

AI Memorisation and Anonymisation under the GDPR Leave Data Protection out of AI Models

Francisco Arga e Lima

Abstract

This paper examines the extent to which the General Data Protection Regulation (GDPR) applies to artificial intelligence (AI) models in light of the inherent memorisation of data during their training. Central to this inquiry is the concept of anonymisation and whether it should be assessed under an “objective” standard, requiring irreversibility for any party, or a “subjective” one, dependent on the means reasonably available to a specific controller.

We do so to demonstrate that using an objective standard of anonymisation to AI models would render them subject to the GDPR, regardless of best measures adopted by developers. If anonymisation is interpreted that way, then developers are met with the impossible choice of, either, adopting privacy-preserving techniques to anonymise data but make it lose its utility or degrade the model in the process, or conform themselves to have their model subject to the GDPR and run the risk of being confronted with the situation of i.e. having to fulfil data subject rights when they cannot, themselves, extract personal data stored in the model. We argue instead for a contextual, subjective assessment of identifiability: if specific third parties that have access to the model lack reasonable means to access memorised personal data, the model should be regarded as anonymised to them.

Introduction

The application of the General Data Protection Regulation (“GDPR”) depends on data falling under the definition of personal data, as enshrined in Article 4(1) of the GDPR. This means the information must relate to a natural person.¹ Crucially, the natural person needs to be identified or identifiable. This is assessed in a broad and contextual way, as Article 4(1) and Recital 26 of the GDPR allow for indirect identification, where the controller can reasonably identify the data subject, with additional information obtained elsewhere.²

¹ Court of Justice of the European Union, ‘Judgment of the Court, Case C-434/16 (Nowak)’ (2017) para 34; Court of Justice of the European Union, ‘Opinion of Advocate General Sharpston, Joined Cases C-141/12 and C-372/12 (YS)’ (2013) paras 43–49; Lee A Bygrave and Luca Tosoni, ‘Article 4(1). Personal Data’ in Christopher Kuner, Lee A Bygrave and Christopher Docksey (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020) 109.

² Bygrave and Tosoni (n 1) 110–111.

Although easy to understand in theory, the line that separates re-identifiable data from that which no longer possesses this link – in other words, anonymous data³ – is not so clear in practice. The 2007 Article 29 Working Party Opinion adopted a subjective approach – what we will call “subjective anonymisation” – holding that an hypothetical possibility of reidentification would not be enough to classify data as personal.⁴ This meant that the risk of identification should be assessed from the perspective of the controller, an arguably feasible standard: if a party had no realistic access to information that would allow re-identification, the data should be considered anonymised to that party.⁵ However, subsequent guidance in 2014 shifted towards a more objective criterion: what we will call “objective anonymisation”. Here, anonymisation was defined as a process that renders reidentification impossible without disproportional efforts for any party, as reinforced in Breyer.⁶ This clash of standards is not yet solved, creating regulatory grey areas where it is not clear whether the GDPR applies, despite controllers’ best efforts to anonymise personal data. A notable recent development in this context came through the CJEU’s judgement in *SRB v. EDPS*,⁷ where it appeared to endorse a subjective standard of anonymisation. But more on that later.

This regulatory grey area creates significant uncertainty for controllers, made worse when one of the standards is unreachable. Indeed, objective anonymisation is often unattainable, since, given sufficient auxiliary data or technological progress, reidentification cannot be ruled out entirely and almost everything becomes, or could become, personal. This unreasonableness is all too clear in the AI sector, just by the inherent way AI models work. Given they memorise data and given recent advances in extraction techniques, it may never be possible to demonstrate that it is objectively impossible to reidentify data stored in the model itself. In other words, if the objective standard is used, the AI model itself is subject to the GDPR, regardless of the anonymisation efforts taken by the developer. This leads to a paradox: controllers are called to minimise and anonymise data, but the GDPR doesn’t allow them to do so in practice.

³ Luca Tosoni, ‘Article 4(5). Pseudonymisation’ in Christopher Kuner, Lee A Bygrave and Christopher Docksey (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020) 135–136.

⁴ Lore Leitner, Gabe Maldoff and Mickey Lee, ‘Anonymisation through Separation: What Recent Cases Teach Us about the EU’s Anonymisation Standards’ (2024) 24 *Privacy & Data Protection* 10, 10; Article 29 Working Party, ‘Opinion 4/2007 on the Concept of Personal Data’ (2007) 15.

⁵ Leitner, Maldoff and Lee (n 4) 10–13.

⁶ Court of Justice of the European Union, ‘Judgment of the Court, Case C-582/14 (Breyer)’ (2016) paras 35–49; Leitner, Maldoff and Lee (n 7) 11–13; Article 29 Working Party, ‘Opinion 05/2014 on Anonymisation Techniques’ (2014) 5, 7.

⁷ Court of Justice of the European Union, ‘Judgment of the Court, Case C-413/23 P (*SRB v. EDPS*)’ (2025).

Therefore, our aim is to show that, given AI models' inherent memorisation, there is a pressing need to relativise the concept of anonymisation. This thesis is grounded on three premises. First, the paper exclusively addresses the processing of personal data related to AI memorisation and whether it is possible to consider the model anonymised. While the concerns raised here may also be relevant to the processing of personal data in other phases (i.e. output generation post-deployment), their specific causes and functioning warrant more specific discussion, which is beyond the scope of this paper. Second, it is assumed that these models are developed through a standardised training process involving exposure to vast quantities of data. Without it, the concerns relating to memorisation would not arise. Third, given the nature of internet-scale data scraping and the use of diverse textual corpora for training purposes, the inclusion of personal data is to be expected.⁸ The relevance of data protection law hinges on this inclusion: if the training corpus was devoid of personal data, the application of the GDPR would not arise in the first place.

