THE COLUMBIA
SCIENCE & TECHNOLOGY
LAW REVIEW

VOLUME 27 STLR.ORG NUMBER 1

ARTICLE

FORGET ME NOT? MACHINE UNLEARNING'S IMPLICATIONS FOR PRIVACY LAW

Jevan Hutson,[*] Cedric Whitney[†] & Jay T. Conrad[‡]

Generative AI systems are increasingly relied on and are already actively reshaping how we think about privacy and data protection law. Models ingest and process vast amounts of personal and sensitive data, challenging assurances of compliance with legal frameworks like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) with increasing intensity. Machine unlearning is an emerging tool in practitioners' attempts to address these challenges: the act of selectively removing or suppressing specific data, such as personal data that a data subject requests be deleted, from AI models as means of complying with legal obligations or policy goals. This Article's much-needed analysis of unlearning's technicalities and uses builds on recent critical scholarship that examines unlearning's limitations at the technical and policy level. It delves deeper into machine unlearning's implications for privacy and data protection law by situating it within privacy law's broader ecosystem and proposing actionable pathways for integrating unlearning into enforcement and policy. Specifically, this Article evaluates whether privacy laws' legal, remedial,

[*] Acting Asst. Professor, University of Washington School of Law; Dir. University of Washington Technology Law & Public Policy Clinic. Earlier versions of this Article were presented at the 2025 Privacy Law Scholars Conference at the UCLA School of Law. We are especially grateful to Alessandra Calvi for presenting the paper, and to Willie Agnew, Rachel Hong, Jennifer King, Emily McReynolds, Eric Null, Peter Swire, Cobun Zweifel-Keegan, and other participants for their thoughtful and insightful comments. We further thank Miranda Bogen, A. Feder Cooper, Tiffany Georgievski, Katherine Lee, and Deirdre Mulligan for their support, insights, and guidance. Finally, we are deeply grateful to the editors of the *Columbia Science and Technology Law Review* for their diligent review and meaningful feedback, with special thanks to Lily Zheng and Dante Zhu.

[†] PhD , School of Information, University of California Berkeley

[‡] PhD Student, University of Washington School of Law; LL.M., Technology Law & Policy, Georgetown University Law Center; J.D., Seattle University School of Law

and normative aspirations can be reconciled with the technical realities of machine unlearning in generative AI systems. It also contributes to the privacy profession by proposing a framework for integrating machine unlearning into broader privacy-preserving interventions. ***.*** *In doing so, the Article positions machine unlearning as both a vital new tool as well as a site of contestation in the evolving landscape of privacy and AI governance while providing a forward-looking roadmap for aligning machine unlearning with privacy law's goals.*

I. INTRODUCTION

Don't you forget about me. The Simple Minds anthem that once closed *The Breakfast Club* framed teen identity as an anxious negotiation with memory and erasure. Four decades later, that refrain reverberates in a very different hallway, one patrolled not by lockers but by large-scale generative AI models that never graduate, never sleep, and never forget.

As machine learning systems become integral to decision-making, personalization, and content generation, a growing chorus of users, regulators, and privacy advocates is voicing an opposite request: *forget about me.*[1] When chatbots can effortlessly reproduce passages from a college student's blog or an X-ray technician's résumé scraped many years ago, "forgetting" is no longer a teenage dread; it has become a normative imperative grounded in the core commitments of privacy law—autonomy, dignity, and control over one's personal information.[2] These values animate legal instruments such as the General Data Protection Regulation's (GDPR) "right to be forgotten" and emerging U.S. data-deletion rights, which reflect the idea that individuals should not be indefinitely defined by digital traces beyond their control.[3] Forgetting, in this sense, is not merely a technical safeguard but a recognition of the individual's ongoing interest in temporal and contextual integrity, the right to have one's past data lose its hold over one's present identity. Yet the tools available to regulators—delete the record, shred the disk, empty the recycle bin—presume data sits in tidy rows, ready to be

[1] *See, e.g.*, Yonghao Tang et al., *Ensuring User Privacy and Model Security via Machine Unlearning: A Review*, 77 COMPUTERS, MATERIALS & CONTINUA 2646, 2646 (2023) (explaining that users increasingly request that AI systems forget specific data to mitigate privacy risks and comply with data-protection laws); A. Feder Cooper et al., *Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice*, ARXIV 1, 2-3 (Oct. 31, 2025), https://arxiv.org/pdf/2412.06966 [https://perma.cc/8CKG-P66N] (observing that unlearning has become central to legal and policy debates over how to operationalize deletion rights in AI systems).

[2] *See generally* Lillian R. BeVier, *Information About Individuals in the Hands of Government: Some Reflections on Mechanisms for Privacy Protection*, 4 WM. & MARY BILL RTS. J. 455, 458 (1995); Jerry Kang, *Information Privacy in Cyberspace Transactions*, 50 STAN. L. REV. 1193, 1202-03 (1998); Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 534 (2006).

[3] *See, e.g.*, Cooper et al., *supra* note 1, at 2-3 (articulating the gap between machine unlearning's technical methods and privacy law's normative goals of autonomy and control); Tang et al., *supra* note 1, at 2646 (describing the GDPR's right to be forgotten as the legal impetus for unlearning research); *see also* Min Chen et al., *When Machine Unlearning Jeopardizes Privacy*, 2021 PROC. ACM SIGSAC CONF. ON COMPUT. & COMMC'NS SEC. 896, 896-97 (Nov. 15-19, 2021), https://dl.acm.org/doi/pdf/10.1145/3460120.3484756 [https://perma.cc/FN7S-4VUC] (linking unlearning to GDPR's and CCPA's right to be forgotten).

vacuumed away.[4] Modern AI is messier: once personal data is baked into billions of parameters, deletion feels less like hitting the backspace key and more like trying to remove one drop of paint from an entire mural.

To address this emerging challenge, researchers in computer science have proposed a set of methods collectively referred to as "machine unlearning."[5] Machine unlearning encompasses various techniques aimed at modifying a trained model to selectively remove or reduce the influence of specific data points.[6] These methods range from structural removal[7], which attempts exact deletion of data and its influences from the model architecture, to approximate retraining and output suppression[8], which seeks to dull the data's influence or make it unavailable to end users. Each method varies in efficacy, complexity, and computational demand, as well as its theoretical compliance with privacy regulations.

While the promise of machine unlearning is compelling, its integration into privacy law and policy has not been thoroughly explored. This leaves critical gaps in both scholarship and practice. Legal literature has only recently begun to grapple with how technical possibilities align or conflict with normative privacy goals.[9] Consequently, privacy regulators, practitioners, and scholars face uncertainty around whether machine unlearning can truly fulfill the rights and obligations set forth in current privacy laws.

This Article addresses that critical gap. It evaluates the technical literature on methodologies of machine unlearning, maps these techniques onto existing legal standards and normative privacy principles, and examines how machine unlearning aligns or conflicts with core privacy law objectives such as lawful collection, purpose limitation, data minimization, rectification, and erasure. In this way, this Article provides a clear, structured analysis of machine unlearning's potential—and its limitations—in addressing privacy concerns raised by generative AI.

[4] *See* Cooper et al., *supra* note 1, at 13-15 (discussing how data-deletion and "right-to-be-forgotten" obligations under privacy law motivate interest in machine-unlearning methods and analyzing the technical and legal challenges of implementing such remedies).

[5] *See* Avinth Thudi et al., *On the Necessity of Auditable Algorithmic Definitions for Machine Unlearning*, 31 USENIX SEC. SYMP. 4007, 4008--09 (August 10-12, 2022), https://www.usenix.org/system/files/sec22-thudi.pdf [https://perma.cc/73MM-XN4F].

[6] *Id.*

[7] Also called "exact unlearning." *See* Hanon Yan et al., *ARCANE: An Efficient Architecture for Exact Machine Unlearning*, 31 PROC. INT'L JOINT CONF. ON A.I. 4006, 4007 (2022), https://www.ijcai.org/proceedings/2022/556 [https://perma.cc/H4RL-QVSQ]; *see generally* Lucas Bourtoule et al., *Machine Unlearning*, 2021 IEEE SYMP. ON SEC. & PRIV. 141, 141-43, https://ieeexplore.ieee.org/document/9519428 [https://perma.cc/9S5T-6Y6U] (introducing and categorizing foundational approaches to machine learning, including structural-removal methods).

[8] *See* H. Yan et al., *supra* note 7, at 4007.

[9] *See, e.g.*, Cooper et al. *supra* note 1; Saskia Keskpaik, *Machine Unlearning*, EUR. DATA PROT. SUPERVISOR: TECHSONAR REP. 2025, at 19 (Nov. 15, 2024), https://www.edps.europa.eu/system/files/2024-11/24-11-15_techsonar_2025_en.pdf [https://perma.cc/CD7Y-9ALG] (discussing emerging privacy implications of machine unlearning techniques).

Importantly, this analysis recognizes that machine unlearning is not a panacea. While certain methods offer strong theoretical guarantees, practical limitations remain significant.[10] This is particularly true in terms of scalability, computational costs, and robustness against adversarial attacks.[11] For instance, structural methods provide rigorous deletion guarantees yet often require substantial computational resources.[12] In contrast, approximate methods and output suppression techniques are more scalable and cost-effective but frequently lack the precision necessary to fully satisfy stringent legal standards, including the GDPR's "right to be forgotten," under which data subjects may request their personal information be deleted. [13]

In addition to machine unlearning's technical constraints, conceptual tensions also emerge. Privacy laws were traditionally crafted with discrete databases in mind; compliant data deletion meant straightforward removal. [14] However, generative AI models do not simply store data; they generalize from it. Successful unlearning, in the technical sense, means that a model's post-unlearning behavior is statistically indistinguishable from that of a model trained without the deleted data.[15] Yet even when this benchmark is met, unlearning may leave latent traces: residual patterns, correlations, or representational artifacts that continue to shape a model's outputs or enable reidentification of the affected individual. [16] These residual influences frustrate the normative goals of privacy law, which center on individual autonomy, control over personal information, and protection from reputational and relational harms. [17] When personal data continues to inform a model's generative behavior even indirectly, individuals lose meaningful control over how they are represented or remembered, thus undermining the spirit of rights like the GDPR's "right to be forgotten" and the CCPA's deletion right.[18] Recent empirical research underscores these risks: even after unlearning, adversaries can infer or reconstruct supposedly "forgotten" data through membership inference,

[10] *See* Thudi et al., *supra* note 5, at 4009 ("However, even after the speedup [of unlearning methods], the costs may still be too high for some.").

[11] *See* discussion *infra* Section II.D.

[12] *See e.g.*, Bourtoule et al., *supra* note 7, at 156 ("Costs Associated with Storage").

[13] *See* Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation) [hereinafter GDPR], art. 17, 2016 O.J. (L 119) 1.

[14] *See* Cooper et al., *supra* note 1, at 13-15.

[15] *See* Tang et al., *supra* note 1, at 2650-51 (defining "exact unlearning" as the condition where a model behaves as though the deleted data were never used for training).

[16] *See* M. Chen et al., *supra* note 3, at 896 (demonstrating that unlearned models may still leak information about deleted data through discrepancies between original and unlearned models).

[17] *See, e.g.*, Cooper et al., *supra* note 1, at 13-16; M. Chen et al., *supra* note 3, at 897 ("[R]emoving information from a model's parameters does not guarantee that this model could never produce related information at generation time.").

[18] *Id.*; *see also* GDPR, art. 17 (establishing a right to erasure of personal data); CAL. CIV. CODE § 1798.105 (2023) (granting consumers the right to request deletion of personal information collected by businesses).

model inversion, or prompt-based exploitation, exposing the limits of current unlearning methods.[19]

To navigate these complexities, this Article advocates for a nuanced regulatory approach. Machine unlearning can meaningfully bolster privacy governance when it is treated as a partial remedy layered alongside data-minimization, purpose-limitation, differential privacy, and rigorous oversight.[20] As such, this Article asserts that machine unlearning should be integrated into a broader privacy governance framework, encompassing both preventive measures and reactive remedies. Specifically, the Article proposes a multifaceted enforcement strategy aligned with the authority of regulatory bodies such as the Federal Trade Commission (FTC), Department of Justice (DOJ), and state Attorneys General. This approach emphasizes proportionality as a guiding legal principle, meaning that unlearning-based remedies should be calibrated to the nature and gravity of the underlying harm or unlawful data use.[21] In other words, the severity of the intervention (for example, targeted unlearning versus full model disgorgement) should correspond to the degree of risk to individuals' rights or the extent of the non-compliance, ensuring remedies are both effective and practicable within existing enforcement structures.[22]

This Article contributes significantly to privacy law scholarship by bridging the technical realities of machine unlearning with legal and regulatory objectives. It provides policymakers, legal practitioners, and scholars with a clear-eyed assessment of machine unlearning's capabilities when situated as a crucial tool within a diversified privacy protection arsenal rather than as a standalone solution. In doing so, it lays the foundation for informed policy development and enforcement strategies that can effectively safeguard privacy in the age of generative AI.

