Forgetting by Design: Testing the Effectiveness of Machine Unlearning in Right to Be Forgotten Data Deletion

Jericka Guy
Computer Science Department
Hampton University
Hampton, VA

The Right To Be Forgotten (RTBF) is a legal requirement that, when implemented, allows individuals to request their information be deleted from digital media. However, with the use of machine learning models in today's modern age, full data deletion can be technically challenging. In this research paper, the author will evaluate the effectiveness of machine unlearning as a complete data-deletion method for complying with RTBF. The research method uses a pre-trained neural network, tested with varying sizes of a forget set and membership interference attacks (MIA), to determine whether the model retains information from deleted data. The results highlight the limitations of machine unlearning and the implications it can have on RTBF.

Keywords: *Cybersecurity, Right to Be Forgotten (RTBF), Machine Learning, AI*

I. Introduction

The Right To Be Forgotten (RTBF) principle allows individuals to request the removal of their personal data from various search engines and online databases. Created by the European Union under Article 17 of the General Data Protection Regulation (GDPR), RTBF is not a global law but a concept that gives individuals the right to control and manage their personal information. Whatever the reason to request removal from digital media, the RTBF offers individuals the freedom and privacy to make the request and secure their personal information. The ability to abide by and honor these requests has been very difficult in this interconnected digital age. However, to avoid the legal and reputational consequences of noncompliance, many organizations have developed and implemented various deletion policies. Although implementation may vary depending on the privacy laws in the many states and countries where RTBF is enforced, it is evident that RTBF has influenced internet usage and the operation of search engines and online databases. Many organizations now voluntarily extend deletion policies to all their users, and some even promise complete deletion on their sites. However, with artificial intelligence (AI) and machine learning (ML), the deletion process is complex and not always straightforward or easy to implement. (Richards, 2020).

The use of AI and Machine-learning pipelines as a significant process in modern society creates a layer of difficulty when deleting personal data. Sites often use

personal data to train models using patterns and relationships to optimize performance. Personal data can be stored in model parameters, feature stores, embeddings, or derived datasets. This unique ability to process personal data makes traditional deletion very difficult, as specific prompts to delete often fall short and sometimes not recognized. In recognized cases, sometimes even after the original dataset is deleted, there may be intermediate artifacts, such as training checkpoints, preprocessing outputs, and batch logs, that retain information derived from the user's personal data. Primary data records do not account for derivative representations embedded in trained models or distributed systems. Residual traces can sometimes be hidden from standard deletion workflows. These technical realities demonstrate that achieving full compliance with RTBF in ML systems is difficult in real-world practice.

The presence of personal data despite deletion and attempts to comply with the requests raises many concerns across privacy, cybersecurity, and data governance. Incomplete deletion can leave individuals vulnerable to internal misuse, external exploitation, and data re-identification. Organizations face legal, reputational, and regulatory risks when they are unable to meet individual requests and expectations. Machine unlearning has emerged as a promising technical approach to address these limitations. Unlike standard deletion, unlearning algorithms aim to selectively remove the influence of specified data from trained machine-learning models without requiring full retraining. The process involves partitioning data into retain and forget sets, then using unlearning strategies to change the models. This results in the models behaving as if the forgotten data was never seen, and supports the goals of RTBF in AI environments. However, the effectiveness of machine unlearning is not absolute and depends on factors such as the size of the forget set, model complexity, and the evaluation techniques used. Membership Inference Attacks (MIA) will test and measure whether a model still retains knowledge of the forgotten data.

This research investigates the reliability of machine unlearning as a mechanism to support RTBF in modern AI systems. By analyzing distributed system architectures, training pipelines, and residual model influence, the study evaluates how well machine unlearning can enforce data deletion in practice. Using experiments that vary the size of the forget set and measuring model performance and vulnerability, this paper provides insights into the capabilities and limitations of unlearning as a privacy-preserving tool. The findings highlight both the promise and the challenges of using AI-based deletion techniques to meet legal, reputational, and regulatory obligations in complex digital infrastructures.

II. Foundational Information

Machine-learning (ML) pipelines create specific challenges for RTBF compliance. The patterns and relationships that develop from the data used to train models may persist in model parameters, embeddings, feature stores, or derivative datasets. Traditional deletion methods often fall short of completely removing the original dataset (Buso, 2020). Intermediate artifacts resulting from preprocessing, feature extraction, and data augmentation can generate additional datasets that retain residual personal information. Therefore, simply deleting raw data does not remove its influence from trained models.

