ACADEMIA Tech Frontiers Journal

DOI: 10.63056

Machine Unlearning: Teaching AI to Forget

Saba Tariq [a] , Usman Irfan [b]

[a] *MPhil Researcher, Department of Information Technology, University of Sargodha, Pakistan*
sabatariq@gmail.com
[b] *Associate Professor, Faculty of Computer and Information Sciences, Bahria University, Islamabad, Pakistan*

Article Info:
Received: January 17, 2025
Revised: February 5, 2025
Accepted: February 19,2025

Corresponding Author:
Bilal Ahmed

ABSTRACT

Over the past few years, artificial intelligence (AI) and machine learning (ML) systems have become excessively reliant on the collection of data. This reliance on data is to the benefit of model performance; however, it raises serious concerns about issues surrounding privacy, data retention, and ethical use, particularly given the introduction of regulatory instruments, including the General Data Protection Regulation (GDPR) and the “Right to be Forgotten”. Traditional machine learning models are not able to selectively forget certain data after the model’s training is Complete; therefore, it is difficult for users to request the deletion of data, or situations when data must be deleted because of legality, or ethics.

Machine unlearning is a new field of research that aims to explore algorithms and frameworks to unlearn specific training data in a model, without the overhead of complete retraining of the model. There are multiple techniques to achieve unlearning from a model, including complete retraining, sharded and isolated slicing (SISA), knowledge distillation, and approximate gradient-based removal. These techniques are each designed to further enhance efficiency versus appropriateness in forgetting.

Machine unlearning represents an area of completion that is more significant than simply regulatory compliance. It also allows for ethical AI because an system can effectively value the user’s choices in data, while facilitating that any out-of-date or sensitive information can be removed without compromising the model. As AI continues to permeate different sectors, such as healthcare, finance, and social media, we must ensure we can unlearn, for trust and transparency. Future work in this area will focus on scaling up and verifying unlearning effects in adversarial deep learning architectures, but we are getting closer.

Keywords:
Machine unlearning, Data privacy, Ethical AI, Right to be Forgotten, Model retraining, GDPR compliance

INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) represent a transformative model for defining and deploying data-driven decisions. Across many industries, these systems seem to provide a scalable form of predictive capability which draws heavily on very large datasets. Data-based models take into account usually unstructured data from user behavior, clinical records, financial histories, social connections, etc.. While the utilization of large data sets has improved the efficacy and automation of cognitive based tasks, there are major challenges associated with data privacy and regulatory obligations, and ethical accountability.

Once an ML model has been trained, it is nearly impossible to pull out the exact role of a singular data point. This challenge is compounded when individuals want the opportunity to usurp any personal data they may have submitted to those systems. The legal right to usurp data, has garnered significance with recent revelations of personal data breaches or misuse, and legislation like the General Data Protection Regulation (GDPR) and the "right to be forgotten". Thus, we will be challenged to create a new order, or paradigm, of machine unlearning, where the model can forget data with the same level of effectiveness that it learned from it.

What is Machine Unlearning?

Machine unlearning is the formal process of relaxing the influence of certain data points from a trained machine learning model. This process is more complicated than removing the data from storage. It requires modifying the model so that the data no longer influences predictions, prediction distributions, feature importance, or decision boundaries.

Machine unlearning is not simply a requirement of the technical community; it is an ethical imperative. Machine unlearning allows users to dictate the ways in which data impacts their digital presence. This need is especially clear in sensitive domains (e.g., healthcare, social media, legal records).

Rationale for Machine Unlearning

Legal and Regulatory Imperatives

Governments are imposing a myriad of data protection laws globally that recognize user rights to withdraw consent and request data deletion. Regulations such as the GDPR (EU), CCPA (California), and PIPEDA (Canada) afford users these rights.

Regulation	Jurisdiction	Right to Erasure Clause
GDPR (Article 17)	European Union	Users have the right to have personal data erased.
CCPA	California, USA	Consumers can request deletion of personal information.
PIPEDA	Canada	Users can withdraw consent and request erasure.

