The Right to be Forgotten in the Age of AI: Legal, Philosophical, and Technical Challenges

Alessandro Bucci[1], Josep Domingo-Ferrer[2,3], and Anna Monreale[1] Francesca Naretto[1]

[1] University of Pisa, Largo Bruno Pontecorvo 3, Pisa, Italy
a.bucci12@studenti.unipi.it, name.surname@unipi.it
[2] Universitat Rovira i Virgili, Av. Països Catalans 26, Tarragona, Catalonia
josep.domingo@urv.cat
[3] LAAS-CNRS, Université de Toulouse, 7 Av. du Colonel Roche, Toulouse, France

Abstract. The increasing availability of Artificial Intelligence systems in everyday life has drawn growing attention to the right to erasure, as defined by the General Data Protection Regulation. This right raises questions about how learned information can be removed from trained Artificial Intelligence models.

In this paper, we examine the problem of machine UNLEARNING from legal, philosophical and technical perspectives. We highlight the misalignment in terminology, as well as between legal expectations, which require data erasure at the individual level, and current technical methods, which typically operate at the class level. In addition, we explore the practical difficulties in evaluating the effectiveness of UNLEARNING, particularly the limitations of using privacy attacks, such as Membership Inference Attack, as evaluation tools.

Considering the several challenges in this setting, the lack of standardized metrics for UNLEARNING, and the difficulty of assessing whether a model has truly forgotten specific data, we see the need for developing a unified evaluation framework at the European level. Such a framework should define robust, transparent, and reproducible criteria for assessing UNLEARNING performance.

1 Introduction

In recent years, the integration of Artificial Intelligence (AI) systems into everyday life, especially in high-impact domains, such as healthcare, finance or education, has attracted considerable attention from researchers, industry stakeholders as well as policymakers. This is reflected in the development of various regulatory frameworks. Among them, the European Union's AI Act (AIA) is one of the most widely discussed. Other notable initiatives include the regulations in the United States [12] and regulations on the AI in various countries, such as China and Japan [3,13]. In the remainder of this work, we focus on the AIA and the GDPR.

Among the many challenges these regulations seek to address, one topic has gained increasing relevance: *machine UNLEARNING*. The core idea is that individuals whose data have been used to train an AI system may later wish to have that data removed. While this issue is not explicitly covered by the AIA, it is clearly addressed in the General Data Protection Regulation (GDPR), which defines the *right to erasure* (or right to be forgotten) as a fundamental element of user privacy.

In the context of modern Machine Learning (ML) models, and particularly Large Language Models (LLM), UNLEARNING is not only a matter of deleting records from a training dataset. In fact, during the training process, AI systems capture and encode patterns and information about the data they process. As demonstrated in a growing body of literature on privacy attacks, such as re-identification and membership inference, it is often possible to extract sensitive information, and even re-identify records, from trained models [20]. Therefore, achieving erasure requires more than simply removing the original data from the training set. In particular, it demands the removal of the AI model's internal representation of that data.

In this paper, we address the problem of UNLEARNING by analyzing it from legal, philosophical and technical perspectives. In Section 2, we introduce the topic considering each of these three domains separately. Then, in Section 3, we present our perspective on the matter by comparing the desiderata emerging from legal and philosophical considerations with the current technical advancements in the field. Section 4 concludes the work.

2 Related works

2.1 Unlearning from a legal perspective

The right to privacy holds a central place in legal systems, particularly within the European Union. This is evident in the importance attributed to the General Data Protection Regulation (GDPR) [6], which provides a comprehensive framework for the protection of personal data. In recent years, the rise of Artificial Intelligence (AI) has brought renewed focus to privacy issues, and among them, the concept of *machine UNLEARNING* has become increasingly relevant.

In particular, legal scholars are focusing on examining the applicability of machine UNLEARNING within existing regulatory frameworks, focusing mostly on the GDPR [6] and the AI Act (AIA) [7]. Their analysis usually regards *"an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier"* ARTICLE 4 (1) GDPR, also called data subject.