This Article proceeds in four Parts. Following this Introduction, Parts II through IV each aim to bridge the technical capabilities of machine unlearning with the

[19] *See* M. Chen et al., *supra* note 3, at 896-97 (reporting that membership inference attacks can outperform classical baselines post-unlearning); *see also* Cooper et al., *supra* note 1, at 12 (noting that generative models can reintroduce forgotten information through prompts, a phenomenon termed "ununlearning"). *See generally* Ilia Shumailov et al., *UnUnlearning: Unlearning Is Not Sufficient for Content Regulation in Advanced Generative AI*, ARXIV 1 (June 27, 2024), https://arxiv.org/pdf/2407.00106 [https://perma.cc/W8UN-CMTG].

[20] *See* Cooper et al., *supra* note 1, at 3 ("Unlearning methods are imperfect and may serve as only one approach of many that could, in some cases, contribute to addressing aspects of issues that are of interest to policymakers.").

[21] *See also* GDPR, recital 129 (noting that remedies and sanctions must be effective, proportionate and dissuasive). *See generally* Alessandro Achille et al., *AI Model Disgorgement: Methods and Choices*, 121 PROC. NAT'L ACAD. SCI. (Apr. 19, 2024), https://www.pnas.org/doi/10.1073/pnas.2307304121 [https://perma.cc/N9V6-NUDG] (explaining that model disgorgement and unlearning can be applied proportionally to the severity of harm or data misuse).

[22] *See* Achille et al., *supra* note 21, at 3-4 (proposing that regulators select among technical remedies such as retraining, selective forgetting, or differential-privacy retraining based on the scale of the data defect and resulting harm).

normative and operational demands of privacy law. Part II synthesizes the emerging computer science literature on machine unlearning and categorizes the leading techniques into families of structural removal methods, approximate retraining, and output suppression. It explains how each method functions, highlights respective trade-offs, and identifies key technical limitations, including persistent counter-privacy risks such as the ability to re-identify data through model outputs.[23] By clarifying both the potential and the limitations of machine unlearning, Part II equips legal practitioners and policymakers with a grounded understanding of the technique's functional contours.

Part III analyzes how machine unlearning aligns with and frequently challenges existing privacy law frameworks. It examines how machine unlearning maps onto foundational legal principles like lawful collection, purpose limitation, data minimization, the right to rectification, and the rights to object or withdraw consent.[24] While machine unlearning may appear to satisfy some legal obligations in form, this Part argues that it often fails to meet their normative intent, particularly when latent data influence or residual outputs continue to implicate individual privacy. Part III also considers how machine unlearning could augment existing remedies (such as full model deletion and algorithmic disgorgement) in highlighting the risks of substituting meaningful accountability with incomplete technical fixes.

Part IV proposes a regulatory and enforcement framework for operationalizing machine unlearning as one component of a broader privacy intervention strategy. Drawing on the authority of the Federal Trade Commission under Section 5 of the Federal Trade Commission Act[25], the Department of Justice's consumer protection toolkit[26], state Attorneys General enforcement powers[27], and emerging global norms[28], this Part outlines how machine unlearning can be implemented alongside preventive or reactive remedies. These include privacy-preserving model design, output filtering, model deletion, and algorithmic disgorgement. It evaluates how such approaches could be tailored to practical realities like computational cost, scalability, and the tradeoff between forgetting efficacy and model performance. Finally, this Part warns against overly technocratic "compliance-by-design" approaches—frameworks that treat privacy protection primarily as a technical or procedural matter to be engineered into systems ex ante rather than as a substantive,

[23] *See generally* Shumailov et al., *supra* note 19.

[24] *See, e.g.*, GDPR; CAL. CIV. CODE §§ 1798.100-.199.100 (2018) [hereinafter CCPA] (as amended by the California Privacy Rights Act of 2020).

[25] *See* Federal Trade Commission Act [hereinafter FTC Act] § 5, 15 U.S.C. § 45 (2023).

[26] *See Consumer Protection Branch*, U.S. DEP'T OF JUST., https://www.justice.gov/archives/civil/consumer-protection-branch [https://perma.cc/E563-SYF7] (last visited Nov. 8, 2025).

[27] *See Center for Consumer Protection*, NAT'L ASS'N OF ATT'YS GEN., https://www.naag.org/our-work/center-for-consumer-protection/ [https://perma.cc/7H9B-DP2U] (last visited Nov. 8, 2025, at 17:33 EST).

[28] *See, e.g.*, GDPR, art. 17; CCPA §§ 1708.100-.199.100.

ongoing obligation—and instead calls for a multifaceted, dynamic governance strategy that centers substantive privacy protections over procedural adequacy.

II. SYNTHESIZING THE COMPUTER SCIENCE LITERATURE ON MACHINE UNLEARNING

Part II synthesizes the emergent computer science literature on machine unlearning.[29] This includes analyses of structural removal methods, approximate retraining methodologies, and suppression techniques, each which have varying levels of effectiveness in removing or obscuring personal data. By clarifying how these techniques function and where they fall short, this Part helps legal practitioners and policymakers understand the possibilities and the limitations of machine unlearning.

In practice, a single deletion tool should not be relied upon to improve data privacy. Providers of machine learning services, and thus corresponding unlearning services, should pair "heavier," that is, exact or certified[30] unlearning routines (those used for legally significant takedown requests) with "lighter" filters and prompts (those used for day-to-day safety). The heavier methods supply a defensible record for potential data protection, privacy, or compliance audits while the lighter weight methods keep daily-used inference latency low, meaning a responsive and seamless "unlearned" data experience on the user end.[31] That providers understand this layered approach is important as subsequent sections in Part II map to individual layers of the real-world technology stack, including structural techniques that rewrite the model, approximate updates that tweak its weights, and suppression methods that guard the interface.[32] The comparative analysis at the end of Part II therefore posits not which method is best but which blend of methods meets the legal and operational needs of a given context.

A. Structural Removal Methods

Structural approaches for removing data start from the premise that the only way to "truly" forget a record is to rebuild the learner itself.[33] Afterwards, the model

[29] Throughout this Article, the terms "machine unlearning" and "unlearning" will be used interchangeably.

[30] *See* Chuan Guo et al., *Certified Data Removal from Machine Learning Models*, 119 PROC. MACH. LEARNING RSCH. 3832 (2020) (describing an approach that has "a very strong theoretical guarantee that a model from which data is removed cannot be distinguished from a model that never observed the data to begin with"); Thudi et al., *supra* note 8, at 4014 (describing an approach that has "the advantage of providing rigorous guarantees at the model level").

[31] *See* Thudi et al., *supra* note 5, at 4007 ("We thus conclude that . . . an entity's only possible auditable claim to unlearning is that they used a particular algorithm designed to allow for external scrutiny during an audit.").

[32] *See generally* Bourtoule et al., *supra* note 7 (describing the SISA framework techniques of sharding, isolating, slicing, and aggregating).

[33] *See, e.g.*, *id.* at 141 ("A naive way to have such models provably forget is to retrain them from scratch."); Thudi et al., *supra* note 5, at 4009 ("Exact unlearning for DNNs is based on

behaves as if the record never existed.[34] This Section explains and proposes various structural removal methods and assesses the strength of their guarantees of the requested data's deletion. It also addresses practical concerns regarding deletion requests' technical impact on the models themselves, such as known impacts on outputs and predictions, the required compute power to execute the deletions, and relevant time constraints.

Researchers first demonstrated structural removal with Sharded, Isolated, Sliced, Aggregated (SISA) training.[35] SISA is a technique that divides the original dataset into many independent pieces called shards.[36] A small sub-model is then trained on each shard before averaging the sub-models to obtain final predictions.[37] With SISA, when a data subject asks for their information to be erased, only the sub-model that ever saw that data is rebuilt; the other sub-models are left untouched.[38] Importantly, SISA significantly decreases the computation required for data deletion as compared with 'starting over' with an entirely and holistically retrained model.[39]

Relevant to privacy concerns, SISA also results in before-and-after snapshots of the model that differ in predictable ways, making comparisons easier for confirming the deletion.[40] Because of this, privacy professionals can compare the two points to confirm the deletion has occurred. This can be done because SISA's update rule is fixed: if you repeat the same deletion on the same shard, you get the same new weights in the final model every time.[41] Thus, an auditor who holds both versions can determine with better-than-chance accuracy whether the record had in fact been present and subsequently deleted before the model was retrained.[42]

retraining. In detail, the model owner needs to discard the old model, remove the data points that are required to be unlearned, and train a new model on the modified dataset.").

[34] *See, e.g.*, Bourtoule et al., *supra* note 7, at 141; Thudi et al., *supra* note 5, at 4009.

[35] *See* Bourtoule et al., *supra* note 7, at 142.

[36] *See id.*

[37] *See id.*

[38] *See id.*

[39] *See id.* at 141-42.

[40] *See* Thudi et al., *supra* note 5, at 4010-11; Thanh Tam Nguyen et al., *A Survey of Machine Unlearning*, ARXIV 1, 5 (Sept. 17, 2024), https://arxiv.org/pdf/2209.02299 [https://perma.cc/6WSE-KTNZ] ("The goal of unlearning verification methods is to certify that one cannot easily distinguish between the unlearned models and their retrained counterparts."). *See generally* Bourtoule et al., *supra* note 7 (explaining how SISA training's implementation can be audited to confirm data removal).

[41] *See* Bourtoule et al., *supra* note 7, at 154 ("One could imagine that authorities relevant to the enforcement of the right to be forgotten could audit the code base to validate the implementation of SISA training").

[42] Thudi et al., *supra* note 5, at 4009 ("Reproducing the alleged computation is synonymous to showing its plausibility."); Nguyen et al., *supra* note 40, at 3 ("[A] verification (or audit) is needed to prove that the model actually forgot the requested data and that there are no information leaks.").

Later scholarly work realized that the shard need not be defined by where the data are stored.[43] Instead, it can be defined by *what* the data describe.[44] Another structural data removal method, ARCANE,[45] builds dozens of narrow "expert" networks, each specializing in one class or topic, and only stitches them together at inference time (when the networks, combined, generate outputs from data-based inferences).[46] If the deletion-requested record lives solely in the expert network of a particular topic, the system can retrain that expert network alone before it re-sews it into the overall network used for inferences (outputs).[47] This does mean, however, that ARCANE relies on preemptive bookkeeping.[48] The machine learning service provider must maintain a precise map of which expert network touched a given record so that the correct sub-networks can be retrained when a new data deletion request arrives.[49]

A benefit to privacy professionals using ARCANE architecture is the increased speed of data deletion while maintaining model accuracy: experiments on benchmark image collections, that is, first-pass collections of images used to train an individual expert network, have shown that ARCANE cuts the selected data's deletion time from nearly three-quarters of an hour for naive retraining to roughly three minutes, while accuracy on the untouched classes falls by less than a percentage point.[50] By aggressively limiting the scope of the privacy-necessary retraining, ARCANE demonstrates that organizations can guarantee data deletion backed with practical and reasonable turnaround times, particularly for everyday takedown requests.[51]

Classical statistical models such as logistic regression and soft-margin support-vector machines have a special property: they are "single-solution" learners.[52] As such, researchers can write an exact algebraic formula that scrubs out a single training record while keeping every other weight intact.[53] For any given training set, there is only *one* set of weights that minimizes the margin of error; thus, the error always rises smoothly as you move away from that point.[54] Because the landscape is so well-behaved, there are no hidden pockets or second-best valleys.[55]

[43] *See, e.g.*, H. Yan et al., *supra* note 7, at 4007.

[44] *See, e.g.*, *id.* ("Instead of uniform division, we divide the dataset by class . . . .").

[45] ARCANE has a confusing formal acronym (Architecture foR exaCt mAchine uNlEarning); the acronym is rather used as a descriptive name for a specific architecture for exact machine unlearning. *See id.*

[46] *See id.*

[47] *See id*. at 4008.

[48] *See id.*

[49] *See id.*

[50] *See id.* at 4010-11 ("When large data unlearning . . . the accuracy of ARCANE would not degrade too much [and] the training and unlearning of ARCANE is much faster than SISA.").

