IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 6, Ser. 1 (Nov. – Dec. 2025), PP 102-110
www.iosrjournals.org

Machine Unlearning
A Qualitative Review of Theoretical, Regulatory, and Ethical Dimensions

Darsh Sai Kiran Velagapudi

Abstract
Machine unlearning—the selective removal of the influence of specific training data from a deployed machine learning model—has become a pressing problem at the intersection of computational efficiency, regulatory compliance, and AI ethics. Legislative frameworks including the EU's General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and Singapore's Personal Data Protection Act (PDPA) now require organisations to demonstrate verifiable erasure of personal data from trained AI systems. This review maps the technical landscape of machine unlearning, examining exact methods such as SISA (Sharded, Isolated, Sliced, and Aggregated) training alongside approximate approaches including influence functions, gradient ascent, and model editing. It then situates these methods within the applicable regulatory frameworks and examines the ethical dimensions—fairness, accountability, adversarial misuse, and individual agency—with attention to tensions that remain unresolved. Four open problems emerge clearly: the verification gap, scalability to foundation models, adversarial robustness, and the absence of standardised evaluation benchmarks. Machine unlearning is not merely a compliance mechanism; it is a foundational capability for trustworthy AI, and real progress on it requires genuine convergence across machine learning research, law, and ethics.

I. Introduction

AI and machine learning systems are woven into everyday life—personalised content recommendations, medical diagnostics, autonomous vehicles, financial decisions. All of them depend on the ability to learn from large quantities of data. That dependency is now in tension with legal, ethical, and technical norms that demand the ability to selectively forget. The question is no longer merely how machines learn, but whether and how they can be made to unlearn.

Machine unlearning refers to removing the influence of specific training data points from a trained model without retraining from scratch. Cao and Yang (2015) introduced the term in the context of statistical query learning. The field has grown steadily since, pushed along by the GDPR's "right to erasure" (Article 17) and by the CCPA. Both regulations establish that individuals can request meaningful, verifiable deletion of their personal data from systems that processed it.

The naive solution—remove the data and retrain—is computationally out of the question for modern deep learning. A model like GPT-4 or ResNet-152 can take weeks of training on specialised hardware; retraining in response to every deletion request is neither economically nor temporally feasible. The challenge of machine unlearning therefore sits squarely at the intersection of computational efficiency, statistical theory, and regulatory compliance.

This paper provides a qualitative review of the machine unlearning literature, covering key theoretical frameworks, methodological approaches, application domains, and open challenges. Sources were identified through searches of Google Scholar, arXiv, and SSRN using terms including "machine unlearning," "right to erasure," "approximate unlearning," and "differential privacy." Priority was given to peer-reviewed conference and journal publications from 2015 onward, with supplementary reference to regulatory documents and legal scholarship. Unlike experimental studies that benchmark specific algorithms, this review aims to provide a conceptual map of the field for researchers and practitioners seeking to understand the state of the art, identify open problems, and situate machine unlearning within the broader landscape of privacy-preserving AI.

II. Technical Landscape of Machine Unlearning

2.1 Defining Machine Unlearning

A precise definition is essential but non-trivial. In the strictest sense, exact unlearning requires that after removing a data point z from training set D, the resulting model M' should be statistically indistinguishable from a model M* retrained from scratch on D \ {z}. No information about z should remain encoded in the model's parameters, loss landscape, or internal representations. Cao and Yang (2015) operationalised this for statistical query learning, where models can be expressed as summations over data points, making algebraic removal of individual contributions possible.

Modern neural networks do not lend themselves to such clean decompositions. Deep learning models are trained via stochastic gradient descent over non-convex loss landscapes, and the influence of any single training example is distributed in a highly non-linear, entangled way across all parameters. This has motivated approximate unlearning, which relaxes the exact criterion to a probabilistic or bounded notion of forgetting—producing a model whose behaviour is close enough to a retrained model according to some distance metric, at a fraction of the computational cost.