One interesting point of view addressed by [14] is the understanding of the legal status of an AI system i.e. if a model could be regarded as personal data or not: ARTICLE 4 (a)-(b) GDPR refers to *personal data* as any information related to an identified or identifiable person, and *processing* as any operation carried out on personal data. For the authors, if the model allows to indirectly identify

data subjects, then the model itself may be considered as personal data. They suggest that if the model does not leak personal data from the training set, a request for erasure could be complied with through an approximation unlearning method rather than an exact unlearning method.

In another work among the published in this field of research, Zhang et al. [4] describe four main individual (or data subject) rights:

- *Right to be Informed*: [ARTICLE 13-14 GDPR] Individuals have the right to be informed when their data is collected and used;
- *Right of Access*: [ARTICLE 15 GDPR] Individuals have the right to request information about their personal data, including whether such data have been processed and, if so, how they have been used;
- *Right to Rectification:* [ARTICLE 16 GDPR] Individuals have the right to request that their personal data be rectified by the controller if the data are inaccurate or incomplete;
- *Right of Erasure* (or Right to be forgotten): [ARTICLE 17 GDPR] The data subject may request the erasure of their personal data without undue delay. This is typically interpreted as within one month, extendable to two in complex cases;

To enforce these four privacy rights, the authors focus on Large Language Models (LLM) and consider two main scenarios: *chat history*, (e.g. the chat and all its history the user has with the LLM) and *in-model data* (e.g. the embedded data inside the LLM). In the first case, for allowing the four privacy rights, the data subject may be granted the capability to explicitly consent to, or refuse, the use of their personal data by the chatbot provider. Regarding the second case, instead, the authors highlights the difficulty of enforcing the four rights in this complex scenario. To overcome it, they propose a possible solution, structured in different components that tackle different rights. Firstly, they suggest the application of privacy preserving machine learning approaches, such as Differential Privacy (DP) [5]. In addition, to tackle the right to rectification and erasure, they propose the usage of machine UNLEARNING as a method to adjust the original model, as well as *band-aid* approaches that filters or correct the model output rather than operating on the dataset or the model itself.

Another interesting work in this setting is presented in [19]. It deals with the topic of privacy and data protection in General Purpose AI, focusing on *balancing test*, sensitive data, data leakage and data minimization. In particular, they refer to the balancing test (ARTICLE 6(1)(f) GDPR), i.e. a test to evaluate whether a controller's legitimate interest in processing data is proportionate and does not override a data subject's rights. In [19], the authors argue that AI training can be justified on the basis of the balancing test, which considers data processing lawful if it is necessary and proportionate. They further note that the application of privacy protection strategies to the data supports compliance under this legal basis (ARTICLE 25 GDPR). As for *sensitive data treatment*, the authors highlights a misalignment between the GDPR and the AIA. In particular, they refer to the ARTICLE 9 GDPR and ARTICLE 10(5) AIA. Technically, the GDPR prohibits the processing of sensitive data, since it can be target by privacy

attacks. On the contrary, AIA allows the exceptional processing of special categories of personal data, such as sensitive data, for bias detection and correction. For this reason, the authors suggest to redefine ARTICLE 9 GDPR. For the *data leakage* the authors tackle the problem of model inversion attacks: privacy attacks during which an attacker tries to infer personal information about a data subject, by exploiting the outputs of a ML model. The presence of such attacks mine the *right to be forgotten* (ARTICLE 17 GDPR), since malicious adversaries may be able to find information hidden in the AI system. For this reason, is necessary to find techniques that allow a correct unlearning of the unwanted information. The authors consider the possibility of treating LLM as personal data: this implications could allow data subjects to invoke their *right to be forgotten* under ARTICLE 17 of the GDPR, asking for the removal of the entire ML model. If the model is not considered personal data, there is still the challenge to remove specific data subject record from an AI model. As a solution they propose machine UNLEARNING methods, even if they acknowledge that they are still in development and may not be universally applicable to all models or data. In [19], the authors also address the principle of *data minimization*, which requires that personal data be collected only when it is relevant and necessary for a specific purpose. This principle is articulated in ARTICLE 5(1)(b)–(c) of the GDPR, and is further reflected in ARTICLE 10(2) of the AIA, which requires that high-risk AI systems be developed using high-quality datasets. As a way to reconcile these requirements, the authors propose the use of DP, which allows for the inclusion of statistical patterns from data while limiting the identifiability of individuals, thereby supporting both legal compliance and technical feasibility.

