



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

The Vanishing Record: Artificial Intelligence, Anonymisation, and the Unexamined Premises of Data Protection Law

Dr. Moatoshi Ao¹, Pradeep Gulrajani²

¹Faculty of Law, University of Delhi, Delhi, India. E-mail: toshibaimchen@gmail.com

²Faculty of Law, University of Delhi. E-mail: pgulrajani@law.du.ac.in

ABSTRACT

Data protection law sits on foundations that it has never seriously examined: that it is possible to locate and destroy a record on request; that it is possible to make a record anonymous without identifiers by performing some steps; that a controller which takes a recognised safeguard thereby discharges its obligations. Each premise was created during a time of registers and released tables. None resists the influence of artificial intelligence: a trained model keeps some traces of the individuals from whom it was trained but keeps no record of them in any place where they can be found. This article is a call for the courts and regulators to recognise that the resulting gap does not need to be filled by an engineering solution, but rather is a doctrinal shortcoming, and that the published scientific record provides the evidence that they require to fill the gap. Building on that record, and six episodes of enforcement and adjudication in four jurisdictions, it suggests that proportionality review must shift from a documentary to an evidentiary discipline and that de-identification safe harbours in the statute should be reconsidered when the data is intended to be used to train a model.

Keywords: Artificial intelligence; data protection; anonymisation; right to erasure; proportionality; differential privacy; General Data Protection Regulation

Introduction

Data protection law was built for the register and the filing cabinet. Its central mechanisms assume a world in which personal information occupies a determinate location: a row that can be found, a file that can be corrected, a record that can be destroyed. The right to erasure assumes that something can be erased. The principle of data minimisation assumes that data comes in countable units, so that fewer of them may sensibly be demanded. The distinction between personal and anonymous data assumes that identifiability is a property capable of being removed by an operation performed upon the data itself. These assumptions were not naive when they were formed. They were accurate descriptions of how personal information was then processed, and the law that grew from them has proved durable and, in most applications, effective.

Artificial intelligence unsettles all three. A machine-learning model is not a repository of the data from which it was built. It is a set of numerical parameters adjusted, over many iterations, until the system reproduces the statistical structure of the examples it was shown. The finished model contains no rows, no files, and no records. Yet neither is it a clean abstraction. Depending upon how it was constructed, and in particular upon the degree to which it has been fitted too closely to its training examples, a model may retain the influence of specific, identifiable individuals in a manner permitting that influence to be detected, and sometimes reversed, by a determined adversary. Researchers have shown that it is possible to determine, from the behaviour of a deployed system alone, whether a particular person's record formed part of its training corpus (Shokri, Stronati, Song, and Shmatikov, 2017). They have shown that a facial recognition system's confidence outputs suffice to reconstruct a recognisable likeness of a named individual (Fredrikson, Jha, and Ristenpart, 2015). They have shown that where several institutions train a model collaboratively without pooling their data, the intermediate updates they exchange may be inverted to recover the underlying examples with startling fidelity (Zhu, Liu, and Han, 2019). The model, in other words, is a channel through which personal information may flow outward, though it contains no personal records in any form the law presently recognises.

The principal technical answer to this difficulty is differential privacy, a formal guarantee constraining how much any single record may influence the output of a computation, expressed through a parameter conventionally written as epsilon. Its intellectual origins lie in theoretical computer science rather than in law (Dwork, McSherry, Nissim, and Smith, 2006; Dwork and Roth, 2014), and its adaptation to modern machine learning is barely a decade old (Abadi et al., 2016). Its structure, however, will be immediately familiar to any lawyer. Choosing epsilon requires balancing the intrusion upon the individual against the usefulness of the system, and justifying the point at which that balance has been struck. That is a proportionality judgment in all but name, and it is at present being made almost entirely by engineers.

That displacement of a proportionality judgment from legal into technical hands is the point of departure for what follows. This article advances three claims. The first is diagnostic: that the difficulty artificial intelligence

poses for data protection law is not a gap in coverage but a defect in premises, and that the doctrines of erasure, minimisation, and anonymisation each presuppose a technical capability which machine learning does not supply. The second is evidentiary: that the scientific literature has already established, with a degree of rigour and replication that would satisfy any tribunal, both what privacy safeguards cost and what they deliver, so that the factual questions upon which these doctrines turn need not be relitigated from first principles in every case. The third is prescriptive: that proportionality review must become an evidentiary discipline, in which a controller's choice of safeguard is justified by reference to published evidence of that safeguard's actual performance, rather than a documentary one in which the invocation of a recognised technique is treated as self-certifying. The argument proceeds through the constitutional foundations of informational privacy, the statutory architecture giving those foundations effect, the scientific record itself, and six episodes in which courts and regulators have already confronted the problem.

2. Informational Privacy As a Constitutional Value

Before that displacement can be corrected, the value which the correction is meant to serve must first be identified, and it is a constitutional one. That privacy is a fundamental right that can be limited only after it has been met on the basis of a structured test is now mainstream in jurisdictions where artificial intelligence (AI) is primarily developed and applied. The modern articulation of such a right is generally attributed to the decision of the German Federal Constitutional Court in a census case from 1983, which invalidated certain parts of a statute on population census and derived, from the constitutional guarantees of human dignity and the free development of personality, a right of informational self-determination: the authority of the individual to decide, in principle, when and within what limits personal facts are disclosed (Bundesverfassungsgericht, 1983). What is so special about that judgment more than four decades later, is its reasoning. The Court's main insight was that the danger of automated processing lies not in any single item of data but in aggregation, and that information which is individually trivial becomes constitutionally significant once collected and combined at scale. That is precisely what a machine-learning system does, and it was written before there was any kind of recognisable commercial machine learning system.

