vignette del pacchetto synthpop. Comprehensive R Archive Network (CRAN), 2016. synthpop: Bespoke Creation of Synthetic Data in R

Panfilo, Daniele, Alexander Boudewijn, Sebastiano Saccani, Andrea Coser, Borut Svara, Carlo Rossi Chauvenet, Ciro Antonio Mami, and Eric Medvet. "A Deep Learning-Based Pipeline for the Generation of Synthetic Tabular Data." *IEEE Access* 11 (2023). IEEE Xplore Full-Text PDF:

Patel, Brijesh, Gary Francis, Indika Wanninayake, Alan Pilgrim, and Ben Upton. *LTI Synthetic Data: Technical Report – Study 3*. Version 1.2. BAE Systems Applied Intelligence Labs. June 8, 2020. ASC_0259_Study3_FinalReport_v1_2.pdf

Patel, Preeti. "Synthetic Data." *Business Information Review* 41, no. 2 (2024): 48–52. https://doi.org/10.1177/02663821241231101.

Rajagopal, Sri, Corrine Bai, and Mark Rowland. *Synthetic Data: Facilitating Innovative Solutions*. Viewpoint. Arthur D. Little, 2024. SYNTHETIC DATA: FACILITATING INNOVATIVE SOLUTIONS

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. arXiv preprint arXiv:1602.04938v3 [cs.LG], August 9, 2016. 1602.04938

Steier, Amy, Lipika Ramaswamy, Andre Manoel, and Alexa Haushalter. *Synthetic Data Privacy Metrics*. arXiv preprint arXiv:2501.03941v1 [cs.LG], January 7, 2025. 2501.03941v1

Susser, Daniel, Daniel S. Schiff, Sara Gerke, Laura Y. Cabrera, I. Glenn Cohen, Megan Doerr, Jordan Harrod, Kristin Kostick-Quenet, Jasmine McNealy, Michelle N. Meyer, W. Nicholson Price II, and

Jennifer K. Wagner. “Synthetic Health Data: Real Ethical Promise and Peril.” *The Hastings Center Report* 54, no. 5 (2024): 8–13. https://doi.org/10.1002/hast.4911.

Tamim, James. “The Brussels Effect and the GDPR: EU Institutions as Catalysts for Global Data Protection Norms.” *European Digital Policy Institute*, June 17, 2024.

Tan, Jasper, Blake Mason, Hamid Javadi, and Richard G. Baraniuk. *Parameters or Privacy: A Provable Tradeoff Between Overparameterization and Membership Inference*. arXiv preprint arXiv:2202.01243v2 [stat.ML], November 30, 2022. 2202.01243

Trovato, Carmine Andrea, and Chiara Rauccio. “L’anonimizzazione è morta? Un’analisi dei dati sintetici come proposta per superare la dicotomia “dato personale-non personale”.” *Ciberspazio e Diritto* 23, no. 71 (2-2022): 235–259

Wang, Echo, Katrina Mott, Hongtao Zhang, Sivan Gazit, Gabriel Chodick, and Mehmet Burcu. “Validation Assessment of Privacy-Preserving Synthetic Electronic Health Record Data: Comparison of Original Versus Synthetic Data on Real-World COVID-19 Vaccine Effectiveness.” *Pharmacoepidemiology and Drug Safety* 33, no. 10 (2024): e70019. https://doi.org/10.1002/pds.70019.

Online and Journalistic Articles:

Aindo. “Enhancing Model Training with Differential Privacy.” *Aindo Blog*, March 5, 2025. Accessed September 14, 2025. https://www.aindo.com/blog/differential-privacy

Aindo. “Unlocking the Future with Simulation.” *Aindo Blog*, September 17, 2025. Accessed September 19, 2025. Unlocking the future with simulation - Aindo AI

Aindo. “What Is Europrivacy and Why Does It Matter?” *Aindo Blog*. June 23, 2025. Accessed September 19, 2025. What is Europrivacy and why does it matter? - Aindo AI

Barker, Tyson. “Europe Can’t Win Its War for Technology Sovereignty. The European Union is running in circles in pursuit of ‘digital sovereignty’.” *Foreign Policy*, January 16, 2020. Accessed April 11, 2025. https://foreignpolicy.com/2020/01/16/europe-technology-sovereignty-von-der-leyen/.