2.2 Unlearning from a philosophical perspective

The topic of UNLEARNING has also been tackled from a philosophical perspective. Among the different contributions in this setting, we report the analysis of Floridi in [16] since they perfectly aligns with our future discussion. In this paper the author highlights the difference between making information *inaccessible* or *unavailable*. In the first case, to make information inaccessible we need to remove it completely. Instead, unavailable refers to the possibility of blocking/masking information, without removing it. Clearly, machine UNLEARNING refers to the process of making information inaccessible. However, there are currently many AI systems that perform procedures of masking for making information unavailable. In this context, the author underlines that the main limit when making information unavailable is that there is always a risk to make information accessible again. Regarding the UNLEARNING, Floridi finds three main challenges that may arise and we should consider when developing regulations and technologies:

- *Misuse:* In the wrong hands, UNLEARNING can be used to censor, spread misinformation, propaganda and vandalism.
- *Overuse:* There is a risk that model developers may prioritise data collection and training without regard for existing legal constraints, assuming that any non-compliant data can be removed later via UNLEARNING.

– *Underuse:* Masking can be always cheaper than UNLEARNING, and unless UNLEARNING is required by law, masking could be preferred for economic reasons.

After the presentation of these challenges, the author concludes its work by writing that *"It has taken millennia to develop a culture of recording. It will not take decades to refine one of deleting, but it seems that MU* (Machine UNLEARNING) *will likely be part of such development."*

2.3 Unlearning in computer science

In this section we provide a high level overview of the machine UNLEARNING methods available from a technical point of view. Given the numerous publications in this setting, in the remaining of this paper we categorise UNLEARNING methods in two categories: `By-design`, which are methods that embeds the UNLEARNING in their structure [2,10]; and *post-hoc*, which refers to methods that apply UNLEARNING after the training of generic AI system. The category of post-hoc procedures is further divided in *ex-novo* (`Post-ExNovo`) and *in place* (`Post-InPlace`). In the first case, we refer to methods that do not modify the original AI system, but create a novel AI system, starting from the original one and the UNLEARNING constraints [21,1,15]. The other category, instead, directly modifies the original AI model [9,11]. An overview of the taxonomy is reported in Figure 1. In the remaining of this section, we refer to the AI systems as Machine Learning (ML) models to be compliant with the computer science terminology.

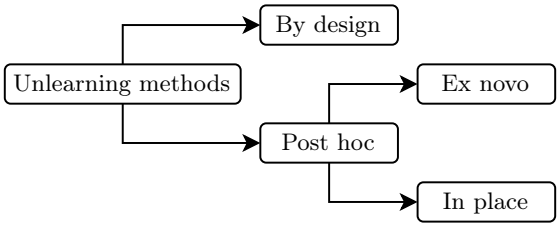


Fig. 1. Unlearning methods taxonomy

By-design methods In this category, we find methods that embeds the UNLEARNING directly in the structure of the ML model. Therefore, they integrate in their architecture the capability to unlearn information in a reliable way [2,10].

One of the most promising techniques in this context is called Sharded, Isolated, Sliced, and Aggregated (SISA) [2]. In this case, training data are split into shards, and each shard is further divided into slices. At this point, on each shard, they train a ML model, exploiting the slices for batch-level training. After the training, all the ML models are aggregated. At running time, if a specific record that needs to be forgotten is inside one slice of a shard, the corresponding ML model is retrained from the last parameters saved before training on the

slice containing that record. The proposed methodology is promising, even if its evaluation is tailored to the computational times. As mentioned in the Section 2.1, validating the correctness of UNLEARNING is a difficult task, but in general requires performing some privacy attacks, that are missing in this work. In the context of `By-design`, we find also the work of Graves et al. [10]. The authors propose Amnesiac Unlearning, in which they store all the batch updates containing sensitive data, then, when there is a request to forget some records, they remove the contribution to the ML model parameters concerning the update of the batches containing the records to forget. It has to be noted that implementing this approach using a pre-existing model to fine-tune is straight-forward, unlike SISA that needs a tailored architecture. They propose an empirical evaluation, comparing the accuracy of their approach w.r.t. the *exact unlearning*, i.e. the ML model trained from scratch without the records to forget. They also consider a privacy-related evaluation, attacking their model with *Model Inversion Attacks* [8] and *Membership Inference Attacks* [20,17].