If the German court supplied the reasoning, the instruments of international law supplied only the aspiration. International instruments replace the value with more general provisions, banning any arbitrary interference with privacy without providing an operative test (United Nations, 1948, art. 12; United Nations, 1966, art. 17). In the particular case of deoxyribonucleic acid (DNA) data, it was the European Court of Human Rights that provided one, in *S. and Marper v. United Kingdom* (2008), which ruled that the indefinite storage of DNA profiles and fingerprints from citizens who were arrested but never convicted was disproportionate to any legitimate law-enforcement purpose. The Court's logic revolved around the unique nature of biometric identifiers: they can be reused, made to refer to more than one context, and used to identify an individual long after they are collected. That logic applies directly to the issue of facial pictures being scraped for the purposes of creating recognition systems, an action that has been litigated in various jurisdictions since then.

From Strasbourg the enquiry travels to Delhi, where it receives its fullest elaboration. The Supreme Court of India makes the most well detailed statement. A bench of nine judges, in *Justice K.S. Puttaswamy (Retd.) v. Union of India* (2017), unanimously found that the concept of privacy, including informational privacy, was an integral part of the constitutional right to life and personal liberty guaranteed under Article 21 of the Constitution. Several members of the Court mentioned risks of aggregation and profiling closely reflecting those of the German census case, saying that the harmlessness of any one piece of data is not a response to the constitutional argument, as the harms are the result of the correlation of many pieces of data. The subsequent challenge to the constitutionality of the Aadhaar identification scheme broke down the standard in four steps: it had to advance a legitimate goal, it had to be rationally connected to that goal, it had to be necessary (in that no less intrusive means would achieve the object), and the benefits obtained must be proportional to the harms suffered by the individual (*Justice K.S. Puttaswamy (Retd.) v. Union of India*, 2019).

What is remarkable about these three lines of authority is their convergence. They stem from different legal traditions, are based on different texts, and were decided over a period of over 30 years. But each ends up with the same structural enquiry: an objective that is clearly legitimate; a logical relationship between that objective and the intervention; a requirement that no gentler means would suffice; and, finally, a weighing of intrusion against benefit. The convergence is not just of academic interest. It is different from an argument over the meaning of specific language in the statute, in that a properly framed argument about the adequacy of a privacy safeguard can be learned by a court in one jurisdiction and then applied to a case in a different jurisdiction. A controller that successfully argued its preferred safeguard was the least intrusive measure it could have chosen to implement that would have achieved the objective, has advanced substantially the same case before the Strasbourg court, Supreme Court of India and a supervisory authority in the European Union (EU). A controller who cannot, has failed in every forum for the same reason. The proportionality enquiry then provides a common measure upon which to spend technical evidence that is discussed below.

Two aspects of this jurisprudence are noteworthy and will be emphasized, as they form the basis for all that follows. The first is methodological. In all of these situations, the determining court ruled that the architecture of the processing system was a question of fact to be examined, not a technical one that was not within its jurisdiction. The German court looked at how the census would be collected, the Strasbourg court looked at

what would happen to a retained DNA profile, and the Indian court looked at how the biometric records would be collected, matched and shared. None treated the system as a black box whose operation the law need not understand. The second feature is doctrinal. The necessity element of the proportionality test (the “no less restrictive alternative” requirement) cannot be met in abstract terms. It requires evidence as to what the alternatives are and what they would cost. A court asking whether a controller might have achieved its purpose by less intrusive means must know what such a system would have looked like and how much worse it would have performed. That evidence is in the published scientific literature and it is not prominent in the reasoning of the decisions where it is most needed.

3. The Statutory Architecture And Its Unexamined Premises

These constitutional foundations do not operate directly upon controllers. They operate through statute, and it is in the statutory detail that their unexamined premises become visible. The General Data Protection Regulation translates the constitutional value into an operative rulebook organised around a small set of principles: lawfulness, fairness and transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality; and accountability (European Parliament and Council, 2016, art. 5). Each is intelligible, administrable, and largely uncontroversial when applied to a database. Each becomes considerably more difficult when the object of the processing is a model rather than a record.

Consider data minimisation, which requires that personal data be adequate, relevant, and limited to what is necessary in relation to the purposes of the processing (European Parliament and Council, 2016, art. 5(1)(c)). Applied to a form, the principle is straightforward: do not ask the applicant’s marital status if marital status is irrelevant to the decision. Applied to the training of a model, it collides with a well-documented empirical regularity, namely that predictive performance improves with the volume and granularity of the training corpus, and that the improvement is not marginal. The principle therefore imposes a real and quantifiable cost, and the controller’s obligation is to justify why that cost ought not to have been borne more heavily. That justification cannot be assessed without evidence as to the shape of the trade-off in the particular case.