Bie, Alex, and Umar Syed. “Generating Synthetic Data with Differentially Private LLM Inference.” *Google Research Blog*, March 18, 2025. Accessed September 14, 2025 https://research.google/blog/generating-synthetic-data-with-differentially-private-llm-inference

D’Acquisto, Giuseppe. “Dati sintetici: cosa sono, le applicazioni e i rischi da gestire.” *Agenda Digitale*, May 6, 2024. Dati sintetici: cosa sono, le applicazioni e i rischi da gestire - Agenda Digitale

Del Re, Enrico. “Servono nuove tecnologie per la protezione dei dati: le sfide post-GDPR per la UE.” *Agenda Digitale*, June 19, 2023. Accessed April 15, 2025. https://www.agendadigitale.eu/sicurezza/privacy/servono-nuove-tecnologie-per-la-protezione-dei-dati-le-sfide-post-gdpr-per-la-ue/.

Grimes, Keltin, Collin Abidi, Cole Frank, and Shannon Gallagher. “3 Recommendations for Machine Unlearning Evaluation Challenges.” *Software Engineering Institute Insights (blog)*, Carnegie Mellon University, August 26, 2024. Accessed September 15, 2025. https://www.sei.cmu.edu/blog/3-recommendations-for-machine-unlearning-evaluation-challenges

J.P. Morgan. “Synthetic Data for Real Insights.” *J.P. Morgan Technology Blog*, accessed August 10, 2025. Synthetic Data for Real Insights

Keskpaik, Saskia. “Machine Unlearning.” *European Data Protection Supervisor – TechSonar*. Accessed September 14, 2025, https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/machine-unlearning_en

O’Brien, Keith, and Amanda Downie. “What Is AI in Insurance?” *IBM*, November 17, 2024. Accessed August 5, 2025. https://www.ibm.com/it-it/think/topics/ai-in-insurance

Pecan AI. “The Rise of Machine Unlearning.” *Pecan Blog*, June 25, 2024. Accessed September 15, 2025. https://www.pecan.ai/blog/the-rise-of-machine-unlearning

Ribeiro, Gonçalo. “Synthetic Data Applications In Finance.” *Forbes*, April 3, 2024. Accessed August 5, 2025. https://www.forbes.com/councils/forbestechcouncil/2024/04/03/synthetic-data-applications-in-finance/.

Tripepi, Tiziana. “Dati e intelligenza artificiale, Aindo è la startup del mese.” *InnLifes*, April 29, 2025. Accessed August 18, 2025 https://www.innlifes.com/startup/aindo-dati-startup

Wikipedia contributors. “Variational Autoencoder.” Wikipedia. Last modified August 27, 2025. Accessed September 19, 2025. Variational autoencoder - Wikipedia

Institutional Sources – (Websites and Documents):

Article 29 Data Protection Working Party. *Opinion 05/2014 on Anonymisation Techniques*. WP216. Brussels: European Commission, April 10, 2014. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Court of Justice of the European Union. *Judgment of the Court (First Chamber) of 4 September 2025, EDPS v SRB, Case C-413/23 P*. ECLI:EU:C:2025:645. EDPS v SRB (Notion de données à caractère personnel)

European Commission, *Shaping Europe's Digital Future: Commission Presents Strategies for Data and Artificial Intelligence*, press release, Brussels, February 19, 2020. Shaping Europe's digital future

European Commission. "A European Strategy for Data." *Shaping Europe's Digital Future*. Last modified April 9, 2025. Accessed April 9, 2025 https://digital-strategy.ec.europa.eu/en/policies/strategy-data.