Machine unlearning has emerged as a technical solution to address the challenges of personal data deletion from search engines and online databases. Unlike traditional deletion methods, unlearning algorithms aim to selectively remove the influence of specified data from trained machine learning models without retraining from scratch. By partitioning datasets into retain and forget sets, machine unlearning techniques attempt to adjust the model parameters so that forgotten data no longer affects predictions. This supports the goals of RTBF in AI environments. However, the effectiveness of unlearning depends on factors such as model architecture, the size of the forget set, and evaluation methods. The success of data removal can be assessed using Membership Inference Attacks, which reveal whether data has been successfully unlearned.

A thorough literature review indicates that achieving full compliance with deletion requests to meet RTBF requirements in modern AI-driven systems is technically difficult. Factors, such as training checkpoints, intermediate embeddings, and derived datasets, can retain residual information even after the primary source data is deleted. These real-world challenges demonstrate the critical need to expand knowledge and find better ways to enforce and achieve compliance with data-deletion practices. Investigating machine unlearning methods to achieve full compliance with data deletion across search engines and online databases could be a reliable privacy protection tool that supports the RTBF.

III. Research Approach

This research investigates the effectiveness of machine unlearning to enable compliance with the RTBF in AI systems. Since direct access to proprietary models is limited, the study relies on publicly available machine learning datasets, pre-trained models, and established unlearning algorithms to simulate deletion requests and evaluate their impact. The goal is to measure how well unlearning techniques remove the influence or behavior of the forgotten data. Safeguarding the inference on the retained data helps promote accuracy in the remaining data.

The analysis focuses on several key aspects of machine unlearning. First, the study defines forget and retain sets within the training data. The forget set represents the data to be erased from the model, while the retain set contains the remaining training data. Models are trained normally and then subjected to unlearning procedures designed to eliminate the contribution of the forget set.

Second, the study evaluates the effectiveness of unlearning using multiple metrics. Accuracy on the retain and forget sets is tracked to ensure that unlearning removes the intended data while maintaining overall model performance. Additionally, Membership Inference Attacks (MIA) are used to determine whether models still reveal information about forgotten data, providing an empirical measure of unlearning success.

Finally, the study examines the impact of forget set size, model architecture, and unlearning method on results. By systematically varying these factors, the research identifies conditions under which machine unlearning is most and least effective. This approach provides a rigorous framework for understanding the capabilities and limitations of machine unlearning in supporting RTBF compliance in AI systems.

IV. Experimental Procedures

The experiment aimed to evaluate the effectiveness of machine unlearning. The study was conducted using a neural network obtained from Github (Pedregosa et al., 2023). The model was pre-trained on a dataset of images.

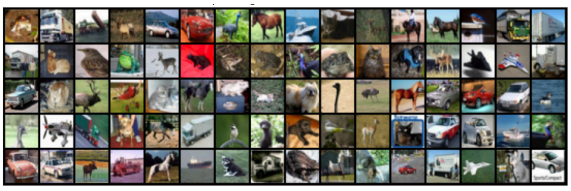

Photo 1: Sample images from CIFAR10 Dataset

The dataset was divided into two subsets, a retain set and a forget set. The retain set was used to represent data the model should continue to learn from. While the forget set images were meant to be removed. To better evaluate how the proportion of forgotten data affects unlearning, the experiment was run with three different forget set sizes. These sizes were defined by manually adjusting the dataset split as shown below:

forget_set, retain_set = torch.utils.data.random_split(train_set, [X, 1-X], generator=RNG)

(Pedregosa et al., 2023)

The *X* represents the fraction of training data assigned to the forget set. The experiment was tested with X = 0.1 (10%), X = 0.2 (20%), and X = 0.4 (40%). Larger values of X were not considered, as the machine still needs a substantial amount of retained data to avoid affecting the models' overall performance.

After defining the forget and retain sets, the model was trained on the full data set. Machine unlearning was then applied to remove the influence of the forget set images. Then creating an updated test model. During unlearning, the model was fine-tuned primarily on the retain set, while the forget set was used to disrupt learned representations associated with the removed data.

