Anomymization in the context of generative AI

Aligning computer science and legal standards

University of Turku
Department of Computing
Master of Science in Technology Thesis
Cyber Security Engineering
July 2025
Mélanie Romano

UNIVERSITY OF TURKU
Department of Computing

Mélanie Romano: Anomymization in the context of generative AI
Aligning computer science and legal standards

Master of Science in Technology Thesis, 93 p., 8 app. p.
Cyber Security Engineering
July 2025

This thesis investigates anonymization in the age of generative artificial intelligence (AI), with a focus on aligning technical approaches from computer science with legal standards, particularly European data protection law. Adopting a multidisciplinary framework, the study explores the evolving notion of personal data, the theoretical and practical mechanisms of anonymization, and the challenges posed by generative AI systems at various levels—including training data, models, inputs, and outputs.

Special attention is paid to the potential of synthetic data generation as a privacy-preserving technique and to differential privacy as a semantic privacy model. The work critically examines whether outputs produced by generative AI can themselves constitute personal data, and to what extent anonymization methods remain effective against modern re-identification attacks. Legal uncertainties surrounding the definition and sufficiency of anonymization, especially in light of the GDPR, the AI Act, and opinions by regulatory bodies like the EDPB, are highlighted.

Ultimately, this study contributes to bridging the gap between legal doctrine and technical realities, offering insights into the strengths, limitations, and necessary evolution of anonymization practices in data-intensive AI systems.

Keywords: anonymization, privacy, regulation, GDPR, synthetic-data, generative AI, EDBP, AI Act, Differential Privacy

Contents

Foreword

This document is the result of an internship conducted in the Security and Privacy (SPICY) team from the IRISA institute in Rennes, France. It is supported by the Cybersecurity, Data Protection and Fundamental Rights Chair hosted at Rennes University (Chaire Cybersécurité, protection des données et droits fondamentaux, Fondation Univ Rennes). Moreover, those researches have benefit from the IPoP Project (Interdisciplinary Project on Privacy, National Cyber funding program : PEPR CYBER). The internship is supervised jointly by Tristan Allard (PhD, HDR) associate professor at the University of Rennes, expert in privacy in data intensive systems, and Margo Bernelin (PhD), a fellow researcher in Law at Law and Social Changes research center (French national Center for Scientific Research/ Nantes University), expert in data protection and digital Law.

The goal of this work is to study anomymization in the context of generative AI by aligning computer science and legal standards. This document adopts a multidisciplinary perspective, situated at the intersection of cybersecurity, data protection law, and artificial intelligence. In order to accurately reflect the complexity and precision required in legal analysis, the writing style incorporates extended quotations, especially to reference legal texts, where preserving the original

wording is necessary to avoid the betrayal of the intended meaning. When official translations of French legal sources were unavailable, translations made, aiming at keeping the text's original meaning.

In parallel, certain technical sections also feature quoted material, where the original formulations offered particular clarity or conceptual significance. By adopting a multidisciplinary approach, this works seeks to accurately present complex, and sometimes unfamiliar concepts. This research started with genuine interrogations on the relationship between Privacy and AI and ended with a critical perspective on it. Such a start was crucial to uncover perhaps overlooked issues on data protection and was kept in the outline of this Master Thesis's chapters answering very direct questions that echoed both legal and computer science interests.

This production represents an effort to build bridges between disciplines and to contribute meaningfully to the ongoing dialogue on data protection, privacy and artificial intelligence through a synthesis of legal and technical analysis.

1 Introduction

Data is surrounding us from all sides and is mandatory for our everyday life. However, in the last decades the amount of data and especially personal data collected has never been this huge, bringing in its wake new threats and attacks. As those risks rose, counterattacks began to appear as well. Since information is collected to be shared, computer scientists created techniques to share only the meaningfully data and to try to mitigate attack risks. Those are known as anonymization techniques. Those techniques and collect are subject to laws and regulated. Among the techniques' multiplicity, this thesis will focus on one specific method : synthetic data. With the rise of generative models, the outputs of such models (also known as synthetic data) gained in popularity and researchers are starting to evaluate this method in order to determine whether or not this can be considered as an anonymization technique. At the same time, synthetic data might appear as a safe way to escape personal data's regulation and is starting to gain in popularity form companies as well as regulatory bodies and states.

1.1 Research questions

But before propagating this method as an anonymization technique, it is important to evaluate its potential and its possible flaws. As such, this thesis will try to bring an answer to the question : is it possible to consider synthetic data as an anonymization technique? And can generative models be considered as personal data ?

As regulation on those might sometimes lack clarity, not unknowingly but more to keep some room to manoeuvre in order to be able to evaluate every system despite the rapid changes of new technologies, it became an increasing need to a useful threshold for companies and regulators to face uncertainties toward evaluating anonymization techniques, especially in the context of synthetic data generation. Upon creating a dialogue between Computer Science and Law, this thesis aims to explore whether or not it is possible to find and fix such threshold (both by looking at existing regulation and questioning technical feasibility). And if not, to find other ways to address this issue.

The research questions this thesis is exploring are the following ones :

- What are the existing legal definitions of personal data under EU law?
- Can data generated by generative models be considered personal data under the EU law ? If so, how is it regulated?
- Can synthetic data be considered a method of anonymization under EU data protection law?

- Is it legally and technically feasible to establish a re-identification risk threshold to determine anonymization? If not, what alternative approaches exist?

1.2 Research methods

This thesis adopts an interdisciplinary and qualitative research methodology, drawing from both legal analysis and computer science. The legal dimension is explored through doctrinal research, including close examination of the GDPR, the AI Act, and regulatory guidance from bodies such as the EDPB and CNIL. Particular emphasis is placed on interpreting legal texts in their historical and regulatory context to assess how anonymization and synthetic data are legally framed. On the technical side, the study investigates anonymization mechanisms through a conceptual and theoretical analysis of privacy models and their vulnerabilities. Techniques such as k-anonymity, differential privacy, and synthetic data generation are critically assessed using academic literature and practical examples with an emphasis on differential privacy. This dual approach enables a comprehensive understanding of how anonymization is conceived, applied, and challenged in the context of generative models, and supports the thesis's goal of aligning legal standards with technical realities.

1.3 Thesis structure

The following chapter (chapter 2) introduces the concept of synthetic data and the technical mechanisms used to generate it, along with their potential privacy implications. The next one (chapter 3) explores the legal framework of anonymization

under the GDPR, the EDPB, and other key regulatory interpretations. Chapter 4 (chapter 4) examines how synthetic data aligns (or fails to align) with legal requirements for anonymization, drawing comparisons with other techniques. The final chapter (chapter 5) provides a synthesis of findings, reflects on challenges, and proposes considerations for future regulatory and technical developments.

2 Personal data and anonymization

This chapter first aims to define the notion of personal data, by drawing a comparative analysis between legal interpretations (primarily those codified in the General Data Protection Regulation (GDPR)) and the technical understanding adopted within computer science. The discussion emphasizes the distinction between direct identifiers, quasi-identifiers, and non-identifying attributes, highlighting the evolving challenges of re-identification in data-rich environments.

It then introduces the theoretical and practical dimensions of anonymization by presenting a taxonomy of anonymization techniques, including k-anonymity, differential privacy, and synthetic data generation, and evaluates their respective strengths and limitations. Particular attention is given to the inherent trade-off between privacy and data utility, as well as to the legal ambiguities surrounding the sufficiency of anonymization under current regulatory frameworks. This foundational analysis serves to contextualize the ensuing chapters, which interrogate the privacy risks and regulatory implications of generative AI systems.

2.1 What is personal data?

Personal data are everywhere, but for most people the notion remains vague. As such, it is important to explain what it is. Before interrogating the regulation or the technique, it seems interesting to start by the definition that can be found in a dictionary. Personal can refer to "relating or belonging to a single or particular person rather than to a group or an organization." or to "private or relating to someone's private life."[1]. And a data can be defined as such : "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer."[2]. However personal data's definition is different : "information held on computers that relates only to you, and that you do not want everyone to know"[3]. This is illustrating something interesting, if we trust the last definition, a concept of the usage of such data appears. Nevertheless, aggregating the two first definitions leads us to understand that personal data are pieces of information (whether facts, numbers, or other details) that relate or belong to a single individual, particularly regarding their private life or identity, and which can be collected, stored, and used for analysis or decision-making, often in electronic form. In this definition there is no notion of purpose. This purpose in the regulation is called 'purpose of processing'. Defining whether something is personal should not be processing dependant, when a user must make a choice whether or not to share its personal data, the user has to be aware of how the personal data will be processed. But in any case, the data remains personal data. The second definition is then the one standing the closest to what the reality is. But how does the regulation define personal data ?

2.1.1 For laws : a protection mechanism

The General Data Protection Regulation (GDPR) Article 4 [4] defines a personal data as :

> "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"

Recital 26 [4] of the same text further precise the notion of personal data and states that :

> "1. The principles of data protection should apply to any information regarding an identified or identifiable natural person.
> 2. 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.
> 3. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly."

The recital then states that the necessary means might be determined by ob-

jective factors such as technology available, time, cost, etc... and that anonymous data is out of the GDPR's reach.

By analysing article 4 and recital 26, we can conclude that personal data refers to the data identifying a natural living person (by opposition to a legal person that would refer to a company, a charity or a public institution), leaving out of the GDPR's scope data on a deceased individual. More interestingly the combined reading of article 4 and recital 26 [4] distinguishes two categories: "personal data" covered by the GDPR and "anonymous information" that are outside of the GDPR's realm and its obligations. Therefore, what qualifies as "anonymous information" and how to get there is crucial. In this regards recital 26 provides that the risk of identification of an individual should be appraised on an "objective basis". The idea is here not to circumvent personal data to only the one that are likely and foreseeably to be identifying an individual for instance because the "amount of time required for identification" is reasonable or/and because the technology to do so is available. However, it is also interesting to understand how technology defines a personal data as well in order to be able to draw a conclusion.

2.1.2 For computer science

Computer scientists, are discussing what a personal data and one major line of work distinguishes between three types of data : identifiers (or ID), quasi-identifiers (qID) and other data.

The identifier (or direct identifier) can be defined by : an "information that directly and uniquely identifies an individual without the need for additional information. Examples include a person's full name, social security number, passport

number, or biometric data (such as fingerprints or facial images)."[5]. In contrast, a quasi-identifier is "a data attribute that does not directly identify an individual on its own, but can potentially identify someone when combined with other quasi-identifiers or external information. Typical examples include date of birth, gender, and ZIP code."[6]. We could simplify by saying that quasi-identifiers are pieces of information that are identifying some individuals but not all of them.

The rest of the information is the information intended to be disclosed, for instance for a study about cancer patients, the information which is not an identifier, nor a quasi-identifier is : whether or not the person has cancer or which type it is. In this case this data is considered as "sensitive data". Those data are the reason for the whole database being shared in the first place.

The Table 2.1 illustrates this notion of ID and qID.

ID number	**Name**	**Date of birth**	**Place of birth**	**Sex**	**Favourite colour**
1234567	Oscar	06.12.1999	England	Male	Purple
8910111	Alice	15.04.1931	France	Female	Orange
1213141	Bob	03.02.1986	Germany	Male	Green

Table 2.1: Example of a small data table illustrating the notion of ID and qID

In this example (Table 2.1), we can say without a doubt that the ID number is an ID. However, when it comes to qID all the orange information might be considered as such since there are not thousands of babies that are born the same day, in the same place with the same sex. With enough information it becomes possible to identify one person among all the others. However, the favourite colour could be deleted from the list of ID and qID. But can it really ?

In the article "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata" Yves-Alexandre de Montjoye and al. [7] explored the re-identification

risks associated with credit card metadata. They found that with just four spatiotemporal data points, 90% of individuals could be uniquely identified in a dataset of 1.1 million people. In another article, Ana-Maria Cretu and al.[8] re-identified people from their energy consumption. Those examples highlight the fact that even data that are not at first glance personal can be used to identify individuals. [1]

The notion of qID and ID has nowadays shown its limits. If an attacker knows that Alice likes Orange then again, the favourite colour becomes a quasi-identifier. As of today, thousands of data points about people have been collected, so the question would be : how many "not relevant information" does it take to be able to identify someone? Multiple studies showed that with those "not relevant information" such as movies ratings [9], or song playlist [10] it was possible to re-identify a person by using those pieces of information with other information in external databases. Not only is it possible to re-identify a person, but also to infer new information about this person. Thus, information as innocent looking as playlist can be used to infer gender, age, demographics, personality traits and so on [10].

It would be important to now consider each and every information about an individual as, at least, a qID. Since in the paradigm of IDs and qIDs both should be protected, it became clear that all personal data must be protected.

Yet, it is possible to argue that all personal data does not bear the same criticality. This is the subject of the GDPR's Article 9(1) [4] :

[1] Another example using non-first glance identifying data is browser fingerprinting, where those innocent looking data combined all together are powerful enough to identify with quasi-certitude a user.

> "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.".

A clear line is drawn between those data and any other personal data. As such we might imagine a paradigm who aims to enforce the privacy of those highly critical data while being a bit more lenient on the protection of the other data (although lenient does not mean no protection at all). Nonetheless, defining in a database what is critical from what is not is still arduous.

In conclusion, the GDPR's definition of personal data seems adequate as technique agrees each data about an individual can and should be considered as personal. However, as stated in the text anonymous data does falls out of GDPR's reach. But what is exactly anonymisation that allows data to not allows the identification of an individual while preserving the data utility ?

2.2 What is anonymization ? Theory, Methods, and Motivations

Anonymization techniques have become crucial in order to comply with privacy regulations and reduce the risk of data breaches while still preserving the data's relevance. It could be quickly defined as the process of removing or altering personal information from data sets so that individuals cannot be identified, either directly or indirectly. It is commonly used to protect privacy when sharing or analysing data.

2.2.1 For computer science : a relative security guarantees

In computer science, anonymization history has been written on the back and forth between attacks and new privacy mitigation techniques. As such, it exists multiple different models (a formal framework that defines how personal data can be protected from unauthorized access or identification during processing or sharing). But before introducing anonymization methods, let break down the anonymization process (Figure 2.1).

On a general scale, data can be categorized into two types: record-level data and aggregate data. Record-level data contains detailed, individual-specific information and presents a higher risk of re-identification. In contrast, aggregate data consists of summarized values (such as totals or averages) which generally pose less risk but may be less suitable for detailed analysis [11]. To achieve anonymization, this data is processed through privacy algorithms (concrete computational mechanisms designed to transform the data in line with the requirements of a given

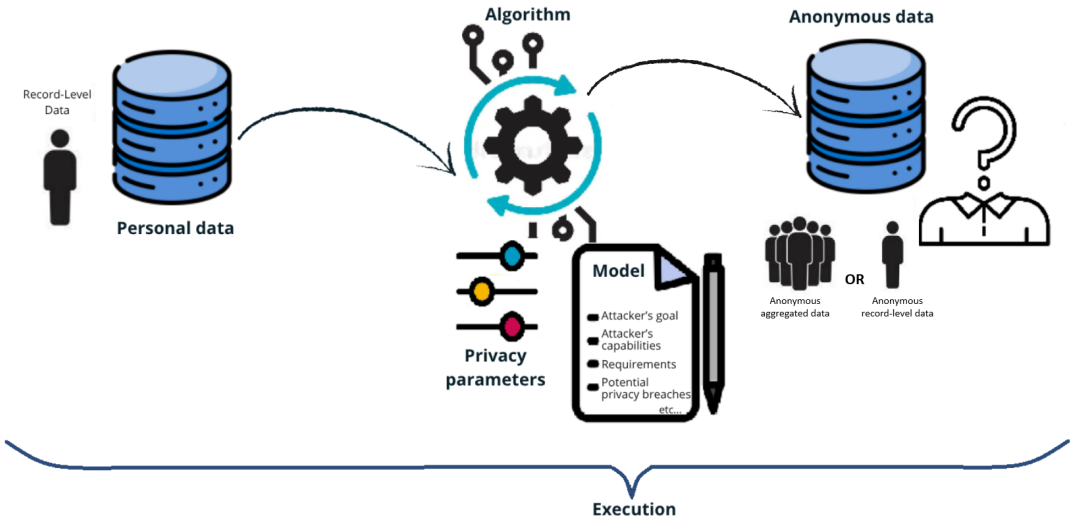

Figure 2.1: Anonymization process

privacy model).

These algorithms use techniques such as suppression, generalization, noise addition, etc.

Their behaviour is governed by privacy parameters, which quantify the level of protection and manage the trade-off between privacy and data utility (as noise for instance will render the data less relevant). These parameters are crucial for making privacy guarantees both measurable and adaptable to various scenarios. The algorithms themselves implement a privacy model, which acts as the foundational framework. This model defines what constitutes a privacy breach and specifies the conditions an anonymized dataset or an anonymization algorithm must meet in order to be considered private. It is built on assumptions about the adversary's goals and capabilities and serves as the benchmark for assessing whether a dataset is sufficiently anonymized.

Nonetheless, anonymization cannot be considered isolated from its context. Factors such as the intended purpose of the data, its nature (e.g., structured or sensitive), the target audience, the availability of auxiliary information, applicable legal frameworks and implemented mitigation strategies all significantly influence the effectiveness and appropriateness of an anonymization method. These contextual elements shape the requirements that a dataset must fulfil before it can be deemed safely anonymized.

Anonymization is a complex mechanism to put in place. This difficulty can be seen in the multitude of different mechanisms that have been created before being attacked and replaced by more effective ones. As in every field in cybersecurity, privacy is driven by the attacks. An algorithm is up to the state-of-the-art until a successful attack is discovered and the cycle starts again. In order to illustrate this never-ending game of attacks / defences, this document Appendix A retrace some of them.

Overall, in the beginning some thought replacing IDs with aliases would suffice, however this technique known as pseudonymisation (and nowadays not enough the meet the GDPR's anonymization need meaning not enough to avoid its reach), quickly shown flaws. In response, some techniques such as : k-anonymity, l-diversity, etc. were created but again, those techniques were proven flawed under certain circumstances. Moreover, those techniques would work only on record-level data. As the need of anonymization for aggregated data grew in parallel, privacy models such as Differential privacy (DP) were also created.

2.2.1.1 Anonymization in the context of DP and synthetic data

In 2006 Cynthia Dwork and al. came with an idea : Differential Privacy (DP) [12]. This model is designed to protect individuals' private information when datasets are analysed or shared. The core principle is that the presence or absence of a single individual's data (say, John's) should not significantly affect the outcome of any analysis. In other words, whether or not John's data is included, the results should appear nearly the same, making it extremely difficult to determine if he was part of the dataset at all.

DP's context is the result of interactive queries (aggregated data) made on a static database. To achieve this security, DP introduces carefully calibrated random noise to the queries' answers, ensuring individual contributions remain hidden while still allowing useful insights from the aggregate data.

Dwork et al.[12] formalized this idea using the Laplace mechanism, which adds noise scaled to the sensitivity (the importance of the information asked) of the function being computed. The result is a statistical cloak—effectively hiding the "tree" in a forest of plausible alternatives. While Differential Privacy provides robust theoretical guarantees, it is most effective on aggregate queries (such as counts or averages) and may be vulnerable to sophisticated attacks if misapplied or used on high-dimensional data.