European Commission. "Data Act Explained." *Shaping Europe's Digital Future*. Last modified January 29, 2025. Accessed April 15, 2025. https://digital-strategy.ec.europa.eu/en/factpages/data-act-explained

European Commission. "Data Act." *Shaping Europe's Digital Future*. Last modified October 10, 2024. Accessed April 5, 2025. https://digital-strategy.ec.europa.eu/en/policies/data-act

European Commission. "Data Governance Act Explained." *Shaping Europe's Digital Future*. Last modified October 11, 2024. Accessed April 5, 2025. https://digital-strategy.ec.europa.eu/en/policies/data-governance-act-explained

European Commission. "European Data Governance Act." *Shaping Europe's Digital Future*. Last modified October 10, 2024. Accessed April 9, 2025. https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

European Commission. "*Press remarks by President von der Leyen on the Commission's new strategy: Shaping Europe's Digital Future*." Bruxelles, February 19, 2020. President von der Leyen on 'Shaping Europe's Digital Future'

European Commission. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Building a European Data Economy*. COM(2017) 9 final. Brussels, January 10, 2017. EUR-Lex - 52017DC0009 - EN - EUR-Lex

European Commission. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Towards a Common European Data Space*. COM(2018) 232 final. Brussels, April 25, 2018. EUR-Lex - 52018DC0232 - IT - EUR-Lex

European Data Protection Board. *Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models*. Adopted December 17, 2024. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models | European Data Protection Board

European Union Agency for Cybersecurity (ENISA). *Data Protection Engineering: From Theory to Practice.* January 2022. Data Protection Engineering | ENISA

European Union Agency for Network and Information Security (ENISA). *Readiness Analysis for the Adoption and Evolution of Privacy Enhancing Technologies*. Version 1.0. December 2015. Readiness Analysis for the Adoption and Evolution of Privacy Enhancing Technologies

G7 Data Protection and Privacy Authorities. *Roundtable Communiqué.* 8 September 2022. Microsoft Word - G7 DPA Roundtable Communiqué 8 sep 2022 final.docx

General Court of the European Union. *Judgment of the General Court (Eighth Chamber, Extended Composition) of 26 April 2023, Single Resolution Board v European Data Protection Supervisor, Case T-557/20*. ECLI:EU:T:2023:219. SRB v EDPS

High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission, April 8, 2019. ai_hleg_ethics_guidelines_for_trustworthy_ai-en_87F84A41-A6E8-F38C-BFF661481B40077B_60419.pdf

Organisation for Economic Co-operation and Development (OECD). *Emerging Privacy-Enhancing Technologies: Current Regulatory and Policy Approaches.* OECD Digital Economy Papers, no. 3512. March 2023. Emerging privacy-enhancing technologies | OECD

Normative Sources:

European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of 13 June 2024 Establishing Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. Official Journal of the European Union L 1689, July 12, 2024. Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio, del 13 giugno 2024, che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n, 300/2008, (UE) n, 167/2013, (UE) n, 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (regolamento sull'intelligenza artificiale)Testo rilevante ai fini del SEE.

European Union. *Charter of Fundamental Rights of the European Union*. 2000/C 364/01. Official Journal of the European Communities, December 18, 2000. EUR-Lex - C:2000:364:TOC - IT - EUR-Lex

European Union. *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 (General Data Protection Regulation). Official Journal of the European Union, L 119, May 4, 2016. Regolamento - 2016/679 - IT - GDPR - EUR-Lex

European Union. *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)*. Official Journal of the European Union, L 151, June 3, 2022. Regulation - 2022/868 - IT - EUR-Lex

European Union. *Regulation (EU) 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, amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)*. Official Journal of the European Union, L 2854, December 22, 2023. Regolamento - UE - 2023/2854 - IT - EUR-Lex