How memorisation works

Before explaining why it is unfeasible for AI models to be objectively anonymised, we need to understand what memorisation is, and for that we must understand how AI models are trained. In simple terms, the process starts with the abstraction of raw text into digital representations capable of being statistically modelled. For this to occur, developers tokenize raw data, turning them into smaller numerical identifiers – tokens – that will be the basis for embeddings.⁹ These embeddings then serve as the model's core representational mechanism, comprising multidimensional vectors derived from token sequences. It is through them that models 'learn' how tokens relate to one another by capturing contextual similarities and differences.¹⁰ They help, i.e. distinguish between meanings of homonyms, serving two primary functions: to process input data and to generate output based on token relations learnt during the training process.¹¹

This means that AI training is essentially a process of statistical abstraction of large corpora of tokenized data into patterns of mathematical significance. This is what we can call encoding,

⁸ Rachel Hong and others, 'A Common Pool of Privacy Problems: Legal and Technical Lessons from a Large-Scale Web-Scrapped Machine Learning Dataset' (2025) arXiv:2506 arXiv preprint 1–2, 23–24.

⁹ Daniel Gervais and others, 'The Heart of the Matter: Copyright, AI Training, and LLMs' [2024] SSRN 2–4.

¹⁰ Confederation of European Data Protection Organizations, 'Generative AI: The Data Protection Implications' (2023) 6–7; Gervais and others (n 12) 3–4; Information Commissioner's Office, 'Explaining Decisions Made with AI' (2022).

¹¹ Lucas Bourtole and others, 'Machine Unlearning', *42nd IEEE Symposium on Security and Privacy* (2021) 1–4; Information Commissioner's Office (n 10).

as the abstraction of high-frequency patterns, with the primary goal of compressing data while retaining enough latent structure to allow the AI model to produce accurate results.¹² However, this feature has limits, since it cannot uniformly discard detail. During training, the model's parameters are iteratively adjusted to weight meaningful correlations between tokens. When the same sequence reappears frequently, parameters stabilise around the values that reinforce its probability, to some extent hardcoding it into the model.¹³ This leads to a second phenomenon, where certain data sequences persist after training in near-verbatim form, because their frequency or overrepresentation skews probability distributions in their favour.¹⁴ This is what we call memorisation.¹⁵

Encoding and memorisation are therefore not independent processes but exist on a spectrum, with encoding being similar to lossy compression, and memorisation being closer to lossless reproduction of content. This behaviour is also not accidental. AI models inherently depend on a balance between general pattern abstraction and selective retention. If not, they would not be able to i.e. spell words correctly. However, while this is very useful in preserving language structures, it creates problems when it leads to the memorisation of personal data.¹⁶

Bringing this back to the data protection realm, the presence of personal data within AI models is thus mainly dependent on the existence of means that reasonably allow access to memorised content. There are two main ways to achieve this.

Firstly, by analysing the model's parameters. However, at the technical level, they are generally not designed for direct retrieval, as these billions of values do not equate to specific data entries. Instead, they work as coefficients within a probabilistic function that maps likely sequences of

¹² Confederation of European Data Protection Organizations (n 10) 6–7; A Feder Cooper and James Grimmelmann, 'The Files Are in the Computer: Copyright, Memorization, and Generative-AI Systems' (2024) arXiv:2404 arXiv preprint 35–38.

¹³ Bourtole and others (n 14) 1–4; Cooper and Grimmelmann (n 15) 35–38, 45–50; Der Landesbeauftragte für Datenschutz und Informationsfreiheit Baden-Württemberg, 'Diskussionspapier: Rechtsgrundlagen Im Datenschutz Beim Einsatz von Künstlicher Intelligenz' (2023) 6–7.

¹⁴ Commission Nationale de l'Informatique et des Libertés, 'Relying on the Legal Basis of Legitimate Interests to Develop an AI System' (2024) <<https://www.cnil.fr/en/relying-legal-basis-legitimate-interests-develop-ai-system>> accessed 25 September 2025; Confederation of European Data Protection Organizations (n 11) 6–7; Bourtole and others (n 12) 1–5; Cooper and Grimmelmann (n 13) 35–38, 45–50; Der Landesbeauftragte für Datenschutz und Informationsfreiheit Baden-Württemberg (n 14) 6–7; Dimitri Staufer, 'What Should LLMs Forget? Quantifying Personal Data in LLMs for Right-to-Be-Forgotten Requests' (2025) arXiv:2507 arXiv preprint 3.

¹⁵ Staufer (n 14) 1.

¹⁶ *ibid.*

tokens.¹⁷ This means that memorised data is not visibly discrete, since it exists mostly as probability peaks within the parameter space.¹⁸

However, a dataset does not cease to contain personal data simply because it cannot be understood in isolation, regardless of the anonymisation standard we use. In fact, and similarly to encrypted data, interpretability is dependent on the existence of tools that can realistically extract or decrypt the underlying content. The CJEU and the European Data Protection Board both point in this direction, by clarifying that identifiability is a risk-based concept, taking into consideration the means that could reasonably be used to reidentify a data subject.¹⁹ This fluidity makes the concept of anonymisation (whether subjective or objective) contingent on technological capacity, as personal data stored today in an unintelligible format might become recoverable in the future. The implication is that, if memorisation occurs, and technological progress allows us to retrieve its content in the future by looking at the parameters, then the model cannot be classified as anonymised *ad eternum* because current tools are insufficient to extract it. The information is there we just can't access it for now.²⁰

Regardless, targeted prompt attacks have succeeded in recovering fragments from the model's training data, showing that it is still feasible to extract memorised data through other means.²¹ While some may argue these attacks as outliers,²² if we use an objective anonymisation standard, this claim does not negate that the model is storing personal data. First, identifiability would depend on the potential use of reasonably available means by any party.²³ Therefore, if adversarial users are able to consistently obtain personal data from AI models, that would establish the storage of personal data. Secondly, whether the user is adversarial or accidental would be immaterial, since the ability to prompt a model to output memorised information would be enough to conclude that the model has retained it.²⁴ Another defence commonly invoked to question this conclusion is the non-deterministic nature of AI outputs. Due to the

¹⁷ Cooper and Grimmelmann (n 12) 52–55; Gervais and others (n 9) 3–4.

¹⁸ Developers are not currently able to fully anticipate or control which data crosses this threshold, since it is not solely determined by the model's design but also by parameter convergence, making the memorisation process partially opaque. See Bourtole and others (n 11) 1–5; Cooper and Grimmelmann (n 12) 45–55.

¹⁹ Court of Justice of the European Union, 'Judgment of the Court, Case C-582/14 (Breyer)' (n 6) para 46; European Data Protection Board, 'Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models' (2024) 14–19; Article 29 Working Party (n 6) 24; Bygrave and Tosoni (n 1) 110–111.