To assess how much the model still remembers the forgotten data, losses were computed separately for the forget set and the unseen test set. These loss values were later used to perform a Membership Inference Attack (MIA). This attack tests whether the model still retains information about the forgotten data by attempting to distinguish samples from the forget set versus unseen test data. High attack accuracy indicates that forgotten data leaves detectable traces in the model, whereas low accuracy near 0.5 suggests that the model treats forgotten and unseen data similarly.

V. Key Findings

The effectiveness of machine unlearning was evaluated using three different forget set sizes: small, medium, and large. For each test, a forget/retain configuration was developed. The model performance was assessed on the retain set and the forget set. Later the Membership Inference Attack (MIA) was applied to quantify the model's residual memorization of forgotten samples. Loss distributions for the forget and unseen test sets were also computed to visualize the impact of unlearning on the model's internal representations.

Prior to applying machine unlearning, the baseline model achieved an accuracy of 99.5% on the retain set and 88.3% on the forget test set. A membership inference attack performed on the original, unaltered model produced an accuracy of 0.576 when distinguishing forgotten samples from unseen test data.

Test	Forget %	Retain %	Retain Accuracy	Forget Accuracy	MIA Accuracy
Small	10	90	98.0%	86.6%	0.570
Medium	20	80	98.0%	86.7%	0.573
Large	40	60	98.0%	86.7%	0.578

Table 1: Forget and retain set ratios and corresponding model performance metrics.

Using the small forget set configuration (10% forget, 90% retain), the model achieved an accuracy of 98.0% on the retain set and 86.6% on the forget set. MIA accuracy was 0.570. Loss values for the forget and test sets were computed and plotted, exhibiting overlap with modest separation (Figure 2).

Using the medium forget set configuration (20% forget, 80% retain), the model achieved an accuracy of 98.0% on the retain set and 86.7% on the forget set. MIA accuracy was 0.573. Loss distributions were plotted, showing modest separation between the forget and test sets (Figure 3).

Using the large forget set configuration (40% forget, 60% retain), the model achieved an accuracy of 98.0% on the retain set and 86.7% on the forget set. MIA accuracy was 0.578. Loss distributions for the forget and test sets were plotted (Figure 4).

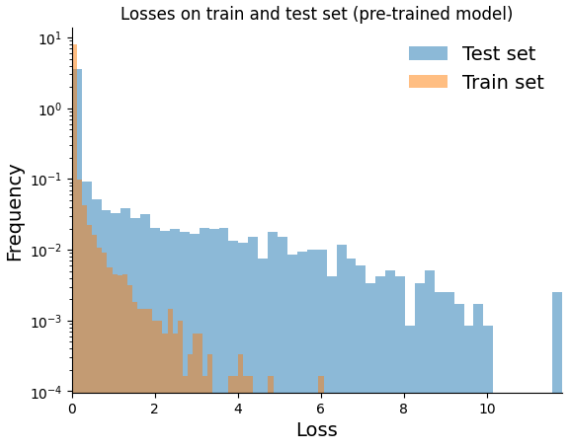

Figure 1: Losses on Train and Test Set Graph (Pre-trained model)