2.2 The Problem of Distributed Memory in Neural Networks

Unlike databases, where individual records can be identified and deleted by a lookup key, a trained neural network distributes the statistical influence of each training example across all of its parameters simultaneously. A single image in a training set of millions does not have a corresponding row that can be removed; its presence has influenced every gradient update during training, and its effects are entangled with those of countless other examples in a way that cannot be cleanly separated after the fact.

Cao and Yang (2015) described this as the problem of transforming learned knowledge back into a form that could be erased. They proposed that for certain model classes with summation-form learning algorithms, it is possible to isolate and subtract the statistical contribution of a specific data subset without full retraining. This insight established the conceptual architecture for subsequent work: machine unlearning is not deletion of data but reversal of influence.

The difficulty intensifies for deep neural networks. Their loss landscapes are so complex and parameter interactions so dense that even small perturbations to one data point's influence can have unpredictable ripple effects on overall model behaviour. The distinction between exact unlearning (provable guarantees) and approximate unlearning (some residual influence accepted in exchange for tractability) is the central organising tension of the field.

2.3 Exact Unlearning: Guarantees at a Cost

Exact unlearning ensures that the post-unlearning model could have been produced by a legitimate training run that excluded the target data. The most straightforward method is complete retraining from scratch on the remaining dataset. It trivially satisfies any notion of exact unlearning, but is computationally prohibitive at the scale of modern AI systems.

More sophisticated exact methods reduce this cost through structural precommitments. The SISA (Sharded, Isolated, Sliced, and Aggregated) training framework, proposed by Bourtoule et al. (2021), partitions training data into isolated shards and trains separate model components on each. When an unlearning request arrives, only the shard containing the target data point needs to be retrained. The final model aggregates predictions from all components.

A related approach exploits differential privacy. If a model has been trained with differential privacy guarantees, the influence of any individual training point is bounded by the privacy budget—and once that budget is exhausted, the point's influence is considered negligible. The catch is that this requires differential privacy to be incorporated at training time, which most existing deployed models never did, and the privacy-accuracy trade-off can degrade model quality substantially.

2.4 Approximate Unlearning: Efficiency with Caveats

Approximate unlearning methods accept that the post-unlearning model may not be perfectly indistinguishable from a retrained one, in exchange for dramatically lower computational cost. They span gradient-based approaches, influence function methods, and fine-tuning procedures.

Gradient ascent is among the most intuitive. Rather than minimising the loss on the full training set, the model is fine-tuned by ascending the gradient with respect to target data points—effectively training the model to perform worse on what it is meant to forget. Jang et al. (2022) demonstrated that gradient ascent on specific training samples can reduce verbatim reproduction of those samples in generated text, though aggressive application risks degrading the model's general capabilities.

Influence function methods estimate the counterfactual effect of removing a training point by computing how model parameters would change if that point had been excluded. They are cheaper than exact retraining but notoriously unreliable for large, non-convex deep networks, where the quadratic approximations underlying them break down (Koh and Liang, 2017).

A more recent and practically promising class of methods is model editing, where targeted layers are modified directly to suppress or alter specific memories. Techniques such as Rank-One Model Editing (ROME) (Meng et al., 2022) have shown the ability to modify specific factual associations in transformer models with minimal collateral damage to other capabilities. Scalability to large numbers of simultaneous unlearning requests, and robustness to adversarial probing, remains an active area of investigation.

III. Methodological Approaches to Machine Unlearning

3.1 Exact Unlearning

Exact unlearning methods provide formal guarantees of complete removal. The most influential framework is SISA training. It partitions the dataset into non-overlapping shards and trains an independent submodel on each. Unlearning a data point z requires retraining only the submodel whose shard contains z, leaving all others intact. The final model aggregates predictions from all submodels, typically via averaging or majority voting.