²⁰ For an opposing view, see Der Hamburgische Beauftragte für Datenschutz und Informationsfreiheit, 'Discussion Paper: Large Language Models and Personal Data' (2024).

²¹ Cooper and Grimmelmann (n 12) 55–60; Staufer (n 14) 5–6; Hong and others (n 8) 5.

²² Der Hamburgische Beauftragte für Datenschutz und Informationsfreiheit (n 20) 6–8.

²³ Bygrave and Tosoni (n 1) 110–111.

²⁴ Confederation of European Data Protection Organizations (n 10) 6-7, 11-13; Cooper and Grimmelmann (n 12) 50–60; Der Landesbeauftragte für Datenschutz und Informationsfreiheit Baden-Württemberg (n 13) 6–7.

stochastic element introduced during generation, the model can yield different responses to the same prompt.²⁵ However, even if replication is not guaranteed, the frequent occurrence of these outputs indicates that the model internally retains structured representations of the training input.²⁶ In this sense, stochasticity does not challenge the retention of personal data, but merely affects the ease of retrieval.

In other words, the inability to pinpoint exactly which parameters encode personal data cannot be used to deny its presence in the model. Even in the absence of direct access to parameters, the existence of memorised content and the possibility of its output will meet the GDPR definition of personal data, regardless of the anonymisation standard used, if no mitigation measures are adopted. But more importantly, even if mitigation measures are adopted, developers will either be unable to reach an objective anonymisation, and thus the model will be subject to the GDPR regardless of best efforts, or, if they are able to reach it, the model will underperform. Let's see why.

Can AI models escape the GDPR?

Since AI models memorise parts of their training data, the next question is how might developers train their models in a way that avoids the memorisation of personal data. To answer that question, we must understand the different phases of AI development.

The ISO 5338 standard defines the AI lifecycle as a series of stages aimed at supporting the development of trustworthy, ethical, and legally sound AI systems.²⁷ While all stages are relevant, the key concern lies in the data collection, pre-processing and model training phases, since these are the points at which large datasets are introduced and used on the model.²⁸ Here we are also introduced to Privacy Enhancing Technologies (“PETs”), meaning technical measures designed to mitigate risks associated with the processing of personal data during the lifecycle of machine learning models.²⁹

Thus, our goal shall be to describe how the most important PETs contribute to the anonymisation of personal data. With this, we will conclude that achieving an objective

²⁵ Cooper and Grimmelmann (n 12) 40–45.

²⁶ *ibid*; European Data Protection Supervisor, ‘Generative AI and the EUDPR - First EDPS Orientations for Ensuring Data Protection Compliance When Using Generative AI Systems’ (2024) 22–23.

²⁷ Enrico Glerean, ‘Training Curriculum on AI and Data Protection Fundamentals of Secure AI Systems with Personal Data’ (2025) 40–41.

²⁸ *ibid*.

²⁹ Alissa Brauneck and others, ‘Federated Machine Learning, Privacy-Enhancing Technologies, and Data Protection Laws in Medical Research: Scoping Review’ (2023) 25 *Journal of Medical Internet Research* 2–4; Fatemeh Mosaiyebzadeh and others, ‘Privacy-Enhancing Technologies in Federated Learning for the Internet of Healthcare Things: A Survey’ (2023) 12 *Electronics* 2703, 2703–2705, 2708.

standard of anonymisation is either effectively impossible or, if attainable, burdensome to an extent that it either compromises the utility of the data and/or model, or so financially and logistically difficult that it becomes prohibitive for any developer that is not a major technological company.

Data collection and pre-processing

One of the first phases of training AI models is the collection of raw data, generally from three sources.³⁰ Firstly, it can be received directly from data subjects. For example, user-generated training data can be collected during beta versions to be logged and reused for fine-tuning.³¹ Alternatively, it can be obtained from commercial providers. In fact, it is quite common for AI developers to use external datasets obtained under licensing schemes or through open data platforms.³² Lastly, developers can also use data querying processes, through which datasets from both first-party and third-party sources are filtered and recombined.³³ Regardless of the method adopted, the GDPR will apply in a rather straightforward way. In fact, more likely than not, the data collected will include personal data.³⁴ This means that it is rather unlikely for us to be facing anonymised data at this stage, whether subjectively or objectively.

We then enter the pre-processing phase, where data is transformed to enhance its quality and to be anonymised.³⁵ This can be done through multiple techniques, such as generalisation,³⁶ suppression,³⁷ perturbation,³⁸ and filtering.³⁹ However, these safeguards are either insufficient to guarantee the objective anonymisation of the training data or, if they do, they degrade the data to a point where it loses its utility, as shown by research that was able to re-identify supposedly anonymous training datasets.⁴⁰

³⁰ Glerean (n 27) 50–53.

³¹ *ibid* 51–53.

³² Giving examples of how we-scraped datasets are created, see Hong and others (n 8) 3–5.

³³ Glerean (n 27) 50–53.

³⁴ For an example on web-scraped datasets, see Hong and others (n 8).

³⁵ Mosaiyebzadeh and others (n 29) 2708–2709.

³⁶ Generalisation reduces the granularity of data, typically by replacing specific attributes with more general or aggregated values. See Glerean (n 27) 53–55.

³⁷ Suppression removes certain data, particularly direct identifiers such as names or national identification numbers. Suppression can be either column-wise, such as dropping names from tables, or more granular using Named Entity Recognition (“NER”) for redacting identifiers from unstructured text. Suppression can also take the form of blurring faces to conceal biometric identifiers. See *ibid*.

³⁸ Perturbation injects randomness to obscure identifiable patterns while trying to preserve the statistical distribution of data. See *ibid*.

³⁹ Filtering involves removing data points based on selected criteria. It serves multiple goals, i.e. discarding outliers, excluding biased or low-quality entries, or removing taboo or irrelevant content. See *ibid* 55–57.