[51] *Id.*

[52] *See* Guo, *supra* note 30, at 3837-38.

[53] *See id*. at 3832-34.

[54] *See id*.

[55] *See id*.

After removing the requested data, the model's predictions on all *other* (undeleted) data stay the same, so the model's overall accuracy is unchanged.[56] Researchers Neel et al. (2021) have expanded the math to cope with thousands of deletion requests in a row without the model's runtime spiraling problematically upward.[57] The resulting deletion guarantee is strong.[58] Given only the final weights, no statistical test, however sophisticated it may be, can tell whether the erased record was ever in the training set.[59] Although this technique applies only to these 'single-solution' learners, it sets the conceptual upper bound for what perfect unlearning looks like.

Contemporary production models, however, are rarely 'single-solution,' thus presenting a different challenge to privacy professionals seeking to comply with deletion requests.[60] Contemporary production models differ from classical models in that they often employ transformer backbones or operate in federated settings where data never leave the user's device.[61] It is important to know that because of these dispersed controllers and data processors, a single data warehouse cannot locate the records requested for deletion.[62] Recent research therefore targets these contexts directly.[63] This area of research is of growing importance because health and finance apps increasingly train models in this federated way.[64]

[56] *See id.*

[57] *See* Seth Neel et al., *Descent-to-Delete: Gradient-Based Methods for Machine Unlearning*, 132 PROC. MACH. LEARNING RSCH. 931, 932 (2021), http://proceedings.mlr.press/v132/neel21a/neel21a.pdf [https://perma.cc/92E2-WKM8].

[58] *See id.* at 932-33.

[59] *See id* at 931.

[60] *See* Ziyao Liu et al., *Privacy-Preserving Federated Unlearning with Certified Client Removal*, ARXIV 1, 1 (2024), https://arxiv.org/pdf/2404.09724 [https://perma.cc/HU29-6L8J] (noting that federated unlearning techniques must work in systems without a single unified model).

[61] *See* Guangsheng Zhang et al., *How Does a Deep Learning Model Architecture Impact Its Privacy? A Comprehensive Study of Privacy Attacks on CNNs and Transformers*, ARXIV 1, 1 (2022), https://arxiv.org/pdf/2210.11049 [https://perma.cc/2AQW-XAR8] (showing that transformers have different privacy vulnerability properties relative to classical architectures); *see also* Sarthak Pati et al., *Privacy Preservation for Federated Learning in Health Care*, PATTERNS, July 12, 2024, at 2 (discussing federated models where client data remains local).

[62] *See* Chunlu Chen et al., *Trustworthy Federated Learning: Privacy, Security, and Beyond*, 67 KNOWLEDGE & INFO. SYS. 2321, 2324 (Nov. 26, 2024) (surveying challenges of decentralized controllers in federated settings); *see also* Joshua C. Zhao et al., *The Federation Strikes Back: A Survey of Federated Learning Privacy Attacks, Defenses, Applications, and Policy Landscape*, ARXIV 1, 1 (Mar. 22, 2025), https://arxiv.org/pdf/2405.03636 [https://perma.cc/C2LC-V99L] (highlighting that federated systems lack centralized data storage).

[63] *See* Z. Liu et al., *supra* note 60; *see also* Lina Ge et al., *A Review of Secure Federated Learning: Privacy Leakage Threats, Protection Technologies, Challenges and Future Directions*, 571 NEUROCOMPUTING (SPECIAL ISSUE) 1, 1 (2023) (exploring methods of unlearning in federated learning contexts).

[64] *See generally* Subhash Nerella et al., *Transformers and Large Language Models in Healthcare: A Review*, 154 A.I. MED. 1 (2024), https://pmc.ncbi.nlm.nih.gov/articles/PMC11638972/ [https://perma.cc/9AXS-C5ZW] (suggesting federated learning paradigm for training healthcare model without divulging patient information); D. T. Braithwaite et al., *Your Spending Needs Attention: Modeling Financial Habits with Transformers*, ARXIV 1 (July 31, 2025),

The transformer backbone context offers deletions completed with low compute costs, maintenance of model integrity, and user-level deletion guarantees that may be appealing to privacy professionals requesting data removal. In transformer backbone contexts, vision-transformer studies have shown that a small low-rank patch applied to a pretrained image classifier can wipe out an entire ImageNet class in under ten GPU-minutes, and the remaining classes scarcely notice the surgery (data removal).[65] In the federated sphere, the FedEraser protocol reconstructs a global model by replaying the history of every client except the deletion-requested one, letting the central server erase a hospital's data contribution, for example, with roughly one quarter of the compute previously required.[66] Similarly, researchers have proposed a client-level protocol that allows individual smartphones to erase local contributions while the distant server replays only uncontaminated updates.[67] The FedEraser protocol reduces central compute by a factor of four, meaning that the same deletion task can be completed using roughly one-quarter of the energy, hardware, and processing time previously required.[68] Such efficiency gains are non-trivial: they make large-scale unlearning more economically and environmentally viable and thus more likely to be adopted by firms and considered proportionate by regulators when evaluating compliance burdens under privacy and data-protection law.[69] For privacy professionals and regulators, these technical advances illustrate that unlearning is no longer purely theoretical. They demonstrate that data-deletion obligations can be operationalized at scale whether at the model layer (through efficient, low-rank updates) or the system layer (through federated unlearning protocols). In other words, the technology is beginning to make selective, legally compliant 'forgetting' feasible and proportionate, offering practical tools to satisfy erasure or withdrawal-of-consent rights without dismantling entire models.

Even with such optimizations, structural unlearning consumes significant hardware.[70] The staggering hardware requirements may make structural unlearning tasks unreasonable to undertake. Machine learning service providers also face a trade-off where the nearer that the method gets to absolute data record erasure, the

https://arxiv.org/pdf/2507.23267 [https://perma.cc/M4LK-PDFG] ("Predictive models form the underpinnings of many systems at financial institutions, such as risk prediction, product recommendations, and fraud detection . . . financial institutions have access to large amounts of user data . . . leveraging this data effectively remains challenging.").

[65] *See* Ikhyun Cho et al., *ViT-MUL: A Baseline Study on Recent Machine Unlearning Methods Applied to Vision Transformers*, ARXIV 1 (Feb. 7, 2024), https://arxiv.org/pdf/2403.09681 [https://perma.cc/WFE5-YEFT].

[66] Gaoyang Liu et al., *Federaser: Enabling Efficient Client-Level Data Removal from Federated Learning Models*, 29 IEEE/ACM INT'L SYMP. ON QUALITY SERV. at 1, 2 (2021).

[67] *Id.*

[68] *See id.* at 1.

[69] *See id.* at 2.

[70] *See, e.g.*, Bourtoule et al., *supra* note 7, at 150 (conducting experiments using high-end hardware, including Intel Xeon Silver 4110 CPUs).

more electricity and scheduling complexity it demands.[71] The original SISA study, for example, showcases the financial and compute-power burdens of unlearning: its largest ImageNet experiment still needed eight Nvidia V100 GPUs (graphics processing units) for *each* retrained shard, a figure that rises quickly when the underlying model is a multi-billion-parameter language model.[72] For added perspective, Eldan and Russinovich (2023) note that their 7-billion-parameter baseline consumed 184,000 GPU-hours (the total time the GPU spends computing) during its initial training run; a full exact retrain of even a modest slice would still land in the same order of magnitude.[73] That reality explains why many machine learning service providers default to using less rigorous, but far cheaper, approximate or suppressive techniques.[74] The next two Sections analyze these techniques.

B. Approximate Retraining Methods

Approximate unlearning methods trade perfect fidelity for speed. They begin with a fully trained network and apply a limited number of weight updates designed to blunt, rather than eliminate, the influence of a chosen record.[75] This is done by adjusting gradients.[76] A gradient is the mathematical direction that most steeply changes the model's error, so climbing the gradient for the offending data makes the network "unlearn" that contribution.[77] The flagship technique is Descent-to-Delete, which performs a handful of carefully scaled gradient-ascent steps on the record marked for removal and an equal number of gradient-descent steps on the records that should remain.[78] Empirical tests show that on logistic and small convolutional models, three to five such steps can mimic a full model retrain while costing less than one percent of the original training time.[79] For privacy lawyers,

[71] *See, e.g.*, *id.* at 142 (requiring greater scheduling complexity for better performance); *see also* Ronen Eldan & Mark Russinovich, *Who's Harry Potter? Approximate Unlearning in LLMs*, ARXIV 1, 1 (Oct. 4, 2023), https://arxiv.org/pdf/2310.02238 [https://perma.cc/QM2A-LSQF].

[72] *See* Bourtoule et al., *supra* note 7, at 150.

[73] Eldan & Russinovich, *supra* note 71, at 1.

[74] *See, e.g.,* OpenAI, *Moderations*, APPLICATION PROGRAMMING INTERFACE REFERENCE, https://platform.openai.com/docs/api-reference/moderations [https://perma.cc/EG62-PG9J] (last visited Nov. 9, 2025) (providing a simple moderation tool to "unlearn"); Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, 36 PROC. INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 27730, 27730 (Nov. 28, 2022), https://proceedings.neurips.cc /paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf [https://perma.cc/WV5A-LWTT] (asserting that bigger models do not necessarily make them better at following user intent).

[75] Zhehao Huang et al., *Unified Gradient-Based Machine Unlearning with Remain Geometry Enhancement,* ARXIV 1, 2-3, (Sept. 29, 2024), https://arxiv.org/pdf/2409.19732 [https://perma.cc/ HW8R-G7KZ].

[76] *Id.*

[77] *Id.*

[78] Neel et al., *supra* note 57, at 4.

[79] *Id.*

the appeal of this method is that the provider can respond to many erasure requests overnight instead of launching a week-long retraining job.

Other research suggests removing the requested information by injecting noise calibrated to each associated weight's importance rather than by adjusting the relevant gradients.[80] Fisher-scrubbing, a technique sometimes called "Eternal Sunshine of the Spotless Net," makes use of the Fisher information matrix, which is a mathematical tool measuring how important each weight is for the model's predictions.[81] The Fisher-scrubbing algorithm then adds just enough random variation to the most sensitive weights determined by the Fisher information matrix to obscure any statistical trace of the targeted record while leaving other predictions intact.[82] In practice this means that, within a tight margin, the post- Fisher-scrub weights behave as though the record had never been present in the model.[83] This is true despite that the network retains over ninety-nine percent of its baseline accuracy on the rest of the data.[84] Because the method needs no access to the original training set, it is attractive for cloud providers that have already deleted or archived raw data pursuant to a retention policy. The drawback is that the privacy guarantee is probabilistic; a sophisticated auditor might still extract faint traces of the deletion-requested record if they run enough queries through the model.[85]

A practical variant to Fisher-scrubbing is targeted fine-tuning, where the provider trains the model for a brief period on "anti-examples" that teach it to down-weight the requested data.[86] Because the rest of the weights stay untouched, the system keeps its performance on unrelated tasks, which may be a key consideration for organizations' achieving proportionality in compliance with privacy laws or remedies.[87] This technique is attractive in that it uses less compute power.[88] To illustrate, recall the earlier mentioned experiment by Eldan and Russinovich, which found retraining a model using structural methods resulted in 184,000 GPU-hours needed for the original training run, which presented an issue were use of that technique to be scaled.[89] Using the Fisher-scrubbing technique instead, they found it possible to force a seven-billion-parameter model to "forget" every line of the Harry Potter books in just *one* GPU-hour, a significant improvement in speed.[90] They first identified the tokens most associated with the

[80] *Id.*

[81] Aditya Golatkar et al., *Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks*, 2020 IEEE/CVF CONF. ON COMPUT. VISION & PATTERN RECOGNITION 9301, 9301-02 (Aug. 5, 2020).

[82] *Id.* at 9302.

[83] *Id.* at 9307.

[84] *Id.*

[85] *Id.*

[86] *Id.* at 9308.

[87] Eldan & Russinovich, *supra* note 71, at 4-8.

[88] *Id.* at 1.