Now coming to another point a purpose limitation is affected differently, and the difference is instructive. The principle permits further processing only for purposes compatible with those for which the data was originally collected, and its enforcement ordinarily depends upon the ability to trace a body of data from collection through to each subsequent use. A model defeats that tracing not by concealing the data but by dissolving it. Once a corpus has been absorbed into a set of parameters, the resulting artefact may be applied to tasks bearing no relation to the purpose for which the underlying records were gathered, and there is no operation upon the model which corresponds to asking whether a given later use is compatible with a given earlier collection. The controller may know what it trained upon and what it now does; it cannot, in the ordinary case, say which parts of the former are doing the work in the latter. The principle is not thereby rendered inapplicable, but its enforcement must shift from tracing to documentation, and from the data to the decisions taken about the data. That shift has consequences for the burden of proof which the case law has not yet worked through.

The right to erasure presents a sharper difficulty still (European Parliament and Council, 2016, art. 17). The practical application is given by the interpretation of the Court of Justice in the *Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* case (2014), which established that a search engine operator is a controller over personal data, and can be required to remove search results of a person in certain conditions. The grounds were based on a fact that was never stated, that the operator had the technical ability to find and delete the particular link about which complaint was made, which happened to be one indexed link, and not interfere with anything else. In a search index that is correct. It is false when the model is trained. The influence of a training record is spread out over the parameters of the system instead of being contained within the system, with the consequence that removing a particular individual’s contribution from an already-trained model is not a deletion operation at all. It is either a retraining exercise, expensive and sometimes not feasible at scale, or an approximation whose adequacy itself requires demonstration. A controller confronted with an erasure request in respect of a deployed model therefore occupies a position the Google Spain Court did not contemplate, and the authority offers no guidance because the question did not arise upon its facts.

If erasure discloses a gap between what the law commands and what the technology permits, the transparency provisions disclose a gap between what the law asks and what the technology is able to say. Article 22, restricting decisions based solely upon automated processing which produce legal or similarly significant effects and requiring meaningful information about the logic involved, addresses a related anxiety concerning opacity (European Parliament and Council, 2016, arts. 13(2)(f), 14(2)(g), and 22). That anxiety is not confined to European law. The Supreme Court of Wisconsin, in *State v. Loomis* (2016), declined to hold that a sentencing court’s consultation of a proprietary recidivism-risk instrument violated due process, but conditioned that holding upon a series of written cautions being supplied to sentencing judges, including a warning that the instrument’s methodology had not been disclosed because the vendor asserted trade secrecy. The two regimes reach the problem by different routes, statutory and constitutional, but converge upon the same difficulty: the legal system requires an account of reasoning which the technology is not structured to give. Article 9’s heightened protection for data revealing racial or ethnic origin, health, or biometric identity compounds the difficulty, since a model trained wholly upon ordinary personal data may nonetheless permit inferences as to

special-category characteristics never deliberately collected. Article 35, requiring a data protection impact assessment before processing likely to result in a high risk, is the procedural vehicle through which such risks are to be identified and addressed, and Article 25's requirement of data protection by design and by default the substantive obligation to build safeguards in from the outset.

The accountability principle deserves separate notice, because it is the provision upon which the argument of this article ultimately rests. Article 5(2) requires the controller not merely to comply with the preceding principles but to be able to demonstrate compliance, and in so doing it allocates a burden. That allocation is unremarkable in ordinary processing, where the evidence of compliance is largely documentary: a retention schedule, a record of processing activities, a lawful basis identified in advance. It becomes onerous, and interesting, when the object of the demonstration is a trained model. To demonstrate compliance with minimisation is to demonstrate that a smaller or less granular corpus would not have sufficed. To demonstrate compliance with the security obligation is to demonstrate that the model does not disclose more than it should. Neither demonstration can be made from the face of the documents. Each requires evidence about the behaviour of the system, which is to say evidence of the kind the scientific literature supplies and the ordinary compliance file does not. The accountability principle therefore already contains, in embryonic form, the evidentiary requirement for which this article contends; what is missing is not the legal hook but the practice of using it.

Everything, however, turns upon the boundary drawn by Recital 26, which places genuinely anonymous information outside the Regulation's scope altogether. If a trained model is anonymous, the obligations described above do not attach to it at all. If it is not, they attach in full. This is the most important development in the area, and the European Data Protection Board answered this question directly in December 2024. The Board concluded that anonymity cannot be presumed merely because a model does not visibly contain training records. It must be assessed case by case, by reference to the likelihood that personal data may be extracted from the model by any means reasonably likely to be used, including probabilistic and adversarial querying (European Data Protection Board, 2024). That is, in substance, a regulatory instruction to consult the adversarial literature: to ask not whether extraction is conceivable in theory but whether it has been demonstrated in practice against systems of the relevant kind.