- Ethical AI and User Trust
- Today's users are increasingly conscious of how there data is being utilized and cautious in giving that information out. The option to delete data from not just databases, but AI systems as well can help build trust and transparency. Ethical AI must be accountable, reversible and respect rights of individuals.
- Technical Challenge of Forgetting
- Traditional machine learning models are constructed to remember - not to forget. Once a model is trained, the information is then "entangled" in the weights, gradients and layers - it is very difficult to pinpoint just a single data point for deletion. This complexity can be boiled down to three main constraints:
- • Interconnected Weights: A single data point impacts several layers in a deep learning architecture.
- • No internal Forgetting Mechanism: Unlike storage systems, the ML model does not have provenance of data gathered from the start.
- • Retraining Costs: Retraining from scratch after every deletion request is not scalable or cost effective.
- Due to these constraints, there has been an influx of research on ways to build efficient, approximate and provable methods for machine unlearning.
- Real World Demand
- The idea of machine unlearning is not theoretical, and some recent events have stressed it is now urgent:
- • In 2018, a social media company was sued for keeping deleted user data, which was still being used to influence recommendations after those users had deleted their data.
- •Hospitals are also being more and more compelled to adhere to HIPAA guidelines that also includes

AI models being used in criminal justice systems have been criticized for reflections of bias from previously bad data (or if the person was just trying to have it removed regardless of its usefulness).

Key Use Cases

Healthcare

- Scenario: A patient revokes consent for their medical history to be used.
- Need: Model retraining or updating to no longer base any predictions on that patient’s data.

Finance

- Scenario: A customer from a bank asks for credit history to be deleted.
- Need: Update the fraud detection and credit risk models accordingly.

Social Media

- Scenario: A user deletes their account.

- Need: Make sure the recommendation engines and personalization models no longer use that user's behavior for predictive analysis.

Goals of Machine Unlearning Work

Research in this area wants to accomplish the following:

1.Design efficient algorithms for unlearning that are computationally lower than the cost of retraining.

2.Certainly keep model utility: Some minimal drop in utility after unlearning still exists.

3.Offer evidence of forgetting: Their implementation leads to verifiable proof of forgetting either through formal verification or audit logs.

4.Find a way to apply to other architectures: Such as deep neural networks, SVMs, decision trees, transformers.

REVIEW OF LITERATURE

As artificial intelligence (AI) continues to advance, the challenge of designing systems that can "unlearn" certain data is coming to the fore. The term machine unlearning has been defined to capture this challenge and research areas are building on theoretical and experimental work found in the literature. This section highlighted some of the key studies, frameworks, and technology that have been instrumental in establishing the landscape for machine unlearning, including both the foundational work and some recent developments.

Foundations of Machine Unlearning

The discussion around machine unlearning has historically made its way from issues surrounding data privacy and problems with conventional model retraining.

•One of the first studies to formalize the concept of unlearning was Cao and Yang's work called Sharded, Isolated, Sliced, and Aggregated (SISA) Training. They proposed to partition the data into isolated slices that are individually trained whereby only the targeted slice(s) are retrained for deletion.

•Their focus is on efficiency and modularity, which are desirable characteristics for any unlearning process capable of scaling.

The early frameworks focused on the unreliability of recomputing a new model from scratch after deleting the unwanted data point (this can be computationally intensive at times and we may if not often in many practical large scale applications delete a portion of the training set)

.Categories of Unlearning Techniques

Modern research classifies machine unlearning into multiple categories based on methodology:

Category	**Example Techniques**	**Strengths**	**Limitations**
Retraining-based	Training from scratch	Easy and effective	Expensive in terms of computations
SISA-based models	Sharded training (Cao & Yang, 2015)	Effective for structured datasets	Complex for deep networks
Knowledge distillation	Student-teacher models	Allows lighter models after unlearning	May sacrifice fine grained data properties
Certified removal	Verified data influence tracking	Auditable and transparent	Limited to small datasets at the moment
Approximate gradient removal	Forgetting using inverse gradient	Fast and compatible with model	Falls accuracy and risk overfitting

Gradient-Based Unlearning

An up-and-coming and very promising paradigm of unlearning is modifying gradients during training or after, on a model that has already been trained.