⁴⁰ OECD, ‘Emerging Privacy Enhancing Technologies: Current Regulatory and Policy Approaches’ (2023) 351 15–19. See generally Arvind Narayanan and Edward W Felten, ‘No Silver Bullet: De-Identification Still Doesn’t

This occurs because the core difficulty in achieving an objective standard of anonymisation at this point lies in anticipating all combinations of data and techniques that may defeat it, while also identifying and removing all relevant features that would render re-identification possible, in a way that does not reduce data's utility. For example, generalisation depends heavily on the application of k-anonymity to ensure that each dataset entry is indistinct from at least k-1 others on identified quasi-identifiers.⁴¹ However, applying k-anonymity quickly leads either to information loss or the impossibility of achieving k. Suppression, on the other hand, does not eliminate all uniqueness since context-specific signals may be enough to render a person identifiable when cross-referenced with other datasets. The same logic occurs with filtering, where it risks omitting significant subpopulations, with the added risk of reinforcing bias instead of mitigating it. Additionally, the specification of "unwanted" information often depends on contextual value judgments, making filtering ethically complex and operationally unreliable as a safeguard.⁴² For perturbation, applying too much noise also degrades data utility, while too little fails to anonymise data.

This then brings us to another problem, related to resource limitations. Complex anonymisation measures and the expertise required to identify quasi-identifiers, balance trade-offs between utility and risk, and quantify re-identification threats goes well beyond ordinary software engineering tasks.⁴³ As a result, even well-intentioned developers may fail to achieve that standard, simply due to its implementation burden.

This means that PETs used at the pre-processing stage fail to pass the high threshold for an objective anonymisation. In other words, if we are to interpret anonymisation as an irreversible process for any party, then most datasets involved in AI development must continue to be treated as including personal data.

In an attempt to - among other things - escape the need to anonymise personal data, developers have increasingly turned to synthetic data, meaning artificially generated datasets that mimic the statistical properties of real data, without directly including actual personal data.⁴⁴ Synthetic

Work' (2014) 8; Jane Heriksen-Bulmer and Sheridan Jeary, 'Re-Identification Attacks - A Systematic Literature Review' (2016) 36 *International Journal of Information Management* 1184; Luc Rocher, Julien M Hendrickx and Yves-Alexandre de Montjoye, 'Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models' (2019) 10 *Nature Communications* 3069.

⁴¹ Glerean (n 27) 53-55.

⁴² *ibid* 55-57.

⁴³ OECD (n 40) 19.

⁴⁴ Alexis Léautier, '[Données Synthétiques] - Dis Papa, Comment on Fait Les Données ? 1/2' (*Laboratoire d'Innovation Numérique de la CNIL*, 2022) <<https://linc.cnil.fr/donnees-synthetiques-dis-papa-comment-fait-les-donnees-12>> accessed 25 September 2025; OECD (n 47) 15-18; Information Commissioner's Office, 'Chapter 5: Privacy-Enhancing Technologies (PETs): Draft Anonymisation, Pseudonymisation and Privacy Enhancing Technologies Guidance' (2022) 35-36.

data has been applied in privacy-sensitive domains, such as healthcare, where real data may not be ethically or legally processed, in order to i.e. simulate edge cases, re-balance underrepresented subgroups, or improve the fairness and accuracy of outputs.⁴⁵

Still, these benefits come with trade-offs, where we are brought back to the anonymisation vs. utility dilemma. On the one hand, synthetic data can resemble real data too closely, particularly if unique traits persist, giving room to linkage or inference risks.⁴⁶ Thus, without control of statistical outliers, data risks falling back within the scope of personal data.⁴⁷ In fact, model inversion attacks and advances in generative modelling increasingly allow malicious actors to reverse-engineer and infer personal data from synthetic datasets.⁴⁸ On the other hand, if developers focus too much on anonymisation, synthetic data can lead to model collapse, as data generated by successive iterations gradually loses connection to the original human-related information.⁴⁹ This decay not only affects model quality but also risks compounding bias, reducing the long-term utility of synthetic data in maintaining fairness and reliability.⁵⁰

This means that synthetic data is not the silver bullet to reach objective anonymisation during data collection or pre-processing phases, as it does not meet this threshold in most real-world scenarios.⁵¹ In fact, regulators have already signalled caution, noting that re-identification remains possible.⁵²

Therefore, at this stage, if we adopt an objective standard of anonymisation, developers face substantial difficulty in escaping the GDPR. It is worth noting that having to apply the GDPR to the data collection and pre-processing stages is not an inherently bad thing. Quite the opposite. Subjecting these phases to data protection requirements can be beneficial, as it

⁴⁵ Glerean (n 27) 68–70; Maria Catarina Batista, ‘Synthetic Data in AI Development: Ensuring Data Protection and Ethics’ (2024) 3948 CEUR Workshop Proceedings 37, 38–39.

⁴⁶ Glerean (n 27) 68–70; Confederation of European Data Protection Organizations (n 10) 20–23.

⁴⁷ Information Commissioner’s Office (n 44) 35–36; Glerean (n 27) 68–70; Confederation of European Data Protection Organizations (n 10) 20–23.

⁴⁸ Information Commissioner’s Office (n 44) 35–36; OECD (n 40) 15–18. See generally Theresa Stadler, Bristena Oprisanu and Carmela Troncoso, ‘Synthetic Data—Anonymisation Groundhog Day’, *Proceedings of the 31st USENIX Security Symposium* (USENIX Security 22 2022).

⁴⁹ Glerean (n 27) 68–70.

⁵⁰ *ibid.*

⁵¹ Information Commissioner’s Office (n 44) 35–36; Confederation of European Data Protection Organizations (n 10) 20–23.

⁵² See, as examples, Office of the Privacy Commissioner of Canada, ‘Privacy Tech-Know Blog: When What Is Old Is New Again – The Reality of Synthetic Data’ (*OPC blogger*, 2022) <<https://www.priv.gc.ca/en/blog/20221012/?id=7777-6-493564>> accessed 25 September 2025; Agencia Española Protección Datos, ‘Datos Sintéticos y Protección de Datos’ (*AEPD Blog*, 2023) <<https://www.aepd.es/prensa-y-comunicacion/blog/datos-sinteticos-y-proteccion-de-datos>> accessed 25 September 2025. See also OECD (n 40) 15–18; Glerean (n 27) 68–70; Confederation of European Data Protection Organizations (n 10) 20–23.

reinforces the principles of data minimisation and security, compelling developers to adopt strong safeguards from the start. The real challenge emerges, however, when the model being trained on these datasets is considered itself as storing personal data, regardless of the safeguards used.