[89] *Id.*

[90] *Id.*

books. They then replaced them with neutral equivalents so that the model's logits—the raw scores before they become probabilities—no longer spiked on 'Potter phrases.'[91]

There are other types of fine-tuning that service providers could use, although with varying degrees of successful data erasure. One newer variant of fine-tuning called Langevin Unlearning adds calibrated noise during this short fine-tune period to provide a probabilistic certificate that the erased data cannot be reconstructed.[92] However, Shumailov et al. (2024) showed that forensic attacks can still recover snippets if the noise budget is too small.[93]

A third approach, Amnesiac Machine Learning, erases classes or individual examples.[94] It does so by first pruning the network into a sparse skeleton by setting many weights to zero, which disentangles the internal representation and reduces the memory capacity for any single record. It then finetunes only that compact core.[95] Once the network is sparse, a brief noise-infused fine-tuning centered on the 'forget' set neutralizes the residual influence.[96] It does so far more effectively than the same procedure applied to a dense model and delivers a reported five-fold speed-up over naive methods while losing less than half a percentage point of accuracy.[97] The upsides of this tactic is that it is a more cost-effective technique, making it appealing to smaller firms.[98] Additionally, service providers can comply with large batches of deletion demands on limited hardware.[99] The legal downside is residual risk.[100] Because the method offers no formal proof of deletion, regulators may still insist on more rigorous post-hoc audits or complementary front-end filters.[101] Consequently, practitioners often pair approximate retraining with routine privacy audits or differential-privacy noise to narrow the exposure window created by partial forgetting.[102]

[91] *Id.* at 4-5.

[92] Eli Chien et al., *Langevin Unlearning: A New Perspective of Noisy Gradient Descent for Machine Unlearning*, ARXIV 1, 1 (Jan. 18, 2024), https://arxiv.org/pdf/2401.10371 [https://perma.cc/P66B-ET8F].

[93] Shumailov et al., *supra* note 19, at 1.

[94] Laura Graves et al., *Amnesiac Machine Learning*, 35 PROC. AAAI CONF. ON A.I. 11516, 11516 (2021) https://ojs.aaai.org/index.php/AAAI/article/view/17371 [https://perma.cc/G3U5-7YKT].

[95] *Id.* at 11518.

[96] *Id*. at 11519.

[97] *Id.* at 11520-21.

[98] *Id*. at 11522.

[99] *Id.*

[100] *Id.*

[101] *Id.*

[102] *See* Jie Xu et al., *Machine Unlearning: Solutions and Challenges*, 8 IEEE TRANSACTIONS ON EMERGING TOPICS COMPUTATIONAL INTEL. 2150, 2164-65 (June 2024), https://ieeexplore.ieee.org/document/10488864 [https://perma.cc/9TZX-UMS4] ("Beyond removing model parameters through machine unlearning algorithms, additional technical and legal steps are required to fully assert this right in practice, such as verifiable proof of unlearning, proof of data ownership, auditing

Differential privacy (DP) often surfaces in policy conversations as an alternative to unlearning, but the two safeguards (unlearning and differential privacy) serve different purposes. It is important to distinguish the two from each other. DP adds carefully calibrated noise during training so that any single record's changes to the model's outputs is limited to be only within a narrow statistical band, thus limiting what can be inferred.[103] However, DP does not remove the record's influence; it merely *masks* it, which is why Ginart et al. (2019) coined the phrase "data removal after learning" to argue that DP alone cannot satisfy a strict erasure order.[104] Recent large language model experiments show that combining DP with a lightweight approximate unlearning step yields stronger privacy than either technique on its own.[105] This is because the noise limits membership inference while the extra gradient steps blunt memorized text.[106] Mattern and co-authors (2023) reach a similar conclusion in their study of client-side voice assistants.[107] Their research showed that a lightweight DP mask followed by a Descent-to-Delete styled two gradient-ascent steps on the 'forget' set cuts membership inference success from 46 percent to near-chance levels while adding under one percentage point of word-error rate.[108]

C. Output Suppression Techniques

Output suppression methods focus on censoring what the model says (output) rather than changing what it has learned.[109] The approach is attractive because it needs a modest amount of additional training time and does not require access to the original data.[110] Notably, its guarantee is purely behavioral, which means it is effective only so long as the refusal policy is not bypassed.[111]

for potential privacy leaks, and employing privacy-enhancing technologies. . . . Most existing approximate unlearning algorithms rely on differential privacy to provide formal unlearning guarantees.").

[103] Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, 9 FOUND & TRENDS THEORETICAL COMPUT. SCI. 211, 211 (2014).

[104] Antonio Ginart et al., *Making AI Forget You: Data Deletion in Machine Learning*, 33 PROC. INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 3518 (Dec. 8, 2019).

[105] Josep Domingo-Ferrer et al., *Efficient Unlearning with Privacy Guarantees*, ARXIV 1, 1 (July 7, 2025), https://arxiv.org/pdf/2507.04771 [https://perma.cc/TCT2-RC8Q].

[106] *See generally* Sijia Liu et al., *Rethinking Machine Unlearning for Large Language Models*, 7 NATURE MACH. INTEL. 181 (Feb. 17, 2025).

[107] Justus Mattern et al., *Membership Inference Attacks against Language Models via Neighbourhood Comparison*, 2023 FINDINGS ASS'N FOR COMPUTATIONAL LINGUISTICS 11330 (July 9-14, 2023), https://aclanthology.org/2023.findings-acl.719.pdf [https://perma.cc/D4T6-NXR7].

[108] *Id.*

[109] Cooper et al., *supra* note 1, at 1-2.

[110] *Id.* at 10-11.

[111] Yukai Zhou et al., *Don't Say No: Jailbreaking LLM by Suppressing Refusal*, 2025 FINDINGS ASS'N FOR COMPUTATIONAL LINGUISTICS 25224, 25224-25 (July 27-Aug. 1, 2025), https://aclanthology.org/2025.findings-acl.1294.pdf [https://perma.cc/Q8B6-A52N].

The most influential technique in this category is reinforcement learning from human feedback, often abbreviated RLHF.[112] In an RLHF workflow, human annotators score model answers and then train a separate reward model on these scores so that the system begins to prefer answers that align with policy goals (such as refusing to reveal personal data).[113] RLHF only teaches the model to produce a safe user-facing refusal message like "I am sorry but I cannot help with that request." Ouyang and colleagues found that an InstructGPT model trained with RLHF reduced toxic or biased language compared with the original GPT-3 and was even preferred over larger, more powerful baselines for helpfulness and truthfulness.[114]

A lighter-weight form of suppression relies on carefully written prompts or instructions that steer the model away from sensitive content at inference time.[115] A system prompt can instruct, "Do not reveal any personally identifying information or copyrighted text," and that single line can be deployed instantly across thousands of replicas without retraining.[116]

The technique is popular for its speed, but it is vulnerable to what researchers call *prompt injection*, an attack that tricks the model into ignoring the safety instruction by embedding conflicting directions in the user prompt.[117] Because the prompt layer has no cryptographic separation from user input, a determined adversary can iterate through phrasing variations until the filter cracks.[118] Zou and collaborators catalogued a library of adversarial prompts that bypassed multiple commercial filters and forced models to reveal disallowed outputs with high reliability.[119] The result is that prompt-based suppression may satisfy low-risk consumer use cases yet offers little comfort where a regulated entity must show to regulators that disclosure is impossible or improbable rather than merely discouraged.

[112] Ouyang et al., *supra* note 74.

[113] *See id.* at 27731; Dzmitry Bahdanau et al., *Learning to Understand Goal Specifications by Modelling Reward*, ARXIV 1-2 (Dec. 23, 2019), https://arxiv.org/pdf/1806.01946 [https://perma.cc/T3GF-GPQL].

[114] Ouyang et al., *supra* note 74, at 27732.

[115] *See,* e.g., *id.* at 27743 (showing inference-time prompts can effectively modulate a language model's behavior, particularly regarding toxicity).

[116] *See*, e.g., *id*. at 27732 (showing models' toxic output generation was immediately and significantly altered by about 25% less toxicity merely by the inclusion of a "respectful prompt" at the time of execution).

[117] *See, e.g.,* Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models*, ARXIV 1, 4 (Dec. 20, 2023), https://arxiv.org/pdf/2307.15043 [https://perma.cc/38RW-EVDQ] (discussing the creation of adversarial suffixes designed to be embedded in the user input to conflict with and circumvent the model's primary safety instruction).

[118] *See, e.g.*, *id.;* Sahar Abdelnabi et al., *Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*, 16 PROC. ACM WORKSHOP ON A.I. & SEC. 79, 79-80 (Nov. 30, 2023); Ahmed Salem et al., *Maatphor: Automated Variant Analysis for Prompt Injection Attacks*, ARXIV 1 (Dec. 12, 2023), https://arxiv.org/pdf/2312.11513 [https://perma.cc/H9KR-HWGK].

[119] Zou et al., *supra* note 117, at 1-3.

External filtering, sometimes called a moderation layer, is a suppression technique that treats the large model like a black box—meaning its internal parameters and decision processes are not directly observable or modifiable—and instead screens its inputs or outputs with a separate classifier.[120] OpenAI's Moderation Application Programming Interface (API) illustrates this design: it receives a candidate response, assigns a probability of violating categories such as "hate" or "sexual," and blocks or edits the text if the score crosses a threshold.[121] Researchers have proposed introducing specialized watchdogs (code that is responsible for making sure that certain particular parameters or standards are obeyed) that recognize personal identifiers or toxic speech with higher recall in order to reliably redact or replace the offending span before it reaches the end user.[122] The strength of external filters lies in their modularity.[123] Filters can be improved or replaced without touching the original model and can be customized for different legal domains by inserting industry- and other data-specific filters (for example, a medical named-entity recognizer).[124] The weakness of external filtering is in its "whack-a-mole" character: the filters may let novel disclosure patterns slip through or they may over-block legitimate content.[125] Importantly, the filters do not eliminate the underlying data.[126] They merely conceal it at the interface level, leaving the model's internal representations unchanged.[127] Shumailov and co-authors showed that if an attacker gains direct access to the model weights, every data the filter hides can still be extracted, confirming that suppression controls do not qualify as unlearning in the strict legal sense.[128] In privacy law terms, unlearning in the strict sense requires that the data and its functional influence be erased, not just hidden so that the model behaves as though the data were never processed at all.[129]

[120] Sarah Ball et al., *On the Impossibility of Separating Intelligence from Judgment: The Computational Intractability of Filtering for AI Alignment*, ARXIV 1 (July 9, 2025), https://arxiv.org/pdf/2507.07341 [https://perma.cc/J2FE-SZBD].

[121] *See* OpenAI, *supra* note 74.

[122] Matthew Pisano et al., *Bergeron: Combating Adversarial Attacks Through a Conscience-Based Alignment Framework*, ARXIV 1, 2 (Aug. 18, 2024), https://arxiv.org/pdf/2312.00029 [https://perma.cc/CK62-54EZ].

[123] Yongzhe Huang et al., *How Good Are the LLM Guardrails on the Market? A Comparative Study on the Effectiveness of LLM Content Filtering Across Major GenAI Platforms*, UNIT 42 (June 2, 2025), https://unit42.paloaltonetworks.com/comparing-llm-guardrails-across-genai-platforms/ [https://perma.cc/8BYB-S9XX].

[124] *Id.*

[125] Jing Cui et al., *Recent Advances in Attack and Defense Approaches of Large Language Models*, ARXIV 1, 2 (Dec. 2, 2024), https://arxiv.org/html/2409.03274v2 [https://perma.cc/47QK-DS7Q].

[126] *See id.* at 13 ("On the other hand, decoding-based approaches do not address the core problem of harmful output from LLMs, since the harmful knowledge remains within the model").

[127] *Id.*

[128] Shumailov et al., *supra* note 19, at 2.

[129] *See, e.g.*, Bjørn Aslak Juliussen et al., *Algorithms that Forget: Machine Unlearning and the Right to Erasure*, 51 COMPUT. LAW & SEC. REV. at 1, 7-9 (Nov. 2023) (explaining that behavioral

D. Effectiveness vs. Practical Constraints

Each variation of unlearning methods has strengths, weaknesses, and trade-offs in its privacy guarantees, computational cost, scalability, and robustness.

1. Strict Compliance and Effectiveness at Removing Personal Data

Of the three types of unlearning techniques discussed above, structural unlearning provides the most convincing evidence that personal data have truly been erased in compliance with a deletion request. When an exact method succeeds, statistical tests cannot distinguish the unlearned model (from which the data was removed) from one trained fresh without the record (in which the data was never inputted). This property was formally proved for certified-removal algorithms, such as that of Guo et al.'s closed-form update for logistic regression.[130] This level of fidelity matters in high-stakes settings; for example, when a hospital must guarantee that a facial-recognition system no longer recognizes a former patient or when firms must prove more definitively to regulators that they have satisfied data subjects' deletion requests.