Post-ExNovo methods In this context we find SCalable Remembering and Unlearning unBound (SCRUB)[15], in which they exploit a standard teacher-student paradigm. Technically, the original model is the teacher, and a new student model is trained to perform correctly on all of the records, a part from the ones that are in the forget set. Hence, we can think of it as a student that is trained to obey the teacher in the majority of the cases, but disobey for the forget set. In addition, to avoid over forgetting they use checkpoints to rewind the model to more plausible outputs. In this paper, the evaluation of UNLEARNING is tackled by considering two main directions. The first one is technical, while the second one is related to the compliance with the GDPR. In particular, they evaluate their method using *Membership Inference Attacks*.

Another contribution in this setting is from Bonato et al., which introduce Selective-distillation for Class and Architecture-agnostic unleaRning (SCAR)[1]. It is again a teacher-student paradigm. In this method, the original teacher model performs unlearning in the feature space by shifting the embeddings of the records to be forgotten toward the nearest incorrect class. This operation introduces confusion specifically for those samples, hence reducing the model's ability to retain information about them. Then, a student model is trained via knowledge distillation using a set of unrelated data. During this process, the teacher model serves as the label provider, guiding the student while excluding direct exposure to the forgotten data. Also in this case, the evaluation is conducted considering both the technical and the privacy point of view, even if only the Membership Inference Attack is considered.

Post-InPlace In this category we find one of the most popular UNLEARNING methods: Fisher Forgetting [9]. The main intuition is that, to forget the data, we can inject noise into the network's weights. While the idea is interesting, the evaluation conducted is merely technical and does not prove, in our opinion, privacy protection.

Another interesting work is Certified Data Removal [11], in which the authors apply a one-step Newton update to the model's weights in order to negate the

influence of one or more data points targeted for unlearning. To further obscure any residual effects that could potentially allow reconstruction or tracing back to the removed data, a small amount of random noise is added to the updated weights. Their method was evaluated on model accuracy, the speedup between their method and exact unlearning and they also tracked the gradient residual norm, which indicates, if small, that the method closely approximates a model on which exact unlearning was applied.

Wang et al.[21] propose Knowledge Gap Alignment (KGA), a method that guides unlearning by minimizing the gap between a smaller auxiliary model's predictions on forget data and the original model's predictions on unseen data. In practice, this approach involves training an external model to steer updates on the original one. Also in this case, evaluation is conducted by considering a technical point of view and a privacy one, with Membership Inference Attack.

Last but not least, Graves et al. [10] introduce an UNLEARNING algorithm that exploits relabeling. If the UNLEARNING regards an entire class, the target class labels are replaced with randomly selected incorrect labels. On the contrary, when we aim at UNLEARNING only a set of records, each of them is replaced with a small number of randomly mislabeled copies of it. After the relabeling, the new dataset is used to fine tune the ML model. Also in this case, they propose an empirical evaluation and a privacy-related evaluation, attacking their model with *Model Inversion* and *Membership Inference Attacks*.

3 Legal and Technical Perspective on Machine UNLEARNING

In this section, we examine the state-of-the-art approaches for machine UNLEARNING in computer science, evaluating which technical directions align with legal requirements, particularly those outlined in the GDPR, and identifying the open challenges that remain.

Before proceeding, we note that some authors argue that privacy-preserving technologies, such as DP, can mitigate the need for UNLEARNING by reducing the differentiability of individuals within a dataset. While these techniques do play an important role in safeguarding personal information, they do not inherently support the right to rectification or erasure. For this reason, and despite some references in the literature equating them with UNLEARNING [4,19], we do not consider privacy protection techniques as valid substitutes for UNLEARNING mechanisms in this discussion. Instead, we argue that a comprehensive and effective framework for UNLEARNING should combine multiple strategies: (i) *opt-out* mechanisms, to prevent inclusion of data in the first place, (ii) *privacy protection* technologies, to reduce exposure risk for all the users in the data, and (iii) dedicated *UNLEARNING protocols*, to address post-inclusion erasure requests. Given this context, in the remainder of this section we focus our discussion on point (iii), namely UNLEARNING strategies. Hence, we assume that the deployment of the ML model is conducted responsibly, incorporating both opt-out mechanisms and privacy protection technologies as preliminary safeguards. The remainder of