SISA substantially reduces unlearning cost: with k equal shards, the unlearning cost drops by a factor of k in the best case. By further slicing each shard into incremental portions and saving model checkpoints, even finer-grained unlearning is possible—only slices trained after the target data point need to be reprocessed. Bourtoule et al. (2021) showed that SISA can achieve unlearning costs several orders of magnitude lower than full retraining while maintaining competitive model utility. The trade-off is model quality: no submodel has access to the full training data, and the relationship between sharding granularity and accuracy requires careful management.

Table 1: SISA Framework — Performance Across Datasets (20 Shards / 50 Slices)

Dataset	Unlearning Requests	Speed-up Factor	Accuracy Degradation (%)	Viability
Purchase	8	4.63x	< 2.0	High
SVHN	18	2.45x	< 2.0	High
ImageNet	39	1.36x	16.14	Low
Mini-ImageNet	4	8.01x	16.70	Low

Source: Bourtoule et al. (2021), arXiv:1912.03817

The results reveal a clear split in SISA's utility. For simpler classification tasks—Purchase (a multi-class consumer dataset) and SVHN (digit recognition)—SISA achieves speed-up factors above 4x with accuracy degradation below 2 percentage points, which is viable in production. On ImageNet and Mini-ImageNet, accuracy degrades by over 16 percentage points relative to a non-sharded baseline—an unacceptable cost for deployment. The likely cause is that complex vision tasks depend on cross-dataset statistical patterns; sharding breaks these patterns by preventing any submodel from seeing the full training distribution. In the Singapore context, this suggests SISA is viable for simpler government classification tasks but inadequate for complex vision systems used in healthcare diagnostics or biometric identification.

3.2 Approximate Unlearning via Influence Functions

For deep neural networks where exact retraining is prohibitively expensive, approximate unlearning uses targeted parameter modifications to scrub the influence of the forget set. Three mathematical paradigms dominate.

Influence Functions: Operationalised for deep learning by Koh and Liang (2017), influence functions estimate the counterfactual effect of removing a training point on model parameters. The parameter shift induced by removing a single data point z is approximated as:

$$\Delta\theta \approx H^{-1} \cdot \nabla L(\vartheta, z)$$

where H is the Hessian of the loss function and the gradient term captures the loss at data point z. The Hessian inversion is expensive—its dimension is the square of the number of parameters—but Kronecker-Factored approximations have made this tractable for moderate-scale models.

Fisher Information Matrices: A related approach uses the Fisher Information Matrix (FIM) to identify which parameters are most sensitive to the forget set, then applies targeted weight perturbations to reduce their influence. This has shown promise in federated settings where the forget set is distributed across clients.

Gradient Ascent and Weight Perturbation: Classical Gradient Ascent applies the inverse of the normal training update to unlearn specific data points. It is conceptually simple but fragile: aggressive ascent can cause catastrophic forgetting of the retain set, while insufficient ascent leaves the forget set's influence intact. Successive Random Relabeling (SRL) attempts to mitigate this by replacing forget-set labels with noise prior to gradient-based updates.

3.3 Gradient-Based and Fine-Tuning Approaches

A distinct family of methods operates by directly manipulating gradient updates. A model trained without data point z would have traversed a different gradient path during optimisation. Approximate unlearning can therefore be achieved by computing gradients with respect to the forgotten data and updating parameters in the ascent direction—sometimes called "unlearning by forgetting" (Graves et al., 2021).

Liu et al. (2022) proposed FedEraser, a federated unlearning mechanism in which clients collaboratively apply historical gradient updates in reverse to approximate the effect of retraining. Chundawat et al. (2023) introduced Bad Teacher unlearning, where a poorly-performing "bad teacher" model distils forgetting into the target model while a "good teacher" simultaneously preserves performance on retained data. Knowledge distillation offers another avenue: an unlearned model can be constructed by distilling knowledge from the original model while explicitly excluding information about the target data.

3.4 Federated Unlearning

Federated learning—training a shared model across decentralised clients without centralising raw data—introduces its own unlearning challenges. When a client exercises a right to erasure, the global model must forget that client's data without accessing other clients' private data. FedEraser (Liu et al., 2022) addresses this by enabling clients to contribute reverse gradient updates corresponding to their data to be forgotten, effectively unwinding their contribution to the global model.