While DP is a powerful privacy model, it is not mutually exclusive with other privacy-preserving techniques. In practice, these models can be applied with different algorithms. One such privacy algorithm, which is gaining increasing attention, is the generation of synthetic data. Those data can satisfy the requirements of DP or other privacy models, offering a flexible and privacy-conscious alternative

to using real data directly.

Synthetic data generation would have the advantage to work both on aggregated data and record-level data. Such data does not originate from direct observations of the real world but is instead artificially created one to mimic the statistical properties of real data, having the advantage of not being actual personal data [13]. For instance, while the fairy tales collected by the Brothers Grimm were written by humans and based on real oral traditions, a story generated by an AI in the style of the Grimm's Brothers could be considered synthetic. It imitates the tone and structure of real stories but is entirely fabricated.

Synthetic data is often produced by artificial intelligence systems. These can include large language models (LLMs), probabilistic graphical models, or Bayesian networks, depending on the type of data and the desired outcome. Broadly speaking, synthetic data can be divided into three types: partially synthetic, fully synthetic, and hybrid.

Partially synthetic data refers to datasets where only certain elements, usually sensitive ones, such as identifiers or quasi-identifiers—have been replaced. These replacements are generated to follow the same probability distributions as the original values, thus preserving some of the dataset's structure while attempting to hide specific identities. However, this method is fundamentally flawed. Because only part of the data is altered, it still carries a risk of disclosure, attackers may infer sensitive information from the unmodified parts. Due to these inherent limitations, this type of synthetic data will not be discussed further in the following sections.

Fully synthetic data, by contrast, is generated entirely by a model that has been trained on real data, which may or may not have undergone privacy-preserving

transformations. This model learns the statistical patterns and relationships within the original dataset and then produces entirely new data points that replicate those patterns. These synthetic records are not tied to any actual individuals, and they omit all identifying parameters. This approach aligns with the idea that synthetic data is an artificial reproduction of real data it aims to represent the properties of the original dataset faithfully, but without directly copying it.

Finally, there is hybrid synthetic data. In this case, fully synthetic data is combined with real data that has been subject to privacy-enhancing techniques. The goal is to find a balance between maintaining the utility of the dataset and ensuring the protection of personal information. This approach leverages the strengths of both real and synthetic data to create a dataset that is both useful and secure.

In essence, synthetic data is a constructed imitation of real-world data, designed to preserve the useful characteristics of the original while omitting the identifying elements that pose privacy risks. This document is focused only on fully synthetic data.

2.2.1.2 Privacy and utility trade-off (and other concerns)

In the majority of those methods (l-diversity, k-anonymity, DP) some privacy parameters appear (l, k ε). Those parameters represent the level of privacy that are intended to be reached by each method and are settled mostly depending on the size of the dataset, the privacy that is expected and the utility that still needs to remain. A dataset where all data are indistinguishable

The trade-off between privacy and utility is a central concern in privacy.

However, privacy and especially privacy achieved with synthetic data generation, might have to face other challenges. Several other dimensions might play

a role in evaluating the effectiveness, fairness, and trustworthiness of approaches such as synthetic data generation and differential privacy. One such dimension is fairness and bias mitigation. AIs used to generate synthetic data can reproduce or even amplify biases present in the original dataset, so it is important to assess whether group-level fairness (e.g., across gender or ethnicity) and individual-level fairness are maintained.

Transparency and accountability are also essential. Users and regulators must be able to understand how the data was generated by the AI, how privacy is protected, and whether the methods are auditable and explainable. Legal and ethical compliance is equally important, especially under frameworks like the GDPR or the EU AI Act. Synthetic data should respect principles like data minimization, purpose limitation, and user autonomy (even if it is technically not considered personal data). Robustness against attacks is another crucial factor. Privacy models must withstand inference and re-identification attacks, such as membership inference or model inversion, which can compromise the privacy of individuals. Social acceptability and trust also play a significant role, systems must be perceived as ethical and respectful of data subjects' rights and expectations, with clear communication about data use. Lastly, scalability and cost-effectiveness must be considered, as methods should be computationally feasible and adaptable to real-world data environments without becoming overly complex or expensive to deploy. These seven dimensions : privacy, utility, fairness, fidelity, transparency, robustness, and trust are metrics that can be used, other than utility, when talking about generative AI and their outputs from a privacy point of view [14] [15].

2.2.2 For laws : between myth and reality

Anonymization holds a privileged position in both European and French legal frameworks as a tool to facilitate the free circulation of data, especially sensitive personal data, without the constraints of data protection regulations. Since there are numerous different texts regulating anonymization, AI and personal data, the Table 2.2 recapitulates the main ones cited in this document.

Under the GDPR, anonymized data is no longer considered "personal data" and thus falls outside the scope of the regulation (Recital 26) [16]. This has promoted anonymisation as a lawful strategy for data sharing and innovation through Europe.

In France, this view was institutionalised for the sharing of data produced within State institutions such as agencies, government departments, local authorities, courts or even hospitals [17]. In the healthcare sector, anonymisation as a data sharing techniques for 'open health data' was introduced by the 2016 Health Act [18], which created a centralised health data base (the Système National des Données de Santé (SNDS)) and enabled access to it firstly on the basis on data anonymisation presuming it would be possible [19]. Similarly, the European Health Data Space (EHDS) encourages anonymisation for broader secondary use across Member States [20].

While anonymisation renders data non-personal under the GDPR, the act of anonymizing remains a processing operation and is therefore subject to the regulation [2].

[2]See Rapport du Conseil d'État, *La donnée au service de l'intérêt général*, 2021.

2.2.2.1 The crucial 2014 WP 29 opinion on anonymisation

Bridging the gap between practices and the Law, the article 29 Working group party (the ancestor of the European Data Protection Board, gathering EU's national data protection authority) issued a guidance on what anonymisation means with regard to the 1995 Data Protection Directive [21] (the GDPR's predecessor). It defined anonymisation as "the result of processing personal data in order to irreversibly prevent identification" and outlined three cumulative criteria: impossibility to single out, to link, and to infer[3].

The opinion discussed several anonymisation techniques, including k-anonymity, but noted that no method guarantees full anonymity under their three criteria. It advocated a contextual, risk-based approach, an argument that has gained weight over time.

However, the 2014 opinion fell short as research showed that methods like k-anonymity are often insufficient for high-dimensional or sensitive datasets. Since the original publication of k-anonymity in 2006, its vulnerabilities have been well-documented, yet the 2014 WP29 opinion was still promoting it as viable.

National data protection authorities such as the French CNIL tried to offer some guidance about anonymization techniques by publishing practical fiches. Recently, the CNIL ruled in its 2024 Cegedim decision that relying solely on k-anonymity to anonymize health records was inadequate due to modern re-identification risks [24].

Even recent legal texts such as the Data Governance Act (DGA) [25] acknowledges these limitations. Recognising the difficulty to obtain anonymous data sets,

[3]Those metrics are however criticized [22] [23].

this text introduced a new legal category of data that even anonymized remains at risk of re-identification when transferred outside the EU the: "non-personal but highly sensitive data" which was latter pick up by the European Health Data Space (EHDS) 2025 text [26].

In light of these developments and its critics [27], the European Data Protection Board (EDPB) is preparing a new opinion, expected in 2025. The 2024 preliminary draft suggests a distancing from static criteria (like k-anonymity) toward contextual and probabilistic assessments, aligned with differential privacy principles. This upcoming guidance addresses synthetic data generation and machine learning-based anonymisation, signalling a shift to model-based risk evaluations that better reflect modern re-identification capabilities.

2.2.2.2 Misconceptions Around Anonymization

Anonymization has long been viewed as a sufficient safeguard for privacy. This belief fostered the expectation that any dataset could, through technical means, be made non-personal and publicly shared. The 2016 French law [18] reflects this assumption. However, technical advances and legal critiques have since challenged this belief.

This legal optimism has been criticized as the 'myth of anonymisation', the idea that technical transformations alone can eliminate re-identification risk. Yet, the GDPR does not define "anonymisation", only pseudonymisation (Art. 4(5)) [4]. As a result, anonymisation is inferred by contrast to the definition of personal data, requiring interpretation on a case-by-case basis. But a rough approach would be to consider them as data that cannot be linked to an identified or identifiable person, considering "all means reasonably likely to be used". This legal vagueness

provides flexibility for the judges and organisations to define the anonymized data depending on the context. But on the other hand, it creates uncertainty for the companies leaving them without clear guidance to ensure compliance, particularly in contexts involving complex or sensitive data. While anonymization is seen, quite mistakenly, as a sound and always effective privacy preserving technique for the sharing of even sensitive data by lawmakers, what count as anonymization is quite left untouched. In this regard, no national law not the GDPR defines anonymisation, creating legal uncertainty around it and the need for guidance.

In conclusion, it is possible to state that synthetic data does not fall under the GDPR as this technique, if used to achieve anonymization, might be considered enough to protect the privacy of such data. However, when talking about synthetic data generation, it is mandatory to interrogate the texts overseeing the generative AIs as they are generated by such systems. Article 50(2) of the AI Act requires that AI-generated synthetic content (e.g., text, audio, video) be clearly marked as such in a machine-readable format [28]. Exemptions apply in limited contexts, such as law enforcement or formatting tasks. The AI Act does not clarify whether synthetic data should be considered personal, non-personal, or "highly sensitive" data, leaning on the GDPR to do the distinction between those. Nor does this act specify whether anonymisation must be applied to generative AI outputs trained on personal data. Currently, synthetic data exists in a legal grey zone. Some regulators treat it as non-personal, while others argue that if derived from personal data using weak models, it may still lead to leakage. While synthetic data is often proposed as a solution to GDPR constraints, some legal scholars warn this trend may serve as a regulatory bypass, especially in sensitive sectors like health [29] [13]. This evolving legal landscape calls for updated regulatory guidance (especially

from the EDPB) to clarify the status of anonymisation and synthetic data within modern data ecosystems.

Text name	**Date**	**Source**	**Subject**	**Type of law**
European Health Data Space (EHDS) [26]	2025	EU	Health data, personal data, data sharing, access, sharing, secondary use, and patient rights.	Hard law
General Data Protection Regulation (GDPR) [4]	2016 (before AI Act)	EU	Personal data (protection), privacy, fundamental rights, obligations, transfers, and enforcement.	Hard
2014 WP29 (Opinion 05/2014 on Anonymisation Techniques) [21] *under revision, awaited in 2025*	2014 (before AI Act)	EU	Anonymisation, personal data, privacy : guidance on anonymisation and re-identification risk	Soft law
Data Governance Act (DGA) [25]	2022 (before AI Act)	EU	Data sharing (via trusted intermediaries), personal data, public sector data	Hard law
2024 EDPB Guidelines on AI [30]	2024 (before AI Act but very close)	EU	Personal data, guidance clarifying GDPR compliance for AI, anonymisation, and complex processing.	Soft law
CNIL Fiches (Guidelines and Reference Frameworks) [31]	Since 2024 (after AI Act)	France	Personal data, privacy, anonymisation, AI	Soft law
AI Act [28]	2024	EU	Artificial Intelligence, fundamental rights, safety, regulating AI with risk-based obligations for providers, users, and high-risk systems.	Hard law
LIL (Loi Informatique et Libertés) [32]	1978, amended 2018 (before AI Act)	France	Personal data, privacy, implementing GDPR with national specifics	Hard law
Health System Modernization Act (Loi de modernisation de notre système de santé) [18]	2016 (before AI Act)	France	Health data, health system reform, health data reuse, reformed governance for health data access.	Hard law

Table 2.2: Index of legal documents mentioned above and below

This chapter explored the foundational concepts of personal data and anonymization, both from technical and legal perspectives. Personal data encompasses any information that can be linked directly or indirectly to an individual, making its protection a central concern in both regulatory and technical domains. Anonymization, as we have seen, is not a binary state but a nuanced process, driven by privacy models, algorithms, and parameters that mediate the delicate balance between data utility and individual privacy. It goal however remains the same : prevent identification of personal data.

However, as data practices evolve, particularly with the rise of synthetic data generation, the boundaries between what is considered personal and non-personal become increasingly blurred. This domain of synthetic data and artificial intelligence challenge traditional definitions of personal data. Moreover, those techniques might answer some questions and concerns while bringing new challenges. Some scholars already highlight the fact that a proper regulation on those techniques is needed [13] [27]. This is bringing new questioning such as : Is synthetic data personal ? And more broadly, is AI personal data ? Are those techniques enough to guarantee privacy ? Such questions are crucial nowadays as more and more people use generative AI [33] and as the regulatory texts provoked some debates as well [30] (the EDPB's new version is especially awaited in the near future). Those interrogations require multi-disciplinary perspectives in order to align computer science requirements and legal standards.

The next chapters, will critically examine the intersection of anonymity, synthetic data, and AI. Questioning where personal data truly ends and whether AI-generated content or decision systems themselves can be considered personal data.

3 Generative AI and personal data

This chapter explores the interplay between generative models and personal data, through both legal interpretation and technical scrutiny. It begins by examining the regulatory definitions provided under the European Union's AI Act, introducing generative models as systems capable of producing synthetic content (such as text, audio, or images) that may or may not reflect underlying personal data. Particular attention is paid to the legal uncertainty surrounding such outputs, especially when models have been trained on datasets containing identifiable or sensitive information.

It then turns to the architecture of generative models, distinguishing between statistical approaches and machine learning-based systems. A typology of known privacy attacks is presented, including membership inference, model inversion, data extraction, and other adversarial techniques that may lead to the disclosure of training data. These technical risks are situated within broader concerns about the potential re-identification from synthetic outputs and the adequacy of current privacy-preserving mechanisms. Ultimately, this section contributes to understanding how generative AI challenges existing definitions of personal data and underscores the limitations of both technical safeguards and legal frameworks.

3.1 Generative AI : definition and privacy challenges

3.1.1 A clear legal definition

The recital 12 of the same text explicit the notion by stating

> "the definition should be based on key characteristics of AI systems that distinguish it from simpler traditional software systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations. A key characteristic of AI systems is their capability to infer. This capability to infer refers to the process of obtaining the outputs, such as predictions, content, recommendations, or decisions, which can influence physical and virtual environments, and to a capability of AI systems to derive models or algorithms, or both, from inputs or data. The techniques that enable inference while building an AI system include machine learning approaches that learn from data how to achieve certain objectives, and logic- and knowledge-based approaches that infer from encoded knowledge or symbolic representation of the task to be solved. The capacity of an AI system to infer transcends basic data processing by enabling learning, reasoning or modelling. The term 'machine-based' refers to the fact that AI systems run on machines. The reference to explicit or implicit objectives underscores that AI systems can operate according to explicit defined objectives or to implicit

> objectives. The objectives of the AI system may be different from the intended purpose of the AI system in a specific context. For the purposes of this Regulation, environments should be understood to be the contexts in which the AI systems operate, whereas outputs generated by the AI system reflect different functions performed by AI systems and include predictions, content, recommendations or decisions. AI systems are designed to operate with varying levels of autonomy, meaning that they have some degree of independence of actions from human involvement and of capabilities to operate without human intervention. The adaptiveness that an AI system could exhibit after deployment, refers to self-learning capabilities, allowing the system to change while in use. AI systems can be used on a stand-alone basis or as a component of a product, irrespective of whether the system is physically integrated into the product (embedded) or serves the functionality of the product without being integrated therein (non-embedded)".

This notion of AI system is important and encompasses all AI systems from LLMs (such as GPT) to classifiers AI (which can for instance classify the importance of emails, images, etc...). Within those AI systems. However, the AI Act does not define what a generative one is.

In its Article 50, the AI Act is imposing further obligations upon AI systems that generate synthetic content (which are generally called generative AIs). This article requires their providers to be transparent about their systems by indicating to its users that the output is a synthetic content. In this sense, article 50(2) provides that "AI systems, including general-purpose AI systems, generating syn-

thetic audio, image, video or text content [...]". However, it is important to note that in this Article, generative AIs can generate only synthetic audio, image, video or text content. It is possible to generate other types of contents such as datasets, code, statistics, positions, consumption habits and so on. The article never mentions those other types of generated data. In the french translation of the article, however, it is written *"systèmes d'IA, y compris de systèmes d'IA à usage général, qui génèrent des contenus de synthèse de type audio, image, vidéo ou texte [...]"* translating into "AI systems, including general-purpose AI systems, generating synthetic content as audio, image, video or text content [...]". In this text, the list of possible generated outputs is given as an example, defining in the generative AI, all AI generating synthetic content. This translation is closer to the possible intended meaning of the text.

This thesis will define a generative AI system (abbreviated as generative AI or simply called AI) as an AI system that creates synthetic data that differs from collected data, also referred to as 'real data' as it is an artificial reproduction and must faithfully represent its properties. A generative AI will be separated into four different phases : the training dataset, the model, the input (in some cases) and the output. The Figure 3.1 illustrates the sequence of those phases.

The generative AI will be trained on one or more chosen data type (might it be images, videos, text, geographic positions, etc...), those data will be treated and cleaned in order to erase abnormalities or biases for example. This is called the training dataset. The AI will then be trained on this dataset (and fine-tuned at the end of the training) to create a model able to perform the task given to the AI. The model is the "core" of the AI. The inputs of an AI (this especially concern Large Language Models (LLMs)) is the way the user communicates with the AI

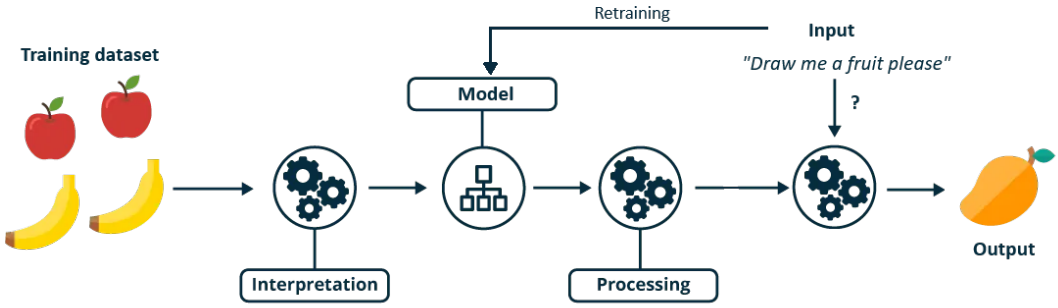

Figure 3.1: The four AI phases : the dataset, the model, the input and the output

model. The outputs of an AI (or synthetic data) are what the AI is producing. Often, the output data type is very similar to the training data type.

3.1.2 Different types of generative AI models

Before delving into the different parts of an AI system, it is important to understand the different types of generative AI models that exists, as different model might not face the same restrictions or problematic. In the world of generative models, two large families coexist: statistical-based approaches and machine learning-based approaches. Both aim to generate new data, but they rely on different philosophies. The Figure 3.2 illustrates those two categories of AIs.

The statistical-based approaches rely on explicit probabilistic models. These models attempt to describe the data distribution using predefined mathematical formulas. For instance, Gaussian Mixture Models (GMMs) [34], Bayesian Net-

works [35] or Latent Dirichlet Allocation (LDA) [36] are statistical-based AIs. In these models, the generative process is typically defined by a set of parameters that control how data is sampled. For example, in a GMM, data is generated by first selecting one of several Gaussian distributions and then sampling from it. These approaches have the advantage of being interpretable and mathematically well-defined. However, they struggle with complex, high-dimensional data (like images or text), where defining an explicit distribution becomes difficult.