Approximate retraining methodologies, on the other hand, simply narrow rather than close the gap. For instance, Fisher-scrubbing eliminates the direct *memorization* of select data so that probing the model's weights reveals no *obvious* trace of the "forgotten" data.[131] Yet influence that has diffused into hidden features can persist and the data can still be identified after the fact. Several studies show that membership inference attacks, which tests whether a specific record was in the training set by querying the model and looking for over-confidence, can guess whether a particular "forgotten" point was in the initial training set with better-than-chance accuracy[132]—a situation that may increase privacy compliance risks. Thus, audits should be common if this method is deployed as means of legal compliance or to satisfy deletion requests. Evidence from the Data Provenance Initiative, which links fifty-four million web documents to which passages have been memorized by state-of-the-art language models,[133] underscores why such audits matter. The study showed extensive verbatim copying of copyrighted and personal text, exhibiting that providers' claims of "forgetting" are hard to verify without provenance-level tracing.[134]

or output-level suppression does not constitute data erasure under Article 17 GDPR and that true compliance requires the model to eliminate both the data and its influence); *see also* Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317, ¶ 93 (May 13, 2014) (holding that erasure requires making personal data inaccessible and preventing further processing); GDPR, art. 17.

[130] *See* Guo et al., *supra* note 30, at 3833-34.

[131] *See* Golatkar et al., *supra* note 81, at 9301-02.

[132] *See* Thudi et al., *supra* note 5, at 4009-4020.

[133] Shayne Longpre et al., *A Large-Scale Corpus for Benchmarking Memorization in Language Models*, 6 NATURE MACH. INTEL. 975, 976-77 (Aug. 2024).

[134] *See id.* at 975-76, 980-83.

Finally, output suppression, by design, never touches the internal weights nor the inputted data.[135] As such, it offers no protection if an attacker colloquially looks "under the hood" of the model or discovers a prompt that slips past filters.[136] As already mentioned, relying on output suppression methods alone to demonstrate compliance with deletion requests is precarious.[137] Output suppression is thus likely best used in conjunction with other and more legally-reliable unlearning techniques.

Given the above considerations, for regulators who interpret the data subject's erasure rights literally, only structural or other certified methods provide a defensible assurance that the data is gone.

2. Computational Cost and Scalability

Exact unlearning asks the provider to repeat a sizeable fraction of the original training job, so its bills scale with model size and with the volume of deletion requests. Privacy professionals will need to weigh the compute and financial costs of structural unlearning against the alternative approximate unlearning and output suppression models to find what best fits their organizations' needs. To reiterate the issue of scalability, remember how in Bourtoule et al.'s ImageNet experiments, a single SISA shard retrain still occupied eight NVIDIA V100 GPUs for several hours; a large vision model might require dozens of shards, each retrained separately before re-aggregating their outputs.[138] Even where Yan et al. report that ARCANE's one-class experts cut runtime dramatically compared with full retraining, purging a mid-size ResNet still meant training twenty-plus sub-networks and transferring gigabytes of checkpoint data back into a central aggregator.[139] And, these compute costs compound in production because the system must also

[135] *See* Yi Dong et al., *Building Guardrails for Large Language Models*, 235 PROC. MACH. LEARNING RSCH. 11375, 11375-76 (2024) (defining guardrails as post-hoc systems that filter the inputs and outputs of trained LLMs, not their internal weights or data).

[136] *See* William Hackett et al., *Bypassing Prompt Injection and Jailbreak Detection in LLM Guardrails*, 1 PROC. WORKSHOP ON LLM SEC. 101, 101--02, 108 (Aug. 1 2025), https://aclanthology.org/2025.llmsec-1.8.pdf [https://perma.cc/VY77-FTKD] (showing that multiple commercial guardrail systems, including Azure Prompt Shield and Meta Prompt Guard, can be easily bypassed by adversarial prompt injection and character-insertion attacks, demonstrating that filtering-based suppression offers no protection once an attacker discovers a prompt that slips past the filters).

[137] *See* Biwei Yan et al., *On Protecting the Data Privacy of Large Language Models*, 5 HIGH-CONFIDENCE COMPUTING J. at 1, 5 (June 2025) (explaining that output filtering merely masks data and cannot constitute true deletion under data-protection law); *see also* Juliussen et al., *supra* note 129, at 7-9 (explaining that behavioral or output-level suppression does not constitute data erasure under Article 17 GDPR and that true compliance requires the model to eliminate both the data and its influence).

[138] *See* Bourtoule et al., *supra* note 7, at 141-142 (reporting that SISA retraining on ImageNet required multiple shard-level retrains, each using eight NVIDIA V100 GPUs for several hours, illustrating the scalability challenges of machine unlearning).

[139] *See* H. Yan et al., *supra* note 7, at 4007-11 (reporting that ARCANE's one-class experts substantially reduce retraining time compared with full retraining, yet unlearning still requires training twenty-plus sub-networks and transferring checkpoint data to a central aggregator).

track which record touched which shard, store historical checkpoints for audit purposes, and schedule GPU time around other product priorities.[140] Consider also how Neel et al. show that certified removal for convex models keeps runtime bounded even after thousands of deletion requests, but they acknowledge that an equivalent procedure for a 70-billion-parameter transformer would need petaflop-days of compute, a workload measured in tens of thousands of dollars on current cloud pricing.[141] These are indeed serious tradeoffs between exact unlearning and the financial and compute realities that privacy professionals must consider.[142]

Approximate retraining, on the other hand, slashes those numbers: recall how the Harry-Potter experiment erased copyrighted text from a seven-billion-parameter model in roughly one GPU-hour, a reduction of five orders of magnitude compared with the 184,000 GPU-hours spent on the original pre-training.[143] Output suppression is cheaper still because a new system prompt or an updated moderation classifier can be pushed to every replica within minutes.[144] Unlike retraining or structural unlearning, this process runs entirely on standard processors (CPU) rather than high-cost graphics processors (GPUs) during model use ("inference"), meaning it adds only minimal computational expense and delay.[145] Additional GPU resources are needed only if the filter itself relies on a separate neural network.[146] This efficiency makes output suppression attractive to companies and regulators alike, since compliance updates can be deployed rapidly without the energy or hardware demands of full retraining.

The hierarchy is therefore stark. At the top are structural removal methods, such as SISA retraining and ARCANE-style modular rewrites, which deliver the highest level of certainty that data and its influence are erased, but they are computationally

[140] *Id.* at 4007-11 (discussing ARCANE's requirement to track sub-datasets, save and reuse training states, and aggregate sub-model outputs, illustrating the systemic overhead that compounds compute costs in production).

[141] *See* Neel et al., *supra* note 57, at 2, 8 (explaining that certified removal methods for convex models keep runtime bounded even after thousands of deletions, but scaling to large non-convex models would require massive compute resources, potentially costing tens of thousands of dollars).

[142] *See* H. Yan et al., *supra* note 7, at 4006 (explaining that exact unlearning introduces substantial computational and time overhead, underscoring the trade-offs that privacy practitioners must balance between rigorous deletion guarantees and feasible resource use).

[143] *See* Eldan & Russinovich, *supra* note 71, at 2-3 (demonstrating that their "approximate unlearning" method erased Harry Potter–related content from the 7B-parameter LLaMA2 model in roughly one GPU-hour, compared with 184,000 GPU-hours for pretraining).

[144] Ruichen Qiu et al., *A Survey on Unlearning in Large Language Models*, ARXIV 1, 11-12 (Oct. 29 2025), https://www.arxiv.org/pdf/2510.25117 [https://perma.cc/5J6H-D67Y] ("[Inference time unlearning] significantly reduces the computational requirements and enables broader applicability across different scenarios."); Sungmin Cha et al., *Towards Robust and Cost-Efficient Knowledge Unlearning for Large Language Models*, ARXIV 1 (Apr. 24, 2025), https://arxiv.org/pdf/2408.06621 [https://perma.cc/CS2Q-MCJR].

[145] Cha et al., *supra* note 144.

[146] *Id.*

expensive and scale poorly to large foundation models.[147] In the middle sit approximate retraining methods, including Fisher scrubbing, targeted fine-tuning, and amnesiac machine learning, which can reach frontier-scale systems at a tolerable cost but leave residual traces of the deleted data.[148] At the bottom are suppression techniques, like RLHF and external moderation filters, that scale effortlessly across replicas and updates yet achieve only behavioral concealment, not true deletion.[149] In short, the trade-off runs from certainty to scalability: the more complete the forgetting, the higher the computational and financial price.

3. Adversarial Robustness

From a security perspective, structural unlearning again leads the pack, provided the protocol includes randomness or differential-privacy noise to mask telltale weight changes.[150] When those safeguards are present, an adversary who inspects the weights cannot tell whether any given record was ever included, which blocks extraction attacks except with negligible probability.[151] Deterministic schemes such as the original SISA are less robust because an attacker can compare (or industry terms, "diff") two model snapshots to detect who was deleted.[152]

Approximate retraining has no formal guarantee, so shadow-model audits often succeed in spotting partial deletions, and careless fine-tunes can even create new privacy leaks by overfitting to the remaining data.[153] Suppression methods fare worst against a motivated adversary. Even state-of-the-art systems such as OpenAI's GPT-4 have proven vulnerable: researchers have reversed its RLHF safety tuning with only a few hundred example pairs, fully restoring disallowed

[147] *See* H. Yan et al., *supra* note 7, at 4006-08 (explaining that exact unlearning methods such as ARCANE and SISA retraining offer the strongest deletion guarantees but impose substantial computational and time overhead, making them difficult to scale to large foundation models).

[148] *See generally* Golatkar et al., *supra* note 81, at 9303-04 (describing Fisher-information-based selective forgetting as an efficient post-training method that reduces retraining cost but leaves measurable residual information in model weights, illustrating the limits of approximate unlearning).

[149] *See* Ouyang et al., *supra* note 74, at 27733 (discussing characteristics of RLFH and its application to aligning language models on distribution tasks).

[150] *See generally* Guo et al., *supra* note 30 (masking residual via random loss perturbation to achieve certified removal and indistinguishability); Dwork & Roth, *supra* note 103 (formalizing DP noise mechanisms that render per-record inclusion undetectable).

[151] *See* Guo et al., *supra* note 30, at 3837-39 (demonstrating that certified removal with loss-perturbation or differential-privacy noise makes the post-unlearning model statistically indistinguishable from one trained without the deleted record, preventing membership-inference or extraction attacks except with negligible probability).

[152] *See* Thudi et al., *supra* note 5, at 4016-17 (showing that deterministic unlearning methods like SISA can be "forged" because an adversary may compare or "diff" two model checkpoints and reconstruct which records were deleted, revealing that such schemes lack robustness against audit or adversarial inspection).

[153] Yuechun Gu et al., *Auditing Approximate Machine Unlearning for Differentially Private Models*, ARXIV 1, 2 (Aug. 26, 2025), https://arxiv.org/pdf/2508.18671v1 [https://perma.cc/ZJ59-UNLF].

outputs.[154] Prompt-injection studies catalog entire libraries of "jailbreak" phrases that bypass static filters, and attackers now automate the search for new exploits using other language models.[155] In short, output suppression assumes a cooperative user, approximate retraining assumes an honest but resource-constrained adversary, and structural removal is the only line of defense that remains credible if the model itself leaks.

Real-world systems therefore mix and match. A provider might run a certified unlearning protocol when a court or regulator orders deletion, then layer prompt rules and an external moderation API on top for everyday safety and speed. The research record shows that no single method solves all privacy problems, but continued progress suggests that the toolbox is widening. Aligning technique to legal requirements is the central design choice: if the goal is strict compliance, structural or at least certified unlearning is mandatory; if the goal is rapid iteration with acceptable risk, approximate and suppressive methods can fill the gap while more rigorous processes run in the background.

A final consideration is the life of the data after it leaves the original model. If that model has been distilled into a smaller clone or incorporated into downstream products, deleting the source weights does not retract the derivative systems.[156] Therefore, any practical unlearning policy must inventory and, if necessary, extinguish all downstream models or issue retuning patches so those derivatives no longer embed the contested information. Otherwise, perfect unlearning at the source leaves a compliance gap the size of the product ecosystem.

III. SYNTHESIZING THE REQUIREMENTS OF PRIVACY AND DATA PROTECTION LAW

Part III examines how these technical mechanisms align with, yet often challenge, requirements of existing privacy laws. Here, the Article critically examines the conceptual and practical gaps between machine unlearning's technical methodologies and privacy law's normative goals. This article considers these gaps in relation to common core levers of privacy law, including lawful collection, purpose limitation, data minimization, the right to correction (rectification), and the rights to object to or withdraw consent for processing.[157] It argues that unlearning may satisfy some legal requirements in letter but not in spirit, particularly when latent (learned) knowledge or model outputs still compromise individual privacy. It also argues how machine unlearning could complement or strengthen emerging privacy remedies advocated by legal scholars and consumer

[154] *See* Hackett et al., *supra* note 136, at 101-03, 105-07.