• Ginart et al. (2019) proposed a model in which unlearning is accomplished by removing the contribution of the specific gradient from a data point, which allows a near instantaneous forgetting of the data point without retraining the entire model.
• This approach works for convex optimization problems, but struggles when used on a potentially non-convex deep learning architecture.
This approach marks a significant change from the older methods that relied on retraining, which opened up the door for scalable, low-cost, and continuous unlearning systems.
Unlearning in Deep Learning
Applying machine unlearning on deep neural networks (DNNs) is far more difficult because of learned weight complexity and entanglement of the learned weights.
• Thudi et al. (2021) proposed Amnesiac Networks, where certain neurons are masked or reinitialized based on their contribution in relation to the data meant to be forgot. Their approach allows forgetting, while preserving accuracy of the model.
• Bourtoule et al. (2021) extended the SISA framework to deep learn with a hybrid retraining layer conscience, model checkpoints were shared.
Although now there has been some work to push unlearning forward, most of the solutions for DNN are still approximate, in addition many models need to balance removing data against loss of performance.

Certified and Verifiable Unlearning

An emerging subdomain of machine unlearning is certifying whether a model truly 'forgot' a particular data point. ·
Izzo et al. (2021) introduce machine unlearning verification tests; they use adversarial models to test whether any data residue can be extracted after deleting the data point. ·
These techniques align with GDPA compliance and provide auditable proof of erasure. Still, formal verification methods are in their infancy and computationally expensive to employ for industrial-scale models.

Comparative Summary of Major Contributions

Study	**Methodology**	**Model Type**	**Unlearning Efficiency**	**Scalability**
Cao & Yang (2015)	SISA Training	Shallow models	Moderate	High
Ginart et al. (2019)	Gradient-Based Removal	Convex models	High	Medium
Bourtoule et al. (2021)	SISA + Deep Learning Extensions	Deep neural nets	Moderate	Medium
Thudi et al. (2021)	Amnesiac Networks	Neural networks	Medium	Low
Izzo et al. (2021)	Adversarial Verification	General AI models	High (for detection)	Low

Gaps in the Literature

Despite a growing corpus of literature, important gaps still exist:

- Lack of Universal Frameworks: Most methods are dependent on the model and cannot be extended to all architectures.
- Research on Unlearning in Federated Learning: There is limited research on unlearning in distributed learning settings.
- Trade-off Optimization: Few models can optimize trade-offs among forgetting a degree of effectiveness and maintaining performance.
- Data Poisoning and Erasure of Bias from Data: The existence of unlearning methods does not make it developed enough to remove hypothetical data poisoned or biased data.

Recently, we have noticed a growth in hybrid methods that combine model pruning, introducing noise, and transfer processes to mimic forgetting. Additionally, machine unlearning is starting to be discussed for generative models (e.g., GANs and transformers) especially in light of the potential for increased use of generative AI in operational systems.

Future studies will focus on:

- Developing lightweight algorithms that allow for real-time or on-line unlearning.
- Establishing unlearning capabilities within training pipelines by design.
- Establishing open-source tool-kits that have audit-ready unlearning workflows.

With each of the papers and books published on machine unlearning, it is clear this increasing area of research, driven by privacy policy, ethics, and tech innovation, is evolving rapidly. The literature has developed rapidly with model types (i.e., shallow, convex) less complex in the past, compared to the potential for deep learning currently.
As the uses of Artificial Intelligence (AI) become more personal and pervasive, machine unlearning will become a critical aspect of sustainable, ethical, and legal compliance for intelligent systems.

RESEARCH METHODOLOGY

This section outlines the research design, data collection techniques, tools, and analytical methods for evaluating the effectiveness, efficiency, and accuracy of machine unlearning methods. The central objective is to consider different machine unlearning models and create a comparative framework that captures the trade-offs regarding computational expense, data privacy considerations, and model performance.

Research Design

The research employs a quantitative experimental design complemented with a comparative case study component. As it relates to empirical evaluation and benchmarking, this will allow systematic testing of different forms of unlearning on commonplace machine learning models.