this section is organised as follows. In Section 3.1, we address the issue of inconsistent terminology used in legal and philosophical contexts compared to the technical setting. In Section 3.2, we argue for the necessity of developing individual UNLEARNING strategies, given that most existing techniques in computer science focus on class UNLEARNING, but the legal frameworks require individual UNLEARNING. In Section 3.3, we analyze the open challenge of evaluating the effectiveness of UNLEARNING strategies.

3.1 Understanding the Terminology

As presented in Section 2, machine UNLEARNING refers to the process by which an AI system forgets specific training data, without compromising its overall performance. This task is particularly relevant given the privacy regulations, in particular the GDPR, which grants under ARTICLE 17 to individuals the right to request the removal of their personal data from datasets used to train AI systems. As already mentioned, the simple removal of the record from the training dataset is not enough to guarantee the privacy of the individual, therefore technical solutions should be developed to support the enforcement of such rights even in the AI and LLM scenarios.

In Section 2, we categorise UNLEARNING methods into three main families: `By-design`, `Post-ExNovo`, and `Post-InPlace`. In this work, to find common definitions from the technical, legal and philosophical perspective, we adopt the term *exact UNLEARNING* for methods that ensure the model behaves as if it had been retrained from scratch without the data to be forgotten. This definition includes both `By-design` approaches, where the model is built with unlearning capabilities by construction and hence has a strong unlearning guarantee, and `Post-ExNovo` approaches, which achieve unlearning by instantiating a new model trained from scratch without the target data. Both aim to guarantee that the influence of the forgotten data is entirely removed, either structurally or through complete re-training.

By contrast, we consider `Post-InPlace` methods, i.e. those that directly modify the parameters of an already trained AI model without full retraining, as *approximate UNLEARNING*. These techniques typically offer only partial removal of learned representations and lack formal guarantees that the resulting model is equivalent to one trained without the forgotten data. While more efficient in practice, they introduce uncertainty regarding the completeness and verifiability of the unlearning process, particularly in the context of complex or large-scale models.

In practice, performing exact UNLEARNING, i.e. retraining a model from scratch without the data to be forgotten, is often infeasible in real-world applications. Contemporary ML models are computationally intensive, and full retraining for each deletion request is impractical.

Although `By-design` and `Post-ExNovo` methods are conceptually better aligned with the objectives of exact UNLEARNING, they still are in their infancy. Current methodologies face significant technical limitations. For example, their performance may degrade when large volumes of data need to be removed, particularly

if the forgotten data influence multiple components or deeply embedded latent representations within the model. As a result, these approaches require further research before they can be considered reliable and scalable in real-world scenarios.

Another important aspect, highlighted by [16], is the distinction between making information *unavailable* and *inaccessible*. The right to erasure does not clarify whether information should be made inaccessible by blocking its output or entirely unavailable by removing it from the AI system using UNLEARNING approaches. Currently, blocking techniques are widely adopted due to their lower computational costs compared to UNLEARNING. These methods also preserve the AI model's overall functionality, as they affect only the output rather than the internal representation. This contributes to an underuse of UNLEARNING, as anticipated by [16], with blocking emerging as the preferred approach, especially in large-scale AI systems like LLM. We remind the reader that blocking strategies tackle the privacy protection by providing a security shield. However, this does not have any privacy guarantees and can be subjected to security attacks.

It is also worth to notice that machine UNLEARNING can be considered as a trade-off between blocking and retraining from scratch. Blocking acts like placing a barrier between the user and the unwanted information, while retraining eliminates that information by rebuilding the AI system without it. UNLEARNING, by contrast, removes the unwanted data from the AI but may leave behind some residual traces. As a result, it may offer a greater resilience than blocking and is in general less computationally expensive than retraining from scratch. For these reasons, in the remaining of this paper, we focus on UNLEARNING techniques and not on blocking ones.