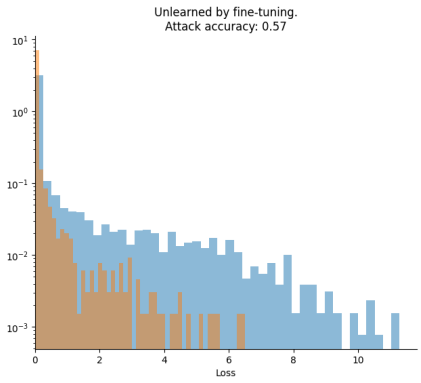

Figure 2: Small Forget Set – Loss Distribution and Attack Accuracy

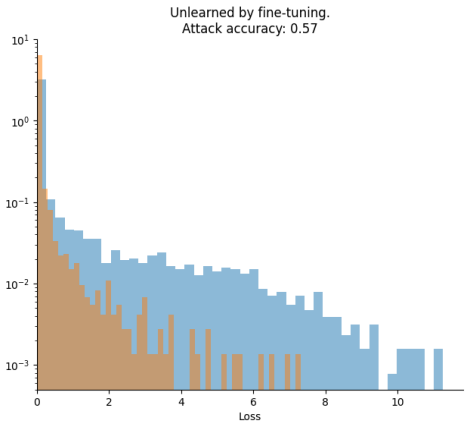

Figure 3: Medium Forget Set – Loss Distribution and Attack Accuracy

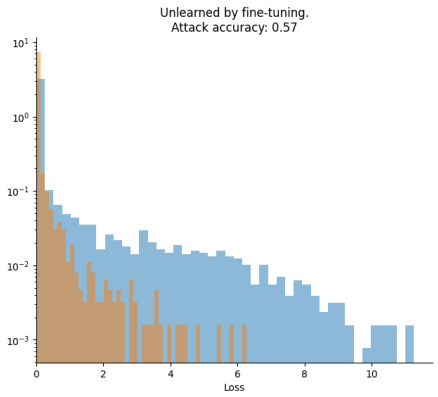

Figure 4: Large Forget Set – Loss Distribution and Attack Accuracy

VI. Interpretation of Results

The results show that machine unlearning reduces the model's knowledge of the targeted forget data, but it does not completely remove it. Across all three forget set sizes, small (10%), medium (20%), and large (40%), the model's accuracy on the retain set stayed consistently high at 98.0%, indicating that unlearning did not harm the model's performance on the data that was set for retention.

In contrast, accuracy on the forget set dropped slightly, and remained around 86.6–86.7%. This small change shows that the model only partially forgot the targeted samples. Even when more data was added to the forget set, the model did not lose much more accuracy on that set. The persistently high forget set performance highlights that traces of the forgotten data remain within the model.

The Membership Inference Attack (MIA) further confirms the accuracy of the data, since the MIA measures how well someone could tell if a sample was in the forget set or not. Scores above 0.5 indicate that the model still retains recognizable information about forgotten samples. In the experiments, MIA accuracy remained between 0.570 and 0.578 even as the forget set size increased. This demonstrates that, despite unlearning, forgotten data leaves detectable patterns in the model's behavior.

Looking at the loss distributions provides additional evidence. Loss values quantify how well the model predicts each sample. When we plotted the losses for the forget and unseen test sets, there was only partial separation. Many forgotten samples still had low loss values, meaning the model still recognized them. Even with larger forget sets, overlap remained, showing that unlearning does not fully erase the influence of the targeted data.

In simple terms, these results show that machine unlearning weakens but does not erase the memory of forgotten samples. The model continues to retain detectable traces of the forgotten data, which could pose privacy risks if the data is sensitive. Overall, while unlearning can reduce the model's reliance on specific data, it cannot guarantee complete removal. This emphasizes the need for stronger techniques and verification methods to truly delete information in trained models.

VII. Conclusion

This study investigated the effectiveness of machine unlearning as a tool to support the Right To Be Forgotten in AI systems. By systematically varying the size of the forget set and evaluating model performance through accuracy metrics, loss distributions, and MIA, the research provides evidence that current unlearning methods reduce but do not fully eliminate the influence of forgotten data. Across all forget set sizes, the model maintained high accuracy on retained data while only slightly decreasing accuracy on forgotten data. MIA results consistently

remained above 0.5, indicating that residual information about the forgotten samples persists even after unlearning. Loss distribution analysis further confirmed that forgotten samples continue to produce low loss values, demonstrating detectable memory within the model.

These findings highlight a fundamental limitation of machine unlearning: while it can weaken a model's reliance on specific data, it cannot guarantee the complete erasure of information. Residual traces of forgotten data pose privacy risks, particularly in applications that handle sensitive information. Therefore, achieving true RTBF compliance in AI systems will likely require complementary strategies, including model verification, more robust unlearning algorithms, and architectural changes to ensure that data can be fully removed. Overall, this research underscores the promise of machine unlearning while emphasizing the need for further development and validation to make AI systems reliably forget sensitive information.