Federated unlearning is particularly relevant in healthcare and financial services, where data is siloed across institutions by regulatory mandate. Singapore's PDPC has highlighted federated approaches as a promising architecture for privacy-preserving AI in its advisory guidelines on AI recommendation systems. As Singapore's AI governance infrastructure develops, federated unlearning is technically viable and regulatorily aligned as a mechanism for maintaining data subject rights without compromising collaborative model development.

3.5 Evaluation Frameworks and Benchmarks

The absence of standardised evaluation metrics is a widely acknowledged impediment to progress. The primary evaluation tool is the membership inference attack: an adversary is given a model and a data point, and attempts to determine whether that point was in the training set. Ideal unlearning reduces the adversary's accuracy to chance (50%) on forgotten points. If the adversary can still correctly identify forgotten data points with above-chance accuracy, unlearning is incomplete.

Beyond membership inference, model utility is measured via standard accuracy metrics on retained data, and computational cost is assessed in FLOPs or wall-clock retraining time savings. Recent benchmarks address the evaluation gap more systematically. TOFU (Task of Fictitious Unlearning) provides a controlled benchmark for evaluating LLM unlearning, using synthetically generated biographies of fictitious authors to isolate memorisation from general knowledge (Maini et al., 2024). MUSE (Machine Unlearning Six-way Evaluation) proposes a multi-dimensional evaluation suite covering forget quality, model utility, and privacy leakage across diverse tasks. Adoption of such benchmarks is essential for reproducible comparisons across methods.

IV. Regulatory and Privacy Imperatives

4.1 The Right to Erasure Under GDPR

Article 17 of the GDPR gives people the right to demand deletion of their personal data—no undue delay, no runaround. When it was drafted, lawmakers had databases in mind: structured tables, identifiable records, things you can delete with a SQL command. Machine learning models were not part of that picture. They are now, and nobody has a clean answer for what deletion actually means inside a trained model.

The definitional problem is real. GDPR's erasure obligations apply to personal data *processing*, and regulators have started interpreting a model's retention of a person's data influence as processing. The European Data Protection Board has noted that data encoded in model parameters may count as personal data if the model can reconstruct or make decisions about identifiable individuals. That's a significant position, even if its implications haven't fully worked through the legal system yet.

Enforcement has moved the conversation along faster than legal theory has. Italy's data protection authority banned ChatGPT temporarily in 2023—ultimately lifted, but the message was clear enough. Regulators are willing to act against AI systems whose data practices can't be explained or controlled, even when the technical and legal standards haven't fully settled. Whether machine unlearning is strictly required for GDPR compliance is no longer a question that lives only in academic papers. It's becoming a practical legal question, and the answer is still being worked out.

4.2 Beyond GDPR: A Global Regulatory Mosaic

GDPR is not the only framework that bears on machine unlearning. The California Consumer Privacy Act (CCPA) and its successor, the California Privacy Rights Act (CPRA), grant California residents deletion rights that parallel GDPR's. Brazil's Lei Geral de Proteção de Dados (LGPD) and China's Personal Information Protection Law (PIPL) contain analogous provisions. Taken together, these frameworks reflect a global convergence toward the principle that individuals retain meaningful control over their data even after it has been processed by AI systems.

The EU AI Act, which entered into force in 2024, adds a further layer by imposing requirements on high-risk AI systems to maintain documentation of training data provenance and to provide mechanisms for correcting or removing erroneous data. The Act does not mandate machine unlearning by name, but its data governance requirements create the practical conditions under which unlearning capabilities become necessary for compliance. The United States has taken a more fragmented approach, with sectoral regulations such as HIPAA and FERPA creating relevant but narrowly targeted data control obligations.