3.2 Individual vs. Class UNLEARNING

Another limitation we encountered while analyzing the state-of-the-art approaches concerns the granularity of data removal. In fact, the GDPR explicitly demands record-level compliance: users may request the erasure of their individual data, not just from the training data but also from the ML model. A well-known example of this principle is the case of Mario Costeja González, who requested the removal of his personal information from the Google search engine [4].

In contrast, most existing work on machine UNLEARNING, as reported in Section 2, has focused on the removal of entire classes of data. To illustrate this distinction, consider a machine learning model developed for healthcare, with the objective of predicting the presence of cancer cells and the corresponding cancer stage (stages from 1 to 6, with increasing gravity). In many computer science settings, UNLEARNING is typically performed at the class level, for example, removing all data associated with *cancer level 4*. In contrast, compliance with the GDPR often requires the removal of a specific individual's data from the training set. While this individual naturally belongs to a class (e.g., cancer level 4), the objective is not to eliminate the class entirely but to remove all traces of that individual's presence in the model. After class UNLEARNING, the model no longer retains information about any member of the removed class. However,

after individual UNLEARNING, the model should still retain the knowledge of the broader class, excluding any influence of the specific record that was deleted.

From a technical standpoint, class UNLEARNING simplifies the task, as it targets coherent and high-volume patterns. In practice, this often means modifying only a few regions of the ML model where that specific class was encoded. However, record UNLEARNING introduces significantly more complexity. First, it is more difficult to trace and isolate the influence of a single data point within the model's parameters. Second, ensuring complete removal without excessive degradation of predictive performance is inherently challenging. If too much information is removed, model performance may suffer; if too little is removed, residual patterns may persist, thereby failing to meet the UNLEARNING objective.

After having clarified the important distinction between class and individual UNLEARNING, we note that some methods in the computer science literature, such as [10,2], introduce individual-specific strategies, including relabeling and selective fine-tuning. These strategies can be considered `Post-InPlace` techniques, since they directly modify the ML model. While these approaches appear promising from a theoretical standpoint, they raise concerns about whether they actually remove learned associations or merely obscure them. In particular, relabeling strategies typically operate as follows: the records we aim at UNLEARNING are assigned incorrect labels, and the ML model is then fine-tuned on this *erroneous* data. Rather than enforcing UNLEARNING through the removal of learned information, this method introduces additional learning paths, such that the final model becomes unable to distinguish between the original record with the incorrect label and the other newly introduced instances. This method is interesting from a privacy standpoint. In fact, relabeling for fine-tuning confuses the AI model by reinforcing incorrect patterns. As a result, the vulnerability of the model to privacy attacks, such as Membership Inference Attacks, may be significantly reduced. However, in real-world applications, where UNLEARNING may be required for hundreds or even thousands of records, this approach can lead to a significant reduction in the overall predictive performance of the model, similarly to the other approaches.

3.3 Evaluation of UNLEARNING Techniques

In recent years, the field of machine UNLEARNING has received significant attention within the computer science community. Numerous techniques have been proposed and presented at top-tier conferences, targeting specific tasks, different kind of data (e.g. images, tabular, time series etc.) or ML architectures (e.g., CNNs, transformers, and even LLMs).

However, despite this growing interest, it is evident from the analysis of the state of the art that the evaluation of UNLEARNING methods remains an open challenge. In particular, determining whether UNLEARNING has occurred effectively is a non-trivial task.

From a technical perspective, most evaluation strategies for class UNLEARNING rely on performance-based heuristics, such as verifying that the model pro-

duces incorrect predictions for the forgotten class. While this may be suitable in controlled or ad hoc scenarios, these heuristics become less reliable as the complexity of the system increases. For example, in the case of complex ML models, such as multi-modal or multi-label classifiers or LLMs, evaluating the model's general performance is already a difficult task and hence the evaluation of UNLEARNING even more so.