On the other hand, machine-learning approaches (most especially deep learning-based models) do not require explicit distribution formulas. Instead, they learn the data distribution directly from the data itself, often using neural networks. Some examples are GANs (Generative Adversarial Networks) [37], VAEs (Variational Autoencoders) [38], Diffusion Models [39] or Large Language Models (LLMs) [40]. These models are highly flexible and powerful for modelling complex, high-dimensional data such as natural images, videos, or natural language. However, they often require vast amounts of data and computing power, and their internal representations are less interpretable than classical statistical-based models. Furthermore, since they learn from the data distribution directly, they are retaining such distribution, making them vulnerable to privacy attacks aiming at getting such data.

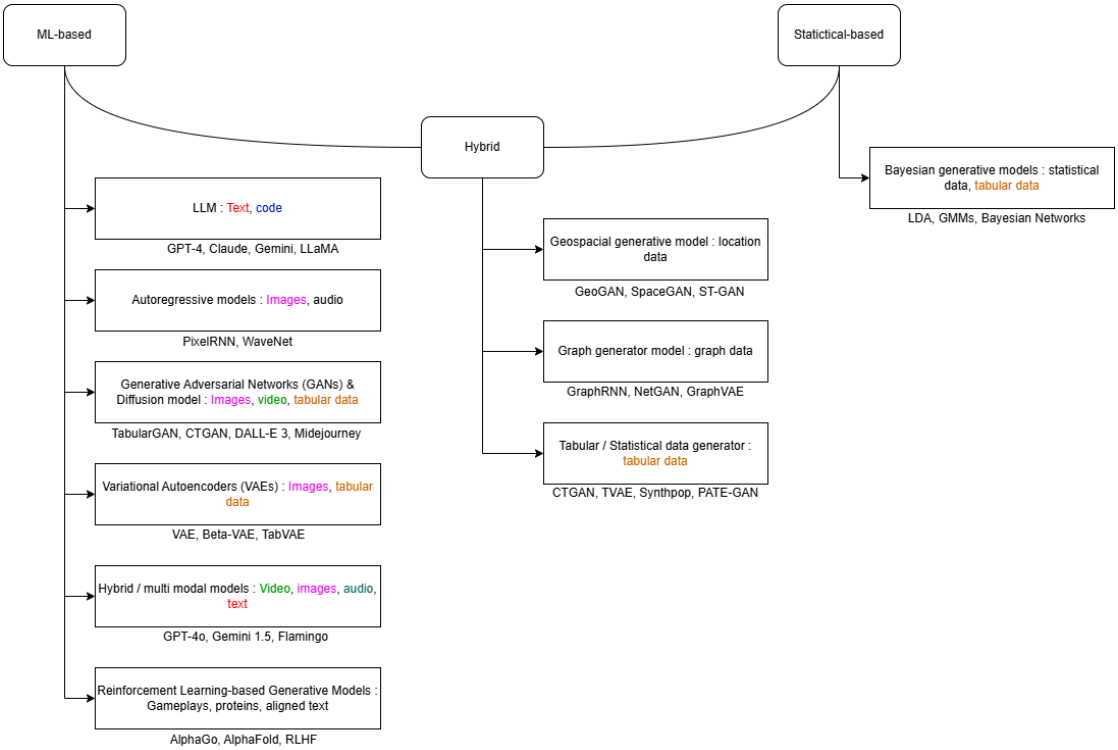

Figure 3.2: An overview of well-known generative AI models (non-subjective as it exists many more)

3.2 Privacy attacks on generative AI

3.2.1 The attacker

The nature of an attacker targeting a generative AI model varies depending on whether privacy-preserving techniques have been implemented. In the absence of such measures, launching an attack requires a solid understanding of AI systems. Conversely, when privacy protections (such as DP) are in place, additional expertise in both AI and the underlying privacy mechanisms becomes essential. Other factors influencing an attacker's capability include the level of access to the model

(e.g., white-box, grey-box, or black-box) and prior knowledge of the input data and its format.

Interestingly, computational power is not a decisive factor in privacy-related attacks. On the other hand, the most dangerous competence of an attacker is its ability the find knowledge. Two dimensions of expertise are particularly relevant. The first involves familiarity with the data—either the training dataset in full or in part, possibly obtained from other sources or via social engineering. The second relates to technical understanding: insights into how the AI system operates and how privacy measures are deployed, including their potential weaknesses. Privacy techniques are therefore designed to complicate the attack process, ideally discouraging attempts altogether. These mechanisms aim to increase the time, effort, and expertise required to yield actionable results, thus raising the likelihood that an attacker will abandon their target in favor of a less protected one.

Attackers differ in their expertise, objectives, and available resources. Understanding who might be attempting to compromise the system is essential for selecting appropriate privacy thresholds. Misidentifying the potential adversary may result in inadequate safeguards and increased exposure to data leakage. The Netflix re-identification case illustrates this risk: developers assumed a low-risk attacker without access to auxiliary data and relied on basic anonymization techniques, which proved insufficient when more capable adversaries exploited hidden vulnerabilities [9]. Similarly, a 2024 study titled "A False Sense of Privacy" revealed that 74% of private content remained inferable after applying popular personal data removal techniques, including synthetic data methods that lacked differential privacy safeguards [41].

In cybersecurity, attacker classification is standard practice, as adversaries vary

widely in motivation, persistence, and competence. Before assessing threats or selecting countermeasures, it is crucial to ask: Who might be the attacker ? What does he know?

Attackers can be broadly categorized into several profiles: casual users exploiting known vulnerabilities, skilled professionals, organized groups (including companies, hacktivists, or criminal networks), and state-affiliated entities. Each category carries distinct risks in terms of resources, goals, and persistence. For instance, while an individual or small team may be discouraged by complexity or resource constraints, well-funded organizations or nation-state actors may persist regardless of these barriers.

Motivations also vary widely, from gaining recognition through public disclosures, to extracting insights for future attacks, monetizing stolen information, or targeting individuals based on sensitive attributes such as political views, religion, or sexual orientation. Once the attacker's profile is understood, it becomes possible to assess which specific threats a given AI system may face and implement appropriate safeguards accordingly.

3.2.2 Membership Inference Attacks

The most known possible attack on AI systems is Membership Inference Attacks (MIAs). In this case, the goal is to determine whether a specific data record was included in the model's training dataset. By making queries with a known data point to the model, the attacker will observe the confidence scores or loss. Overfitted models[1], often behave differently on seen data compared to unseen data.

[1]Models that learned the training data too closely, including its noise and outliers, resulting in poor generalization to new, unseen data.

A good illustration of this is imagining an hospital setting, if an attacker suspects Alice was in a diabetes study, they can query the model with Alice's medical data. If the model returns high confidence, it may suggest Alice's record was in the training set.

3.2.3 Model Inversion Attacks

The aim of this attack is to reconstruct input features (even sensitive ones) by leveraging the model's outputs. The attacker will iteratively query the model and adjust inputs to maximize output confidence for a given class, effectively reverse-engineering inputs. For instance, in facial recognition systems, attackers have reconstructed approximate face images of people used during training. This might lead to partial or full recovery of private attributes like health status, facial features, or biometric data.

3.2.4 Training Data Extraction Attacks

In this context the attacker will try to extract entire examples or text sequences from a trained model. LLMs (like GPT or BERT) trained on large corpora may "memorize" rare or unique phrases. By prompting carefully, attackers can retrieve some of the training data. In 2020, Carlini and al. showed that GPT-2 could output real email addresses and names from its training data using targeted prompts [42].

3.2.5 Property Inference Attacks

In order to infer global properties of the training data (not specific records), the attacker will use multiple shadow models and analyse patterns in model behaviour to deduce facts like "Was the dataset skewed toward a specific ethnicity?" In this example the attacker is inferring that the facial recognition model was primarily trained on light-skinned individuals. This attack can compromise statistical privacy, reveal dataset bias or confidential population traits.

3.2.6 Feature Inference Attacks

Last, the attacker can try to infer missing or hidden features of an individual's record based on known features and model outputs. If a model takes many inputs (e.g., age, zip, diagnosis), and the attacker knows some of them, he can infer the rest. It might be possible to guess a patient's HIV status from age, ZIP code, and symptoms using a medical prediction model. This attack can allow the attacker to reconstruct sensitive or protected attributes even if they are not directly exposed.

Table 3.1 synthesizes the different attacks explained above.

Attack name	**Technical knowledge**	**Expertise (knowledge on the AI)**	**What the attacker learns**	**Prerequisites** [2]
Membership Inference (MIA) [43] [44]	AI	Some knowledge of dataset characteristics and the dataset itself	Whether or not a specific data point was part of training set	Black-box or white-box access; ability to query the model
Model Inversion [45] [46]	AI	Some knowledge of model architecture or dataset (how it was constructed)	Reconstruct features of training data	White-box or black-box access; multiple queries
Training Data Extraction [47] [48] [3]	AI and prompting	Deep understanding of model behavior/data	Actual training samples (especially in language models)	White-box (preferred) or black-box access; massive query volume
Property Inference [49] [50]	AI and shadow models	Moderate knowledge of training dataset	Properties/statistics about the dataset (biases, confidential population traits, etc...)	White-box or black-box access; access to some model outputs, ability to query the model
Feature Inference [51] [52]	AI	Parts of the dataset (some of the attributes)	the unknown attributes	black or white box, ability to query the model

Table 3.1: Index of the different attacks on generative AI models

2. A white-box attack assumes the attacker has full access to the internal workings of a model or system. In contrast, a black-box attack assumes the attacker can only interact with the system by observing its inputs and outputs, without knowing anything about how it works internally.

3. This attack is working on all AI type but especially on LLMs

4 Privacy challenges

As generative AI systems become more prevalent and sophisticated, they increasingly intersect with concerns around data protection and privacy. While previous chapters explored the foundational concepts of personal data, the mechanisms of anonymization, and the nature of generative AI, this chapter delves deeper into the practical challenges that arise at each stage of an AI system's lifecycle. Specifically, it examines how datasets, models, inputs, and outputs each pose unique privacy risks (despite, or sometimes because of, applied anonymization techniques).

By analysing real-world attack scenarios and technical vulnerabilities, this chapter aims to critically assess the extent to which synthetic data generation can effectively safeguard personal information. It also interrogates whether current anonymization practices are sufficient to prevent re-identification or attribute inference in the face of modern adversarial capabilities. Special attention is paid to differential privacy, model perturbation, and dataset sanitization as methods for mitigating these risks.

4.1 Privacy challenges : generative AI dataset

> "A dataset in machine learning and artificial intelligence refers to a collection of data that is used to train and test algorithms and models. These datasets[...] provide the necessary input and output data for the algorithms to learn from.
>
> Datasets of all kinds, including both structured and unstructured data, can be employed in machine learning and AI. Data that has been organized in a certain manner, such as a spreadsheet or database table, is referred to as structured data. [..] Unstructured data, on the other hand, describes the information that isn't set out in a particular format, such as text or images. [...]
>
> The quality and size of a dataset can also impact the performance of machine learning and AI systems. A dataset that is too small may not be representative of the problem that the system is trying to solve and may result in poor performance. On the other hand, a dataset that is too large may be difficult to process and may require additional resources, such as computing power and storage."[53].

In most of the cases, if the dataset has not been anonymized, the data contained inside can be considered as personal data.

4.1.1 Non-anonymized dataset

Such a non-anonymized dataset, falls under the scope of the GDPR, triggering a set of complex obligations for data processors developing or deploying AI systems.

For example, they must comply with strict data retention limits and define specific, explicit purposes for processing. This can be particularly challenging when the same dataset is used to train or operate multiple AI systems, each potentially serving different purposes, thus requiring a clear identification of multiple processing purposes depending on the specific deployment context. In order to avoid this constraint, data processors often remain very vague as of the usage of the collected data, allowing them to use them in multiple AIs if needed.

Moreover, the GDPR [4] creates some rights for the data subject (the person) : the Right to Information[1] (Art 12-14), Right of Access[2] (Art 15), Right to Data Portability[3] (Art 20), Right to Object[4] (Art 21), Right to Rectification[5] (Art 16), Right to Erasure ("Right to be Forgotten")[6] (Art 17), Rights Related to Automated Decision-Making and Profiling[7] (Art 22). Those 3 last rights (Right to Rectification, Right to Erasure, Rights Related to Automated Decision-Making and Profiling) are bringing challenges in practice.

Concerning the right to rectification and erasure, the data subject can ask for its information to be rectified, however as it will be explained in the next section about the AI model, once the model has retained information a retraining

[1]The right to be informed about how personal data is collected, used, and processed, including the purposes and legal basis.

[2]The ability to request access to personal data held by a controller, along with details about how and why it is being processed.

[3]Possibility to retrieve personal data in a machine-readable format and to transfer it to another controller.

[4]Individuals can object to the processing of their personal data, especially for direct marketing or when processing is based on public interest or legitimate interests.

[5]Individuals can request correction of inaccurate or incomplete personal data concerning them.

[6]Individuals can request the deletion of their personal data under certain conditions, such as when it is no longer necessary or processed unlawfully.

[7]Individuals have the right not to be subject to decisions based solely on automated processing, including profiling, that significantly affect them, with some exceptions.

is necessary in order to modify or erase a data. But in the case where the dataset has been suppressed because the AI did not need further training this retraining is impossible. Moreover, re-training an AI has a significant cost both in time and resources. Regarding the Rights Related to Automated Decision-Making and Profiling it means a certain transparency of the model, but this is hard to achieve.

4.1.2 Anonymized dataset

Once a dataset is fully anonymized, it is no longer subject to the General Data Protection Regulation (GDPR). However, before applying any anonymization technique, it is crucial to understand the nature of the data in question.

Data can take many forms, including tabular data (structured as rows and columns, such as in Excel files), textual data (unstructured written content), graphical data (e.g. social networks or survey diagrams), and semi-structured data (formats like XML or JSON). Each type poses unique challenges in terms of privacy protection and therefore requires tailored anonymization strategies.

A commonly used distinction in privacy research is record-level data and aggregate data. Record-level data, also known as microdata or individual-level data, contains information about specific individuals, often with multiple attributes per entry. Because of its granularity, it is particularly vulnerable to privacy breaches, especially linkage attacks, where an adversary can re-identify individuals by combining quasi-identifiers with external datasets. To mitigate such risks, syntactic privacy models have been developed. Those privacy methods are a family of techniques that protect privacy by modifying the structure of the data, often through generalization or suppression, to secure identifiable information. These methods

aim to make individuals indistinguishable from others in the dataset. Well-known examples include k-anonymity and l-diversity, which ensure that each individual is hidden within a group of similar entries. These approaches are well-suited for structured, tabular datasets. However, they face significant limitations in more complex or high-dimensional datasets, where maintaining both privacy and data utility becomes difficult (for instance in texts). Furthermore, syntactic methods are vulnerable to several types of attacks, such as homogeneity and minimality attacks, and often fail to maintain privacy guarantees when datasets are combined or queried repeatedly. They are also generally ineffective for textual data, where identifying sensitive elements is far more ambiguous.

Aggregate data, by contrast, summarizes information across groups rather than focusing on individuals. It includes statistical outputs like averages, frequencies, regression coefficients, or the results of trained machine learning models. While aggregate data may seem less risky, it is not inherently anonymous. Research has shown that it is susceptible to a variety of inference attacks, including membership inference (determining whether an individual's data was included in the analysis), attribute inference (predicting unknown attributes about individuals), and reconstruction attacks (rebuilding parts of the original dataset from summary statistics).

To address these risks, semantic privacy provides a more robust and theoretically grounded approach. Unlike syntactic methods, semantic privacy does not rely on the format or structure of the data. Instead, it aims to limit what an attacker can learn from the data, regardless of what auxiliary information they possess. Differential privacy, the most widely used model in this category, ensures that the inclusion or exclusion of any single individual has a minimal impact on the

overall output of a data analysis. This is achieved by adding mathematically calibrated noise to query results or model parameters. DP is particularly well-suited for statistical data analysis and machine learning applications, where maintaining the privacy of individuals while extracting useful aggregate insights is critical. However, it is generally not applicable to unstructured data like raw text, where defining and applying such protections is much more complex.

In summary, the choice between syntactic and semantic privacy approaches depends largely on the type of data and the intended use. Record-level data typically requires syntactic protections, though these are increasingly limited in complex data environments. Aggregate data, while more abstract, still demands strong safeguards such as DP to prevent indirect disclosures. Therefore, a careful assessment of the data type and privacy risks is essential before selecting an appropriate anonymization strategy.

This thesis is principally focused on DP, since this privacy scheme is stronger against attacks. However, in order to apply DP on an AI, the dataset is not suitable. The DP algorithm is then applied mainly on the model. It is still possible in theory to apply local DP on the training dataset (the Figure 4.1 is illustrating its functioning). In the local model, each data is locally perturbed before being used. The aggregator (in our case the final training dataset) only receives noisy data, eliminating the need to trust the aggregator with raw data [54]. The problems of this method are first the cost (each data must go through the privacy mechanism all the data are not going at once as in the centralized model) but also the need for a very large dataset since each data is losing accuracy while, with centralized DP it is the aggregation that loses accuracy, resulting in less overall loss. Thus, this method is context dependent and not often used when talking about generative

AI since it is constraining and costly.

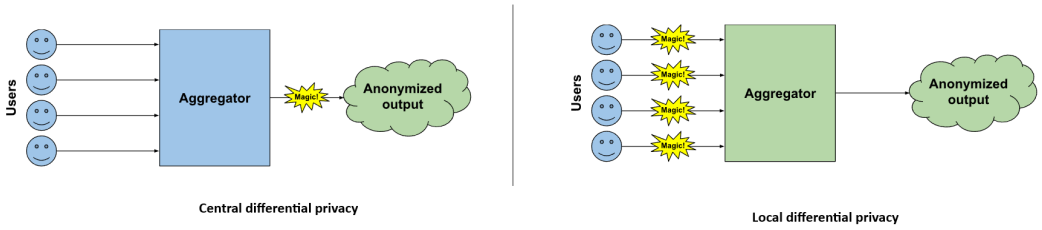

Figure 4.1: Illustration of central Differential privacy versus local Differential Privacy. Source : Damien DesFontaines [54]

4.2 Privacy challenges : generative AI model

An AI model can be defined as the result of a training process; for instance, a model trained to generate new data is considered a generative model. As the CNIL states, "An AI model is a mathematical construct that generates deductions or predictions from input data. The model is derived from annotated data during the learning (or training) phase of the AI system." [55]. However, the model is only one part of a broader AI system, which also includes interfaces, data flows, and other components that enable the system's full operation and deployment [56].

The Figure 4.2 illustrates how the AI model (in this case a deep neural network) is a succession of nodes which are related depending on their weights (probabilities) of apparition of the next one depending on the previous one. As for the dataset, it is important to distinguish two cases depending on the anonymization (or not) of the dataset the model has been trained with.