Model training

Our next stop is at the training phase, where the fundamental challenge is to ensure that it does not lead to the retention of personal data. To address this, developers turn to a variety of PETs: decentralizing the structure of AI training, adding noise to the training data, model parameters, or outputs, employing cryptographic techniques. While all of them are important, they are not, separately or jointly, enough to guarantee that the information embedded within the model is objectively anonymised. Let us look at the main ones and understand why.

Federated Learning

Unlike conventional centralised machine learning processes, Federated Learning (“FL”) is an architecture in which a model is trained collaboratively across multiple decentralized nodes without transferring raw data to a central location.⁵³ Each local data holder trains a local model on its own private dataset and shares parameters either with a central server or with peers, depending on the setup.⁵⁴ These updates are then aggregated to form a final model, in a loop that continues until it achieves the desired performance.⁵⁵

FL remains one of the most promising techniques for privacy-focused AI training. The application of FL in healthcare shows this potential, where hospitals jointly train AI models without sharing their data.⁵⁶ FL has also been deployed in consumer contexts, for example, in speech processing applications to personalise models (i.e. for predictive text or voice assistants) without uploading i.e. raw voice data to a central server.⁵⁷

⁵³ Kanishka Ranaweera and others, ‘Federated Learning with Differential Privacy: An Utility-Enhanced Approach’ (2025) arXiv:2503 arXiv preprint 1–2; Agencia Española Protección Datos and European Data Protection Supervisor, ‘Federated Learning’ (2025) 6–9.

⁵⁴ FL can be implemented in one of two ways. In centralised FL, a central server orchestrates model distribution and aggregation. It sends an initial model to each node, collects locally trained updates, and aggregates these updates into a global model. Fully decentralised FL eliminates the central server: nodes exchange parameters directly among themselves and achieve consensus collaboratively. See Agencia Española Protección Datos and European Data Protection Supervisor (n 53) 6–9.

⁵⁵ Ranaweera and others (n 53) 1–2; Agencia Española Protección Datos and European Data Protection Supervisor (n 53) 6–9.

⁵⁶ Agencia Española Protección Datos and European Data Protection Supervisor (n 53) 12–13.

⁵⁷ *ibid.*

However, FL is not an easily-implementable architecture. First, since the training is based on separate data sources that cannot be cross-compared, it becomes difficult to assess and ensure the integrity, quality, and consistency of the training data, hampering the ability to conduct normalisation, deduplication or credibility checks. This means that if local models rely on inconsistent formatting, outdated data, or incomplete records, training can become skewed or unstable and the downstream result will likely be non-optimal global model.⁵⁸ Differences in device capabilities make matters worse. For example, in cross-device FL, IoT devices have varying computational power and storage capacity and therefore limited ability to run long-term training iterations at acceptable processing speeds. This makes their participation intermittent and error-prone, leading to issues like stragglers (slow or non-responding clients) and dynamic participation (clients joining or leaving mid-training) that reduce model convergence.⁵⁹

This brings us to a second problem: communication costs. Local models must frequently transmit parameter updates to a central server or to peers. This repeated communication rapidly consumes network resources and contributes to inefficiencies.⁶⁰ Compounded by the dependence on stable connectivity, model updates from unreliable or low-bandwidth devices can delay global model aggregation, degrade the model's performance, or cause training failures.⁶¹

But even if developers manage to get through these challenges there is still an unresolved trade-off between anonymisation and resource expenditure. As we will see, techniques used to harden FL against adversarial attacks significantly increase the computational resources needed to train the model, since more iterations will be needed to ensure accurate results.⁶² Added to this, and crucially, FL does not guarantee the objective anonymisation of the global model. On the contrary, one of the central concerns is that while raw training data remains local, model updates may still leak personal data.⁶³ This is due to the logic we explained before on memorisation: local models are still AI models and may memorise personal data included in their training corpus.⁶⁴ As such, this risk of identifiability may be transferred to the global one.

⁵⁸ *ibid* 13–14; Glerean (n 27) 67–69.

⁵⁹ Agencia Española Protección Datos and European Data Protection Supervisor (n 53) 13–14.

⁶⁰ Glerean (n 27) 67–68.

⁶¹ OECD (n 40) 23.

⁶² Mosaiyebzadeh and others (n 29) 2713–2714.

⁶³ Ranaweera and others (n 53) 1–2; Agencia Española Protección Datos and European Data Protection Supervisor (n 53) 14–16.

⁶⁴ Brauneck and others (n 29) 6.

In fact, several studies confirm that adversaries can derive training samples from model updates or gradients transmitted during FL rounds.⁶⁵ In other words, even if raw datasets are never shared, it may still be possible for the final model to have memorised parts of them.

This means that while global models are generally understood as non-personal data – assuming sufficient aggregation and abstraction of parameter input⁶⁶ – the soundness of this assessment is conditional upon the robustness of the aggregation process and the absence of identifiability risks. This brings us back to the GDPR’s standard for anonymisation. If we adopt an objective standard of anonymisation, even a low probability of re-identification, when feasible through inference attacks, may be sufficient to regard the data as personal.⁶⁷ In the context of FL, such reversibility cannot be excluded, particularly given variations in implementation quality, attacker capabilities, and the complexity of data structures. Given continued academic demonstrations of training data extraction from FL models, the required objective threshold of irreversibility will likely remain unmet.⁶⁸

Differential Privacy

Differential Privacy (“DP”) is a mathematically defined model designed to protect data during computational processes, by injecting random noise into it, ensuring that the presence or absence of personal data has a limited impact on the model’s output.⁶⁹ The goal is to make it statistically implausible for an observer to infer whether any personal data was included in the dataset. To that end, the result of DP is controlled by two parameters: epsilon (ϵ), which quantifies the privacy loss or ‘budget’ (in other words, how similar the output of a computation will be when a single individual’s data is present versus when it is not), and delta (δ), which defines the maximum probability that the privacy guarantee is violated.⁷⁰

Despite its theoretical advantages, DP also has limitations. First, DP’s implementation requires structured and controlled data inputs. However, many real-world training datasets include

⁶⁵ These risks intensify when limited numbers of users participate, making it easier to reverse-engineer the process. See Agencia Española Protección Datos and European Data Protection Supervisor (n 53) 14–16; Glerean (n 27) 67–68; Ranaweera and others (n 53) 1–2; Mosaiyebzadeh and others (n 29) 2707–2708.