A possible approach to evaluating UNLEARNING techniques involves the use of *privacy attacks*. In recent years, several privacy attacks have been developed to assess the extent to which ML models retain information about their training data. These attacks typically assume the presence of an adversary who leverages some public information to extract private details from the model. The underlying intuition is that ML models often encode patterns and characteristics of the data they were exposed to during training. One of the most widely studied techniques is the *Membership Inference Attack* (MIA) [20], in which the objective is to determine whether a specific record was included in the training set. While MIA is the most known attack in this setting, other types of privacy attacks exist that aim to infer sensitive attributes or reconstruct parts of the training data. In this landscape, it is difficult to determine whether the application of a single privacy attack is sufficient to validate the goodness of the UNLEARNING, or whether it is necessary to perform a broad range of available privacy attacks to adequately assess the effectiveness of UNLEARNING from different perspectives.

However, regardless of whether one or multiple privacy attacks are used, this type of evaluation strategy provides only partial validation. We argue that privacy attacks offer *necessary but not sufficient* conditions for assessing the effectiveness of UNLEARNING. A commonly proposed evaluation pipeline [10,15,1,21] involves (i) performing a privacy attack against the ML model before UNLEARNING, (ii) applying the UNLEARNING process, and then (iii) repeating the same attack on the unlearned ML model. Ideally, the target record or class to be forgotten should no longer be identifiable after UNLEARNING.

While this approach is conceptually sound, privacy attacks are not perfectly accurate. In particular, we identify two main problems that may arise in this setting:

1. *Missed Detection Before UNLEARNING*: It may occur that the record intended to be forgotten is not detected by the privacy attack even before the UNLEARNING process, resulting in inconclusive outcomes. In such cases, the evaluation pipeline fails to provide any meaningful signal regarding the effectiveness of UNLEARNING.
2. *Uncertain Result After UNLEARNING*: Even if the initial attack successfully identifies the target data, the failure of a subsequent attack does not necessarily imply that UNLEARNING has taken place. This outcome may instead indicate a weakness in the attack technique itself or unrealistic assumptions about the attacker's prior knowledge.

Another problem that we identify is that the effectiveness of privacy attacks as an evaluation strategy is constrained by the current state of the art. There is no

guarantee that future, more powerful attacks will not reveal residual information presumed to be forgotten since they were evaluated on different and less powerful attacks [18].

After a careful analysis of the state of the art, we argue that relying on privacy attacks as an evaluation strategy presents substantial limitations and may not represent a reliable nor comprehensive solution. However, if privacy attacks are to be adopted as a valid and standardized method for evaluating UNLEARNING, the research community must first reach an agreement on several key aspects, such that: which types of attacks to consider for the evaluation, the assumptions under which they should be conducted, and the datasets on which they should be tested. Without such consensus, each UNLEARNING method may appear effective only within a narrow and selectively constructed experimental setting, and therefore cannot be regarded as a generally reliable or robust solution.

For these reasons, we advocate for the establishment of a *common evaluation framework* at the European level, ideally overseen by an independent authority. The purpose of such a framework would be to define robust, transparent and reproducible standards for evaluating the effectiveness of UNLEARNING techniques. At present, no official common evaluation framework exists at the European level for assessing machine UNLEARNING methods. However, the European Data Protection Supervisor has acknowledged the relevance of machine UNLEARNING in the context of data protection and the right to be forgotten. In particular, it has emphasized the need to develop reliable techniques for removing data from ML models without compromising their overall performance[4].

4 Conclusions

Machine UNLEARNING has gained increasing attention with the rise of powerful ML models and especially LLM. The topic lies at the intersection of legal rights, philosophical principles, ethical concerns and technical feasibility. While the GDPR provides a normative foundation for the right to be forgotten, current technical approaches often do not meet these expectations, especially with respect to granularity, practical effectiveness and verifiability.

We have first analyzed the terminology, showing that there are various terms used in UNLEARNING for different purposes. In particular, we emphasize the distinction between blocking and UNLEARNING. We argue that UNLEARNING is currently underused, as defined by Floridi [16], due to the presence of blocking approaches, which are less computationally expensive and do not decrease the performance of the AI model. However, we have also observed that UNLEARNING offers a less costly alternative compared to full retraining.

Our analysis has shown that most existing technical methods focus on class removal, which is insufficient to satisfy legal demands for individual control over

[4] Link to the article titled Machine UNLEARNING from the European Data Protection Supervisor: `https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/machine-unlearning_en`

personal data. In addition, the lack of standardized evaluation criteria, combined with the limitations of existing privacy attacks as proxies for unlearning verification, limit the actual assessment of these techniques.