[155] *See* Zou et al., *supra* note 117, at 10-12.

[156] *See* Cooper et al., *supra* note 1, at 8-13 (explaining that unlearning cannot propagate to fine-tuned, distilled, or downstream models, so deleting the source weights does not retract derivative systems built from them).

[157] *See* GDPR, arts. 5-7, 16.

protection enforcers, such as model deletion and algorithmic disgorgement[158], by providing a more granular, technically feasible mechanism for removing the influence of unlawfully obtained or inaccurate data without requiring the destruction of an entire model.

A. Overview of Privacy Law's Normative Goals

Modern privacy laws rest on a set of core principles governing the collection, use, and management of personal data.[159] At the point of collection, data must be gathered lawfully and fairly, typically referred to as a "legitimate" legal basis for collection[160] (e.g., the data subject's informed consent, contract necessity, or other legally permitted ground).[161] There must also be transparency about how the data will be used and processed.[162] Once collected, personal information should be used only for these specific and explicitly stated purposes, and not repurposed in any ways incompatible with those original objectives ("purpose limitation").[163] Hand-in-hand with purpose specification and limitation is data minimization: the

[158] *See* Tiffany C. Li, *Algorithmic Destruction*, 75 SMU L. Rev. 479 (2022); Jevan Hutson & Ben Winters, *America's Next 'Stop Model!' Model Deletion*, 8 Geo. L. Tech. Rev. 124 (2024); *see also* Daniel Wilf-Townsend, *The Deletion Remedy*, 103 N.C. L. Rev. 1809 (2025); Achille et al., *supra* note 21; Fed. Trade Comm'n, Statement of Commissioner Rohit Chopra In the Matter of Everalbum and Paravision Commission, at 1, 2 (Jan. 8, 2021), https://www.ftc.gov/system/files/documents/public_statements/1585858/updated_final_chopra_statement_on_everalbum_for_circulation.pdf [https://perma.cc/67ER-EQKK] ("It will be critical for . . . regulators . . . to pursue additional enforcement actions [beyond algorithmic disgorgement] to hold accountable . . . technology [providers] who make false accuracy claims and engage in unfair, discriminatory conduct.").

[159] Note that the entity or individual which collects, determines use and processing, and makes decisions regarding or otherwise controls the personal data collected is often referred to as the "data controller" or "controller" in alignment with the language of the GDPR. This is reflected herein. *See* GDPR, art. 5.

[160] *See* GDPR, recital 40; GDPR, recital 41.

[161] *See* GDPR, art. 6; Fed. Trade Comm'n, *supra* note 158, at 2 ("Commissioners have voted to enter into scores of settlements that address deceptive practices regarding the collection, use, and sharing of personal data. There does not appear to be any meaningful dispute that these practices are illegal").

[162] *See* Shumailov et al., supra note 19; GDPR, art. 5(1)(a); GDPR, recital 58; Rebecca Kelly Slaughter, *Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission*, 23 Yale J.L. & Tech. (Special Issue 1), Aug. 2021, at 1, 40, https://yjolt.org/sites/default/files/23_yale_j.l._tech._special_issue_1.pdf [https://perma.cc/8CE2-RZA3] ("The [FTC] can also use its deception authority . . . where marketers of products or services represent that they can use machine-learning technology in unsubstantiated ways . . . ."); Joshua A. Goland, *Algorithmic Disgorgement Destruction of Artificial Intelligence Models as the FTC's Newest Enforcement Tool for Bad Data*, 29 Rich. J. L. & Tech. 1, 9, https://jolt.richmond.edu/files/2023/03/Goland-Final.pdf [https://perma.cc/2PFC-MTMJ] ("In most states, companies can use, share, or sell any data they collect . . . without notifying . . . that they're doing so." (quoting Thorin Klosowski, *The State of Consumer Data Privacy Laws in the US (And Why It Matters)*, N.Y. Times: Wirecutter (Sept. 6, 2021), https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/ [https://perma.cc/X8XN-NE7U])).

[163] *See* GDPR, arts. 5(1)(b), 6(3)(2), 6(4); FTC Act § 5(a)(1) ("[D]eceptive acts or practices in or affecting commerce[] are hereby declared unlawful.").

principle that data controllers should collect and retain only the minimum data necessary to achieve the stated purpose rather than stockpiling personal information indefinitely or for a yet undetermined use.[164] Privacy regimes impose storage limitations, requiring that data not be kept longer than needed, and mandate appropriate security safeguards against unauthorized access.[165] Crucially, modern privacy laws empower individuals with rights to control their information and its use. Data subjects can access the data held about them, request its correction, seek its deletion (colloquially, the "right to be forgotten") and object to or withdraw consent for certain processing.[166] One privacy pillar, for example, is data accuracy and quality. This is captured in individuals' right to correction (also called "rectification"): if the personal data is used in the controller's or in algorithmic decision making, data subjects have the right to update or correct inaccuracies in their own identifying information (for instance, an incorrect spelling of or out-of-date name).[167]

These principles, often termed Fair Information Practice Principles ("FIPPs"), are embedded in comprehensive frameworks worldwide (note: even where there is no U.S. national privacy standard, e.g. The Department of Homeland Security's privacy policy framework).[168] For example, the EU's General Data Protection Regulation (GDPR) codifies lawful collection and fairness, purpose limitation, data minimization, accuracy, storage limitation, and integrity and confidentiality in its Article 5 principles, and it grants robust individual rights including access, rectification, erasure, portability, and the right to object or restrict processing.[169] The GDPR further operationalizes these norms by requiring Data Protection Impact Assessments and "privacy by design and default" measures.[170] In the United States, the California Consumer Privacy Act (CCPA) and follow-on state laws have embraced many of these concepts by providing rights to know, delete, and opt out of certain data uses—even if the U.S. generally has historically relied on upfront notice and choice instead of broad purpose limitations.[171] Overall, despite variations in scope and enforcement, global privacy laws share the normative goals

[164] *See* Shumailov et al., *supra* note 19; *see also* GDPR, art. 5(1)(c).

[165] *See* Shumailov et al., *supra* note 19; *see also* GDPR, art. 5(1)(e).

[166] *See* Shumailov et al., *supra* note 19; *see also* GDPR, arts. 7(3), 15-17.

[167] *See, e.g.,* GDPR, arts. 5(1)(d), 16; GDPR, recital 65.

[168] *See, e.g.,* DEP'T OF HOMELAND SEC., THE FAIR INFORMATION PRACTICE PRINCIPLES (Dec. 29, 2008), https://www.dhs.gov/publication/privacy-policy-guidance-memorandum-2008-01-fair-information-practice-principles [https://perma.cc/23KG-H43X]; DEP'T OF HOMELAND SEC., PRIVACY POLICY GUIDANCE MEMORANDUM 2008-02, DHS POLICY REGARDING PRIVACY IMPACT ASSESSMENTS (Dec. 30, 2008), https://www.dhs.gov/sites/default/files/publications/privacy_policyguide_2008-02_0.pdf [https://perma.cc/68VQ-ZAQ7].

[169] *See* GDPR, art. 5.

[170] *See, e.g.*, GDPR, arts. 25, 35. For more on "privacy by design," see WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES (Harvard Uni. Press, 2018).

[171] *See* California Consumer Protection Act of 2018, CAL. CIV. CODE §§ 1798.115, 1798.105, 1798.120.

of giving individuals control over personal data and ensuring organizations handle that data in a limited, fair, and accountable manner.

B. Conceptual & Practical Tensions with Unlearning

Machine unlearning has emerged as a technical approach to align AI models with the above privacy principles.[172] In theory, unlearning allowing a model to "forget" specific personal data that should no longer be used, making it appealing as a privacy-preserving mechanism.[173] In practice, however, unlearning techniques face profound conceptual and practical limitations that complicate their alignment with privacy law's aims.

One major challenge is that removing data from a training dataset does not fully erase its influence on a trained model.[174] Once a model has been trained, the data's imprint remains entangled in the model's parameters and learned patterns. Technical studies confirm that simply excising one person's data after the fact is insufficient to scrub all traces of it from a complex model's knowledge.[175] The model may have abstracted general latent patterns or rules from that data, especially if the information overlaps with other training examples.[176] Accordingly, current unlearning methods can target the observed data (the exact records used in training) but struggle to remove more diffuse latent knowledge that the model inferred from those records.[177] Illustratively, a generative model might be "unlearned" on a specific document containing a person's private facts, yet the model could still reproduce parts of those personal facts or answer questions about the data subject by relying on residual patterns, synonyms, or related context learned elsewhere. As one group of researchers put it, there is "no clear way to remove" higher-level concepts that a model has generalized from the data; a small removal cannot reliably make the model unknow a broader idea.[178] In sum, a model that has "consumed" personal data cannot simply de-digest as if it were never there; that data, once digested into the model's weights, is akin to an irretrievable ingredient in a recipe.[179] This reality creates tension with privacy regimes' clear expectations

[172] *See, e.g.*, Li, *supra* note 158; Hutson & Winters, *supra* note 158, at 129 (2024); Wilf-Townsend, *The Deletion Remedy*, *supra* note 158, at 1854; Achille et al., *supra* note 21, at 2.

[173] Slaughter, *supra* note 162, at 39 (discussing algorithmic disgorgement).

[174] *See* Rishav Chourasia & Neil Shah, *Forget Unlearning: Towards True Data-Deletion in Machine Learning*, ARXIV 1 (Feb. 14, 2024), https://arxiv.org/pdf/2210.08911 [https://perma.cc/B27N-CM2Z].

[175] Shumailov et al., *supra* note 19, at 1; M. Chen et al., *supra* note 3, at 896.

[176] Kostantinos Papadamou et al., *Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children*, 14 PROC. INT'L AAAI CONF. ON WEB & SOC. MEDIA 522, 523 (2020), https://ojs.aaai.org/index.php/ICWSM/article/view/7320/7174 [https://perma.cc/563T-ZGP8].

[177] Cooper et al., *supra* note 1, at 2.

[178] *Id.*

[179] *See* Ken Ziyu Liu, *Machine Unlearning in 2024*, STAN. A.I. LAB'Y BLOG (May 2024), https://ai.stanford.edu/~kzliu/blog/unlearning) [https://perma.cc/D8EB-AYDE].

that a person's data can be erased, including eliminating its influence, upon request.[180]

Moreover, machine unlearning does not guarantee that the model's outputs will never again reflect the removed information. Even after purging a data point from the training set and adjusting the model to remove or diminish its influence, the system might still generate content that reveals sensitive details by coincidence or through other knowledge. Or, it might produce information that so closely resembles the forgotten data that it risks identifying the data subject.[181] Generative AI models are probabilistic and combinatorial; they can "transcend the information exactly contained in their training data" by synthesizing pieces of knowledge into new outputs.[182] Thus, as a recent study noted, even if all instances of, for example, a copyrighted image or a person's documents have been successfully removed from a model's training corpus, it may not be "impossible for the model to generate outputs that resemble" that image or text later on.[183] A clever user prompt can sometimes reintroduce the ostensibly unlearned information and coax the model to produce it.[184] In the privacy context, this means that a model might still divulge a person's data, or a close approximation of it, even after an unlearning procedure purports to forget that data. One analogy is that a human "forgetting" a fact may still recall it later when prompted differently; the knowledge is not truly gone, just not immediately accessible. This undermines the spirit of data privacy erasure rights: if a model can regenerate someone's personal information despite deletion, has it really been erased from the processing ecosystem?[185]

In addition to these completeness problems, unlearning techniques encounter practical feasibility issues that can dilute their privacy value. Fully retraining a large AI model from scratch to forget a handful of data subjects can be computationally prohibitive. Modern models have hundreds of billions of parameters, and retraining them even once costs enormous time and resources.[186] Consequently, many machine unlearning methods are instead approximate: they try to estimate and

[180] *See, e.g.*, Wilf-Townsend, *supra* note 158, at 1854; FED. TRADE COMM'N, DISSENTING STATEMENT OF COMMISSIONER ROHIT CHOPRA *IN RE* GOOGLE LLC & YOUTUBE, LLC, at 1, 2 (Sept. 4, 2019), https://www.ftc.gov/system/files/documents/public_statements/1542957/chopra_google_youtube_dissent.pdf [https://perma.cc/BW8V-GSPS] ("I believe [that in allowing Google to keep its algorithms] the Commission is contravening clear Congressional intent to substantially penalize violators of children's privacy beyond their ill-gotten gains."); GDPR, art. 17; GDPR, recital 66 ("[T]he right to erasure should also be extended in such a way . . . to erase any links to, or copies or replications of those personal data.").