4.3 Singapore's Regulatory Landscape

Singapore is a particularly instructive case study in machine unlearning governance. The Personal Data Protection Act (PDPA), administered by the Personal Data Protection Commission (PDPC), establishes data subject rights including the right to withdraw consent for data use. As AI systems become more central to data processing, this increasingly implicates machine unlearning as a compliance mechanism.

In January 2026, Singapore launched the world's first Model AI Governance Framework for Agentic AI, signalling the government's intent to remain at the frontier of AI governance. While the framework does not prescribe specific unlearning techniques, it establishes data accountability and corrective capability requirements that make unlearning infrastructure operationally necessary. The PDPC has also issued Advisory Guidelines on the Use of Personal Data in AI Recommendation and Decision Systems, which explicitly address mechanisms for correcting or removing data influencing automated decisions.

Singapore's role as a regional AI hub—hosting major research centres at NUS and NTU, both of which have contributed to the machine unlearning literature—makes it a critical jurisdiction for the development and standardisation of unlearning methods. Research at NUS Computing, including work by Low et al. (2022) on Markov Chain Monte Carlo-based machine unlearning, and NTU's work on surgical machine unlearning, reflects the region's growing investment in this area.

4.4 Intellectual Property and Copyright Dimensions

A distinct but related regulatory pressure comes from copyright law. The training of large language models on internet-scraped text and code has generated substantial litigation, with rightsholders arguing that training constitutes reproduction of copyrighted works without authorisation. The New York Times Company's lawsuit against OpenAI and Microsoft, filed in 2023, alleges that defendants' models can reproduce substantial portions of copyrighted articles, constituting copyright infringement.

If courts hold that training on copyrighted material without a licence is infringing, machine unlearning becomes the technical mechanism by which organisations might remedy past infringement and prevent future liability. Demonstrating that specific copyrighted works have been effectively removed from a model's knowledge could become a litigation defence or a precondition for licensing agreements between AI companies and content rightsholders.

V. Ethical Dimensions of Machine Unlearning

5.1 Fairness and Bias: Does Forgetting Make Models More or Less Fair?

One of the more counterintuitive tensions in machine unlearning concerns fairness. The ability to remove specific data points appears to offer a remedy for historical biases embedded in training data: if a model was trained on data that reflects discriminatory patterns, selective removal of the most biased examples might improve fairness on downstream tasks.

The picture is more complicated. Cao et al. (2023) investigated the fairness implications of machine unlearning across several fairness-aware learning scenarios and found that naive unlearning procedures—particularly those based on gradient ascent—could worsen fairness outcomes in some cases. Underrepresented groups tend to be associated with fewer, more distinctive training examples; removing or perturbing these examples therefore has a larger proportional effect on the model's behaviour for those groups relative to majority groups. Unlearning in the name of privacy or accuracy can inadvertently deepen the inequities that responsible AI frameworks are trying to address.

The implication is that machine unlearning is not a neutral technical operation—it is an intervention with distributional consequences. Any practical unlearning system must be designed with awareness of those consequences, and evaluated not only for its efficacy in removing targeted data but for its effects on downstream users more broadly.

5.2 The Tension Between Forgetting and Accountability

Machine unlearning creates a genuine conflict between the right to be forgotten and the imperative to maintain accountability for past AI decisions. AI systems are increasingly used to make consequential decisions in criminal justice, healthcare, and financial services. The ability to audit and reconstruct the basis for those

decisions—including the training data from which the model's behaviour derived—is necessary for holding organisations accountable when those decisions cause harm.

If an individual's data is erased from a model after that data informed a discriminatory decision, the erasure may destroy evidence relevant to a legal claim or regulatory investigation. The right to erasure, understood as an individual right, could in principle be exploited by organisations as a mechanism for evading accountability. Wachter, Mittelstadt, and Russell (2017) anticipated this tension, arguing that the opacity of AI systems poses fundamental challenges to meaningful recourse for individuals affected by automated decisions. Machine unlearning could deepen that opacity if not accompanied by robust audit trail requirements—a regulatory gap that has not yet been adequately addressed.