⁶⁶ Brauneck and others (n 29) 6.

⁶⁷ Agencia Española Protección Datos and European Data Protection Supervisor (n 53) 14–16.

⁶⁸ Mosaiyebzadeh and others (n 29) 2707–2708; Agencia Española Protección Datos and European Data Protection Supervisor (n 53) 19–23.

⁶⁹ Mosaiyebzadeh and others (n 29) 2709; Evgenia Novikova and others, ‘Analysis of Privacy-Enhancing Technologies in Open-Source Federated Learning Frameworks for Driver Activity Recognition’ (2022) 22 Sensors 2983, 2987–2988.

⁷⁰ Novikova and others (n 69) 2987–2988; Information Commissioner’s Office (n 44) 30–34.

unbounded or messy data, outliers, and multiple contributions per individual.⁷¹ This is problematic, since DP often assumes bounded numerical inputs inside a predefined range.⁷² Defining these bounds can introduce either high bias (if too restrictive) or high variance (if too permissive), creating a tension between statistical accuracy and data protection guarantees.⁷³ Second, there is a lack of universally accepted thresholds for what constitutes an appropriate privacy budget.⁷⁴ In practice, ϵ values differ significantly across use cases, ranging from values below 1 to values as high as 100, with no authoritative norm to define inadequacy or sufficiency.⁷⁵ Where DP has been applied on a large scale it has often attracted criticism for using high privacy budgets that can lead to re-identification risks, by allowing adversaries to combine information across multiple queries or with auxiliary data sources.⁷⁶ Not just that, but real-world model training involves iterative experimentation, parameter tuning, and diagnostic queries, all of which can contribute to cumulative privacy loss, rendering DP's formal guarantees void.⁷⁷ This privacy degradation is difficult to quantify and track, making the actual effectiveness of DP unclear.⁷⁸ On the other hand, there is the degradation of the model with the introduction of too much noise. Here, we have the same anonymisation versus utility trade-off we met before: increasing noise improves anonymisation efforts but simultaneously reduces the accuracy and generalisation capabilities of the model.⁷⁹ This limitation is particularly evident in deep learning models where learning is highly sensitive to small changes in parameters.⁸⁰

Third, DP requires expert configuration and case-by-case tuning, which significantly raises the cost of implementation. Since data sensitivity, query types, record contribution patterns, and adversarial threat models vary across domains, fixed configurations rarely suffice. As a result, deploying DP involves numerous judgment calls – on how to set ϵ , how to define query neighbourhoods, how to preprocess the data – which limits its accessibility and usefulness to large organisations with specialised teams.⁸¹ Additionally, many real-world decisions required

⁷¹ Kareem Amin, Alex Kulesza and Sergei Vassilvitskii, 'Practical Considerations for Differential Privacy' (2024) arXiv:2408 arXiv preprint 1–2.

⁷² *ibid.*

⁷³ *ibid.*

⁷⁴ OECD (n 40) 15–18.

⁷⁵ Information Commissioner's Office (n 44) 30–34; OECD (n 40) 15–18.

⁷⁶ Information Commissioner's Office (n 44) 30–34; OECD (n 40) 19.

⁷⁷ Amin, Kulesza and Vassilvitskii (n 71) 2–3.

⁷⁸ *ibid.*

⁷⁹ Ranaweera and others (n 53) 1–2; Glerean (n 27) 65–66.

⁸⁰ Ranaweera and others (n 53) 1–2.

⁸¹ Information Commissioner's Office (n 44) 30–34.

to use DP – like selecting bounds or interpreting identifiers – are not determined by DP.⁸² This leads teams to resort to intuition, heuristics, or conventions to bridge this gap between theory and practice. As a result, DP forces new subjective assessments, often with even less transparency or scientific agreement.⁸³

This means that the deployment of DP does not ensure *per se* that the model is objectively anonymised. In fact, the effectiveness of DP’s anonymising function depends highly on the context of use – particularly the level of ϵ . If ϵ is set too high, noise is minimal, reducing anonymisation as subsequent interactions with the trained model could be used to extract memorised data.⁸⁴ If set too low, the usefulness of the data or model severely degrades.⁸⁵

Cryptography

The final category of PETs consists of cryptography, which aims to ensure that personal data remains inaccessible during computation or collaborative training. Developers have different options in this regard, but two stand out.⁸⁶

Firstly, we have Secure Multi-Party Computation (“SMPC”) as a cryptographic protocol that enables two or more parties to jointly compute a function over their respective datasets without revealing them to each other.⁸⁷ This way, one of the key features of SMPC is its ability to remove the need for a trusted third party in collaborative computations.⁸⁸

Then, we have Homomorphic Encryption (“HE”), another privacy-enhancing cryptographic technique that is also used to enable computation on encrypted data.⁸⁹ In simple terms, it allows participants to perform functions on ciphertext that, once decrypted by the data subject or controller, yield results equivalent to those that would be obtained had the operations been performed on plaintext data.⁹⁰ The main characteristic of HE is therefore that the processing party never has access to the underlying unencrypted data, reducing the risk of exposing

⁸² Amin, Kulesza and Vassilvitskii (n 71) 1–2.

⁸³ *ibid* 1, 4.

⁸⁴ Glerean (n 27) 65–66.

⁸⁵ Information Commissioner’s Office (n 44) 30–34; Ranaweera and others (n 53) 1–2; Amin, Kulesza and Vassilvitskii (n 71) 1–2.

⁸⁶ For a detailed analysis see OECD (n 40) 15–22; Information Commissioner’s Office (n 44) 12–23, 26–28.

⁸⁷ This process relies on techniques such as secret sharing, which divides data into cryptographic fragments distributed among parties. By doing so, no individual party has access to complete, intelligible data, which mitigates the risk of exposure from malicious insiders or external attacks. See Mosaiyebzadeh and others (n 29) 2709; Information Commissioner’s Office (n 44) 15–19.