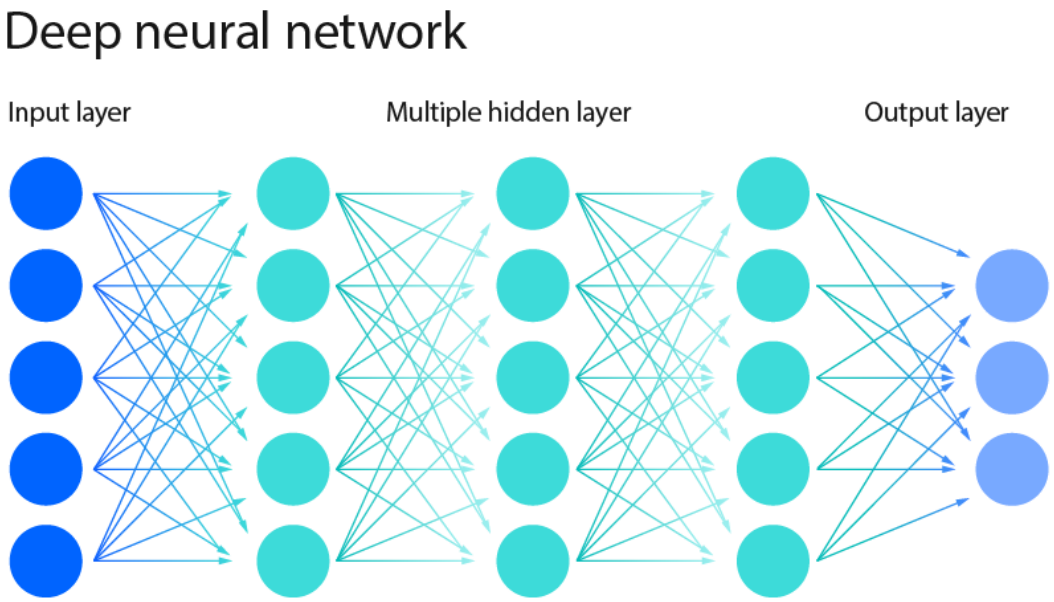

Figure 4.2: Illustration of a deep neural network model. Source : IBM website [57]

4.2.1 Model trained with personal data

If the model has been trained with personal data, then it can be possible to consider the model as personal data itself. It can be surprising but the model while trained on such data can retain information from its training dataset. An AI model (especially neural networks and large language models) learn by identifying patterns and relationships in its training data. By adjusting internal parameters (weights) based on provided data it can achieve this learning. These learned parameters essentially encode the statistical relationships and knowledge extracted from the data.

Each weight adjustment during training helps the model predict outputs closer to the actual training examples. After extensive training, these weights become

stable and contain embedded knowledge of the patterns found in the training data. Thus, complex information, such as phrases, facts, or even personal details, can be indirectly encoded within these parameters if the data contains recurring or distinctive elements. This embedding is not a simple "memory" like a database but is a statistical encoding of learned associations and patterns.

Even after the AI training, certain data details might persist. Models sometimes memorize exact examples, particularly if the information appears frequently or distinctively or if the data is repeated multiple times during training. In this case, if the model contains personal data, it might fall under the GDPR as some might argue that the model in itself is a personal data. In this regard, according to the European regulation it should be possible for the data holder to perform their rights to : Information (Art 12-14), Access (Art 15), Data Portability (Art 20), Object (Art 21), Rectification (Art 16), Erasure (Art 17) and Related to Automated Decision-Making and Profiling (Art 22). However, those rights regarding an AI are challenging to guarantee.

Regarding the rights to information, access and related to automated decision-making and profiling, it means that the AI would be able to show some transparency. However, transparency is difficult to achieve when talking about AI systems. The first reason is that it would mean to keep a track of all the nodes the AI went through in order to give an answer and to be able to explain why it went thought this node instead of another. In addition, for the same request the AI might not give twice the same answer and not go through the same nodes. Moreover, giving the explanation would mean giving away a part (or all) of the model, but when talking about an AI the real product is the model. It is what companies sell, and the result of the AI's training, thus companies are not keen on

divulging it, even to the rightful data subject.

Now, about the rights to rectification, object and erasure, it is nearly impossible to achieve those without retraining the model (which means still having the training dataset in hand). In the EDPB 2024 opinion [30], it is mentioned that a possible way out to avoid retraining would be post-training. However, the post training is still an ongoing research field. Post-train means trying to identify the node(s) that has retained a specific information. As mentioned previously, this is challenging because transparency is hard to achieve in an AI model. Thus, even the identification of the involved nodes might already fail. After the identification, the goal is now to remove all the concerned nodes. Nonetheless, since all the nodes are connected, the question is the magnitude of erasure that must be achieved in order to completely erase the data from the AI. Too little would mean the information is still somewhat present. For instance, if the goal is to erase a card number, while asked for two first numbers of a card the AI would still give the first two numbers of this card. While erasing too much would mean to lose information in the AI, when balancing the training cost of an AI this might not be acceptable for the companies either. Thus, researchers are still investigating this technology [58] [59].

Not only those problems are real concerns, but the GDPR enforce a certain time limitation to the storage of personal information. However, in the case where the data are directly imbued inside the model, even if the dataset in itself is deleted, personal data remains accessible in a way. And the only option to "erase" the now too old data would be to retrain the whole model. In this case, even post-training might not be achievable. As following the regulation, the data collector might already have erased the old data from the training dataset. In its AI fiches, the

CNIL states "The retention of data must be planned in advance and monitored over time. The defined retention periods must also be applied to the data concerned, regardless of their medium." [60]. Highlighting the importance to consider this potential problem.

4.2.2 Model trained through an anonymized dataset

If the model is trained with data that has been anonymized beforehand (such as data coming from local DP mechanism). Then the model can be considered anonymous. However, it is important to note that the security of the model, the output and the dataset is also depending on the security scheme used. Using k-anonymity is better than using no anonymisation technique but since the technique has been proven weak to some attacks, it is important to use the best algorithm possible to minimize the possible risks since retraining a model is costly and so creating a training dataset.

4.2.3 Model anonymized through model perturbation

One of the most used technique in order to anonymize a model is model perturbation. Model perturbation techniques aim to introduce randomness (or noise) into the training process or the model itself to limit the leakage of sensitive training data. The goal is to prevent an attacker from inferring whether a particular data point was in the training set, even if they have access to the model's outputs or internals. Since DP can only be used on statistics, instead of using it on dataset (which might be statistics but also tabular, textual data or even graphs, etc...) DP algorithms are used on the model itself which is composed of weights that are

the apparition's probability of each node. By adding noise, the model is secured. Two main techniques might be used.

The first one is DP-SGD (Differentially Private Stochastic Gradient Descent) [61], it adds noise to the gradients (a numerical signal that tells the model how to adjust its weights to make better predictions) during training. This technique is commonly used in training LLMs and other large models.

The other one is Post-hoc Noise Injection or Perturbation [62] [63], the goal here is to add noise after training, for example in model weights or outputs, but is less commonly used in practice due to lower utility. Overall, multiple other model perturbation techniques exist. The type of technique to be used mainly depends on : the type of training dataset, the type of synthetic data the AI will output but also on the type of AI model (LLM, balesian...).

However, model perturbation is often used with a training dataset composed of synthetic data. Since DP can only be used on statistics it is often used on the model (during its training or right before the output of the synthetic data), but the synthetic data generated through this process are often not considered anonymized enough and thus, a first model using perturbation is trained on the non-anonymized dataset and the outputs (synthetic data) of this first model are used to train or fine tune another one which is also perturbed. The second model can then be used as a product (a company sells its trained model to another company as a product) or the synthetic data then generated can be exchanged or sold as anonymized. The Figure 4.3 illustrates this process.

However, this thesis tends to argue that synthetic data might be worth considering as personal data, the question is then if a model trained over synthetic data can generate synthetic data that are anonymized enough to not be consid-

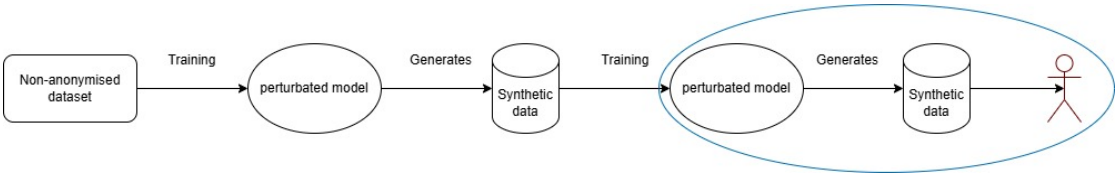

Figure 4.3: Model perturbation's usage

ered as personal. To our knowledge, no research has been done about how many models the synthetic data must go through to be less vulnerable to attacks (or which privacy factor each generation by a new model is bringing). However, those methods, while still being subject to attacks, are the ones with the better efficiency when it comes to generative AI anonymization. It is still important to consider the type of data and the type of generative AI to be protected in order to choose the best mechanism to preserve privacy. It is to be acknowledged that multiple privacy schemes can be used together to achieve a better protection, while keeping in mind the balance between privacy and utility and the potential costs (monetary, computational, in time, etc.) those methods might need.

4.3 Privacy challenges : generative AI input

The input, or prompt, is a "natural language text describing the task that an AI should perform. A prompt for a text-to-text language model can be a query, a command, or a longer statement including context, instructions, and conversation history."[64]. Usually, prompts are used in order for the user to communicate with a text-to-content AI (might it be text, image, speech, etc...). Prompts are mostly (if not only) used in LLMs.

Most of the LLMs (such as GPT) are using prompts to loop and train further

on the prompts provided by the users (cf Figure 3.1). Since users do not use privacy methods in order to make their prompts anonymous, and since the model is then re-trained on these, the same problems regarding the GDPR's rights are appearing as in the non-anonymized dataset (subsection 4.1.1).

It might be possible to argue that before looping (and be used for the AI's re-training), those data could be anonymized, however anonymizing text is a difficult challenge. First, DP does not work with the usual LLMs inputs (such as text, images or videos). Nonetheless, it might be possible to use local differential privacy if the prompt is integrated to the dataset locally. However, this method is working only if the dataset is small enough to be able to be fully stored locally (which is not the case with really large LLMs such as Gemini) and is subject to a high computational cost as each row of the dataset has to be anonymized using DP. A second option for this method would be for the user to anonymize itself its prompt or to trust a third party to do so. Since the goal is to not share the prompt information with the AI's provider, the anonymization must happen before the data arrives on their server. By doing such a thing, the accuracy of the AI's provided answer would be very diminished.

Another possible method is using filters, instead of DP or other anonymization technique, in order to identify what is personal data from what is not before training the AI on the prompts. E-mail addresses, names, addresses, phone numbers and so, are easily identifiable. However, other data might slip through this filtering. It is important to understand that in this particular case, it is not possible to filter everything since, for instance, a text would lose its meaning if filtered too much. As such it is possible only to scale down the filtering, not up, to keep the utility of the prompt. In addition, the unfiltered data might still be personal

enough to identify the sender if combined all together.

4.4 Privacy challenges : generative AI output

4.4.1 When "non-real" data ...

The output of the generative AI is called synthetic data. Before exploring the challenges, those data are creating, lets remind the definition of a synthetic data. It refers to information that does not originate from direct observation of the real world. Instead, it is artificially created to imitate the statistical properties and behaviour of real data, without reproducing any actual records [13].

They are often produced by artificial intelligence systems. These can include large language models (LLMs), probabilistic graphical models, or Bayesian networks, depending on the type of data and the desired outcome. Broadly, synthetic data can be divided into three types: partially synthetic, fully synthetic, and hybrid.

The synthetic data we are studying, are generated entirely by a model that has been trained on real data, which may or may not have undergone privacy-preserving transformations. This model learns the statistical patterns and relationships within the original dataset and then produces entirely new data points that replicate those patterns. These synthetic records are not tied to any actual individuals, and they omit all identifying parameters. This approach aligns with the idea that synthetic data is an artificial reproduction of real data that aims to represent the properties of the original dataset faithfully, but without directly copying it. In essence, synthetic data is a constructed imitation of real-world data,

designed to preserve the useful characteristics of the original.

Then the question would be : how an imitation of a real data could be personal? It might make no sense that a fully crafted data is personal since the "person" it is about simply does not exist. The problem is that it is possible to attack those synthetic data in order to retrieve the training data (which are not synthetic and thus personal).

4.4.2 ... rhymes with personal

While synthetic data is often praised for its privacy-preserving potential, it is not immune to privacy attacks. The core challenge lies in maintaining a delicate balance: synthetic data must be statistically similar enough to the real data to be useful, yet dissimilar enough to prevent leakage of individual-level information. This trade-off introduces significant ambiguity in defining and measuring privacy risks, as well as in assessing whether a synthetic dataset is "too close" to its source.

Attacks on synthetic data can largely be categorized into membership inference attacks, attribute inference attacks, and data extraction attacks. These reflect the broader privacy vulnerabilities seen in AI systems.

Membership inference attacks occur when an attacker can determine whether a specific data point was included in the model's training dataset. This becomes a concern when synthetic data is generated by models trained on sensitive datasets. If synthetic samples are too reflective of training examples, attackers can infer membership, which may be enough to compromise individual privacy.

Attribute inference attacks aim to uncover missing or hidden attributes of an individual based on partial information and access to a trained model or syn-

thetic dataset. These attacks exploit correlations learned by the generative model and become increasingly feasible when the synthetic data retains high fidelity to original statistical dependencies.

Data extraction attacks represent one of the most direct threats. In some cases, language models or data generators unintentionally reproduce exact or near-exact training samples. A notable case is discussed in Carlini et al.'s article "Extracting Training Data from ChatGPT" [65], who demonstrated that ChatGPT could sometimes output sequences from its training data. This kind of attack raises important questions: is the model merely generalizing based on broad patterns, or is it overfitted and leaking specific training instances? The ambiguity is particularly problematic when seemingly personal information, like a common name or address, is generated. Determining whether this represents a privacy violation depends on whether such information is likely due to general population statistics or a direct leak from the training data.

Measuring the privacy risk associated with synthetic data is far from straightforward. One common metric involves comparing the statistical similarity between synthetic and real datasets, either globally or at the level of task performance (e.g., machine learning accuracy on both sets). However, such metrics do not provide a direct measure of privacy risk. Instead, researchers often resort to evaluating models based on their susceptibility to privacy attacks, treating privacy risk as a function of attack success. The paper "Let the Privacy Games Begin" [66] Suriyakumar and al., emphasizes this game-theoretic perspective, where synthetic data privacy is evaluated by simulating realistic attacker models. However, measuring privacy-risks based on the attacks is again a difficult task mainly depending on the metrics used in order to determine if the attacks was successful or not (this

is discussed in detail in section 5.3).

There is no clear threshold that defines when synthetic data becomes "too similar" to the original data. Indeed, high similarity is desirable for utility, but beyond a certain point, it may indicate overfitting to individual training records. The boundary is ill-defined: how close is too close? This remains a grey area in synthetic data research. As a result, the quality and safety of synthetic data must be judged not only on their utility but also through rigorous privacy evaluations and adversarial testing.

4.4.3 Synthetic data challenges

Upon holding a lot of hopes, synthetic data are answering some problems and challenges while having their own. It is first important to define in which context they are used and for which usage. As demonstrated in the previous section, synthetic data and more generally AIs can be targeted by various attacks, allowing the attacker to retrieve information from their training dataset. Thus, synthetic data should not be used without another privacy-preserving technique in order to try to mitigate those risks.

As such, the tempting promise of having accurate data while answering the dilemma between privacy and utility, does not hold. It is still mandatory to use privacy preserving techniques, but it is possible to think about less strict measures that might be enough to guarantee privacy, as synthetic data might bring some privacy or at least make the attacks harder.

Using synthetic data also means answering a need, this can be separated in two different scenarios : the need to enrich a dataset or the need to use it as privacy

mean to escape the GDPR's reach. Those scenarios can be illustrated by taking an hospital that has cancer patients' data for those age intervals : [30-49] and [50-69]. First, the hospital wants to infer the data for the interval [70-89] or to create bigger dataset for the intervals they already have, they can use synthetic data. While doing so, they have to keep in mind that for the interval [70-89] they in fact do not have the data and thus cannot check if the statistical distribution is matching reality. The second scenario, the hospital in order to be able to share more freely the data (or to sell them) will generates synthetic data on the intervals [30-49] and [50-69] that they already possess and argue that the data has been anonymized and thus do not fall under GDPR.

This is not the only problem, AIs tend to hallucinate sometimes [67] [68], if generating entire new dataset, nothing is ensuring the AI (one time out of hundreds) might not be hallucinating. This one occurrence might be drowned inside the amount of generated data, but if the generated dataset is small, then the impact of one hallucination might be consequent. Moreover, AIs tends sometimes to amplify the biases [69] [70] [71], thus, the generated data might amplify a characteristic, and then the generation might not follow the statistical representation of the real-data characteristics. In regard of those two possible problems, the solution would be to be able to notice when such things are happening by asking explanations to the AI on the conditions of their output generation (answering the questions : why, how and with which path). However, this problem, known as the explainability problem and rejoining the transparency one, is a nearly impossible problem to solve. These problems are hard to solve because AI models are often very complex, with millions of parameters, making their inner workings too intricate for humans to easily follow or explain. Moreover, being able to erase the

path taken by an AI can lead to other problems : first, blocking one path does not ensure this problematic output cannot be reached by another one (the bounds of the possible data generation are hard to set). Second, being able to give the path of an AI as an explanation means giving (maybe without the weights) some information about the model. However, the model is the core product of an AI company, worth sometimes millions. Giving even hints on it, might lead to maybe new attacks. In addition, the path might be really hard to read and comprehend, and in the end does not give enough information to allow the owner of the model to ensure the problematic is solved. Balancing performance and interpretability remain, then, a core challenge.

Not only does the AIs might face some challenges, but privacy techniques also come with their own. DP face a serious one : fairness. Recent studies have shown that even when privacy techniques are correctly applied (especially DP), they can unintentionally lead to unfair results [72]. For example, models trained with DP, which aims to hide individual information by adding randomness to the data, can perform worse for minority groups. This is because the added noise may affect smaller or more unique groups more than others. In their example, Bagdasaryan and al. showed that a private sentiment analysis model was much less accurate for African-American users compared to others [73] (Figure 4.4).

This problem also appears in how models are trained. Many private training methods, rely on techniques that limit the influence of each data point to protect privacy. But if certain groups tend to have stronger or more unique signals in the data, these protections can end up weakening their impact on the model, making the model less accurate for them [75]. So, while DP is good at hiding personal information, it can also create or worsen unfairness between groups. For

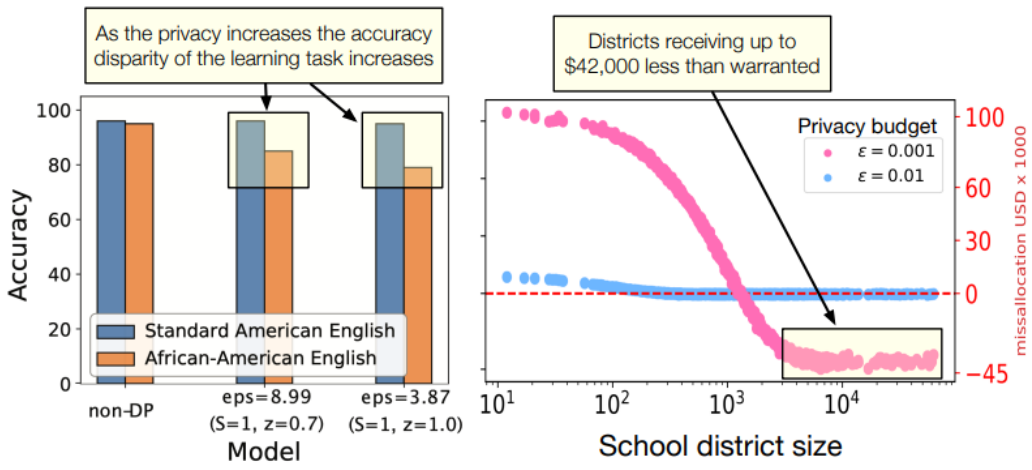

Figure 4.4: Left: Disparities arising in DP sentiment analysis tasks (image from Bagdasaryan et al. [73]). Right: Disparity arising in fund allocations to school districts (image from Tran et al. [74]).

this reason, it is important not only to think about how private the data is, but also how fair the outcomes are. Some recent research have proposed ways to fix this, such as adjusting how training is done for different groups, or making the model more sensitive to differences in accuracy between groups [76] [77] [78]. These approaches aim to balance privacy and fairness, so that protecting individuals does not come at the cost of treating them unequally. But still, choosing the right one in order to balance between privacy and fairness can be a challenge. This possible DP unfairness, while knowing AIs can amplify biases might lead to some other new problematics as well. However, to our knowledge, no study has yet explored this. Either the AI worsen the unfairness or correct it, but it might be interesting to study.