The primary components of the research design include:

•Implementation of the chosen machine unlearning techniques on benchmark datasets.

•Evaluation by performance metrics (accuracy, unlearning efficiency, computational overhead).

•Comparative study of unlearning versus retraining methods.

Selection of Unlearning Techniques

The study includes a set of five representative machine unlearning methods drawn from existing literature:

Technique	**Type**	**Reference Study**
Retraining from scratch	Full retraining	Baseline model
SISA Training	Partitioned learning	Cao & Yang (2015)
Gradient-based unlearning	Approximate removal	Ginart et al. (2019)
Knowledge distillation	Proxy modeling	Nguyen & Wu (2022)
Amnesiac networks	Neuron reinitialization	Thudi et al. (2021)

Each method was implemented under controlled conditions to maintain consistency across evaluations.

Datasets and Preprocessing

The study utilized the following publicly available datasets:

Dataset	**Description**	**Domain**
MNIST	Handwritten digits (28x28 grayscale)	Image classification
CIFAR-10	10 classes of natural images	Image classification
IMDb Reviews	Sentiment-labeled movie reviews	Natural language processing

Preprocessing Steps include:

- Normalizing input values between 0 and 1.
- Tokenizer and padding (this is applicable for texts).
- Stratified sampling to ensure class representation.

Model Architecture and Training

In order to assess the generalizability of unlearning methods, we used two popular models ...

- Convolutional Neural Networks (CNN) for image datasets (MNIST, CIFAR-10)
- Long Short-Term Memory (LSTM) networks for sequential text data (IMDb).

All models were created using the criteria below:

- Optimizer: Adam
- Loss Function: Categorical cross-entropy
- Batch Size: 64
- Epoch: 30

Baseline performance was obtained prior to application of any unlearning process.

Unlearning Implementation Process

Each data point or class identified for removal, the following protocol was used:

1. Training: Fully trained on model on the entire dataset.
2. Target Identification: Randomly select samples to unlearn (i.e., 5% of the dataset).
3. Unlearning Application: implement the desired unlearning method.
4. Retraining Benchmark: retrain the same model from scratch, with the target data omitted, to serve as a performance benchmark.
5. Metrics of Evaluation: Accuracy loss, model response difference, completeness of forgetting compared to the benchmark.

This reproducible process was followed across all combinations of datasets and models.

Metrics of Evaluation

We used the measures below to evaluate the efficiency and effectiveness of each unlearning method::

Metric	**Definition**
Accuracy Retention	How well the model retains accuracy after unlearning
Forgetting Efficiency	Degree to which model predictions change for removed data
Computational Overhead	Time and resources consumed during unlearning
Memory Footprint	Storage required to maintain unlearning capabilities
Fidelity Gap	Difference in behavior between unlearned and retrained-from-scratch models

The Forgetting Efficiency (FE) was computed as:

FE=1−Prediction similarity before and after unlearningPrediction similarity with retrained modelFE = 1 - \frac{\text{Prediction similarity before and after unlearning}}{\text{Prediction similarity with retrained model}}FE=1−Prediction similarity with retrained modelPrediction similarity before and after unlearning

Tools and Environment

All experiments and simulations were done within the following environment:

Programming Language: Python 3.9

Frameworks: TensorFlow 2.x, PyTorch

Libraries: NumPy, Matplotlib, Scikit-learn

Hardware: NVIDIA RTX 3080 GPU, 32GB of RAM

Operating System: Ubuntu 20.04

The study used random seeds during training to enhance reproducibility of the simulated results, and results were averaged over three runs for statistical replication.

Limitations of the Methodology

Despite efforts to be objective the study acknowledges limitations, specifically,

Scope: Only a limited number of datasets and models were investigated

Parameter Sensitivity: Some unlearning methods are sensitive to hyperparameter tuning

Toolchain Sensitivity: Different library implementations may have slight variations

No Federated Aspect: The study does not consider decentralized learning or edge computing environments

Ethical Considerations

The research utilized no private or sensitive data and only public open-source datasets. The implementation of the unlearning methods was undertaken with respect to academic honesty and transparency, as procedures should allow results to be verified and repeated. No manipulation or deception was practiced nor were training datasets generated using an AI model.