Comparative frameworks confirm that the difficulty is structural rather than a peculiarity of European drafting. India's Digital Personal Data Protection Act, 2023 organises its obligations around consent and the figure of the Data Fiduciary, with exemptions for research and statistical purposes, and its application to model training awaits elaboration in subordinate rules. The United States regulates sectorally: the Health Insurance Portability and Accountability Act permits disclosure of health information from which eighteen enumerated categories of identifier have been removed (Health Insurance Portability and Accountability Act, 1996), a statutory safe harbour resting upon the same generalise-and-suppress logic as the k-anonymity model developed in the computer science literature (Sweeney, 2002). California's consumer privacy statute confers rights to opt out of the sale or sharing of personal information without imposing any general minimisation duty (California Consumer Privacy Act, 2018). The European Union's Artificial Intelligence Act tiers obligations by risk rather than by data type, requiring that training, validation, and testing datasets for high-risk systems be relevant, sufficiently representative, and to the best extent possible free of errors, and carries its own penalty structure reaching seven per cent of worldwide turnover for the gravest infringements (European Parliament and Council, 2024, arts. 10 and 99). China's Personal Information Protection Law requires separate consent for automated decision-making with significant effects and channels cross-border transfers through state security assessment rather than an independent supervisory authority (Personal Information Protection Law, 2021). Four regulatory traditions, four institutional designs, and in each the same unresolved question: what does it mean to protect a person whose data has been absorbed into a model?

4. What The Scientific Record Establishes

The doctrinal difficulties described above turn upon questions of fact. It is worth setting out with some care what the published scientific literature actually establishes, because the answers are more settled than the state of the case law would suggest. Four bodies of evidence are relevant: the cost of formal privacy guarantees, the demonstrated success of extraction attacks, the specific vulnerabilities of collaborative training, and the long record of re-identification research.

Upon cost, the foundational study remains the paper which first rendered differential privacy practical for modern machine learning. Its authors reported that a system trained under a strict formal guarantee achieved roughly ninety-five to ninety-seven per cent accuracy upon a simple handwriting recognition benchmark, but only some seventy-three per cent upon a harder photographic classification task, and that even the latter figure depended upon a portion of the system having been pre-trained upon public data without any privacy guarantee whatever (Abadi et al., 2016). Later refinements have narrowed but not eliminated the gap. For the lawyer, there are two outcomes that follow: the first is that a cost of a particular privacy parameter is not necessarily portable: it depends quite strongly on the difficulty of the underlying task, and so cannot be employed to make a proportionality argument in another deployment. The second is at a more basic level. The cost that remains is not a fault of the engineer for which a fix is needed. Bassily, Smith, and Thakurta (2014) have shown that any non-trivial differential privacy guarantee on a finite database requires some loss of

accuracy. A controller who represents that improved technique will shortly render the trade-off costless is therefore asserting something the mathematics forbids. Upon extraction, the evidence is equally concrete and considerably more alarming. The study which introduced membership inference against machine learning models trained auxiliary systems to imitate the behaviour of a target and used them to determine whether given records had formed part of the target's training data. Tested against genuine commercial prediction services rather than laboratory constructions, the attack achieved a median precision of ninety-four per cent (Shokri et al., 2017). Yeom, Giacomelli, Fredrikson, and Jha (2018) subsequently established the mechanism formally: the vulnerability tracks the degree to which a model performs better upon its training data than upon unseen data, so that the generalisation gap operates as a measurable proxy for privacy risk. That finding is of considerable practical value to a regulator, because it identifies an observable property of a deployed system correlating with the very risk the law is concerned to address. More recently, Carlini et al. (2022) have argued that the conventional method of evaluating such attacks understates them, because what matters to an affected individual is not an adversary's average success rate but whether the adversary can identify some individuals with high confidence; re-evaluated at a stringent false-positive threshold, their proposed attack proved roughly an order of magnitude stronger than its predecessors. Read against the Board's insistence that anonymity be assessed by reference to all means reasonably likely to be used, this methodological point carries direct legal consequence: a controller resting an anonymity claim upon average-case figures has not addressed the standard.

Upon collaborative training the position is more counter-intuitive still. Federated learning is frequently advanced as a structural solution to privacy and to cross-border transfer difficulties alike, upon the footing that raw data never leaves the contributing institution (McMahan, Moore, Ramage, Hampson, and Arcas, 2017). Two lines of research qualify that claim severely. Zhu, Liu, and Han (2019) demonstrated that an observer of a single exchanged update might reconstruct the underlying training images to near pixel-level fidelity, a result extended by Geiping, Bauermeister, Dröge, and Moeller (2020) to more realistic architectures and larger aggregations, though with the attack's effectiveness diminishing as more examples are bundled together and as calibrated noise is introduced. The second line comes from the researchers who devised the differentially private variant of federated training, and who reported candidly that securing a meaningful guarantee at the level of an entire contributing institution, rather than an individual record, demands substantially more participants and more communication rounds than an unprotected system, precisely because the protected unit is larger (McMahan, Ramage, Talwar, and Zhang, 2018). A comprehensive survey of the field lists this problem among its principal unsolved challenges (Kairouz et al., 2021). The legal implication is direct: an institution cannot treat the adoption of a federated architecture as discharging its transfer obligations, because upon the published evidence the exchanged updates may serve as an adequate proxy for the data behind them unless properly calibrated noise has been added and independently verified.

The fourth body of evidence after the above stated is the oldest and most thoroughly replicated. Sweeney (2000), analysing United States census data, found that the combination of five-digit postal code, sex, and full date of birth uniquely identified an estimated eighty-seven per cent of the American population, and demonstrated the point by purchasing publicly available voter registration records and using those three attributes to isolate the hospital discharge record of the sitting Governor of Massachusetts from a dataset the State had released as de-identified. A careful re-analysis by Golle (2006), using more recent census data, placed the figure lower, at sixty-one to sixty-three per cent, while confirming the underlying phenomenon. A methodologically distinct demonstration extended the lesson from demographic attributes to behavioural records: Narayanan and Shmatikov (2008) substantially de-anonymised the Netflix Prize dataset, released stripped of identifiers and containing only film ratings and dates, by cross-reference against publicly posted reviews upon a separate site, upon the strength of the observation that a handful of approximately dated ratings constitutes a highly distinctive signature.