5.3 The Misuse Risk: Adversarial Unlearning

Unlearning techniques could be exploited for harmful purposes. Just as they can remove personal data or copyrighted content, they could in principle suppress safety-relevant training examples. A model trained to refuse harmful requests might be subjected to targeted unlearning of the refusal-inducing examples in its training set, effectively removing its safety guardrails.

This adversarial application—sometimes called "safety unlearning" in the alignment literature—is a concern AI safety researchers have begun to take seriously. Existing work on large language model jailbreaking demonstrates that adversarial fine-tuning can rapidly degrade safety behaviours in aligned models (Wei et al., 2023). Machine unlearning provides a potentially more principled mechanism for achieving the same end. As unlearning techniques become more accessible, the barrier to adversarial deployment decreases, which argues for treating machine unlearning as a security-critical capability rather than purely a privacy tool.

5.4 Consent, Agency, and the Limits of Individualised Forgetting

There is a significant gap between the formal right to erasure and the practical ability to exercise it. In practice, most individuals whose data was used to train AI systems have no knowledge that their data was used, no practical means of identifying which systems trained on it, and no realistic ability to assess whether an unlearning request has been honoured.

Meaningful consent infrastructure would require at minimum: transparency mechanisms by which individuals can discover which AI systems have trained on their data; standardised request portals through which erasure can be formally exercised; technical certification processes by which unlearning can be independently verified; and regulatory enforcement capacity sufficient to impose consequences for non-compliance. None of these exist at scale.

This does not mean machine unlearning is without value. Its ethical significance lies less in the individual erasure case and more in the systemic capability it represents. The development of robust unlearning infrastructure signals that AI systems can be governed, corrected, and held accountable—a commitment whose institutional and symbolic value may be as important as its technical execution.

VI. Synthesis and Discussion

The four dimensions examined in this paper—technical, methodological, regulatory, and ethical—do not merely coexist. They actively conflict with one another in ways that any complete account of machine unlearning must confront.

The most fundamental conflict is between technical exactness and regulatory demand. Exact unlearning methods such as SISA offer the strongest formal guarantees and are therefore most attractive to regulators seeking verifiable GDPR compliance. Yet the data in Table 1 shows that exact methods impose unacceptable accuracy costs on complex tasks, making them unsuitable for the high-stakes AI applications—healthcare, biometric identification, financial services—that are precisely the focus of regulatory concern. Approximate methods fill this gap computationally but undermine the verifiability that regulation requires. The result is a structural mismatch: the methods regulators can rely on are the ones practitioners cannot afford to use.

A second conflict operates between fairness and efficiency. The approximate methods that are computationally tractable—particularly gradient ascent—are also the ones most likely to cause disparate harm to underrepresented groups, as Cao et al. (2023) demonstrated. An organisation that unlearns data in good-faith compliance with privacy law may inadvertently worsen its AI system's fairness outcomes, creating a new liability even as it resolves an old one. Fairness-aware unlearning methods are not yet mature enough to resolve this tension reliably.

A third conflict operates between individual rights and systemic accountability. The right to erasure is an individual right, but its exercise has systemic consequences: it can destroy audit trails, impair the detection of discriminatory patterns, and reduce a model's performance for groups other than the requester. Regulatory frameworks have not yet worked out how to balance these competing interests, and the technical literature has largely set aside the accountability dimension in its focus on the privacy one.

What this synthesis reveals is that machine unlearning cannot be understood as a purely technical problem. The verification gap—the inability to reliably confirm that unlearning has been effective—is simultaneously a technical problem, a regulatory problem, and an ethical one. Closing it requires not just better algorithms but better institutions: independent auditing infrastructure, standardised benchmarks such as TOFU and MUSE, and regulatory frameworks that translate abstract erasure rights into concrete, technically grounded compliance requirements. Singapore, with its combination of advanced AI research capacity and proactive AI governance, is well positioned to contribute to this institutional development.