⁸⁸ OECD (n 40) 19–21; Information Commissioner’s Office (n 44) 15–19.

⁸⁹ Mosaiyebzadeh and others (n 29) 2709; Information Commissioner’s Office (n 44) 12–15.

⁹⁰ OECD (n 40) 19–21; Information Commissioner’s Office (n 44) 12–15.

personal data during computations.⁹¹ Additionally, HE supports scenarios where the models used must remain confidential alongside the data inputs – for example, in competitive industries or FL settings.⁹² More than that, the integrity of outputs is maintained since the decrypted result of HE computation is mathematically identical to the result that would be achieved if the underlying unencrypted data were used.⁹³

As expected by now, these cryptographic processes have challenges similar to the ones we saw previously. Firstly, there is a fundamental computational burden due to the encryption and decryption overhead. SMPC computations are consistently slower, and require more infrastructure than non-encrypted data processing, particularly when combining data from several entities or when the operations involve complex models.⁹⁴ The same occurs with HE, as operations that take milliseconds in plaintext may require seconds or minutes when operating on ciphertext, depending on the complexity of the operation.⁹⁵ The lower efficiency limits its use at scale, unless substantial computing power is available.⁹⁶

Secondly, since the actual data is encrypted, the analysts conducting SMPC-based computations or using HE-encrypted data cannot see or clean the inputs. This means that standard practices in AI model development – such as data cleaning and validation – become unfeasible after encryption.⁹⁷ If errors or corrupt inputs exist in the data before then, they will persist through the computation process, potentially distorting the model’s output. For example, if the data is not properly prepared before encryption, the computations done through SMPC will produce incorrect or non-converging outputs, and analysts lack the tools to correct them *post hoc*.⁹⁸

Thirdly and crucially, the outputs of these computations can still leak information. For SMPC, if a function reveals an output based solely on a single observation, the content of that observation becomes inferable by default. This means that even if data are encrypted and only statistical outputs are shared, the final result may still relate to identifiable individuals depending on composition and granularity.⁹⁹ For HE, although the data remains encrypted

⁹¹ Mosaiyebzadeh and others (n 29) 2709; OECD (n 40) 19–21.

⁹² OECD (n 40) 19–21.

⁹³ Information Commissioner’s Office (n 44) 12–15.

⁹⁴ Mosaiyebzadeh and others (n 29) 2709; OECD (n 40) 19–21; Information Commissioner’s Office (n 44) 15–19.

⁹⁵ OECD (n 40) 19–21.

⁹⁶ *ibid.*

⁹⁷ *ibid.*

⁹⁸ *ibid.*

⁹⁹ Information Commissioner’s Office (n 44) 15–19.

during processing, there is no inherent guarantee that the outputs of those computations do not themselves allow for the re-identification of data subjects.¹⁰⁰ For example, if the function being computed reveals results based on small data clusters, it may still leak information about the original personal data, especially if attackers repeat operations or combine outputs with auxiliary data sources.¹⁰¹ Additionally, since HE only prevents direct access but not inference from certain outputs or the potential future re-identification if keys are compromised, the output remains personal data under Article 4(1) GDPR.¹⁰²

Bringing this together

We can therefore say that these PETs, by themselves, do not reach the threshold for data to be considered objectively anonymous, or, at least, reach it in a way that can be implemented without damaging the model's performance or the developers' financial well-being. As such, they are frequently coupled together, in an attempt to construct an architecture where the memorisation of personal data is mathematically inexistent.

Nonetheless, we argue that these results are also likely not reached in practice. Even when FL is strengthened with DP, SMPC, or HE, it does not eliminate the possibility of re-identification, model leakage, or adversarial inference. Additionally, we are still met with utility trade-offs, since these PETs affect the accuracy of the model, which in practice limit its development on a wide scale.

Starting with anonymisation techniques, multiple studies have attempted to combine FL with syntactic anonymisation methods, such as k-anonymity, to mitigate the risks posed by potential data leakage during model training and aggregation.¹⁰³ For instance, Grama et al.¹⁰⁴ used k-anonymity to preserve privacy over healthcare data. Their framework showed a favourable balance between data protection and utility, assuming that the datasets had a sufficient number of samples.¹⁰⁵ However, this assumption is crucial. Anonymisation techniques are most effective when data pools are large and homogeneous – as we have seen, conditions not always met in real-world FL settings.

¹⁰⁰ OECD (n 40) 19–21.

¹⁰¹ *ibid.*

¹⁰² Information Commissioner's Office (n 44) 12–15.

¹⁰³ Mosaiyebzadeh and others (n 29) 2709–2710.

¹⁰⁴ Matei Grama and others, 'Robust Aggregation for Adaptive Privacy Preserving Federated Learning in Healthcare' (2020) arXiv:2009 arXiv preprint.

¹⁰⁵ Mosaiyebzadeh and others (n 29) 2709–2710.

Other combinations reinforce this conclusion. For example, federated PF-NMF applies multiple local privacy filters during training to screen out sensitive features.¹⁰⁶ Another technique, FeARH, introduces randomisation in parameter sets shared between devices, reinforcing anonymisation efforts even in the presence of an untrustworthy central analyser.¹⁰⁷ Despite improved outcomes, approaches like these are still subject to the general trade-off inherent to anonymisation: increased de-identification leads to increased information loss. This affects prediction performance, especially in environments where small variations in data have significant implications, such as medical diagnostics. Furthermore, they generally do not offer mathematical guarantees regarding re-identifiability, leaving them insufficient under an objective standard of anonymisation.¹⁰⁸

FL architectures, especially in the healthcare sector, have also been coupled with HE, SMPC, and functional encryption.¹⁰⁹ Zhang et al.¹¹⁰ and Ma et al.¹¹¹ deployed masking and HE schemes on skin cancer and elderly fall detection datasets, respectively, protecting model updates during FL aggregation phases.¹¹² Similarly, CKKS-based HE frameworks have been proposed to support secure brain age prediction without performance loss.¹¹³ While methods such as these allow computation on protected data, they may not fully prevent the exposure of personal data in the final model's outputs. Moreover, if participants collude or the coordinator is malicious, encryption offers little help.¹¹⁴ And then we have increased computation costs, that limit their scalability. SMPC and HE require expensive modular arithmetic and exponentiation operations involving long keys, dramatically increasing processing time and communication overhead.¹¹⁵ While batching and GPU-based acceleration have been proposed to alleviate these downsides, they remain largely infeasible for cross-device FL systems consisting of heterogeneous and resource-constrained endpoints, especially in mobile or IoT settings.¹¹⁶

¹⁰⁶ *ibid.*

¹⁰⁷ *ibid.*

¹⁰⁸ Brauneck and others (n 29) 11.