5 Applied Privacy Mechanisms

This chapter offers a forward-looking reflection on the current state of anonymization techniques in the context of generative models, building on the interdisciplinary foundations laid throughout this thesis. It begins by questioning whether privacy should continue to be treated as a cryptographic problem and moves on to examine the possibility of defining measurable privacy thresholds. Alongside those, it also considers the role of standards and best-effort strategies in guiding future developments. These proposals are not intended as final answers, but rather as potential directions worth exploring further. By evaluating them from both legal and technical perspectives, this chapter aims to contribute to the ongoing dialogue around how anonymization practices might evolve to better respond to the risks posed by modern AI systems.

5.1 The argument for a cryptographic solution

According to the GDPR [4], Art 5.1(c) "Personal data shall be: (c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');". While this is a founding principle, it is still important to understand that having multiple companies respecting it and owning

some data about a person, even as small as two or three pieces of information, can in fine render the efforts of the GDPR less effective. If those piece of information were to be aggregated all together, then the amount of data available on a person would be on a totally other scale. Thus, even if the minimisation principle is a good idea, in reality, it might not be as effective as it was originally planned (but it is still necessary). In case of a problem, it is up to the regulatory authority to judge if this principle has been respected. However, if a company's business is to sell personal data about individuals, is it then respecting the minimization principle to collect all the possible data on and about a person ? This principle is relying on the best-effort from the companies which come with a lot of uncertainties about what is allowed or not. In addition, it is making any judgment harder to make (and longer) as companies might define a blurry purpose in order to have some room to manoeuvre.

In respond to this problem, the CNIL [79] mention that an encrypted dataset might also be able to bring some privacy guarantees. It is important to understand that even if encryption is working towards security as well, it is by definition another field of security with its own algorithms, security models and challenges. Privacy and encryption might be a good match to create a more secure model, but in no case the use of an encryption model should replace the use of a privacy one. Those two techniques are connex, but by essence different.

> "In cryptography, encryption (more specifically, encoding) is the process of transforming information in a way that, ideally, only authorized parties can decode. This process converts the original representation of the information, known as plaintext, into an alternative form known as

> ciphertext. Despite its goal, encryption does not itself prevent interference but denies the intelligible content to a would-be interceptor."[80].

A metaphor of the difference between those two paradigms could be a house that is the data to protect (dataset and model). The aim of encryption would be to close the door that can still be opened by whoever has the key (legitimate person or an attacker that stole it) where privacy aims to immure the door in order to forbid the access to everyone. This thesis will not study the question of encryption as a privacy mechanism as we consider it out of the scope because at some point the data needs to be encrypted and decrypted leaving them accessible at those moments and vulnerable to attacks.

5.2 The argument for a privacy threshold

The world we are living in right now needs data, we all need data in our everyday life : "What weather is it going to be today?", "How much can I spend for my vacations, taking into account the sum I have on my bank account and the activities I want to do?", etc... While data is mandatory to make everyday life's decisions, we are increasingly dependent on digital data and their automatic processing, even to make basic decision such as booking an hotel by using price comparison on websites for instance. In this case, aiming at the hotel options' comparison, the website will collect various data that will include the number of people to find a room for, the age of the children traveling, whether we need to park a car, etc. The collection of personal data is thus mandatory and inescapable. Indeed, while we can opt to access an online booking site, we might not have the possibility to opt-out when talking about state-lead services that require data collection (such

as employment services or hospitals' check-in platforms). Accordingly, individuals can only hope that their personal data is protected and that their privacy will be safeguarded. Thus, the privacy discussion revolves around the adequacy and performance of privacy-enhancing technologies that can be used to protect our data and, particularly, the level of acceptability of the remaining privacy risks. But how to define what is acceptable and what is not? More than just information collection, it is important to question the usage of that information. AI models can be attacked (subsection 3.2.1) and there is no universal answer to mitigate them (the last chapter discusses which method is best for which part of the AI system). Hence, as for companies but also for the data subjects, it would greatly simplify their life if a threshold existed in order for them be able to make the difference between what is acceptable as a risk compared to what is not.

By finding a relevant threshold, which represents a clear and quantifiable criterion, it would be possible to determine what is a sufficient privacy protection from what is an inadequate measure. The significance of such a threshold is twofold. First, it would bring legal clarity since a clearly defined threshold would help judges and regulators discern compliant behaviours from non-compliant ones. But it would also provide operational guidance to organizations aiming to comply with privacy regulations by clearly defining the minimal acceptable privacy protection standards.

From a legal point of view, it is possible to find the threshold notion in the AI Act (regarding AI) [28] that introduces risk-based thresholds, categorizing AI systems into "high-risk," "limited-risk," and "minimal-risk" categories, effectively setting implicit thresholds for acceptable risk [81]. Even if it can sometimes lack clarity, as an AI system might be used in different environments (in a "high-risk"

one and in a "limited-risk" one for instance) [82]. Similar approaches could inspire thresholds for privacy techniques, explicitly clarifying under what conditions anonymization or synthetic data generation can be considered sufficient.

But in order to define in which category an AI system is in, each AI must be examined independently. Their classification depends on multiple factors; some AI system can even be in multiple classifications depending on the client's usage. This classification, which depends on a lot of parameters, makes it extremely hard to classify an AI system. In addition, the threshold problem adds even more parameters to consider (this will be discussed in the section 5.3).

A legally enforceable numerical or mathematical threshold would simplify compliance by offering an effective and justifiable numerical value for privacy guarantees. Occurrences of such thresholds exist in a neighbouring field : cryptography. Cryptography regulation historically adopted explicit thresholds: clear security parameters (e.g., key lengths, bit-strength) defining acceptable security [83] [84] [85] [86]. Such thresholds are attractive because they provide a clear enforceable benchmark. Those thresholds are not only complex to define, but they are also bound to be redefined from time to time in order to follow the latest technological discoveries. For example, as of today, cryptography faces new challenges, especially from quantum computing which is demultiplying the adversarial computational power. If a fixed threshold is defined in privacy in the future, it would mean that, as for cryptography, this threshold could rapidly become obsolete due to emerging threats or technologies and thus, would require methods to offer a rapid adaptation (depending both on the technology to find a new one and on regulation to enforce its usage).

5.3 Setting privacy threshold : a technical challenge

Regulation is already dealing with thresholds in cryptography. But which threshold do we choose, and for what? Those are the questions only technique can answer. In fact, since cryptography and privacy are different, the first question would then be : is it doable? Technique does not provide all the answers to those questions and some problems are still on-going research fields.

Such a threshold would be set to check how robust the privacy scheme is when under attack (it is possible to see it as how good or bad the attacks results are). Before explaining all the challenges, a threshold might create let's explore the resources needed in order to run an attack. Before even evaluating whether the threshold has any significance it is important to consider the attacks and attackers. In order for an attacker to retrieve information they must have a combination of resources : background knowledge, computational power, access to the model (black box attacks or white box attacks) and access to data (we will leave the expertise of the attacker aside even if we could maybe classify the attacks by difficulty). Since, many possible attackers can be found, it is important to understand the risks (in this case the attacker) before choosing a privacy model to protect the information against.

As demonstrated in subsection 3.2.1, there is a multiplicity of possible attackers that can threaten the privacy of the data. Their most powerful weapon is their knowledge (both technical or of the data). In order to try to mitigate this risk DP models are using a potentially omnipotent adversary scenario. A common critic

about privacy (especially DP) is that the attacker is nearly all-mighty : he knows all the training dataset except for one line of it (depending on the scenario, it can be less), as such some might think that this is highly improbable that this could happen in real-life. However, with aggregated databases that can be found nowadays, it has begun to be a non-null probability. Moreover, being prepared for the worst-case scenario might help navigate all the other possible scenarios. A good example of the law's evolution is the regulation adaptation to the potential all-mighty (or at least mightier) opponent is the emergence of a quantum-computing attacker in cryptography. Cryptography has its wild range of attackers as well, but the only representative parameter in those attacks is the computational power of the attacker. Then, the worst possible adversary is one with a "super" computer : a quantum-computing one. On the other hand, for privacy, the attacker can be defined with multiple characteristics and objectives which make him harder to envision upon applying a privacy scheme. Thus, a threshold would at least facilitate the companies' work when talking about privacy, removing the burden of setting it by themselves.

5.3.1 Numerous metrics

Nonetheless, a threshold is a complex problem. First, the metrics are not easy to understand, and scholars do not even agree on them. Not only the metrics for those thresholds, but even the metrics on the attacks' impact. But how is it possible to set a leakage threshold when it is not even clear if an attack succeeded at retrieving information ?

Since, there is a multiplicity of data types, this leads to many possible privacy

schemes (for instance more than 200 algorithms are satisfying DP). Not only this but depending on the datatype of the training dataset and the intended usage, the number of possible different generative AI is tremendous as well. On top that, as seen in the previous section, it is important to understand the possible attackers one might be facing. However, this is not easy task as it means knowing the means and goals of such adversary.

Moreover, a privacy-preserving algorithm might be optimal for one type of dataset but ineffective for another and the same goes for the attackers. As a result, companies often need to experiment with multiple approaches before finding one that fits their specific data. This trial-and-error process is both time-consuming and costly, introducing a new trade-off: privacy versus cost (time can be counted as monetary cost too). To evaluate which algorithm offers the best protection, companies must simulate real-world attacks on their own data. Theoretically, once they identify the algorithm that minimizes data leakage (or ideally eliminates it) they have found the right solution. To address this issue, initiatives like the IPOP project and CNIL [87] are working to develop libraries of privacy attacks that companies can use to test their systems. Other resources also exist, such as the TAPAS Library of Attacks [88]. However, their usefulness has limits. Building and maintaining such libraries requires time, money, and expert knowledge. Moreover, no library can ever fully capture the range of possible attack strategies and at best, they offer a representative sample. Lastly, running all those attacks takes a tremendous amount of time, and the companies might not even know those resources exist. In this context, assessing privacy becomes more about approximation than certainty.

All of this reinforces a deeper issue: the process would be far simpler if a

clear, legally enforceable privacy threshold existed. Without it, companies are left navigating a complex, open-ended space of risks, choices, and costs. But it also brings another crucial question : what exactly constitutes a successful attack?

As mentioned, DP can be implemented via multiple algorithms that are best suited for different types of data, some attacks might perform better depending on the algorithm that has been used or the data type. Moreover, discussions are still ongoing in order to define the best metric to assess the performance of an attack. Upon launching an attack there are several metrics to measure the success of an attack : the actual true negative and true positive are the results when the attack functioned (either by saying the data was not in the database when it was not or saying it was when it actually was), then come the false negative (when the attack says the data was in the database when it was not in reality) and the false positive (the attack pretends the data was not in the database when it actually was).

Intuitively, we could say that a good way to determine if an attack was successful or not would be to calculate how many times the attack was "right" (true negative and true positive) compared to how many times the attack was "wrong" (false negative and false positive). However, in their paper "Membership Inference Attacks From First Principles", Nicholas Carlini and al. came up with an interesting approach. They proposed to grant a bigger importance to the cases where the attacker actually learn something useful, meaning the true positive cases. To do so, they are using a metric they call TPR for True-Positive Rate [89]. But we could think as many other metrics. In their article, they are comparing theirs to a bunch of other (5.1). In order to define the accuracy of an attack, some possible metrics can be the average accuracy (how many times in average does the attack succeed), the TPR, or other metrics (those metrics can be as "home-made" as the

Avatar one used by Octopize for instance [90]). For sure, the metrics are depending on the success and failure measure, as such it is interesting to question what a success is : is it having true negative/positive? Is it having positive (whether false or true) ? Is it when the attacker learns something ? When is all the information retrieved? Some part of it? Only a useful part ? It might be possible to argue that it is dependent on what the attacker goals are. If we are thinking that the AI has been trained with political opponents, then it would make sense to consider only true positives and false negatives when creating an attack. But in reality, it might be all the positive results, even the false positive ones that could catch the attention. On the other hand, if it is a whole hospital database then the positive results are not enough, the attacker needs to retrieve the disease name for instance. The problem of defining whether an attack is successful or not is definitely a non-trivial one, however without an answer to this question (what is a successful attack?) it is out of the picture to even think about the possibility to define a relevant threshold who would then define the good, the bad, and the ugly when talking about privacy protections.

5.3.2 Is there a threshold in the room ?

Regarding data that would not have gone through any privacy mechanism, finding a threshold is a challenge. Since, a various number of data, attackers, algorithms exist and considering the companies do not have the same means, the threshold should be either general or specific to a situation. Having a general threshold would mean to lower all the standards in order for even the smallest company to be able to reach them, or to accept that some of them might not be able to comply

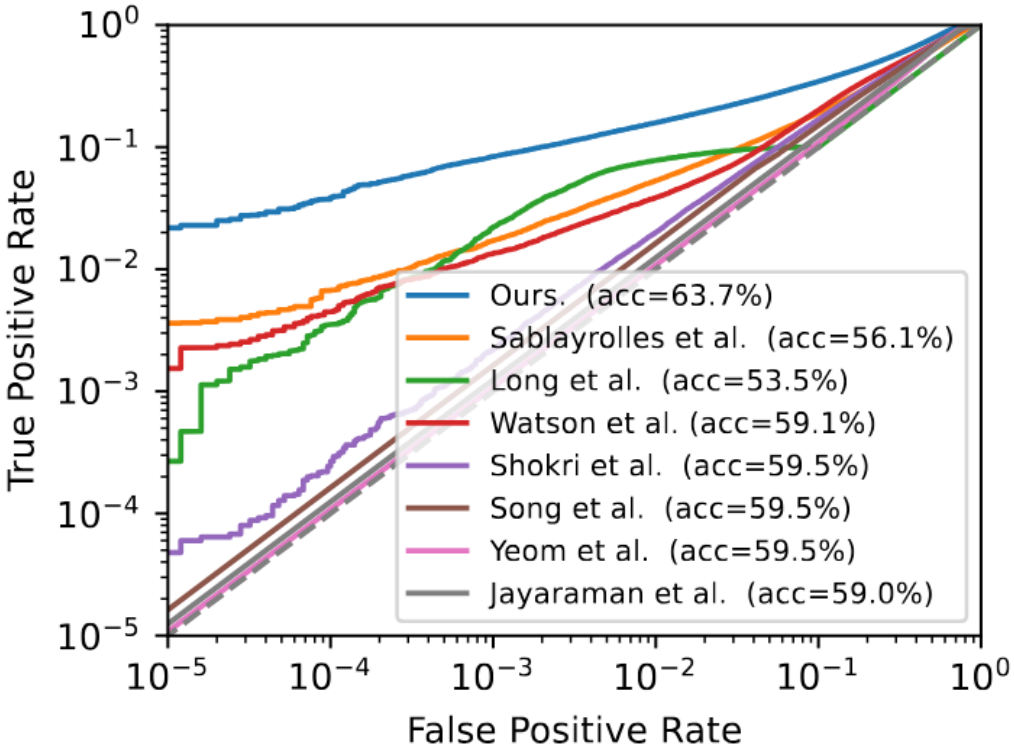

Figure 5.1: Comparing the true-positive rate vs. false-positive rate of prior membership inference attacks reveals a wide gap in effectiveness. An attack's average accuracy is not indicative of its performance at low FPRs. Source : "Membership Inference Attacks From First Principles", Carlini and al. [89]

with the regulation. This seems out of question. Thus, the other possibility would be to arrange a threshold depending on the situation. However, this would mean agreeing on such threshold. This agreement would come from the technique but needs to be investigated. In addition, one crucial question would be the granularity of each sector (situation). Is it a type of data? A type of data correlated to an industry? Correlated to a specific attack? Moreover, this would lead to the values of those thresholds coming from the technique to be applied and implemented by the regulation who would have to be reactive enough to change its texts when the

technical threshold is modified.

While talking about data that has undergone privacy processes there is one process that interest us in this thesis : DP. In DP, there is a few measures that can be associated with thresholds. Let's remember the DP formula :

Let ε be a positive real number and $\mathcal{A}$ be a randomized algorithm that takes a dataset as input (representing the actions of the trusted party holding the data). Let $im\mathcal{A}$ denote the image of $\mathcal{A}$.
The algorithm $\mathcal{A}$ is said to provide (ε, δ)-differential privacy if, for all datasets D_1 and D_2 that differ on a single element (i.e., the data of one person), and all subsets S of $im\mathcal{A}$:

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\varepsilon} \Pr[\mathcal{A}(D_2) \in S] + \delta$$

where the probability is taken over the randomness used by the algorithm. This definition is sometimes called "approximate differential privacy", with "pure differential privacy" being a special case when $\delta = 0$. In the latter case, the algorithm is commonly said to satisfy ε-differential privacy (i.e., omitting $\delta = 0$) [91].
We see here two parameters : ε and δ whose DP depends on, however they are not defined, or at least not with fixed values. Those two parameters (that define how "strong" of a private output we want) are in fact to the discretion of the technician to fix. δ is the chance that the privacy mechanism has to fail (the attacker or someone else learn something). ϵ is called privacy budget and represents a way to measure how much private information is allowed to be leaked when analysing data. It controls how much noise is added to protect people's privacy. Often, we choose an $\varepsilon = 1$, although a bigger ε means less protection but utility while

a smaller one tends to offer stronger privacy protection but less usability of the results.

Typically, privacy budgets (ε) are chosen based on the assumption that an attacker could be very powerful, having access to external data, making repeated queries, or performing advanced analysis. A big ϵ means a small noise and vice versa. However, it is possible sometimes to assume that the attacker is weak: they lack background knowledge, computational resources, or access to multiple versions of the dataset. In these cases, the noise introduced by even a small epsilon would already be enough to block any meaningful attack. As a result, it may seem unnecessary to use a very small epsilon (as a smaller epsilon means stronger privacy but more noise in the data). But paradoxically, it can actually be safe and beneficial to lower epsilon. Since the attacker cannot realistically exploit the data anyway, increasing the noise does not significantly harm utility, and it strengthens privacy guarantees even further. This can help protect against future threats, satisfy regulatory expectations, and promote ethical data practices, even if the immediate risk is low. Then, when the potential attacker is weak, lowering epsilon is a low-cost way to boost long-term privacy.