VII. Open Problems and Future Directions

7.1 Verification and Auditing

The most fundamental open problem in machine unlearning is verification: how can one determine, with confidence, that a given unlearning procedure has actually worked? Current evaluation approaches each have limitations. Membership inference tests may fail to detect subtle forms of memorisation. Output distribution comparisons depend on the choice of test data. Differential privacy certificates provide worst-case guarantees that may be overly conservative.

There is growing interest in "unlearning audits"—third-party verification procedures analogous to financial audits—that provide independent assurance of effective erasure. Such audits would require standardised benchmarks, agreed-upon metrics, and access to model internals. Developing this infrastructure is a priority identified by Thudi et al. (2022) and Nguyen et al. (2022), and is necessary for machine unlearning to fulfil its regulatory function.

7.2 Scalability to Foundation Models

Contemporary AI development is dominated by foundation models—large, general-purpose models such as BERT, GPT-4, and CLIP, trained on massive datasets and fine-tuned for downstream applications. With hundreds of billions of parameters and training datasets drawn from billions of web pages, pinpointing and removing the influence of specific data points is technically formidable.

Parameter-efficient fine-tuning (PEFT) techniques, such as LoRA (Low-Rank Adaptation), offer a promising avenue: by localising updates to low-rank perturbations of the weight matrices, PEFT may enable targeted unlearning without perturbing the full model. This hypothesis is supported by emerging mechanistic interpretability research, which aims to localise specific knowledge representations to identifiable circuits within neural networks (Meng et al., 2022).

7.3 Unlearning Under Adversarial Conditions

Machine unlearning in adversarial settings presents additional complexity. An adversary may attempt to exploit unlearning requests to degrade model performance by strategically requesting removal of data points that are most informative for specific classification tasks. This "unlearning-as-attack" threat model has been studied in the literature, with findings suggesting that coordinated deletion requests can substantially degrade a model's accuracy on specific subgroups. Robust unlearning mechanisms must incorporate safeguards against such exploitation.

Conversely, adversaries may attempt to prevent effective unlearning by poisoning training data in ways that make removal more difficult—embedding targeted information so deeply into model representations that standard unlearning methods cannot adequately address it. The interplay between data poisoning, backdoor attacks, and machine unlearning is a nascent but important research direction (Goldblum et al., 2022).

7.4 Fairness and Unintended Consequences

As discussed in Section 5.1, removing data from a model may have disparate impacts on the model's performance for different demographic groups. Addressing this requires fairness-aware unlearning methods that monitor demographic impact throughout the unlearning process. The right to erasure may also interact with competing social interests: removal of data contributing to criminal pattern recognition models, fraud detection systems, or public health surveillance could impair legitimate societal functions. Resolving such tensions requires sustained engagement across technology, law, ethics, and policy.

7.5 Multimodal Models

The existing unlearning literature has focused mostly on unimodal systems—text-only language models or image classifiers. Contemporary AI systems are increasingly multimodal, integrating text, image, audio, and video representations within a shared embedding space. In models such as CLIP and GPT-4V, the influence of a specific training example may be distributed across modalities in ways that are even harder to attribute and remove. A user's photograph, for instance, may influence both image and text representations in ways that

unimodal unlearning methods cannot adequately address. Developing unlearning methods suited to multimodal architectures is an underinvestigated research direction.

VIII. Conclusion

Machine unlearning has matured from a theoretical concept into a field of active practical importance, driven by the convergence of regulatory mandates and the ubiquitous deployment of data-hungry AI systems. The field has developed a substantial portfolio of methods—from exact unlearning via SISA to approximate unlearning via influence functions, gradient manipulation, federated approaches, and knowledge distillation—each with distinct strengths and limitations.

The central insight unifying these methods is that the goal of unlearning is not mere data deletion but model transformation: the removal of the computational and statistical influence of specific data from a trained system. This framing reveals that machine unlearning is fundamentally a problem of attribution and causality—asking not merely "Is the data gone?" but "Is the model's knowledge of the data gone?" These are distinct questions, and the latter is considerably harder to answer.