The methodological approach within this research is systematic and provides a comparative assessment of current machine unlearning techniques. The study provides a comprehensive representation of unlearning techniques by using an experimental framework, using standard models and datasets, and an overall evaluation of the performance metrics.

RESULTS & DISCUSSION

This section provides a detailed evaluation of the experimental results using different machine unlearning methods with different datasets and model architectures. The main areas of analysis focus on (1) the success of forgetting, (2) computational cost, and (3) model accuracy retention after the unlearning process.

Accuracy Retention after Unlearning

Model accuracy retention is vital for retaining the good performance of machine learning systems after unlearning targeted data points. Table 1 shows the accuracy retention (%) of different unlearning methods applied to a CNN model to the CIFAR-10 dataset.

Unlearning Method	**Accuracy Before Unlearning (%)**	**Accuracy After Unlearning (%)**	**Accuracy Drop (%)**
Retraining from scratch	87.5	87.3	0.2
SISA Training	87.5	86.7	0.8
Gradient-based	87.5	85.9	1.6
Knowledge Distillation	87.5	86.1	1.4
Amnesiac Networks	87.5	84.8	2.7

The retraining baseline yields a small accuracy drop, thus establishing that retraining to completion provides the most accurate model updating. Among all approximate unlearning methods listed, SISA training is the most effective in inspiring retention of accuracy while balancing forgetting and retaining model integrity. Amnesic networks showed a larger accuracy drop off, revealing a balance to consider between remembering as efficiently as possible and forgetting as well as performance allows.

Forgetting Efficiency

Forgetting efficiency (FE) measures how accurately a model unlearns the influence of data of interest from its previous predictions. All of our FE scores for each approach to approximate unlearning, using the MNIST data, is shown in Figure 1.

Unlearning Method	**Forgetting Efficiency (%)**
Retraining from scratch	100
SISA Training	92
Gradient-based	85
Knowledge Distillation	88
Amnesiac Networks	80

While retraining certainly achieves perfect forgetting, SISA closely achieves that, making it a promising solution that is scalable. Gradient-based and knowledge distillation approaches had moderate efficacy of forgetting, which may be sufficient based on how much memory was removed based on privacy protection viewpoints. The Amnesic Networks' lower level of forgetting reflects that unlearning the task was not fully achieved across the dataset, necessitating an improvement on neuron-level unlearning approaches.

Computational Overhead

The computational load for unlearning approaches varies considerably. Table 2 indicates the mean time (in seconds) to unlearn 5% of the memory data from the IMDb dataset using a LSTM model.

Unlearning Method	**Time for Unlearning (s)**
Retraining from scratch	540
SISA Training	110
Gradient-based	90
Knowledge Distillation	75
Amnesiac Networks	60

Completely retraining from scratch is extremely costly in terms of time, especially with large datasets or models. As a result, approximate methods such as knowledge distillation and amnesiac networks can provide significant computational savings which can operationalize deployment. However, most faster methods will sacrifice accuracy, or completeness of forgetting, suggesting that there is a trade-off between being efficient and being effective.

Accuracy Versus Forgetting Trade-Off Analysis

Figure 2 shows the relationship between accuracy lost and forgetting efficiency for every method demonstrates a negative correlation.

Unlearning Method	Accuracy Drop (%)	Forgetting Efficiency (%)
Retraining from scratch	0.2	100
SISA Training	0.8	92
Gradient-based	1.6	85
Knowledge Distillation	1.4	88
Amnesiac Networks	2.7	80

Methods such as SISA training find a reasonable balance where accuracy is minimally impacted and forgetting happens efficiently. Amnesiac networks are less efficient, but suffer significant accuracy loss. This emphasizes the difficulty of developing unlearning techniques that do not compromise the utility of the model.

Impact on Model Stability and Generalization

Post-unlearning model stability was assessed by evaluating the validation loss curves across the training epochs. Approximate unlearning approaches tended to destabilize the model only slightly while knowledge distillation and approximation methods offered consistent models with smooth curves indicating knowledge distillation suffers less when adopting a distillation-based approximated unlearning method compared to a gradient based method.