References

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine Unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*. https://doi.org/10.1109/sp40001.2021.00019

Buso, F. (2020, July 24). *MLOps with a Feature Store*. Medium. https://medium.com/data-for-ai/mlops-with-a-feature-store-8dabc845584a

Caravà, M. (2020). An exploration into enactive forms of forgetting. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-020-09670-6

Carlini, N., Chien, S., Milad Nasr, Song, S., Terzis, A., & Florian Tramèr. (2022). *Membership Inference Attacks From First Principles*. https://doi.org/10.1109/sp46214.2022.9833649

Farzanehfar, A., Houssiau, F., & de Montjoye, Y.-A. (2021). The risk of re-identification remains high even in country-scale location datasets. *Patterns*, *2*(3), 100204. https://doi.org/10.1016/j.patter.2021.100204

Geraldine O. Mbah. (2022). Data privacy and the right to be forgotten. *World Journal of Advanced Research and Reviews*, *16*(2), 1216–1232. https://doi.org/10.30574/wjarr.2022.16.2.1079

Laney, D. (2023, June 13). The Data Purge: An Era Of Defensible Retention And Data Minimization. *Forbes*. https://www.forbes.com/sites/douglaslaney/2023/06/13/the-data-purge-an-era-of-defensible-retention-and-data-minimization/

Lauradoux, C., Curelariu, T., & Lodie, A. (2023, February 6). *Re-identification attacks and data protection law*. MIAI. https://ai-regulation.com/re-identification-attacks-and-data-protection-law/

Netzer, T. (2023, September 21). *The Right to be Forgotten - What Makes it Tough?* k2view. https://www.k2view.com/blog/gdpr-right-to-be-forgotten

Pedregosa, F., Triantafillou, E., & Kowshik, B. (2023, September 2). *[Starting kit for*

the NeurIPS 2023 Machine Unlearning Challenge]. GitHub. https://github.com/unlearning-challenge/starting-kit

Politou, E., Michota, A., Alepis, E., Pocs, M., & Patsakis, C. (2018). Backups and the right to be forgotten in the GDPR: An uneasy relationship. *Computer Law & Security Review*, *34*(6), 1247–1257. https://doi.org/10.1016/j.clsr.2018.08.006

Richards, N., & Hartzog, W. (2020, February 20). *Why Europe's GDPR magic will never work in the US*. Wired. https://www.wired.com/story/us-version-gdpr/

Saxena, N., & Voris, J. (2011). Data remanence effects on memory-based entropy collection for RFID systems. *International Journal of Information Security*, *10*(4), 213–222. https://doi.org/10.1007/s10207-011-0139-0

Schneider, J., Lautner, I., Moussa, D., Wolf, J., Scheler, N., Freiling, F., Jaap Haasnoot, Henseler, H., Malik, S., Morgenstern, H., & Westman, M. (2021). In Search of Lost Data: A Study of Flash Sanitization Practices. *Digital Investigation*.

Sha, A., Nunes, B., & Haller, A. (2018). *"Forgetting" in Machine Learning and Beyond: A Survey*. Arxiv.org. https://arxiv.org/html/2405.20620v1

Shrestha, A., & Mahmood, A. (2019). Review of Deep Learning Algorithms and Architectures. *IEEE Access*, *7*, 53040–53065. https://doi.org/10.1109/access.2019.2912200

Stackpole, B. (2025, March 3). *Bringing transparency to the data used to train artificial intelligence | MIT Sloan*. MIT Sloan; MIT. https://mitsloan.mit.edu/ideas-made-to-matter/bringing-transparency-to-data-used-to-train-artificial-intelligence

State of California Department of Justice. (2024). *California Consumer Privacy Act (CCPA)*. State of California Department of Justice Office of the Attorney General. https://oag.ca.gov/privacy/ccpa

Tirosh, N. (2016). Reconsidering the "Right to be Forgotten" – memory rights and the right to memory in the new media era. *Media, Culture & Society*, *39*(5), 644–660. https://doi.org/10.1177/0163443716674361

What Is The Right to Be Forgotten? How Can Organizations Respond? (2025, January 25). Alation.com. https://www.alation.com/blog/right-to-be-forgotten-compliance-guide/

Why is Data Deletion so Complex? (2025, February 14). Data Privacy Hub. https://www.dataprivacyhub.io/why-is-data-deletion-so-complex/

Wolford, B. (2025). *What is GDPR, the EU's new data protection law?* GDPR.EU; Proton AG. https://gdpr.eu/what-is-gdpr/

Zhang, D., Finckenberg-Broman, P., Hoang, T., Xing, Z., Pan, S., Staples, M., & Xu, X. (2024). *Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions*.