Read together, these four bodies of evidence exhibit a coherence which no one of them displays alone, and it is the coherence which carries the doctrinal argument. Each establishes, from a different direction, that the protective effect of a safeguard is specific to the threat it was designed to meet and specific to the circumstances of its deployment. The cost literature shows that a privacy parameter cannot be transplanted between tasks. The extraction literature shows that residual risk depends upon a property of the particular model rather than upon the technique's name. The federated learning literature shows that an architectural choice may relocate a risk without reducing it. And the re-identification literature shows that a safeguard may perform its intended function faithfully while leaving an adjacent risk wholly untouched. The common lesson is that generic assurances are uninformative. A statement that a controller has 'applied differential privacy', or has 'anonymised the training data', or has 'adopted a federated architecture so that no data leaves the hospital', conveys almost nothing about the residual exposure of any individual whose record was used, because in each case the protective effect depends upon parameters and circumstances the statement omits. It is precisely such statements, however, which presently satisfy regulators and persuade tribunals.

That conclusion should not, however, be pressed further than the evidence will carry it. Candour requires acknowledgment of what this literature does not establish. It is drawn overwhelmingly from image classification and structured tabular data, and from systems considerably smaller than the large generative models now attracting regulatory attention. Whether attack success rates measured against a classifier trained

upon photographs transfer to a language model trained upon a corpus of text is a question upon which the evidence is developing rather than settled. Nor does the literature speak directly to the great majority of deployments, which are neither the commercial prediction services tested in the early attack papers nor the frontier systems examined in the most recent ones, but ordinary systems built by ordinary institutions with ordinary resources. A tribunal should therefore treat published figures as establishing that a risk is real and as indicating its approximate magnitude for systems of a comparable kind, not as supplying a determinate answer for the system before it. That is, however, precisely how technical evidence functions in every other area of adjudication, and it is a considerable improvement upon the present position, in which the risk is neither quantified nor, frequently, adverted to at all.

A related caution concerns not the reach of that evidence but the use which is commonly made of it. The doctrinal significance of the re-identification literature is frequently misstated, and the misstatement matters. That literature establishes that generalise-and-suppress techniques afford weaker protection against record linkage than their proponents supposed. It does not establish, because it was never designed to test, that such techniques afford any protection at all against a model trained upon the treated data subsequently disclosing information through its own behaviour. Those are distinct threats directed at distinct objects: one at a released table, the other at a deployed system. A controller who applies a statutory de-identification safe harbour to a training corpus and infers that the resulting model is anonymous has committed a category error, and the membership inference literature demonstrates the inference to be unsound. The same thing can be expressed in the opposite manner. Even if a system is well-designed and is audited to prevent membership inference, it is still possible for a different method to be used to reconstruct a sensitive attribute: Fredrikson et al. (2015) demonstrated that a facial recognition model's confidence outputs are enough to render a recognisable likeness of a named subject. Protection of privacy is not a scalar value and a safeguard audited against one threat certifies nothing as to another.

5. Six Episodes In Enforcement And Adjudication

The propositions so far offered are not speculative. They have faced the courts, they have faced regulators, and in each case, the problems that arose were instructive because the evidence just described was either not introduced into the record, or was introduced and then presumed, but not examined.

The first episode is also the most direct engagement yet between a supervisory authority and the training of a generative model. The Italian supervisory authority's action against OpenAI is the longest-running one under general data protection law, concerning the training of a generative model. On 30 March 2023, the Garante imposed an immediate temporary limitation on processing of Italian users' data considering the absence of an identified legal basis for the training on personal data scraped from the internet, insufficient age verification and that the breach was not reported (Garante per la protezione dei dati personali, 2023). One month later, following the implementation of an information notice, an opt-out and age-verification, services began once more. The investigation was then resumed and in November 2024 the Authority issued a fine of fifteen million euros, stating that the training had been carried out without a proper legal basis and without adhering to the obligations on transparency and breach notification (Garante per la protezione dei dati personali, 2024). In March 2026, the Tribunale di Roma reversed the decision, but for a jurisdictional reason, not because the training was in itself unlawful, this Court holding that competence over the conduct in question had passed to the Irish authority as lead supervisory authority under the cooperation mechanism (Tribunale di Roma, 2026). The real issue, therefore, has been left completely unresolved after three years of litigation. The fact statement of the initial decision is indisputable: the memorisation risk the Garante has pointed out is exactly the one described in the membership inference literature.