[181] Cooper et al., *supra* note 1, at 13.

[182] *Id.*

[183] *Id.* at 2.

[184] *Id.* at 8.

[185] *See, e.g.*, GDPR, art. 17; GDPR, recital 66 ("[T]he right to erasure should also be extended in such a way . . . to erase any links to, or copies or replications of those personal data.").

[186] Achille et al., *supra* note 21, at 2.

subtract a data point's *influence* on the model without rebuilding the model entirely as a means of saving cost.[187]

But with such approximations come uncertainties about whether the data's influence is truly gone. Indeed, scholars note that there is still no agreed upon metric or test to confirm that a model has completely forgotten a given datapoint.[188] An "unlearned" model might behave almost identically to a model that has been retrained from scratch without the data included—the gold standard for forgetting. Yet subtle differences could still persist between an unlearned and re-trained model; these differences can furthermore create new privacy risks.

Beyond structural retraining, several approximate unlearning techniques aim to reduce residual influence without rebuilding a model end-to-end.[189] One family adds calibrated noise to parameters most sensitive to the targeted record (often guided by the Fisher information), thereby blurring the record's statistical footprint while preserving overall utility. Although attractive where raw training data are unavailable (e.g., due to retention limits), these probabilistic guarantees mean faint traces may persist under intensive probing—useful in practice, but not a perfect substitute for erasure.[190] A pragmatic variant is targeted fine-tuning on "anti-examples" that down-weight specific facts or classes, limiting collateral effects on unrelated tasks—often critical for proportional compliance when only narrow content must be forgotten.[191] Related amnesiac approaches prune networks to sparse "cores" and then run brief, noise-infused updates centered on the forget set, yielding material speed-ups over naïve retraining with modest accuracy trade-offs.[192] These methods illustrate a spectrum: greater efficiency and responsiveness, but weaker formal assurances against leakage. Finally, differential privacy (DP) and unlearning serve distinct roles.[193] DP constrains any single record's marginal influence ex ante; unlearning removes influence ex post.[194] Each alone is incomplete for strict deletion rights, but used together they can reduce membership-inference risk (DP) while blunting memorized content (lightweight unlearning).[195]

[187] *See* Jiawei Liu et al., *Efficient Machine Unlearning via Influence Approximation*. ARXIV 1, 5 (July 31, 2025), https://www.arxiv.org/pdf/2507.23257 [https://perma.cc/4YLZ-TX2X].

[188] *See* Tang et al., *supra* note 1, at 2649-53.

[189] *See, e.g.*, Neel et al., *supra* note 57 (creating a data deletion system by leveraging techniques from convex optimization and reservoir sampling); Golatkar et al., *supra* note 81, at 9302 (proposing a method of data deletion through shifting the weights of a model's probing function); Eldan & Russinovich, *supra* note 71, at 2 (introducing a novel technique for unlearning a subset of training data without retraining from scratch); Graves et al., *supra* note 94 (presenting unlearning and amnesiac unlearning as alternatives to training new models from scratch).

[190] *See* Neel et al., *supra* note 57; Golatkar et al., *supra* note 81.

[191] *See* Eldan & Russinovich, *supra* note 71, at 8-10.

[192] *See* Graves et al., *supra* note 94, at 11518, 11522.

[193] *See* Ginart et al., *supra* note at 104, at 3519; *see also* Dwork & Roth, *supra* note 103, at 214-15.

[194] *See* Ginart et al., *supra* note at 104, at 3519.

[195] *See* Cooper et al., *supra* note 1, at 3 (arguing that differential privacy and machine unlearning are complementary but individually insufficient to satisfy strict data-erasure rights, since each addresses distinct aspects of privacy risk).

For lawyers and regulators, DP should be treated as complementary to—not interchangeable with—unlearning, with combined deployment improving practical privacy outcomes while acknowledging residual uncertainty.

Researchers have demonstrated membership inference attacks that exploit the before-and-after difference in a model's predictions to detect that a particular person's data was in the original training set but removed in the newer version.[196] In fact, such an attack can, in some cases, more confidently identify that a person was included (and then removed) than an attack on the original model alone.[197] Analogously, it is like identifying someone who has not donated DNA by their family members who have. In other words, the act of unlearning can leave a telltale "shadow" of the data, inadvertently flagging that individual's data as existing in the training set.[198] Clearly, this is a perversely counterproductive outcome for privacy. These findings reinforce that machine unlearning, as currently conceived, often cannot fully align with the absolute notion of erasure envisioned by privacy laws.[199] If there remain lingering data imprints and leakage avenues, then a person's data is not entirely forgotten. This raises difficult questions about what it means to comply with obligations like the GDPR's "right to be forgotten" in an AI context.[200]

C. Applying Machine Unlearning to Core Privacy Levers

Despite its limitations, machine unlearning is frequently discussed as a way to bolster compliance with specific provisions of data privacy laws. It is useful to examine how unlearning might apply to several core legal requirements and whether it truly fulfills them or merely offers a partial workaround.[201]

1. Lawful Collection

Privacy laws require that personal data be collected and used on a lawful basis and prohibit using data in ways that violate those conditions.[202] Suppose a dataset was gathered without a valid legal basis (for example, scraped from a website in violation of terms or without required consent, as was the case, for example, in

[196] *See* M. Chen et al., *supra* note 3, at 896.

[197] *See id.* at 906.

[198] *See, e.g.*, *id.* at 896; Shumailov et al., *supra* note 19, at 2.

[199] *See* Cooper et al., *supra* note 1, at 2.

[200] *See, e.g.*, Cooper et al., *supra* note 1, at 4-6; GDPR, art. 17; GDPR, recital 66.

[201] *See* Slaughter, *supra* note 162, at 58 ("The FTC's tools [including algorithmic disgorgeme nt] are still capable of addressing some of the problems posed by algorithms and AI . . . [b]ut confronting the challenges of algorithmic decision-making will also require new tools and strategies.").

[202] *See, e.g.*, GDPR, art. 6; OFFICE OF THE PRIV. COMM'R OF CANADA, PIPEDA FAIR INFORMATION PRINCIPLES (2025), https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/ [https://web.archive.org/web/20250908054302/https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/].

Canada's *Joint Investigation of Clearview AI*).[203] Unlearning theoretically remediates illegally obtained data from the model and stops further unlawful use of the unlawfully collected information. Indeed, regulators and courts might demand disgorgement of ill-gotten data from AI models as a remedy, akin to an order to "forget" data that was unlawfully collected.[204]

The critical question is whether unlearning, as a remedy, actually cures the original harm of unlawful collection. On one hand, unlearning can be an attempt to "rewind" the model to a state where the improper data had never been included, thereby preventing the violator from continuing to benefit from the "fruits" of an illegal data grab.[205] On the other hand, if traces of that data (or its influence) remain in the model even with the data 'unlearned,' the model owner may still indirectly profit from the illicit collection, undermining the deterrent purpose of data protection rules.[206] Accordingly, providers should adopt a layered deletion stack: "heavier" exact or certified unlearning for legally significant takedowns (creating a defensible audit trail), paired with "lighter" front-end suppression and prompt filters for day-to-day safety and latency needs, so the system both withstands audits and delivers responsive user-facing "forgetting."[207]

Scholars have warned of this exact dynamic. If an AI system has already learned from improperly obtained data, simply deleting the source data (or even naively claiming to unlearn it) may have "no impact on an already trained model," leaving behind an "algorithmic shadow:" a persistent imprint of the misused data in the

[203] *See* OFFICE OF THE PRIV. COMM'R OF CANADA, PIPEDA FINDINGS #2021-001: JOINT INVESTIGATION OF CLEARVIEW AI, INC. (2021), https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2021/pipeda-2021-001/ [https://perma.cc/XD7S-MRAL] (last visited Nov. 24, 2025).

[204] *See, e.g.*, Cooper et al., *supra* note 1, at 11; Decision and Order, Everalbum, Inc., FTC Docket No. C-4743 (2021), https://www.ftc.gov/system/files/documents/cases/1923172_-_everalbum_decision_final.pdf [https://perma.cc/4HH3-GFVD]; Slaughter, *supra* note at 162, at 39.

[205] *See* Slaughter, *supra* note 162, at 39; FED. TRADE COMM'N, FTC REPORT TO CONGRESS ON PRIVACY AND SECURITY at 1, 4 (Sept. 13, 2021), https://www.ftc.gov/system/files/documents/reports/ftc-report-congress-privacy-security/report_to_congress_on_privacy_and_data_security_2021.pdf [https://perma.cc/FRJ4-7EFT]; FED. TRADE COMM'N, DISSENTING STATEMENT OF COMMISSIONER ROHIT CHOPRA *IN RE* GOOGLE LLC & YOUTUBE, LLC, *supra* note 180, at 6 ("[The settlement] does not consider . . . any ill-gotten gains from data being used by Google's other properties, the increased value of its predictive algorithm trained by ill-gotten data (which will not be reversed), and other considerable benefits from lawbreaking.").

[206] *See* Shumailov et al., *supra* note 19, at 2-3; FED. TRADE COMM'N, DISSENTING STATEMENT OF COMMISSIONER ROHIT CHOPRA *IN RE* GOOGLE LLC & YOUTUBE, LLC, *supra* note 180, at 4-8.

[207] *See, e.g.*, Guo et al., *supra* note 30, at 3832 (defining certified removal as "a very strong theoretical guarantee that a model from which data is removed cannot be distinguished from a model that never observed the data to begin with"); Thudi et al., *supra* note 5, at 4014 (noting that certified removal methods "provide rigorous guarantees at the model level" and concluding that "an entity's only possible auditable claim to unlearning is that they used a particular algorithm designed to allow for external scrutiny during an audit"); Bourtoule et al., *supra* note 7, at 141.

model's parameters.[208] In such a scenario, the developer retains an unfair advantage via the model's enhanced capabilities or accuracy even after nominally purging the offending data.[209] This creates a lack of incentive for companies to avoid unlawful data collection in the first place: the benefit becomes "baked into the algorithm," so a company might still competitively flourish even when or if later forced to remove or stop using the data.[210]

In summary, unlearning in this context operates as a technical workaround to remediate past legal violations: it can support the goal of lawful collection by preventing continued use of the specific illicit data, but it does not erase the fact that data was illegally collected in the first place, nor can it always ensure the model is free of the taint of that data.[211] Thus, while unlearning can be part of a compliance strategy and may be ordered by regulators as a remedy, it functions more as a mitigation measure than a guarantee that the law's demand for ex ante lawful collection of data has been fully honored.

2. Purpose Limitation & Data Minimization

The principles of purpose limitation and data minimization require that personal data be collected and used only in line with specific, legitimate purposes and that only the data actually needed for those purposes is collected, used, and retained. Machine learning development often strains these principles, especially with regard to large-scale models. AI companies tend to vacuum up enormous datasets (many of them simply because they are available, and others because they are accessible because the organization collected them for other purposes) and then repurpose this data to train models for open-ended tasks. This "collect everything just in case" approach is fundamentally at odds with privacy principles that reject the use of data for new, incompatible purposes or the collection of more data than a given purpose necessitates. The GDPR, for instance, requires a fresh legal basis or a showing of compatibility with prior bases to use personal data in training a general AI model, assuming that goes beyond the original purpose for which the data was gathered.[212]

Machine unlearning provides a possible way to reconcile AI practices with these purpose limitation principles. If data used in training turns out to be beyond the scope of the allowed purpose, or not actually necessary for the purpose, the model developer could unlearn it from the model after the fact. From a developer and business perspective, this is also often more computationally feasible and cost-

[208] Li, *supra* note 158, at 490, 498; *see also* Daniel J. Solove & Woodrow Hartzog, *The Great Scrape: The Clash Between Scraping and Privacy*, 113 CALIF. L. REV. 1521 (2025) (discussing the ramifications of scraping); Shumailov et al., *supra* note 19, at 2.

[209] *See* Shumailov et al., *supra note* 19, at 2-3; FED. TRADE COMM'N, DISSENTING STATEMENT OF COMMISSIONER ROHIT CHOPRA *IN RE* GOOGLE LLC & YOUTUBE, LLC, *supra* note 180, at 4-8.

[210] *See* Shumailov et al., *supra note* 19, at 2-3; FED. TRADE COMM'N, DISSENTING STATEMENT OF COMMISSIONER ROHIT CHOPRA *IN RE* GOOGLE LLC & YOUTUBE, LLC, *supra* note 180, at 4-8.