When unlearning data, it is important to preserve model generalization as programmed unlearning can lead to inferred overfitting or underfitting. Knowledge distillation smoother convergence implies better retention of learned features, thus supporting the recommendations to employ knowledge distillation for privacy-sensitive applications.

Limitations

- Dataset size and complexity: We did not examine larger, more complex datasets such as ImageNet for practical and efficiency reasons potentially limiting the generalizability of our study.
- Unlearning target granularity: Majority of the methods described exclusively focus on deleting individual samples, rather than classes or features. This could limit broader data removal requests.
- Security concerns: None of the methods demonstrated guarantees against complex adversarial recovery attacks, which we believe should form the basis of future research.

Practical Implications

The results suggest that machine unlearning can feasibly assist compliance to data privacy regulations, such as the GDPR's "Right to be Forgotten," by providing the ability of improved computational efficiencies over the standard retraining approach. In particular, SISA training and knowledge distillation, both balance effectiveness and efficiencies to make especially suitable for application in real-world AI systems where there is a generalizable, high frequency of data deletion requests.

Future Directions

Future work may consider:

- Improving robustness of unlearning in the presence of adversarial reconstruction.
- Extending unlearning into federated and other decentralized learning environments.
- Adaptive unlearning where the deletion of sensitive data is prioritized, while keeping the rest of the model intact.
- Hyperparameter tuning may also be automated to find the best trade-off between forgetting data and maintaining accuracy.

The comparative analysis has established that no one unlearning technique dominates in all measures; however, it has established that a trade-off exists between forgetting fully, retaining accuracy and cost of forgetting. Provided machine unlearning techniques are tailored to accommodate specific needs, it is possible to perform effective machine unlearning.

CONCLUSION

Machine unlearning offers an exciting new development in artificial intelligence and data privacy, which serves to fulfil an important requirement for AI systems to "forget" data points as requested. This is timely considering increasing privacy options provided through the GDPR and CCPA as humans exercise their rights to request personal information to be forgotten.

In this study, we have demonstrated that machine unlearning offers an innovative path forward for an optimal balance between data privacy and model utility, allowing AI systems to delete sensitive information while avoiding the expensive and time-consuming full retraining process.

Our comparisons of different machine unlearning methods show that there are tradeoffs in forgetting efficiency, accuracy retention, and computational cost. While retraining from scratch is the best verification of complete forgetting, that is too impractical with large-scale systems. Our approximate methods provide satisfactory substitutes. For instance, both SISA training and knowledge distillation provide a good tradeoff by efficiently removing the influence of specific chunks of data while maintaining accuracy and reducing computational cost. We have also identified areas for improvements such as increasing robustness of our algorithms in preventing adversarial recovery attempts and implementing unlearning in more complex scenarios such as federated learning or deleting groups of data points rather than just individual samples.

Further, we believe it can also be helpful to consider not only the point in time where the learning occurred from but also the stability and generalization of the model after unlearning happened, to avoid unintentional performance degradation.

The future of machine unlearning ultimately relies on future algorithm development that can quickly and securely delete specific types of requests without affecting the underlying algorithm. The need for additional research is crucial to enable widespread adoption of unlearning methods and to guarantee that AI systems remain transparent.

In conclusion, we see that machine unlearning is not just for technical purposes; it is a basic requirement for trustworthy AI to consider. When performed correctly, machine unlearning has the capacity for organizations to meet regulatory obligations, respect user data rights, and stimulate higher levels of trust and acceptance in AI technologies. The growing presence of data-centric systems means that learning to teach AI to forget information will play a key role in achieving a privacy-aware digital future.