The second episode turns from the legality of the training itself to the reach of the authority asked to police it. The Clearview AI proceedings are a case of a failure of jurisdictional reach rather than of substantive uncertainty. The company has built a facial recognition database through billions of images from public websites without consent from the faces pictured. In 2022, the Italian and French authorities imposed fines of twenty million euros each. The United Kingdom's Information Commissioner imposed a monetary fine of £7.5 million, and an enforcement notice to delete (Information Commissioner's Office, 2022). In October, 2023, the First-tier Tribunal allowed Clearview's appeal, holding that processing directed exclusively to foreign law-enforcement clients fell outside the material scope of the United Kingdom regime (Clearview AI Inc. v. Information Commissioner, 2023). That was overturned in October, 2025 by the Upper Tribunal, which ruled that the Tribunal had erred in law, and remitted the case (Information Commissioner v. Clearview AI Inc., 2025). It has been over three years since the original penalty, and it's not settled yet. Running through the litigation is the company's argument that images already available were in the public domain and did not need any further justification for use, an argument that has been rejected by several authorities, who have invoked the strong protection afforded to biometric identifiers, this time in relation to a training corpus.

Where the first two episodes closed without any ruling upon the merits, the third produced one. The transfer of some 1.6 million identifiable patient records from the Royal Free London National Health Service (NHS) Foundation Trust to Google DeepMind for use in building an app to alert staff to cases of acute kidney injury illustrates the necessity limb operating as intended. The Commissioner found that the Trust was in breach of the then applicable legislation as there was inadequate disclosure of the data to patients; insufficient evidence

that the amount of data transferred was required for the clinical safety testing the Trust had conducted; and no privacy impact assessment carried out prior to the arrangements. A monetary fine was not issued, but the Trust gave a formal undertaking for the way it would be managed in future (Information Commissioner's Office, 2017). The reasoning is interesting because it never argued that collaboration between a hospital and a technology developer is inherently forbidden, but rather that the amount and identifiability of what was being transferred was more than was needed for the stated purpose. The absence of a proper necessity assessment would have been much more difficult to sustain had the parties been able to show, from the published evidence, what a more restrictive architecture would have cost in clinical performance.

The fourth episode concerns a different obligation altogether, and it matters here for what it implies about the architectures now proposed in order to escape that obligation. The Irish Data Protection Commission's fine of 1.2 billion euros and order to suspend transfers of data of European users to the United States from Meta Platforms Ireland (Data Protection Commission, 2023) is based on the Court of Justice's decision in *Data Protection Commissioner v. Facebook Ireland Ltd and Maximillian Schrems* (2020) which invalidated the previous transfer framework and established that additional measures are necessary where the law of the destination State does not provide essentially equivalent protection. The Commission determined that Meta's additional measures were inadequate. It is a decision about transfer not model training, but it bears directly on the architecture increasingly proposed in order to avoid transfer altogether. Consequently, there are good reasons for an authority to consider the update traffic as a transfer in substance if federated training is implemented specifically to ensure that only updates are sent across a border, and if those updates can be inverted to recover a close approximation of the underlying data. The engineering option doesn't wash away the legal issue, it moves it.

The rest of the story is completed by two more episodes. The Court's reasoning in *Google Spain* was based on a right whose technical premise does not survive translation to trained models. The Indian Supreme Court in *Puttaswamy* set out a test whose necessity limb cannot be applied without the very evidence that the reported cases so regularly fail to have. All six episodes have a similar structure. In each, the legal resolution was based on a technical proposition: whether data remained identifiable or how much of an individual's data survived a system, whether there was an empirical basis for a claim of minimisation or anonymisation. In each, the determining authority either examined that question, assumed an answer to it, or avoided it by deciding the case on other grounds. In none, indeed, was the tribunal given a systematic exposition of what the science literature already shows. That is a remediable defect, and it is remediable by the ordinary means of proof.

The institutional variety of these episodes is itself significant. They comprise a constitutional court reviewing a national identification programme, a supranational court construing a directive, a specialised supervisory authority applying a regulation to a technology its drafters did not contemplate, first-instance and appellate tribunals differing upon the reach of that regulation, a sectoral regulator proceeding by undertaking rather than penalty, and a lead authority applying an already-settled transfer question to a multinational undertaking. The difficulty is therefore not an artefact of any particular forum, procedural posture, or standard of review. It recurs wherever an adjudicator is required to characterise a processing architecture without first establishing what that architecture does. That observation also suggests where a remedy might most efficiently be introduced. Because the difficulty is evidentiary rather than interpretive, it can be addressed without legislative amendment, by supervisory guidance as to what a controller must demonstrate and by ordinary judicial insistence upon proof of assertions which are, at present, too often accepted upon their face.

6. Proportionality As An Evidentiary Discipline

What those episodes lacked, and what the literature surveyed above supplies, is a means of testing the technical propositions upon which each of them turned. The argument ultimately comes down to one proposition: selecting an epsilon is a proportionality assessment and should be treated as one. The structure of the two exercises is the same. The smaller the epsilon, the more protection it affords the individual and the more damage is dealt to the utility of the system; the opposite is true for large epsilons. Saying that a particular value of epsilon was justified is equivalent to asserting, first, that no more protective value would have served the legitimate purpose, and second, that the intrusion remaining at that value is proportionate to the benefit obtained. Those are, respectively, the necessity limb and the balancing limb of the proportionality enquiry, articulated in language familiar to an engineer.