¹⁰⁹ Mosaiyebzadeh and others (n 29) 2710–2711.

¹¹⁰ Li Zhang and others, 'Homomorphic Encryption-Based Privacy-Preserving Federated Learning in IoT-Enabled Healthcare System' (2022) 20 IEEE Transactions on Network Science and Engineering.

¹¹¹ Jing Ma and others, 'Privacy-Preserving Federated Learning Based on Multi-Key Homomorphic Encryption' (2021) arXiv: 2104.06824 arXiv preprint.

¹¹² Mosaiyebzadeh and others (n 29) 2710–2711.

¹¹³ *ibid.*

¹¹⁴ Brauneck and others (n 29) 11; Novikova and others (n 69) 6–8.

¹¹⁵ Novikova and others (n 69) 6–8.

¹¹⁶ *ibid.*

Lastly, DP can be integrated into FL systems in either a local (DP-SGD) or global (DP-FedAvg) fashion, where it suppresses identifiability either by perturbing local gradients during training or by injecting noise into model parameters during aggregation.¹¹⁷ DP-SGD provides stronger guarantees by perturbing gradient updates before they leave local devices. However, this generally results in reduced model accuracy, extended training times, and increased computational requirements.¹¹⁸ Alternatively, DP-FedAvg applies noise during model aggregation, reducing local client load and improving training efficiency, albeit with lower guarantees against inference attacks targeting the global model.¹¹⁹ This has been tested on use cases like early disease detection or genomic risk profiling that rely on detailed datasets that DP modifications degrade. For instance, in the ADDetector system for Alzheimer’s diagnosis and the FedGAN blockchain-based COVID-19 detection framework, DP application was realised through Gaussian noise injection during training and aggregation.¹²⁰ These and other methods, while having performance benefits, do not entirely eliminate the possibility of data leakage.¹²¹ Additionally, the application of DP is largely limited to horizontally partitioned datasets. Practical support for vertically partitioned data, where each party holds different features for the same record, remains underdeveloped and typically requires intermediary synthetic data or encryption-hybrid methods.¹²²

This means that while PETs can cumulatively reduce memorisation risks, they cannot fully eliminate them since, even under optimal deployment, residual exposure from the model’s output, insignificant data fractions, or edge use cases precludes the certainty of objective anonymisation.¹²³

Conclusion

It is precisely because of this that the CJEU’s decision in *SRB v. EDPS* is so significant. The judgment recognised, when dealing with pseudonymised data, that while the entity holding both the identifying information and the pseudonymised dataset can still link the two (thus the data remains personal for them), third parties who do not have access to the linking information

¹¹⁷ *Ranaweera and others* (n 53) 3–4.

¹¹⁸ *ibid.*

¹¹⁹ *ibid.*

¹²⁰ *Mosajyebzadeh and others* (n 29) 2711–2712.

¹²¹ *Ranaweera and others* (n 53) 3–4.

¹²² *Novikova and others* (n 69) 5–6.

¹²³ *Agencia Española Protección Datos and European Data Protection Supervisor* (n 53) 14–16; *Brauneck and others* (n 29) 11; *Novikova and others* (n 69) 6–8.

may not be able to identify data subjects.¹²⁴ The position of the Court is therefore very simple: identifiability must be assessed from the position of the specific controller, taking into account the reasonable means that could be used, by that party, to re-identify individuals.¹²⁵ As the Court noted, if we were to understand pseudonymisation and anonymisation otherwise, then these concepts would be void of practical relevance, as nearly all data would likely fall under the definition of personal data.¹²⁶

Bringing this to the context of AI memorisation, the information retained by a model should therefore be regarded as anonymised by reference to whether, for a specific party, there are reasonable means of retrieving it. As we saw, there are two main ways of doing so. On the one hand, the analysis of a model's parameters presupposes access to them. Therefore, the possibility of re-identification by model inspection will be relevant mainly for open-source or shared models. However, and given the current state of technology, the possibility of extracting personal data this way remains remote for most controllers, something that developers must nevertheless ensure before sharing their models. On the other hand, we have the model's outputs. Here, to conclude that the model outputted memorised data, it is generally required confirmation by reference to the original training datasets. If these are irreversibly deleted – and, crucially, not accessible to third parties – this confirmation likely becomes impossible. The developer might be able confirm it if it maintains the original data, but for others, the link cannot be established. Therefore, the problem stops being about memorisation specifically but about the general output of personal data by a trained model, regardless of whether it was derived from memorisation or i.e. hallucination.

Looking at memorisation through this lens has significant implications for AI developers, since AI models would be deemed anonymised in circumstances where developers can demonstrate that third-parties – and arguably themselves – cannot reasonably re-identify data stored within, regardless of fringe and unlikely cases where access may be possible with specialised know-how. This is not to advocate for relaxed technical safeguards, but rather to encourage developers in adopting best anonymisation efforts and to ensure that the obligations controllers are subject to can actually be respected. In fact, having a different understanding, such an absolute anonymisation, would result in absurd regulatory outcomes, where developers would be expected to uphold data subject rights – such as erasure – in relation to data they may not

¹²⁴ Court of Justice of the European Union, ‘Judgment of the Court, Case C-413/23 P (SRB v. EDPS)’ (n 7) paras 76–77.

¹²⁵ *ibid* 69–75, 78–80.

¹²⁶ *ibid* 80.

themselves be able to identify or extract from the model, thus rendering compliance impossible. Instead, if a model's architecture and training protocols show that extraction is not feasible without disproportionate effort or in very unlikely scenarios, these models ought to be considered anonymised under the GDPR, at least for the parties to whom this extraction is unfeasible. Exceptionally, should concrete evidence arise of specific third parties succeeding in accessing memorised personal data, those instances can be tackled directly, rather than focusing on hypothetical worst-case scenarios.