A complete unlearning solution would combine the exactness guarantees of SISA-style architectural precommitments with the computational efficiency of PEFT-based approximate methods; it would be verifiable by independent audit using standardised benchmarks; it would incorporate fairness monitoring to detect and mitigate distributional impacts; and it would be robust to adversarial exploitation. No current method achieves all of these simultaneously. The most promising near-term direction is PEFT-based unlearning for foundation models, particularly as mechanistic interpretability research advances our ability to localise specific knowledge representations within neural networks.

The importance of machine unlearning extends beyond regulatory compliance. It is a necessary component of trustworthy AI: a mechanism by which AI systems can be made responsive to the evolving preferences, rights, and circumstances of the individuals they affect. Machine unlearning is, ultimately, an expression of a broader principle—that the power to learn carries with it the responsibility to forget. As AI systems grow more capable and more consequential, establishing this responsibility in technical, legal, and institutional terms is among the defining challenges of our time.

References

[1]. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Traumer, A., Zhang, B., & Papernot, N. (2021). Machine unlearning. arXiv:1912.03817. https://arxiv.org/pdf/1912.03817

[2]. Cao, Y., & Yang, J. (2015). Towards making systems forget with machine unlearning. Proceedings of the IEEE Symposium on Security and Privacy, 463-480.

[3]. Cao, X., Fang, M., Liu, J., & Gong, N. Z. (2023). FedRecover: Recovering from poisoning attacks in federated learning using historical information. IEEE Symposium on Security and Privacy.

[4]. Chundawat, V. S., Tarun, A. K., Mandal, M., & Kankanhalli, M. (2023). Bad teacher: Machine unlearning via null space calibration. arXiv:2205.12374.

[5]. Goldblum, M., Schwarzschild, A., Recht, B., & Goldstein, T. (2022). Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), 1563-1580.

[6]. Graves, L., Nagisetty, V., & Ganesh, V. (2021). Amnesiac machine learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(13), 11516-11524.

[7]. Jang, J., Kwon, T., & Choi, J. (2022). Knowledge unlearning for mitigating privacy risks in language models. arXiv:2210.01504.

[8]. Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. Proceedings of the 34th International Conference on Machine Learning (ICML), 1885-1894.

[9]. Liu, G., Ma, X., Yang, Y., Wang, C., & Liu, J. (2022). FedEraser: Enabling efficient client-level data removal from federated learning models. Proceedings of IEEE/ACM IWQoS 2021, 1-10.

[10]. Low, B. K. H., Tran-Dinh, Q., Ng, T., Jaillet, P., & Seah, C. W. (2022). Pretrained language model-based moral acceptability classification. ACM ASIACCS.

[11]. Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., & Kolter, J. Z. (2024). TOFU: A task of fictitious unlearning for LLMs. arXiv:2401.06121.

[12]. Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems (NeurIPS), 35.

[13]. Nguyen, T. T., et al. (2022). A survey of machine unlearning. arXiv:2209.02299.

[14]. Thudi, A., Deza, G., Chandrasekaran, V., & Papernot, N. (2022). Unrolling SGD: Understanding factors influencing machine unlearning. Proceedings of the 7th IEEE European Symposium on Security and Privacy.

[15]. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box. Harvard Journal of Law & Technology, 31(2), 841-887.

[16]. Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? arXiv:2307.02483.

[17]. European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council: Artificial Intelligence Act. Official Journal of the European Union.

[18]. Personal Data Protection Commission (PDPC), Singapore. (2023). Advisory guidelines on the use of personal data in AI recommendation and decision systems. https://www.pdpc.gov.sg

[19]. Singapore Infocomm Media Development Authority (IMDA). (2026). New model AI governance framework for agentic AI. https://www.imda.gov.sg
[20]. Concordia AI. (2025). State of AI safety in Singapore 2025. https://concordia-ai.com/wp-content/uploads/2025/07/State-of-AI-Safety-in-Singapore-2025.pdf