The Regulation already supplies the vehicle. Article 35(7) specifies the required content of a data protection impact assessment: a systematic description of the processing, an assessment of its necessity and proportionality, an assessment of the risks to data subjects, and the measures envisaged to address those risks (European Parliament and Council, 2016, art. 35(7)). The submission advanced here is that an assessment which does not engage with the published evidence the accuracy cost of the chosen guarantee for a task of the relevant difficulty, the demonstrated success rate of extraction attacks against comparable systems, the established limits of any de-identification step applied to the corpus has not performed the statutory exercise. It has recited the statutory headings. The distinction between reciting and performing is one with which lawyers are entirely familiar in other contexts, and there is no reason its application here should be thought novel.

Three objections may be anticipated to a substantiation requirement of this type.

(1) That tribunals don't have the competence to review the sort of technical evidence involved. The reply to this is that tribunals do regularly review evidence of similar and greater complexity in patent, medical negligence and competition law cases, and the constitutional cases discussed above do show the capacity in precisely this area.

(2) That requiring empirical evidence would place too much of a burden on smaller controllers. This is a reasonable concern, though it suggests that supervisory authorities should provide general testing guidelines which controllers must meet at relatively low cost, rather than foregoing substantiation entirely; a testing protocol is of course cheaper to implement as standard across all cases than expert testimony for each case, and a public protocol reduces rent-seeking.

(3) That the science will outrun the law; any testing required will be out of date as soon as adopted. But those underlying findings have proved remarkably stable: the association between overfitting and membership disclosure is now more than a decade old and has survived many attempts at refutation; the impossibility of a costless privacy guarantee is of similar vintage; and the inadequacy of identifier removal against a determined re-identification attempt is older still. The underlying conclusions, upon which any doctrinal argument is likely to rely, have proved to be rather robust.

None of these objections defeats the requirement. Each rather indicates the form which it ought to take. Three consequences follow, for three audiences. For supervisory authorities, the Board's case-by-case, likelihood-of-extraction approach to model anonymity is doctrinally correct but presently are not operationalised. It should be given effect through published testing expectations: at minimum a membership inference audit for any model trained upon personal data, and a reconstruction audit for any architecture in which intermediate updates are exchanged. The alternative is to leave the assessment to controller self-certification, which the evidence surveyed above suggests will systematically understate risk, because each of the intuitive proxies for safety unremarkable accuracy, a recognised technique, a parameter borrowed from an unrelated deployment has been shown to mislead in a specific and documented way.

For courts, the implication concerns the treatment of technical evidence. A model's generalisation gap, the published success rates of attacks against comparable architectures, and the established limits of a de-identification method applied to the training corpus are relevant, admissible, and in most cases obtainable facts bearing directly upon whether an erasure obligation has been discharged or a necessity claim substantiated. The methodological stance of the German census judgment, of *S. and Marper*, and of *Puttaswamy* that the architecture of a processing system is a matter for judicial examination rather than deference supplies ample authority for so treating them. What ought not to be accepted, without more, is a bare assertion that a system was trained with differential privacy or upon anonymised data. Such an assertion identifies a category of technique; it says nothing of the parameter chosen, the threat model addressed, or the residual risk retained. For legislatures, and particularly those presently elaborating subordinate rules under recently enacted frameworks, the implication concerns statutory safe harbours. Provisions modelled upon the removal of enumerated identifiers were designed to prevent linkage between a released table and an external identified source. The evidence establishes that they perform that function imperfectly but intelligibly. It equally establishes that they were never addressed to, and do not answer, the distinct question of what a model trained upon the treated data may subsequently disclose. A safe harbour sound for data release and unsound for model training ought not to be drafted, or retained, in terms which fail to distinguish the two.

A final observation concerns the division of institutional labour which this proposal implies. It is sometimes suggested that the difficulty described here should be met by legislation: that a new instrument, specific to artificial intelligence, should displace the general data protection framework in its application to trained models. The argument of this article points the other way. The doctrines are not defective in their formulation; they are defective in their application, because they are being applied upon an assumed rather than an established factual footing. A new instrument drafted upon the same assumptions would reproduce the same defect in fresh language, and the recently enacted risk-tiered regime supplies no reason for confidence to the contrary, since its data-governance obligations presuppose exactly the kind of demonstration that is presently not being made. Instead the answer is to be found in the normal institutional powers already held amongst supervisory bodies that could delineate what need be demonstrated, controllers responsible for demonstrating, and tribunals that could simply reject statements not backed up by evidence. None of these requires new powers; each requires only that an existing power be exercised with an awareness of what the technology really does the awareness the scientific literature is capable of supplying, and which this article has set out to gather together.

7. Conclusion

The difficulty described in this article is not that data protection law failed to foresee the advent of artificial intelligence. Normally, the usual interpretive processes will suffice and technologies that are not anticipated by the drafters of the laws frequently find their way into the legal landscape. The challenge is more specific and less tractable: three of the doctrines on which the framework is largely based, erasure, minimisation and the personal-anonymous distinction, do not simply fail to deal with the issue of machine learning. They presuppose, as a condition of their intelligible operation, a technical state of affairs which machine learning does not produce. A record cannot be erased from a model in the sense that the right to erasure contemplates. The data

operations that make a table anonymous do not render the data anonymous for the purposes of a system trained on it. The cost of minimisation is real, measurable, and is mathematically irreducible, rather than a cost that can be made to disappear by better engineering.