Among all the possible values of the different parameters of DP, badly chosen ones can lead to several attacks.

A badly chosen privacy budget (ε), or none at all, can lead to two main attacks. The first one is composition attacks, occurring when there is no privacy budget. This is using multiple queries over time, and add them up to degrade the privacy guarantees. The problem is that when adding up the sum of acquired knowledge, the attacker can unravel a few data [92] [93].

Another one is the Reconstruction attack, which is leveraging the fact that even

if DP is adding noise through a Laplace mechanism, with enough noisy queries, an attacker can reconstruct a close approximation of the underlying dataset [94]. Those two attacks can be mitigated with a privacy budget and should not happen if a correct DP algorithm is used.

The attacks can also leverage the poor noise addition which is an implementation flaw. This is the case of the side-channel attacks can use a non-random seed in the noise generator which could allow attackers to reverse-engineer outputs [95].

Overall, those attacks can be mitigated by choosing wisely the parameters of the DP algorithm and checking if its implementation is not flawed. For instance, a bigger ϵ means a smaller noise weakening the privacy guarantees. In (ε, δ)-DP, a δ too high also allows a meaningful probability of strong leakage. In any case, DP does not guarantee complete anonymity, it guarantees that the presence or absence of any individual in a dataset changes query outputs only slightly.

However, technical thresholds are already used and are already subject conventions. For instance, it is common to have an $\epsilon = 1$ in academic research. Thus, it is possible to say that technical conventions already exist. But to know if those conventions could be transposed into regulation is another story, since it is important in some cases to be able to weaken those protections (it can be because the dataset is too small to introduce too much noise, because the data are losing too much accuracy, etc...). Setting these parameters (ε, δ) correctly remains debated and unclear in practical scenarios. Moreover, in some specific cases such as when talking about graphs, other DP-algorithms are used.

5.4 Setting privacy threshold : standards' contribution

Nevertheless, one advantage of cryptography is its standardization. This would be great to have a standard that determines how, when and on what to use privacy measures. However, as previously stated, since not all data behave in the same way in front of a privacy measure, fixing a threshold and later a standard might be complicated (a standard might need to fix a threshold or the threshold is the open door to a standard or can even be the standard). Those two questions are then parallel : answering one might mean answering the other.

Cryptographic standards (e.g., AES [96], RSA [97], post-quantum cryptographic algorithms [98]) serve as universal references, significantly simplifying compliance. Companies rely confidently on standardized, rigorously tested cryptographic libraries, reducing ambiguity around acceptable security practices. Such standardization also provides a valuable framework for swiftly adapting to emerging threats, such as quantum computing, by adjusting thresholds within an established and recognized standard. Conversely, the absence of standardization makes adaptation slower, more costly and legally uncertain.

Given the complexity involved in defining universally applicable privacy thresholds, an alternative strategy may be to standardize privacy-preserving methods themselves (e.g., differential privacy, federated learning). This approach shifts the complexity and responsibility of threshold-setting from individual organizations to recognized standard-setting bodies (CEN-CENELEC [99], ISO [100], IEC [101], JTC-1 [102], SC-24 [103], IEEE [104], or regulatory authorities). In this sense,

standardization does not eliminate the need for thresholds, it relocates it. Standards inherently incorporate thresholds, either explicitly or implicitly.

Another approach might be to set sector-based standards and thresholds. However, again, depending on the data type the algorithm to use might differ. In addition, depending on the size of the dataset, the content of the data (e.g. medical data or favourite colour), but also of the company and its location, the deployment cost of such standard can be completely different. For instance back when video platform exploded both YouTube and Dailymotion (its French equivalent) were developed at the same time, however, Dailymotion had to struggle adjusting to the potential legal problematics, losing time to establish itself on the market and ultimately not taking the lead when it comes to video platforms.

Having sector-based standards would also mean that each sector needs to take the matter into their own hands in order to set such important parameters. It is then crucial to question if they have enough knowledge but also time to do so. In fact, if researchers from the computer science background are working on setting such standards, they might not fully understand the sector for which they are trying to set the threshold, and if it is someone from the specific sector then, they might not have the technical perspective in order to make an enlightened decision. Moreover, research might take a while, while companies might want to rush things up to be able to use the techniques directly.

Standardization complements efforts to define thresholds by providing unified frameworks such as standardized benchmarks or testing toolkits for privacy attacks. Additionally, it establishes a recognized, regularly updated framework capable of dynamically guiding thresholds to adapt to evolving threats. However, standardization also introduces challenging questions: Are regulatory frameworks

adequately prepared to handle increasingly sophisticated adversaries (like quantum attackers) who could bypass current standards? Can existing standards evolve quickly enough to keep pace with rapidly changing technological landscapes and emerging vulnerabilities? And is it not vain to try to define a standard if no threshold can be found?

The European Commission already called for the CEN-CENELEC [99] to publish in Autumn 2025 standards applicable to detail article 15. In this regard, the briefing document published indicates that "Standards on cybersecurity should define technical and organisational measures to achieve a level of cybersecurity that is appropriate to the risks of AI systems". Given the software nature of AI, some controls in existing standards will be applicable, such as those in the ISO/IEC 27000 family [105]. These may be most relevant for the security of the infrastructure underlying AI systems. However, AI-specific vulnerabilities, such as data poisoning, model poisoning, model evasion and confidentiality attacks. Those last two vulnerabilities pose new challenges that will require specific coverage in standards in order for these to fully cover legal requirements in Article 15 of the Regulation [106].

5.5 Setting privacy threshold : a complex transfer of responsibilities

From what seems to appear, neither the threshold nor the standard might answer perfectly (as of right now but since researches are still on-going there are high chances that in the future those might) to the question : What is enough protection

when it comes to personal data?

Since privacy is closely related to other cybersecurity fields, examining approaches from those fields' regulation might offer valuable insights.

In cybersecurity, regulation typically does not define an explicit threshold for security compliance. Instead, it demands companies to adopt state-of-the-art practices to mitigate vulnerabilities [4] [107] [108]. For instance, article 24 of the GDPR [4] requires

> "1. Taking into account the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons, the controller shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation. Those measures shall be reviewed and updated where necessary. 2. Where proportionate in relation to processing activities, the measures referred to in paragraph 1 shall include the implementation of appropriate data protection policies by the controller.".

Article 32 provides further guidance by stating that

> "1. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, [...] 2. In assessing the appropriate level of security account shall be taken in particular of the risks that

> are presented by processing, in particular from accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to personal data transmitted, stored or otherwise processed.".

Recital 7 of The Data Governance Act 2022 [25], a European Union regulation aiming at creating a safe data circulation environment in the EU by creating new rules for data exchanges, provides that

> "There are techniques enabling analyses on databases that contain personal data, such as anonymisation, differential privacy, generalisation, suppression and randomisation, the use of synthetic data or similar methods and other state-of-the-art privacy-preserving methods that could contribute to a more privacy-friendly processing of data. Member States should provide support to public sector bodies to make optimal use of such techniques, thus making as much data as possible available for sharing. The application of such techniques, together with comprehensive data protection impact assessments and other safeguards, can contribute to more safety in the use and re-use of personal data and should ensure the safe re-use of commercially confidential business data for research, innovation and statistical purposes. In many cases the application of such techniques, impact assessments and other safeguards implies that data can be used and re-used only in a secure processing environment that is provided or controlled by the public sector body".

Accordingly, article 5 (3) and (4) of the same text requires public bodies to grant access to public data for re-use preferably once anonymized. However, this thesis aims to highlight the fact that synthetic data shall not be considered as a

privacy protection in itself since it requires another anonymisation technique to be compliant. It is more a tool to enrich datasets than a protection measure.

The AI Act [28] also tackle state-of-the-art privacy methods by asking that high-risk AI systems' providers develop their model in a way to achieve cybersecurity. Article 15 adds that "The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks." Those reglementations, are advocating for a case-by-case risk analysis in order to define the best privacy practices. The texts emphasize on the fact that privacy shall be appropriate to the risks the companies might face. This is not absolute, but more an assessment depending on menaces. This is really similar to the penetration testing (or pentest) usage in cybersecurity, pentest is a central tool in this paradigm, where cybersecurity professionals mimic hackers to find and document vulnerabilities, which organizations then address proactively. Analogously, in privacy, specialized entities (privacy auditors) already exist, conducting audits similar to pentests, revealing vulnerabilities and suggesting remediations. Such audits demonstrate due diligence ("obligation of means") but raise critical questions: from a legal point of view, is demonstrating due diligence without a clear threshold (or guidelines) sufficient? In fact, court will have more difficulty to consider that all preventative action were taken when there is no threshold, because it will require more effort to evaluate actions. In France, the CNIL (with its documents [31]) is trying to give those guidelines in order to help companies navigate those complex questions.

What is important to realise is that, whether the chosen solution is a threshold, a standard or a pentesting-like obligation, the problem does not disappear, it is just "moved". The responsibility of the choice is transferred from an actor to an-

other. In the case of a standard, the responsibility and complexity are transferred from companies to standardization entities or regulatory bodies or comes from companies to regulation bodies. In the case where the standard comes from regulation authorities, companies simply adhere to accepted standards, demonstrating compliance without individually defining their privacy parameters. In case where the standard comes from the companies, then it is up to them to convince regulation authorities to recognize this standard, otherwise, the company will be the only one using it. This might create a relative chaos, if all companies have their own "home-made" standard.

In the case of a threshold the responsibility of the choice lies either on the scholars or jointly on researchers and professionals in the case of a sector-specific threshold.

Lastly with the pentesing-like regulation, it is up to the judge to decide whether the company did all its possible to protect the personal data or not.

It is also interesting to note that depending on the method, the compliance assessment intervenes at different timelines. Standardization and the threshold definition are taking place before publishing or using any algorithm, whereas pentesting and other similar methods are waiting for attacks to occur before assessing whether or not the protections were sufficient. In cybersecurity those two different timelines are defined by two different notions : security that is happening after the attacks and safety where the assessment is done before any possible attack. Pentesting is an attack-based security method, where standardization and threshold definition might be called safety measures. Moreover, those standards / thresholds might not be possible to create to suit every situation. When looking at other fields such as a nuclear power plant, it is a risk assessment-based method which is used

and then if a problem occurs, the responsibility of the judgement comes to the judge, only him will decide whether the owners did all they could to prevent such problems. It might be because humans and natural catastrophes can occur, no one can imagine all the scenarios and think about measures to mitigate them. The only possible thing is to apply at-best methods. On the other hand, in medicine, when putting a new drug on the market, the risks are (for most of them) already assessed, the drug has some risks thresholds that it should not get crossed to be called a drug and is tested beforehand. Those are (except in unexpected cases with unintended interactions) well defined risks that are bound "only" by molecular interactions, no human or natural catastrophe. In those examples, privacy is more like a power plant, subject to human errors and attacks and catastrophes whereas cryptography is more like a drug assessment. This might partly explain the difficulty to create standards and thresholds. Nowadays, it is the state-of-the-art practices to mitigate vulnerabilities that are encouraged. This is putting the responsibility's burden on the regulation authorities, needing them to inspect each system and assess whether or not they are up to the state-of-the-art regarding privacy.

In navigating the challenge of ensuring privacy, in an era of pervasive AI and massive data exploitation, it becomes clear that no single solution (threshold, standard, or auditing mechanism) can universally guarantee adequate protection. Each approach brings its own strengths, limitations, and implications in terms of responsibility, governance, and practicality.

Thresholds offer the promise of clear, measurable guarantees. Like cryptographic key sizes, they could theoretically offer regulators and companies a shared baseline for compliance and protection. Yet, defining such thresholds in privacy is

a deeply complex task since they must account for an unpredictable diversity of attackers, data types, use cases, and evolving technological capabilities. Worse still, they risk becoming obsolete as new threats (e.g. quantum computing) emerge.

Standardization, on the other hand, shifts the burden of assessing and regulating privacy risk from individual organizations to recognized institutions, offering predictability and uniformity. It is a powerful tool for sector-wide alignment, particularly when paired with flexible mechanisms that allow standards to evolve alongside the state of the art. However, its success depends on the clarity, expertise, and agility of standard-setting bodies, qualities not guaranteed across all domains.

Responsibility-based models, such as the "state-of-the-art" requirement in cybersecurity, suggest a third path: rather than prescribing exact methods, regulators demand demonstrable diligence. This opens the door to adaptive, risk-based compliance (via privacy audits or impact assessments for instance). However, in the absence of fixed thresholds or binding standards, it is up to the courts and regulators to bear the heavy burden of post-hoc evaluation, increasing uncertainty for all actors.

What emerges, then, is a hybrid vision. A combination of adaptable thresholds, evolving standards, and demonstrable due diligence may together offer a more resilient and realistic privacy governance model. But any such model must remain aware of its limitations: privacy is not a static endpoint but a moving target, shaped by technological innovation, human error, attacks, and social norms. Whatever framework is chosen, will not eliminate risk entirely, but aims to define and distribute it fairly, transparently, and responsibly.

While synthetic data is often presented as a viable alternative to circumvent

privacy constraints in data-driven research and innovation, it is not the only possible route. An under-explored but potentially valuable alternative is the use of personal data belonging to deceased individuals. This possibility deserves attention, especially in light of their more flexible legal regulation. In the GDPR, the scope of personal data protection applies exclusively to living individuals. Recital 27 of the GDPR explicitly states that: "This Regulation does not apply to the personal data of deceased persons." [109].

However, this absence of European-level protection does not mean there are no applicable rules. In some member states, national laws extend data protection rights beyond death. Notably, France's "Loi Informatique et Libertés" (LIL), amended to align with GDPR, introduces provisions that allow individuals to determine the fate of their personal data post-mortem. According to Article 85 of the LIL : "Individuals may define, during their lifetime, directives regarding the retention, erasure, and communication of their personal data after death." [32]. [1] However, rights are not transferred to the family unless directives state so, meaning the control over one's data ends at death, barring prior instructions.

Despite legal provisions, using deceased persons' data raises questions of legitimacy, particularly when inferred data could affect the memory or dignity of the individual. A classic example is the public disclosure of former French President François Mitterrand's cancer diagnosis after his death [110], highlighting that inferred data remains personal and can impact the honour of the deceased, even when legal protection no longer applies. It is also important to understand that from the data of the deceased some data on the living can be inferred, in this case,

[1] This legal mechanism, while only valid in France, offers a window of legitimacy and clarity in managing data post-mortem, a domain usually under regulated at the European level.

the inferred data are indeed personal and fall under the GDPR jurisdiction. A good example would be some genetic mutations of a deceased person would probably be transmitted to its descendants, thus, the fact that the descendants might suffer from this genetic mutation is inferred and a personal data. Hence, any use of such data must balance ethical sensitivity and public interest. Where harm or reputational damage could occur posthumously, courts can be seized if damages are proven. However, not all jurisdictions treat posthumous harm similarly. This is underscoring, again, the need for clearly defined, anticipatory directives.

Instead of fabricating synthetic data that may lack real-world fidelity, deceased individuals' data (when ethically sourced and legally processed) can serve as a rich, reliable dataset for: medical research (e.g. studying rare diseases or treatment effects), historical and sociological studies, training AI models on real (yet non-infringing) data, etc... With adequate safeguards (such as anonymization because of the possible inference risk, ethical oversight, and respect for existing directives) this approach could complement synthetic data without the same degree of privacy risks. However, it is important to note that those data might be outdated for medical usage. For example, the statistics might not align with the reality anymore.

6 Conclusion

In the past decades, the demand for high-quality data has significantly increased across sectors, especially in AI development. This need has prompted innovations in privacy-preserving data generation, with synthetic data becoming a focal point. This thesis explored the legal and technical status of synthetic data under the EU data protection framework, aiming to assess whether synthetic data can qualify as anonymized data under the GDPR, and if data produced by generative models may be treated as personal data. The research questions this thesis asked in the beginning can now be answered :

What are the existing legal definitions of personal data under EU law?

EU law adopts a broad and inclusive approach to personal data. According to Article 4(1) of the GDPR, personal data refers to any information relating to an identified or identifiable natural person. This includes both direct identifiers (such as names or ID numbers) and indirect ones (such as location data or online identifiers). While talking about re-identification, account should be taken of technical

means, time, cost, and likelihood of identification. Consequently, the qualification of information as personal data is not solely determined by its content but by the risk it poses in a given environment.

Can data generated by generative models be considered personal data under EU law? If so, how is it regulated?

Data generated by generative models may indeed fall under the definition of personal data where it permits, directly or indirectly, the identification of a data subject. This is particularly the case when models have been trained on personal data and retain enough structure to reproduce or infer individual-level characteristics. In such cases, these outputs trigger the full application of GDPR, including obligations related to lawful basis, transparency, purpose limitation, and data subject rights. However, qualification must be done on a case-by-case basis, and regulators currently rely on contextual analysis rather than formal criteria to assess whether generated data should be considered personal.

Can synthetic data be considered a method of anonymization under EU data protection law?

Synthetic data, by nature, is not automatically anonymized. While it may offer enhanced protection when compared to raw or pseudonymized datasets, its qualification as anonymized under the GDPR depends on the residual risk of re-identification. As discussed, small-scale synthetic datasets with no direct correlation to real individuals might be seen as anonymized under certain conditions.

However, when synthetic data is generated in large volumes, particularly via open models trained on personal data, the risk of memorization or indirect inference increases. Regulatory authorities emphasize that anonymization must be effective and irreversible, considering all means "reasonably likely to be used" for re-identification. Therefore, synthetic data must be accompanied by other privacy-preserving mechanisms (such as differential privacy or access controls) to fulfil legal requirements for anonymization.

Is it legally and technically feasible to establish a re-identification risk threshold to determine anonymization? If not, what alternative approaches exist?

As of today, there is no explicit legal threshold that defines an "acceptable" level of re-identification risk. Instead, EU data protection law relies on a qualitative, context-dependent assessment of identifiability. Technically, attempts have been made to define probabilistic thresholds or confidence intervals to quantify risk, but those efforts are challenged by methodological variability and lack of consensus across disciplines. Moreover, the dynamic nature of re-identification threats complicates the establishment of static metrics. While imperfect, these "best effort" practices currently offer the most pragmatic balance between legal expectations and technical feasibility.

In conclusion, this thesis demonstrates that synthetic data should not be understood as a legal silver bullet, but rather as one component within a broader privacy strategy. Its potential to enhance privacy is significant but must be paired

with rigorous anonymization methods. In addition, its legal qualification under the GDPR is still questionable with no definitive answer. In this sense, synthetic data occupies a regulatory grey zone, whose clarity depends on both evolving legal interpretations and advancements in privacy-preserving technologies. As generative models become more powerful and ubiquitous, their regulation under data protection frameworks will only become more critical, calling for multidisciplinary collaboration between legal scholars, engineers, and policymakers in order to define clear legal guidance for regulatory authorities, companies and data subjects.

6.1 Future work

Nearly a decade after its adoption, the GDPR remains the central legal framework governing personal data protection in the European Union. While it continues to offer robust principles (particularly data minimisation, purpose limitation, and accountability), its relevance and adequacy are increasingly questioned and questionable in light of technological advances, notably in AI. The AI Act, delegates personal data governance to the GDPR, but this delegation assumes that the GDPR, a text born in a pre-generative-AI world, remains fit for purpose. As AI capabilities evolve and intersect deeply with privacy concerns, leading to a pressing need to re-evaluate and possibly adapt existing regulatory texts to meet these novel challenges.