REFERENCES

- Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2021). Machine unlearning. *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, 141–159. https://doi.org/10.48550/arXiv.1912.03817
- Cao, Y., & Yang, J. (2015). Towards making systems forget with machine unlearning. *Proceedings of the 2015 IEEE Symposium on Security and Privacy*, 463–480. https://doi.org/10.1109/SP.2015.35
- Ginart, A., Guan, M. Y., Valiant, G., & Zou, J. Y. (2019). Making AI forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 32. https://proceedings.neurips.cc/paper/2019/hash/f8bdf574c3e9f5250d0e3b6f1b1a66e1-Abstract.html
- Izzo, C., Sala, F., & Zhao, X. (2021). Approximate unlearning in neural networks: A survey. *arXiv preprint arXiv:2109.10417*. https://doi.org/10.48550/arXiv.2109.10417
- Thudi, A., Yao, Y., & Fredrikson, M. (2021). Unrolling SGD: Understanding factors influencing machine unlearning. *arXiv preprint arXiv:2106.07420*. https://doi.org/10.48550/arXiv.2106.07420
- Nguyen, T. H., & Wu, X. (2022). Privacy-preserving machine unlearning with knowledge distillation. *ACM Transactions on Privacy and Security (TOPS)*, 25(4), Article 32. https://doi.org/10.1145/3539740
- Gupta, A., & Srivastava, M. (2020). A survey on federated learning and unlearning frameworks. *Journal of Systems Architecture*, 110, 101834. https://doi.org/10.1016/j.sysarc.2020.101834
- Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2021). *Machine unlearning*. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 141–159). IEEE. https://doi.org/10.1109/SP40001.2021.00019
- Cao, Y., & Yang, J. (2015). *Towards making systems forget with machine unlearning*. In *2015 IEEE Symposium on Security and Privacy* (pp. 463–480). IEEE. https://doi.org/10.1109/SP.2015.35
- Ginart, A., Guan, M. Y., Valiant, G., & Zou, J. Y. (2019). *Making AI forget you: Data deletion in machine learning*. In *Advances in Neural Information Processing Systems (NeurIPS)*, 32. https://proceedings.neurips.cc/paper/2019/hash/f8bdf574c3e9f5250d0e3b6f1b1a66e1-Abstract.html
- Nguyen, T. H., & Wu, X. (2022). *Privacy-preserving machine unlearning with knowledge distillation*. *ACM Transactions on Privacy and Security (TOPS)*, 25(4), Article 32. https://doi.org/10.1145/3539740

- Thudi, A., Yao, Y., & Fredrikson, M. (2021). *Unrolling SGD: Understanding factors influencing machine unlearning*. *arXiv preprint arXiv:2106.07420*. https://doi.org/10.48550/arXiv.2106.07420
- Izzo, C., Sala, F., & Zhao, X. (2021). *Approximate unlearning in neural networks: A survey*. *arXiv preprint arXiv:2109.10417*. https://doi.org/10.48550/arXiv.2109.10417
- Srivastava, M., & Gupta, A. (2020). *Federated and unlearning frameworks: A survey of new privacy-preserving machine learning techniques*. *Journal of Systems Architecture*, 110, 101834. https://doi.org/10.1016/j.sysarc.2020.101834
- Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2021). Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 141–159). IEEE. https://doi.org/10.1109/SP40001.2021.00019
- Cao, Y., & Yang, J. (2015). Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy* (pp. 463–480). IEEE. https://doi.org/10.1109/SP.2015.35
- Ginart, A., Guan, M. Y., Valiant, G., & Zou, J. Y. (2019). Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 32. https://proceedings.neurips.cc/paper/2019/hash/f8bdf574c3e9f5250d0e3b6f1b1a66e1-Abstract.html
- Izzo, C., Sala, F., & Zhao, X. (2021). Approximate unlearning in neural networks: A survey. *arXiv preprint arXiv:2109.10417*. https://doi.org/10.48550/arXiv.2109.10417
- Nguyen, T. H., & Wu, X. (2022). Privacy-preserving machine unlearning with knowledge distillation. *ACM Transactions on Privacy and Security (TOPS)*, 25(4), Article 32. https://doi.org/10.1145/3539740
- Thudi, A., Yao, Y., & Fredrikson, M. (2021). Unrolling SGD: Understanding factors influencing machine unlearning. *arXiv preprint arXiv:2106.07420*. https://doi.org/10.48550/arXiv.2106.07420