Recognising this does not imply giving up on the doctrines, and nothing here is intended to advocate that course. It requires instead that their application be grounded in evidence rather than assumption. Materials for that grounding are available. The price of formal privacy guarantees has been evaluated for tasks of different difficulty, and has been shown to depend upon that difficulty in ways which forbid the portable citation of figures. Extraction attacks have been shown to be successful on real commercial systems and to track an observable property of the model. Collaborative architectures are known to be vulnerable to a degree that poses a threat to their leading advantage. The boundaries of de-identification have been documented, repeated and developed over 25 years. What has been missing is not evidence but its assembly, and the interpretation of it into the language of a tribunal.

The enforcement record indicates this issue will not be academic for long. The Regulation's cumulative penalties now total over seven billion euros across more than two and a half thousand decisions, the vast majority since the start of 2023 (CMS Law, 2026; DLA Piper, 2026); and a new layer of enforcement is coming with the full implementation of the Artificial Intelligence Act for high-risk systems. The first substantial penalty imposed anywhere against a generative model developer for its training practices has already been imposed and then annulled on jurisdictional grounds, leaving the substantive law precisely where it stood. It is to be hoped that the tribunal which finally decides the question will have before it something more substantial than competing assertions as to what a named technique is supposed to achieve. Supplying that is a problem for legal scholarship as well as technical research, and it starts with the observation that the accountability principle in Article 5(2) of the Regulation requires a controller to know, and to be able to demonstrate, what its chosen safeguards actually do.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308-318.
2. Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th IEEE Annual Symposium on Foundations of Computer Science*, 464-473.
3. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. (2022). Membership inference attacks from first principles. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy*, 1897-1914.
4. CMS Law. (2026). *GDPR enforcement tracker report* (7th ed.).
5. DLA Piper. (2026). *GDPR fines and data breach survey: January 2026*.
6. Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 265-284.
7. Dwork, C., and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211-407.
8. Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322-1333.
9. Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. (2020). Inverting gradients: How easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, 33, 16937-16947.
10. Golle, P. (2006). Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM Workshop on Privacy in the Electronic Society*, 77-80.
11. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., and others. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2), 1-210.
12. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., and Arcas, B.A.y. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273-1282.
13. McMahan, H.B., Ramage, D., Talwar, K., and Zhang, L. (2018). Learning differentially private recurrent language models. In *International Conference on Learning Representations*.
14. Narayanan, A., and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, 111-125.
15. Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, 3-18.
16. Sweeney, L. (2000). Simple demographics often identify people uniquely (Data Privacy Working Paper No. 3). Carnegie Mellon University.
17. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557-570.

18. Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of the 31st IEEE Computer Security Foundations Symposium*, 268-282.
19. Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, 32.
20. LEGISLATION, CASE LAW, AND REGULATORY DECISIONS
21. Bundesverfassungsgericht. (1983). Volkszählungsurteil, BVerfGE 65, 1, judgment of 15 December 1983, 1 BvR 209/83 (Germany).
22. California Consumer Privacy Act of 2018, Cal. Civ. Code § 1798.100 et seq., as amended by the California Privacy Rights Act of 2020 (United States).
23. Clearview AI Inc. v. Information Commissioner, [2023] UKFTT 819 (GRC) (United Kingdom).
24. Data Protection Commission. (2023). Decision in the matter of Meta Platforms Ireland Limited, Inquiry IN-20-8-1 (12 May 2023) (Ireland).
25. Data Protection Commissioner v. Facebook Ireland Ltd and Maximilian Schrems, Case C-311/18, ECLI:EU:C:2020:559 (CJEU, 16 July 2020).
26. Digital Personal Data Protection Act, No. 22 of 2023 (India).
27. European Data Protection Board. (2024). Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models (17 December 2024).
28. European Parliament and Council. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union, L119/1.
29. European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.
30. Garante per la protezione dei dati personali. (2023). Provvedimento del 30 marzo 2023 (Italy).
31. Garante per la protezione dei dati personali. (2024). Provvedimento n. 755 del 2 novembre 2024 (OpenAI) (Italy).
32. Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, Case C-131/12, ECLI:EU:C:2014:317 (CJEU, 13 May 2014).
33. Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191; 45 C.F.R. § 164.514(b)(2) (United States).
34. Information Commissioner v. Clearview AI Inc., Upper Tribunal, 7 October 2025 (United Kingdom).
35. Information Commissioner's Office. (2017). Undertaking given by Royal Free London NHS Foundation Trust (3 July 2017) (United Kingdom).
36. Information Commissioner's Office. (2022). Monetary penalty notice and enforcement notice issued to Clearview AI Inc. (18 May 2022) (United Kingdom).
37. Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1 (Supreme Court of India).
38. Justice K.S. Puttaswamy (Retd.) v. Union of India, (2019) 1 SCC 1 (Supreme Court of India).
39. Personal Information Protection Law of the People's Republic of China (adopted 20 August 2021, in force 1 November 2021).
40. S. and Marper v. United Kingdom, App. Nos. 30562/04 and 30566/04 (ECtHR, 4 December 2008).
41. State v. Loomis, 881 N.W.2d 749 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017) (United States).
42. Tribunale di Roma, Sezione Diritti della Persona e Immigrazione. (2026). Judgment of 18 March 2026 annulling Provvedimento n. 755/2024 (Italy).
43. United Nations. (1948). Universal Declaration of Human Rights, UNGA Res. 217 A(III).
44. United Nations. (1966). International Covenant on Civil and Political Rights, 999 UNTS 171.