A big restrain to a widespread usage of state-of-the-art privacy techniques is the invisible cost of privacy failures. Unlike traditional damages, breaches of privacy often leave no immediate or visible trace, making them hard to quantify and, therefore, easy to deprioritize for companies that do not see the potential cost

of such attacks. This invisibility can reduce the companies' incentive to implement robust privacy safeguards, especially in the absence of strong enforcement or reputational consequences.

Perhaps, from a legal standpoint, these situations could be seen as "force majeure" events (events unpredictable and irresistible) justifying an exemption from liability. However, such legal solutions are rare and take years to build. Most of the time those types of situations will be hard to even prove, especially when it comes to the casual link between a privacy harm (such as identity theft) and its cause (a privacy violation). Facing uncertainties, courts have developed techniques in order to ascertain the existence of a causal link even when it is hard or impossible to demonstrate. The case regarding the Hepatitis B vaccine and the recognition in some instance of its causal link with Multiple Sclerosis [111] is a good example of it.

Despite all the AI's capabilities and the undeniable usefulness of the synthetic data it is important to highlight the fact that AI systems are not all-mighty and possess some flaws [112] [113] [114]. AI models trained on synthetic data may suffer from inaccuracies, biases, or hallucinations, all of which carry legal and ethical implications. Synthetic data might help avoiding using real individuals' data, but its validity, fairness, and utility are far from guaranteed. The fragile balance between privacy and utility becomes even harder to maintain in practice. It is important to underline the fact that education on those possible biases and flaws, whether for companies exploiting those data or for the users, is of the upmost importance. Moreover, synthetic data usage might need some clarification as some companies might use it for the only purpose of evading the GDPR [13]. It is important to highlight the necessity of using this method with other privacy-means

while training the AI that will generate the synthetic data.

Other than that, multiple technical questions remain unanswered. Does repeatedly passing synthetic data through multiple AI models enhances privacy? It remains unclear whether this process ensures anonymization or simply layers additional uncertainties. Another possible problem lies in the algorithm fairness of DP. While DP is powerful, it can disproportionately affect minority groups by injecting noise in ways that obscure small but significant sub-population patterns. In AI systems trained under these conditions, the risk of algorithmic unfairness might increase, undermining both technical and ethical goals or it might also erase those fairness problems by correcting them.

Yet, research is still ongoing on those questions and many others, and important progress are being made. From improved synthetic data generation to context-sensitive DP implementations and evolving standards for AI transparency, the field is far from stagnant. This forward momentum is crucial because non-compliance with privacy regulations, even when it causes no immediate harm, undermines legal norms and public trust. The absence of clear injury today does not mean the absence of violation or future risk.

Beyond individual rights, it would be good to question the amount of data that is collected. Are they all needed? Data is collected across countless platforms, yet aggregation threatens the minimisation principle enshrined in Article 5 of the GDPR. While companies may individually respect the principle, aggregated datasets created through merges or third-party sharing can result in detailed individual profiles far beyond the scope of the original purpose and raising legitimate doubts about whether the minimisation principle, as currently framed, still offers meaningful protection. But even if questionable in its current form, this principle

is critical as the situation would be far worse without such principle.

In parallel, it is important to consider the true necessity of an AI system. Even if it is indeed true that the AI is mandatory in some cases it also seems it has become a socio-technical imaginary [115] [116] driven by economic investment and political momentum rather than concrete societal benefit. The proliferation of AI products, research chairs and public-private partnerships often suggests inevitability, but technological hype and trends should not replace critical evaluation. Some applications of AI may be more fashionable than functional.

However, it is important, while answering technical and regulation problems, not to overlook the importance of education and to continue to raise awareness about not only AI but also GDPR. Both public and professional education around the GDPR, AI, and their intersection is crucial. Without clear understanding, neither individuals nor companies can fully engage with their rights or obligations. Education enables not only compliance but empowerment, turning data subjects into active participants in privacy governance. For instance, even for a text published in 2016, some people are still not aware of even the content of the GDPR nowadays, let alone of the rights this text has created for them.

Finally, some might advocate for an AI de-regulation. What happens if all the legal framework is erased? Without regulatory oversight, the protection of personal data would collapse. The absence of law would also mean the abdication of responsibility, leaving companies to self-regulate, a solution that has repeatedly proven insufficient [117] [118] [119]. In a world where personal data can be monetized, weaponized, or lost in opaque systems, dropping legal safeguards would be a social and ethical failure.

References

[1] Cambridge dictionary, *Personal*, https://dictionary.cambridge.org/fr/dictionnaire/anglais/personal, Accessed: 2025-05-27, 2025.

[2] Cambridge dictionary, *Data*, https://dictionary.cambridge.org/fr/dictionnaire/anglais/data, Accessed: 2025-05-27, 2025.

[3] Cambridge dictionary, *Personal data*, https://dictionary.cambridge.org/fr/dictionnaire/anglais/personal-data, Accessed: 2025-05-27, 2025.

[4] European Parliament and of the Council, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, https://eur-lex.europa.eu/eli/reg/2016/679/oj, Accessed: 2025-05-27, 2016.

[5] U.S. Department of Health and Human Services (HHS), Office for Human Research Protections (OHRP), *Guidance on research involving coded private information or biological specimens*, https://www.hhs.gov/ohrp/

regulations - and - policy / guidance / research - involving - coded - private-information/index.html, Accessed: 2025-05-27, 2008.

[6] L. Sweeney, "K-anonymity: A model for protecting privacy", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002. DOI: 10.1142/S0218488502001648. [Online]. Available: https://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf.

[7] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata", *Science*, vol. 347, no. 6221, pp. 536–539, 2015. DOI: 10.1126/science.1256297. [Online]. Available: https://www.science.org/doi/10.1126/science.1256297.

[8] A.-M. Cretu, M. Rusu, and Y.-A. de Montjoye, "Re-pseudonymization strategies for smart meter data are not robust to deep learning profiling attacks", CODASPY '24, pp. 295–306, Jun. 2024. DOI: 10.1145/3626232.3653272. [Online]. Available: http://dx.doi.org/10.1145/3626232.3653272.

[9] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets", *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pp. 111–125, 2008. DOI: 10.1109/SP.2008.33. [Online]. Available: https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf.

[10] P. P. Tricomi, L. Pajola, L. Pasa, and M. Conti, ""all of me": Mining users' attributes from their public spotify playlists", 2024. arXiv: 2401.14296 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2401.14296.

[11] A. Gadotti, L. Rocher, F. Houssiau, A. M. Creţu, and Y. A. de Montjoye, "Anonymization: The imperfect science of using data while preserving privacy", *Science Advances*, vol. 10, no. 29, eadn7053, 2024, Epub 2024 Jul 17. DOI: 10.1126/sciadv.adn7053. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.adn7053.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis", in *Theory of Cryptography Conference*, Springer, 2006, pp. 265–284. [Online]. Available: https://people.csail.mit.edu/asmith/PS/sensitivity-tcc-final.pdf.

[13] M. Bernelin, "Les données synthétiques en santé, nouveaux enjeux pour le droit ?", 2025.

[14] TAILOR Project, *Ethical and legal framework: High-level expert group guidelines*, http://tailor.isti.cnr.it/handbookTAI/main/Ethical_Legal_Framework/HLEG.html, Accessed: 2025-05-29, 2021.

[15] TAILOR Project, *Tailor handbook of trustworthy ai*, http://tailor.isti.cnr.it/handbookTAI/TAILOR.html, Accessed: 2025-05-29, 2021.

[16] European Parliament and Council, *General data protection regulation (gdpr), recital 26*, Accessed: 2025-05-20, 2016. [Online]. Available: https://gdpr-info.eu/recitals/no-26/.

[17] République Française, *Article r322-3 - code des relations entre le public et l'administration*, https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000031910401, Accessed: 2025-05-27, 2015.

[18] République Française, *Loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé*, https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000031912641, Accessed: 2025-05-27, 2016.

[19] République Française, *Loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé*, Accessed: 2025-05-20, 2016. [Online]. Available: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000031912641/.

[20] European Commission, *Proposal for a regulation on the european health data space*, COM(2022) 197 final, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0197.

[21] Article 29 Data Protection Working Party, *Opinion 05/2014 on anonymisation techniques*, WP216, 2014. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

[22] A. T. P. Boudewijn, A. F. Ferraris, D. Panfilo, *et al.*, "Privacy measurements in tabular synthetic data: State of the art and future research directions", in *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023. [Online]. Available: https://openreview.net/forum?id=D08YT1pt4L.

[23] A. Bkakria, F. Cuppens, N. Cuppens, and A. Tasidou, "Information theoretic-based privacy risk evaluation for data anonymization", *Journal of Surveillance, Security and Safety*, vol. 2, no. 3, 2021, ISSN: 2694-1015. DOI: 10.20517/jsss.2020.20. [Online]. Available: https://www.oaepublish.com/articles/jsss.2020.20.

[24] Commission Nationale de l'Informatique et des Libertés (CNIL), *Décision de la cnil dans l'affaire cegedim*, Accessed: 2025-05-20, 2024. [Online]. Available: https://www.cnil.fr/fr/decisions.

[25] European Parliament and Council of the European Union, *Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance (Data Governance Act)*, https://eur-lex.europa.eu/eli/reg/2022/868/oj, Accessed: 2025-05-27, 2022.

[26] European Parliament and Council of the European Union, *Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847*, https://eur-lex.europa.eu/eli/reg/2025/327/oj, Accessed: 2025-05-27, 2025.

[27] J. Sénéchal, "Publication de l'avis de l'edpb du 17 décembre 2024 sur le traitement des données personnelles dans le contexte des modèles d'ia", *Dalloz Actualité – IP/IT et Communication*, Jan. 17, 2025, Consulté le 13 juin 2025. [Online]. Available: https://www.dalloz-actualite.fr/flash/publication-de-l-avis-de-l-edpb-du-17-decembre-2024-sur-traitement-des-donnees-personnelles-da.

[28] European Parliament and Council, *Regulation (eu) 2024/xxx on artificial intelligence (ai act)*, Article 50(2), 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[29] L. Bernelin, *Synthetic data and the myth of anonymity: Legal and technical challenges*, Preprint in Revue Lamy Droit de l'Immatériel, Forthcoming in Revue Lamy Droit de l'Immatériel, 2023.

[30] European Data Protection Board (EDPB), *Guidelines 05/2024 on Anonymisation Techniques under the GDPR*, https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052024-anonymisation-techniques-under-gdpr_en, Accessed: 2025-05-27, 2024.

[31] Commission Nationale de l'Informatique et des Libertés (CNIL), *Fiches pratiques sur l'intelligence artificielle (IA)*, https://www.cnil.fr/fr/les-fiches-pratiques-ia, Accessed: 2025-05-27, 2024.

[32] Commission Nationale de l'Informatique et des Libertés (CNIL), *Article 85 - loi informatique et libertés n° 78-17 du 6 janvier 1978 modifiée*, https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037420626/, Consolidated version including the GDPR-aligned amendments. Governs post-mortem data rights in France., 2018.

[33] F. Duarte, "Number of chatgpt users (march 2025)", *Exploding Topics Blog*, Mar. 2025, Accessed June 2025; demographics, growth, and usage statistics compiled using Semrush and OpenAI data. [Online]. Available: https://explodingtopics.com/blog/chatgpt-users.

[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977, ISSN: 00359246. [Online]. Available: http://www.jstor.org/stable/2984875 (visited on 07/16/2025).

[35] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988, ISBN: 1558604790.

[36] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[37] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks", 2014. arXiv: 1406.2661 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1406.2661.

[38] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", 2022. arXiv: 1312.6114 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1312.6114.

[39] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics", 2015. arXiv: 1503.03585 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1503.03585.

[40] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[41] Z. Xu, B. Kaszás, M. Cenedese, G. Berti, F. Coletti, and G. Haller, "Data-driven modelling of the regular and chaotic dynamics of an inverted flag from experiments", *Journal of Fluid Mechanics*, vol. 987, May 2024, ISSN: 1469-7645. DOI: 10.1017/jfm.2024.411. [Online]. Available: http://dx.doi.org/10.1017/jfm.2024.411.

[42] N. Carlini, F. Tramèr, E. Wallace, *et al.*, "Extracting training data from large language models", in *30th USENIX Security Symposium (USENIX Security 21)*, USENIX Association, Aug. 2021, pp. 2633–2650, ISBN: 978-1-

939133-24-3. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[43] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models", 2017. arXiv: 1610.05820 [cs.CR]. [Online]. Available: https://arxiv.org/abs/1610.05820.

[44] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks", 2021. arXiv: 2007.14321 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2007.14321.

[45] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures", in *ACM Conference on Computer and Communications Security (CCS)*, 2015, pp. 1322–1333. [Online]. Available: https://rist.tech.cornell.edu/papers/mi-ccs.pdf.

[46] Z. Zhou, J. Zhu, F. Yu, *et al.*, "Model inversion attacks: A survey of approaches and countermeasures", 2024. arXiv: 2411.10023 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2411.10023.

[47] N. Carlini, F. Tramer, E. Wallace, *et al.*, "Extracting training data from large language models", in *USENIX Security Symposium*, 2021. [Online]. Available: https://www.usenix.org/system/files/sec21-carlini-extracting.pdf.

[48] S. Ishihara, "Training data extraction from pre-trained language models: A survey", 2023. arXiv: 2305.16157 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2305.16157.

[49] M. P. M. Parisot, B. Pejo, and D. Spagnuelo, "Property inference attacks on convolutional neural networks: Influence and implications of target model's complexity", 2021. arXiv: 2104.13061 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2104.13061.

[50] M. Chase, E. Ghosh, and S. Mahloujifar, "Property inference from poisoning", 2021. arXiv: 2101.11073 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2101.11073.

[51] Y. Chen *et al.*, "Feature inference attack on shapley values", *arXiv preprint arXiv:2407.11359*, 2024. [Online]. Available: https://arxiv.org/abs/2407.11359.

[52] X. Luo, Y. Jiang, and X. Xiao, "Feature inference attack on shapley values", CCS '22, pp. 2233–2247, Nov. 2022. DOI: 10.1145/3548606.3560573. [Online]. Available: http://dx.doi.org/10.1145/3548606.3560573.

[53] Encord, *Datasets - Definition and Explanation*, https://encord.com/glossary/datasets-definition/, Accessed: 2025-05-27, 2024.

[54] D. Desfontaines, *Local vs. central differential privacy*, https://desfontain.es/blog/local-global-differential-privacy.html, Ted is writing things (personal blog), Jun. 2019.

[55] Commission Nationale de l'Informatique et des Libertés (CNIL), *Modèle IA - Définition*, https://www.cnil.fr/fr/definition/modele-ia, Accessed: 2025-05-27, 2024.

[56] M. BERNELIN, *Droit de l'intelligence artificielle: Applications, enjeux et perspectives en santé*, Course materials, 2025.

[57] IBM. "Neural networks". Accessed: 2025-06-12. (), [Online]. Available: https://www.ibm.com/think/topics/neural-networks.

[58] S. Cavuoti, D. De Cicco, L. Doorenbos, *et al.*, "Identification of problematic epochs in astronomical time series through transfer learning", *Astronomy amp; Astrophysics*, vol. 687, A246, Jul. 2024, ISSN: 1432-0746. DOI: 10.1051/0004-6361/202450166. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/202450166.

[59] Q. Lyu, K. Shridhar, C. Malaviya, *et al.*, "Calibrating large language models with sample consistency", 2024. arXiv: 2402.13904 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2402.13904.

[60] Commission Nationale de l'Informatique et des Libertés (CNIL), *Tenir compte de la protection des données dans la collecte et la gestion des données - Fiche pratique IA n°8*, https://www.cnil.fr/fr/tenir-compte-de-la-protection-des-donnees-dans-la-collecte-et-la-gestion-des-donnees, Accessed: 2025-05-27, 2022.

[61] M. Abadi, A. Chu, I. Goodfellow, *et al.*, "Deep learning with differential privacy", in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS'16, ACM, Oct. 2016, pp. 308–318. DOI: 10.1145/2976749.2978318. [Online]. Available: http://dx.doi.org/10.1145/2976749.2978318.

[62] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization", 2011. arXiv: 0912.0071 [cs.LG]. [Online]. Available: https://arxiv.org/abs/0912.0071.

[63] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis", in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284, ISBN: 978-3-540-32732-5.

[64] Wikipedia, *Prompt engineering*, https://en.wikipedia.org/wiki/Prompt_engineering, Accessed: 2025-05-27, 2025.

[65] N. Carlini, F. Tramer, E. Wallace, *et al.*, "Extracting training data from large language models", 2021. arXiv: 2012.07805 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2012.07805.

[66] A. Salem, G. Cherubin, D. Evans, *et al.*, "Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning", 2023. arXiv: 2212.10986 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2212.10986.

[67] Z. Ji, N. Lee, R. Frieske, *et al.*, "Survey of hallucination in natural language generation", *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2022. DOI: 10.1145/3571730.

[68] L. Huang, W. Yu, W. Ma, *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions", *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, Jan. 2025, ISSN: 1558-2868. DOI: 10.1145/3703155. [Online]. Available: http://dx.doi.org/10.1145/3703155.

[69] M. Hall, L. van der Maaten, L. Gustafson, M. Jones, and A. Adcock, "A systematic study of bias amplification", 2022. arXiv: 2201.11706 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2201.11706.

[70] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies", *Sci*, vol. 6, no. 1, p. 3, Dec. 2023, ISSN: 2413-4155. DOI: 10.3390/sci6010003. [Online]. Available: http://dx.doi.org/10.3390/sci6010003.

[71] T. Sharot, "Bias in ai amplifies our own biases", *Nature Human Behaviour*, 2024, Study finds AI can increase user bias in perceptions of gender and status. [Online]. Available: https://www.ucl.ac.uk/news/2024/dec/bias-ai-amplifies-our-own-biases.

[72] F. Fioretto, C. Tran, P. Van Hentenryck, and K. Zhu, "Differential privacy and fairness in decisions and learning tasks: A survey", in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, ser. IJCAI-2022, International Joint Conferences on Artificial Intelligence Organization, Jul. 2022, pp. 5470–5477. DOI: 10.24963/ijcai.2022/766. [Online]. Available: http://dx.doi.org/10.24963/ijcai.2022/766.

[73] E. Bagdasaryan and V. Shmatikov, "Differential privacy has disparate impact on model accuracy", 2019. arXiv: 1905.12101 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1905.12101.

[74] F. Fioretto, C. Tran, and P. V. Hentenryck, "Decision making with differential privacy under a fairness lens", 2024. arXiv: 2105.07513 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2105.07513.

[75] C. Tran, M. H. Dinh, and F. Fioretto, "Differentially empirical risk minimization under the fairness lens", 2022. arXiv: 2106.02674 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2106.02674.

[76] K. Tran, F. Fioretto, I. Khalil, M. T. Thai, and L. T. X. P. N. Phan, "Fairdp: Certified fairness with differential privacy", 2025. arXiv: 2305.16474 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2305.16474.

[77] Z. Yang, Y. Ge, C. Su, D. Wang, X. Zhao, and Y. Ying, "Fairness-aware differentially private collaborative filtering", WWW '23, pp. 927–931, Apr. 2023. DOI: 10.1145/3543873.3587577. [Online]. Available: http://dx.doi.org/10.1145/3543873.3587577.