We argue that a robust and trustworthy framework for machine UNLEARNING must combine opt-out mechanisms, privacy-preserving technologies and dedicated UNLEARNING protocols. To support its development and deployment, we see the need for the establishment of a common evaluation framework at the European level, ideally supervised by an independent authority. This framework should define transparent, reproducible and legally grounded criteria for measuring the effectiveness of UNLEARNING techniques.

References

1. Bonato, J., Cotogni, M., Sabetta, L.: Is retain set all you need in machine unlearning? restoring performance of unlearned models with out-of-distribution images. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 1–19. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-73232-4_1
2. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C.A., Jia, H., Travers, A., Zhang, B., Lie, D., Papernot, N.: Machine unlearning. In: 2021 IEEE Symposium on Security and Privacy (SP). pp. 141–159 (2021). https://doi.org/10.1109/SP40001.2021.00019
3. China: The china regulation: New generation artificial intelligence development plan. Link (2017)
4. Dawen, Z., Pamela, F.B., Thong, H., Shidong, P., Zhenchang, X., Mark, S., Xiwei, X.: Right to be forgotten in the era of large language models: implications, challenges, and solutions. AI and Ethics (2024). https://doi.org/10.1007/s43681-024-00573-9
5. Dwork, C.: Differential privacy. In: International colloquium on automata, languages, and programming. Springer (2006)
6. European Parliament and Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (Apr 2016)
7. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Jun 2024)
8. Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., Chen, B., Wang, X., Xia, S.T., Xu, K.: Privacy leakage on dnns: A survey of model inversion attacks and defenses (2024), https://arxiv.org/abs/2402.04013
9. Golatkar, A., Achille, A., Soatto, S.: Eternal sunshine of the spotless net: Selective forgetting in deep networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9301–9309 (2020). https://doi.org/10.1109/CVPR42600.2020.00932
10. Graves, L., Nagisetty, V., Ganesh, V.: Amnesiac machine learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11516–11524 (2021)

11. Guo, C., Goldstein, T., Hannun, A., Van Der Maaten, L.: Certified data removal from machine learning models. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3832–3842 (2020). https://doi.org/10.5555/3524938.3525297
12. Jakob Mökander, Prathm Juneja, D.S.W., Floridi, L.: The us algorithmic accountability act of 2022 vs. the eu artificial intelligence act: what can they learn from each other? In: Minds & Machines. vol. 32, pp. 751–758 (2022). https://doi.org/https://doi.org/10.1007/s11023-022-09612-y
13. Japan: Japan: the development of artificial intelligence. Link (2019)
14. Juliussen, B.A., Rui, J.P., Johansen, D.: Algorithms that forget: Machine unlearning and the right to erasure. Computer Law & Security Review **51**, 105885 (2023). https://doi.org/https://doi.org/10.1016/j.clsr.2023.105885
15. Kurmanji, M., Triantafillou, P., Hayes, J., Triantafillou, E.: Towards unbounded machine unlearning. Advances in neural information processing systems **36**, 1957–1987 (2023). https://doi.org/10.5555/3666122.3666217
16. Luciano, F.: Machine unlearning: Its nature, scope, and importance for a "delete culture". Philosophy & Technology **36**(2), 42 (2023). https://doi.org/10.1007/s13347-023-00644-5
17. Monreale, A., et al.: Agnostic label-only membership inference attack. In: 17th International Conference on Network and System Security. Springer (2023). https://doi.org/10.1007/978-3-031-39828-5_14
18. Naretto, F., et al.: Evaluating the privacy exposure of interpretable global and local explainers. Trans. Data Priv. (2025)
19. Novelli, C., Casolari, F., Hacker, P., Spedicato, G., Floridi, L.: Generative ai in eu law: Liability, privacy, intellectual property, and cybersecurity. Computer Law & Security Review **55**, 106066 (2024). https://doi.org/https://doi.org/10.1016/j.clsr.2024.106066
20. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP) (2017)
21. Wang, L., Chen, T., Yuan, W., Zeng, X., Wong, K.F., Yin, H.: KGA: A general machine unlearning framework based on knowledge gap alignment. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13264–13276. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.acl-long.740