[78] S. Kim, Y. Roh, G. Heo, and S. E. Whang, "Pfguard: A generative framework with privacy and fairness safeguards", 2025. arXiv: 2410.02246 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2410.02246.

[79] Commission Nationale de l'Informatique et des Libertés (CNIL), *Tenir compte de la protection des données dans la conception du système*, Online practical guide, CNIL "Fiches IA" series, Accessed via CNIL website; outlines steps to integrate data protection in AI system design, 2024. [Online]. Available: https://www.cnil.fr/fr/tenir-compte-de-la-protection-des-donnees-dans-la-conception-du-systeme.

[80] Wikipedia, *Encryption*, Wikipedia, The Free Encyclopedia, Accessed April 2025, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Encryption.

[81] M. Ebers, "Truly risk-based regulation of artificial intelligence how to implement the eu's ai act", *European Journal of Risk Regulation*, pp. 1–20, 2024. DOI: 10.1017/err.2024.78.

[82] N. Rangone and L. Megale, "Risks without rights? the eu AI act's approach to ai in law and rule-making", *European Journal of Risk Regulation*, pp. 1–

16, 2025, Published online 13 March 2025. DOI: 10.1017/err.2025.13. [Online]. Available: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/3AD4822C291C6591BAFD26524CD44C12/S1867299X25000133a.pdf/risks-without-rights-the-eu-ai-acts-approach-to-ai-in-law-and-rule-making.pdf.

[83] National Institute of Standards and Technology (NIST), "Transitions: Recommendation for transitioning the use of cryptographic algorithms and key lengths", U.S. Department of Commerce, Tech. Rep. SP 800-131A Rev. 2, 2022, Specifies minimum acceptable key lengths and algorithm transition schedules. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-131Ar2.

[84] National Institute of Standards and Technology (NIST), "Announcing the advanced encryption standard (aes)", Federal Information Processing Standards Publication, Tech. Rep. FIPS PUB 197, 2001, Establishes AES and key sizes of 128, 192, and 256 bits. [Online]. Available: https://csrc.nist.gov/publications/detail/fips/197/final.

[85] U.S. Bureau of Industry and Security (BIS), *Export administration regulations, category 5—part 2: Information security*, Specifies encryption strength thresholds for export control, 2023. [Online]. Available: https://www.bis.doc.gov/index.php/documents/regulations-docs/2336-category-5-part-2/file.

[86] Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI), *Référentiel général de sécurité (rgs), version 2.0 — annex b: Algorithms and key lengths*, Defines acceptable cryptographic parameters for French public sec-

tor use, 2014. [Online]. Available: https://www.ssi.gouv.fr/uploads/IMG/pdf/RGS_B1.pdf.

[87] CNIL, Inria, IMT Atlantique, INSA Lyon and Université de Rennes, *Ipop project – privacy impact assessment and privacy-preserving tools*, Collaborative French research project aiming to evaluate and strengthen privacy guarantees through attack simulation and formal tools, 2023. [Online]. Available: https://www.cnil.fr/fr/le-projet-ipop-un-projet-pour-evaluer-la-protection-des-donnees-personnelles.

[88] TAPAS Project Contributors, *Tapas library of privacy attacks*, Open-source collection of privacy attacks for evaluating AI models and systems, 2024. [Online]. Available: https://tapas-privacy.readthedocs.io/en/latest/library-of-attacks.html.

[89] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles", 2022. arXiv: 2112.03570 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2112.03570.

[90] Octopize, *Octopize – ethical anonymization with avatar-based synthetic data*, French start-up offering GDPR-compliant anonymous synthetic data; CNIL-assessed technology, 2025. [Online]. Available: https://www.octopize.io/.

[91] Wikipedia, *Differential privacy*, https://en.wikipedia.org/wiki/Differential_privacy, Accessed June 10, 2025, 2025.

[92] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy", 2015. arXiv: 1311.0776 [cs.DS]. [Online]. Available: https://arxiv.org/abs/1311.0776.

[93] T. Steinke and J. Ullman, *Why privacy needs composition*, Explains interactive and non–interactive composition in DP, 2020. [Online]. Available: https://differentialprivacy.org/privacy-composition/.

[94] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting", 2018. arXiv: 1709.01604 [cs.CR]. [Online]. Available: https://arxiv.org/abs/1709.01604.

[95] J. Jin, E. McMurtry, B. I. P. Rubinstein, and O. Ohrimenko, "Are we there yet? timing and floating-point attacks on differential privacy systems", 2024. arXiv: 2112.05307 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2112.05307.

[96] National Institute of Standards and Technology, "Federal information processing standards publication 197: Advanced encryption standard (aes)", U.S. Department of Commerce, Gaithersburg, MD, FIPS Publication 197, Nov. 2001, Updated May 9, 2023. [Online]. Available: https://csrc.nist.gov/publications/detail/fips/197/final.

[97] U. Jonsson and B. Kaliski, "PKCS #1: RSA Cryptography Specifications Version 2.2 (RFC 8017)", Internet Engineering Task Force, RFC 8017, Nov. 2016. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc8017.

[98] National Institute of Standards and Technology (NIST), "Draft fips 203: Module-lattice-based key-encapsulation mechanism standard", U.S. Department of Commerce, Draft FIPS Publication 203, Aug. 2023. [Online]. Available: https://csrc.nist.gov/publications/detail/fips/203/draft.

[99] CEN–CENELEC, *Cen–cenelec*, Official website of the European Committee for Standardization and European Committee for Electrotechnical Standardization, Accessed June 2025, 2025. [Online]. Available: https://www.cencenelec.eu/.

[100] Geneva, Switzerland. [Online]. Available: https://www.iso.org.

[101] Geneva, Switzerland. [Online]. Available: https://www.iec.ch.

[102] Standardization in the field of information technology. [Online]. Available: https://www.iso.org/isoiec-jtc-1.html.

[103] Subcommittee for Computer Graphics, Image Processing, and Interaction. [Online]. Available: https://www.iso.org/committee/45382.html.

[104] New York, USA. [Online]. Available: https://www.ieee.org.

[105] International Organization for Standardization and International Electrotechnical Commission, *Iso/iec 27000-family – information security management systems (overview and vocabulary)*, Online resource, Overview of the ISO/IEC 27000 family of information security standards, 2018. [Online]. Available: https://www.iso.org/fr/standard/iso-iec-27000-family.

[106] J. Soler Garrido, S. De Nigris, E. Bassani, *et al.*, "Harmonised standards for the european ai act", European Commission, Joint Research Centre, Seville, Spain, JRC Reference Report JRC139430, 2024, see pages 5–6. [Online]. Available: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC139430/JRC139430_01.pdf.

[107] European Parliament and Council, *Directive (eu) 2016/1148 of the european parliament and of the council of 6 july 2016 concerning measures for a high common level of security of network and information systems across the union*, Official Journal of the European Union, L 194/1–30, 19 July 2016, First EU-wide cybersecurity directive ("NIS Directive"), 2016. [Online]. Available: `https://eur-lex.europa.eu/eli/dir/2016/1148/oj`.

[108] European Parliament and Council, *Directive (eu) 2022/2555 of the european parliament and of the council of 14 december 2022 on measures for a high common level of cybersecurity across the union, repealing directive 2016/1148 (nis 2)*, Official Journal of the European Union, L 333/80–152, 27 December 2022, Repeals and replaces the original NIS Directive; to be transposed by 17 October 2024 :contentReference[oaicite:1]index=1, 2022. [Online]. Available: `https://eur-lex.europa.eu/eli/dir/2022/2555/oj`.

[109] European Parliament and Council of the European Union, *Recital 27 - regulation (eu) 2016/679 of the european parliament and of the council (general data protection regulation)*, `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, Recital 27 clarifies that the GDPR does not apply to the personal data of deceased persons., 2016.

[110] C. Gubler and M. Gonod, *Le Grand Secret*. Paris: Éditions Plon, 1996, Reveals details about President François Mitterrand's concealed prostate cancer diagnosis during his presidency., ISBN: 9782259185802.

[111] European Court of Justice, *Case c-621/15, n.w. and others v. sanofi pasteur msd snc and others*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:62015CJ0621, Judgment of the Court (Second Chamber) of 21 June 2017, 2017.

[112] Z. Li, "The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination", 2023. arXiv: 2304.14347 [cs.CY]. [Online]. Available: https://arxiv.org/abs/2304.14347.

[113] S. M. Williamson and V. Prybutok, "The era of artificial intelligence deception: Unraveling the complexities of false realities and emerging threats of misinformation", *Information*, vol. 15, no. 6, 2024, ISSN: 2078-2489. DOI: 10.3390/info15060299. [Online]. Available: https://www.mdpi.com/2078-2489/15/6/299.

[114] M. Alabi, "Ethical implications of ai: Bias, fairness, and transparency", Nov. 2024. [Online]. Available: https://www.researchgate.net/publication/385782076_Ethical_Implications_of_AI_Bias_Fairness_and_Transparency.

[115] A. Markelius, C. Wright, J. Kuiper, N. Delille, and Y.-T. Kuo, "The mechanisms of ai hype and its planetary and social costs", *AI and Ethics*, vol. 4, pp. 727–742, 2024. DOI: 10.1007/s43681-024-00461-2. [Online]. Available: https://doi.org/10.1007/s43681-024-00461-2.

[116] Z. Sramek and K. Yatani, *Research as resistance: Recognizing and reconsidering hci's role in technology hype cycles*, 2025. arXiv: 2504.08336 [cs.HC]. [Online]. Available: https://arxiv.org/abs/2504.08336.

[117] World Privacy Forum, *Many failures: A brief history of privacy self-regulation*, Online, https://www.worldprivacyforum.org/2020/10/report-many-failures-conclusion/, 2020.

[118] The Guardian, "Ftc: Social media companies' surveillance practices raise privacy concerns", *The Guardian*, 2024, Accessed May 2025. [Online]. Available: https://www.theguardian.com/technology/2024/sep/19/social-media-companies-surveillance-ftc.

[119] Electronic Privacy Information Center (EPIC), *Hearing on big data: Privacy risks and needed reforms in the public and private sectors*, Online, https://epic.org/documents/hearing-on-big-data-privacy-risks-and-needed-reforms-in-the-public-and-private-sectors/, 2023.

[120] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity", in *2007 IEEE 23rd International Conference on Data Engineering*, 2007, pp. 106–115. DOI: 10.1109/ICDE.2007.367856.

Appendix A : A quick anonymization history

A.1 Pseudonymization

Pseudonymization is one of the earliest data protection techniques, where direct identifiers are replaced with artificial ones (pseudonyms), so that data can no longer be attributed to a specific individual without additional information. According to Article 4(5) of the GDPR [4], "Pseudonymisation means the processing of personal data in such a manner that the data can no longer be attributed to a specific data subject without the use of additional information." It is important to note that under the GDPR, pseudonymized data is still considered personal data, unlike data processed through stronger anonymization techniques.

A simple example of pseudonymization would be replacing a name like "Alice" with "person1". To enable re-identification, the mapping between real identities and pseudonyms is stored separately—ideally on a different server—to minimize the risk in case of a data breach. Unlike anonymization, pseudonymization allows for the re-identification of individuals through the use of a key or mapping table, making it a reversible process.

Pseudonymization began to be discussed as a privacy-preserving technique in the late 1980s. However, in 2016, the GDPR [4] did not recognized it as such, thus, pseudonymized data are still considered as personal data.

However, its limitations became apparent early on. A notable example is the case from the mid-1990s involving the Massachusetts Group Insurance Commission (GIC), which released anonymized health data of state employees, including then-Governor William Weld. Although direct identifiers like names and addresses were removed, the dataset retained quasi-identifiers such as ZIP code, birth date, and gender. Researcher Latanya Sweeney demonstrated that 87% of the U.S.

population could be uniquely identified using just those three attributes [6]. By cross-referencing the dataset with publicly available voter registration records, she successfully re-identified Governor Weld's medical information.

This case is a seminal example of the vulnerabilities inherent in pseudonymization and demonstrates how powerful re-identification attacks can be when auxiliary data is available. Even if a dataset appears secure in isolation, the abundance of external data sources today can make re-identification trivial.

While pseudonymization reduces the risk of identification, it does not offer complete anonymity and must be used carefully, especially when other data sources can be exploited.

A.2 K-anonymity

To improve upon the limitations of pseudonymization, researchers developed techniques such as k-anonymity, which aims to make individual records indistinguishable from at least k-1 others with respect to a set of quasi-identifiers (QIDs). Quasi-identifiers are combinations of attributes that may not directly identify an individual but can do so when combined (such as ZIP code, age, and gender).

The principle behind k-anonymity involves a multi-step transformation of the dataset. First, all direct identifiers like names or social security numbers are removed. Then, the records are grouped so that each group contains at least k entries sharing similar values for the quasi-identifiers. Finally, these quasi-identifiers are either generalized or suppressed to make individuals within a group indistinguishable from one another.

Consider the patient database: Table A.1.

Name	**ZIP**	**Age**	**Gender**	**Disease**
Bob	02138	29	*Male*	colon cancer
Georges	02138	29	*Male*	colon cancer
Thomas	02139	45	*Male*	colon cancer
Alice	02140	65	*Female*	Blood cancer

Table A.1: Basic patient dataset

To achieve 2-anonymity, the data must be modified so that each combination of quasi-identifiers appears in at least two records. This can be done by generalizing the ZIP codes, ages, and even genders, depending on the distribution of the data. The resulting dataset might look like the Table A.2.

ZIP	**Age**	**Gender**	**Disease**
02138	29	*Male*	colon cancer
02138	29	*Male*	colon cancer
[02139, 02140]	[45, 65]	[*Male*, *Female*]	colon cancer
[02139, 02140]	[45, 65]	[*Male*, *Female*]	blood cancer

Table A.2: 2-anonymous version of the dataset

While this version of the dataset satisfies the definition of 2-anonymity, it still presents vulnerabilities. For example, if someone knows that Bob is 29 years old and lives in ZIP code 02138, then it is still possible to confidently deduce that he has HIV, even though the record cannot be directly linked to him. This scenario illustrates a type of privacy breach known as a homogeneity attack.

In a homogeneity attack, all the records within a k-anonymous group share the same value for a sensitive attribute. As a result, merely knowing that someone belongs to the group is enough to reveal the sensitive information. K-anonymity alone is not sufficient when the sensitive attribute lacks diversity within a group.

Another well-known vulnerability of k-anonymity is the background knowledge attack. In this scenario, an attacker leverages auxiliary information to reduce

uncertainty about an individual's record. Suppose an anonymized group of four individuals contains one person who does not smoke, while the rest do. If the attacker already knows that Alice does not have colon cancer, they can immediately infer which record corresponds to her, defeating the purpose of anonymization.

These types of attacks expose two fundamental weaknesses of k-anonymity : the lack of diversity in sensitive attributes within groups and the model's sensitivity to external or background information. In both homogeneity and background knowledge attacks, anonymity fails because quasi-identifiers are not sufficiently varied and because attackers can use additional knowledge to bypass the protection.

Due to these limitations, k-anonymity is often combined with stronger privacy models, such as l-diversity or t-closeness, which aim to address these specific weaknesses by ensuring greater variation in sensitive values and accounting for background knowledge.

A.2.1 L-diversity

To address the weaknesses of k-anonymity, particularly its vulnerability to homogeneity and background knowledge attacks, researchers proposed another paradigm l-diversity. This model aims to ensure that within each group of k-anonymous records, the sensitive attribute contains at least l "well-represented" distinct values. The idea is that even if an attacker can isolate a group of records based on quasi-identifiers, they should not be able to confidently infer the sensitive attribute of any individual in the group. For example, in a medical dataset, each group should contain multiple differing values (here diseases) such as "colon cancer" and "blood cancer" thereby reducing the certainty of inference.

While l-diversity offers a stronger privacy guarantee than k-anonymity, it is not immune to attacks. One notable limitation is the so-called similarity attack, which exploits what is known as the semantic weakness of l-diversity. In this scenario, a group may appear diverse because it contains multiple distinct values for the sensitive attribute, yet those values may be semantically very similar. For instance, a group containing the diagnoses "colon cancer" and "blood cancer" technically satisfies 2-diversity. However, an attacker still learns that the person likely has some form of cancer. Because l-diversity focuses purely on the number of distinct values and not on their semantic distance or meaning, it cannot distinguish between genuinely diverse information and superficially diverse but semantically close data. Avoiding such leakage requires the data handler to have a deep understanding of the dataset and to ensure that the sensitive attributes included in each group are truly diverse in meaning, not just in form.

ZIP	**Age**	**Gender**	**Disease**
$[02138, 02140]$	$[29, 60]$	$[Male]$	colon cancer
$[02138, 02140]$	$[29, 60]$	$[Male]$	colon cancer
$[02138, 02140]$	$[29, 60]$	$[Male]$	colon cancer

Table A.3: 3-diverse dataset, where a line had to be deleted to protect Alice's data (a line for blood cancer could also have been added)

The problem with our example (Table A.3) is that the utility of our dataset is really limited now. Moreover, the disease probability repartition has been impacted : we could think that the patients have 100% chances to have either colon cancer, when in reality, it is less than 10% (for instance).

Which leads to another significant vulnerability : the skewness attack. This attack arises from the distribution of sensitive values within the population and exploits situations where a sensitive attribute is rare in the general population but

disproportionately represented in a specific anonymized group. For example, if only 1% of the population has HIV, and one out of three individuals in a supposedly 3-diverse group has HIV, the conditional probability that any member of that group has HIV becomes much higher than it would be in the general population. In this case, even though the group technically satisfies the requirement of l-diversity, the presence of the rare value creates a strong inference channel. This happens because l-diversity only considers diversity within each group and does not consider the overall distribution of sensitive values across the dataset.

Therefore, while l-diversity significantly improves upon k-anonymity by enforcing per-group diversity in sensitive attributes, it still suffers from important limitations. The effectiveness of this model depends not only on achieving numeric diversity but also on ensuring semantic independence and accounting for global distribution. As such, more advanced models like t-closeness [120] have been proposed to further reduce the risk of disclosure.

A.3 Differential Privacy

In 2006 Cynthia Dwork and al. came with an idea : Differential Privacy (DP) [12]. This model is designed to protect individuals' private information when datasets are analysed or shared. The core principle is that the presence or absence of a single individual's data (say, John's) should not significantly affect the outcome of any analysis. In other words, whether or not John's data is included, the results should appear nearly the same, making it extremely difficult to determine if he was part of the dataset at all. However, DP does not work on static data (such as tabular data etc...) but on aggregated data (like query answers : mean, sum,

count, etc...). Therefore, it is not used in the same context as k-anonymity or l-diversity.

To achieve this, DP introduces carefully calibrated random noise to the queries' output, ensuring individual contributions remain hidden while still allowing useful insights from the aggregate data. Dwork et al. [12] formalized this idea using the Laplace mechanism, which adds noise scaled to the sensitivity (whether the data is highly personal such as sexual orientation, personal, or non-personal) of the function being computed. The result is a statistical cloak—effectively hiding the "tree" in a forest of plausible alternatives. While Differential Privacy provides robust theoretical guarantees, it is most effective on aggregate queries (such as counts or averages) and may be vulnerable to sophisticated attacks if misapplied or used on high-dimensional data.