[211] *See* Shumailov et al., *supra note* 93, at 5.

[212] GDPR, art. 5(1)(c) ("limited to what is necessary") & (e) ("for no longer than is necessary for the purposes").

effective than an absolute retraining of the model, which may encourage taking privacy-preserving actions before remedial requirements demand so. Consider if a language model was trained on users' email data collected for the purpose of providing an email service (and not for training a separate-use AI model). The purpose limitation principle might be violated in this case. Thus, the service provider might respond by unlearning those emails from the model once that incompatibility is recognized. Similarly, complying with data minimization principles might call for culling other extraneous personal data from the training set and retroactively minimizing privacy-risking data exposure by unlearning any other data that was not truly needed.

The question is whether such ex-post unlearning sufficiently meets the legal standard. The essence of large AI models is purpose-agnostic or omni-purposeful by design, meaning they are intentionally created to be capable of many tasks and uses. This directly conflicts with the idea of collecting minimal data for a singular, limited purpose. Privacy laws envision purpose limitation and minimization as *proactive* constraints, that is, instructive on how one designs a data processing activity. By contrast, unlearning is *reactive* and partial; it occurs after the model has already ingested the data (and, notably, after the processor has likely derived some benefit from it). In most cases, the "horse has left the barn": the model has generalized from the data in a way that cannot easily—or possibly—be re-contained. Studies of generative AI have observed that it is "nearly impossible to perform any meaningful purpose limitation (or data minimization)" once data has trained into a broad AI model given that the model's utility comes from mixing and generalizing data in unpredictable ways. Unlearning a dataset that was used out-of-purpose does not necessarily limit what the model can do with the other input data; the model might still be capable of the same broad range of functions, perhaps just slightly less or differently so. Likewise, removing some data points as a data minimization step only marginally reduces the vast corpus the model holds; the overall practice of training on maximal data remains.

To counter this problem, enforcement authorities have begun requiring algorithmic disgorgement, that is, algorithmic destruction, in certain cases by ordering the deletion of not only the data itself but also any models or algorithms derived from it.[213] For example, in the FTC's *Everalbum* case, a company that built facial recognition models on unlawfully retained user photos was required to delete "any models or algorithms" developed with that data.[214] This kind of remedy goes beyond technical unlearning as means of addressing any ill-gotten gains that must be eliminated.[215]

[213] Decision and Order, Everalbum, Inc., *supra* note 204.

[214] *Id.* at 2 (defining "affected work product"), 5 (detailing order compliance requirements).

[215] FED. TRADE COMM'N, *AI Companies: Uphold Your Privacy and Confidentiality Commitments*, OFFICE OF TECHNOLOGY BLOG (Jan. 9, 2024), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/01/ai-companies-uphold-your-privacy-confidentiality-commitments [https://perma.cc/7G3X-RZWR] ("[T]he FTC has required businesses that unlawfully obtain

Even where ex post unlearning is necessary, ex ante minimization can materially reduce remedial burden. Incorporating differential privacy during training bounds per-record influence and, when coupled with lightweight unlearning on the out-of-purpose subset, can lower re-identification and memorization risk while aligning with purpose-limitation goals.[216] The two safeguards are functionally distinct—DP masks influence; unlearning removes it—but combined, they better approximate the spirit of minimization than either alone and make subsequent removals less destabilizing.

However, blanket deletion of models can be a severe measure. If only a small portion of a model's training data was collected unlawfully (say one individual's data among billions of training points), deleting the entire model would be an extreme and arguably disproportionate step. Unlearning offers a more tailored alternative: rather than throwing out the proverbial barrel of wine because of one drop of poison, developers can remove just that drop and adjust the brew. If unlearning can reliably eliminate the influence of unlawfully obtained data, it might serve as a legally acceptable remedy that effectively "cleans out" the data from the model so that it is fit for lawful use moving forward. The challenge is ensuring that the cure is effective.[217] Given the difficulty of precisely removing all traces of a datum, one could argue unlearning is at best a partial fix that mitigates ongoing unlawful processing but is not a true absolution of the initial violation.[218] The initial unlawful collection or use still occurred; depending on the jurisdiction, the controller may still face penalties for that unlawful act regardless of unlearning as an available, albeit partial, remedy.[219]

That said, unlearning could still demonstrate a good-faith effort to honor purpose limitation and minimization on a micro level. Consider if a user withdraws consent for a secondary use of their data. The company might unlearn that individual's data from any models, thereby ceasing that incompatible use going forward. This could align with purpose limitation requirements to cease using the data for unconsented purposes, and could, in theory, be done relatively quickly after the objection is raised. From a regulatory perspective, such targeted unlearning might be deemed better than doing nothing. However, it is important to note that it still may not fully satisfy the spirit of the law.[220]

Regulatory scrutiny may remain. Even if a model is later forced to 'forget' data due to purpose limitation requirements, the unlearning could be deemed an after-

consumer data to delete any products—including models and algorithms . . . ."); *see, e.g*., Decision and Order, Everalbum, Inc., *supra* note 204.

[216] *See, e.g.,* Dwork & Roth, *supra* note 103; Ginart et al., *supra* note 104.

[217] Slaughter, *supra* note 162, at 58.

[218] *Id.*

[219] *See, e.g.,* Decision and Order, Everalbum, Inc., *supra* note 204; FED. TRADE COMM'N, STATEMENT OF COMMISSIONER ROHIT CHOPRA IN THE MATTER OF EVERALBUM AND PARAVISION COMMISSION, *supra* note 158, at 2 ("[T]he FTC needs to take further steps to trigger penalties, damages, and other relief [in addition to algorithmic disgorgement]").

[220] FED. TRADE COMM'N, DISSENTING STATEMENT OF COMMISSIONER ROHIT CHOPRA *IN RE* GOOGLE LLC & YOUTUBE, LLC*,* *supra* note 180.

the-fact bandage on what may have already been a violation of lawful collection, use, retention, or other privacy principles. Complicating regulators' compliance assessments is that a partially unlearned model might still retain insights that originated from now-disallowed data (this is because, as discussed, the influence can linger in latent form). In strict legal theory, continuing to use those lingering insights could be viewed as continuing the data processing and use beyond the allowed purpose despite any unlearning efforts made. Each of these scenarios will differ in the facts, and it is likely that regulators will ultimately need to determine when unlearning is or is not an appropriate remedy on a case-by-case basis. In practice, regulators might be pragmatic: if the model no longer directly reproduces or targets the disallowed data, and the developer can show they made substantial efforts to remove it, then unlearning may be deemed sufficient to meet the controller's purpose limitation or minimization obligation.

Yet tensions remain. Unlearning is a clunky fit for privacy principles that were meant to guide data usage from the ground up. The very nature of unlearning highlights a need for more dynamic interpretations of those principles when it comes to AI. This may include treating unlearning as a partial compliance mechanism while recognizing that truly purpose-limited, minimal-data AI development would require a very different approach (perhaps using smaller, more context-specific training datasets or other privacy-preserving training techniques rather than mass-scraping online data and later trying to forget some pieces of it).[221]

3. Rights to Correction & Erasure

Two of the most powerful individual rights in data protection law are the right to rectification (correction of inaccurate data) and the right to erasure (deletion of data, often called the right to be forgotten).[222] Although applying these rights in the context of machine learning models presents unique challenges, unlearning could be a key method for complying with such requests.

First, consider the right to rectification. If a traditional database contains an incorrect birthdate for a user, as an example, rectification means updating that field to the true value. But if a large language model has learned an incorrect fact about someone, there is not a single "field" that can be edited to fix the error. The incorrect information is diffused across the model's neural weights. One approach to correction would be to supply and train the model outright with the truthful information. Or, developers could fine-tune the model so that it unlearns the false data and "relearns" (is input with) the correct data. This is essentially a combination of unlearning and new learning, sometimes referred to as *model editing*.[223] As illustration, developers might design a special fine-tuning step that makes the model forget a defamatory statement about a person and replace it with accurate information. Because rectification-aimed unlearning that is executed only by

[221] *See* Shumailov et al., *supra note* 19 at 5.

[222] *See* GDPR arts. 16, 17.

[223] *See* Liu, *supra* note 179.

removing the tainted data and then retraining the model might not guarantee the model's outputs are subsequently correct, rectification in AI often requires affirmative correction mechanisms beyond 'just forgetting.' This currently remains an active research area with some experimental successes in "model editing" algorithms. In practice, model-editing pipelines often blend targeted fine-tuning (anti-examples that demote the erroneous content) with amnesiac steps (structured pruning + brief noise-infused updates) to damp residuals and limit collateral drift.[224] While these edits do not guarantee perfect rectification, they offer measurable, localized corrections that better effectuate rectification than deletion alone. Legally, this raises the question of, if a model continues to output an incorrect statement about someone, whether the controller is in violation of the obligation to rectify. Furthermore, if the only way to fix this issue is to retrain or significantly alter the model, whether that might be deemed necessary for compliance. Thus, while unlearning can assist with rectification by "wiping out" incorrect datasets (e.g., telling the model to forget everything it learned from a particular erroneous document), it may need to be paired with additional training on corrected data to truly comply with the right to rectification. Summarily, unlearning is one tool in that toolbox, but it is not a complete solution to "correcting" a model's knowledge.

The right to erasure (or "right to be forgotten") is where machine unlearning has been most directly invoked. When an individual exercises the right to erasure, a data controller must delete that person's personal data and cease further processing of it, barring some exceptions. In the context of a trained AI model, this implies the person's data should also no longer have any effect on the model's outputs—essentially, the model should forget the person. As discussed above, one brute-force way to honor such a request is to delete the model entirely since the model is, in part, a product of that personal data.[225] Deleting an entire model because one individual wants to be forgotten is usually impractical and arguably beyond what the law requires in most cases.[226] Scholars and regulators have stated that interpreting the right to be forgotten to require complete model erasure would be "extreme" and could unduly impair the rights of others or the utility of the model.[227] This is precisely why machine unlearning research is coming into prominence: it is a less draconian way to selectively remove one person's influence on a model without either retraining from scratch or throwing out the whole system.[228] If unlearning works as intended, the model after unlearning should perform as if it never saw the forgotten individual's data in the first place.[229] In other words, unlearning aims to produce a model functionally equivalent to a fresh model trained on a dataset with that person's data omitted.[230] If achieved, this

[224] *See* Eldan & Russinovich, *supra* note 71, at 1-5; *see also* Graves et al., *supra* note 94, at 11516.

[225] Decision and Order, Everalbum, Inc., *supra* note 204.

[226] Cooper et al., *supra* note 1, at 9.

[227] *Id.*

[228] *Id.*

[229] *Id.* at 2.

[230] *Id.* at 9.

would legally satisfy both the letter and the spirit of the right to erasure: the individual's data is not just eliminated from the training database, but also has no appreciable impact on the AI's behavior going forward.[231]

The efficacy of unlearning vis-à-vis erasure rights, however, must be scrutinized. A core requirement of the right to be forgotten is that "any influence of the data on the model disappears."[232] Unlearning methods strive for this, but as we have seen, complete disappearance is hard to guarantee. There is a legal gray area here. If a model cannot absolutely guarantee that none of its outputs or internal representations reflect a deleted individual's data, is the controller in compliance with an erasure request? Or is a reasonable best effort enough? Where full retraining is infeasible, Fisher-guided "scrubbing" can probabilistically obscure a record's footprint by perturbing the most sensitive weights.[233] This approach is appealing when original training data are no longer accessible, but its probabilistic nature complicates claims of total erasure; a sophisticated auditor may still extract faint signals, underscoring the need for verification protocols discussed below.

Data protection authorities have not yet provided definitive guidance on how perfect the "forgetting" must be within AI contexts. It is conceivable that regulators may accept a standard of "reasonable effort"—e.g., the controller used state-of-the-art unlearning methodologies and the model no longer deliberately or predictably outputs the person's information—even if a remote chance that the model could reveal traces of information remains. Conversely, if it is shown that an AI model can still produce someone's reportedly deleted personal data (say, the person's exact home address as equivalent to what was included in the training data) after an unlearning process takes place, regulators would likely deem that non-compliant as the data clearly persists in the model. One empirical complication is that proving a model has forgotten something is difficult (which ties into verification issues discussed later[234]). The controller might argue the model will not output "X," but how can the individual or a regulator be sure?

Comparing unlearning with outright model deletion highlights a trade-off between precision and certainty. Deleting the entire model guarantees that the individual's data can no longer influence any outputs (full certainty of compliance) but sacrifices all the useful knowledge gleaned from other data. Conversely, unlearning surgically tries to remove only the forbidden data and its influence, preserving the rest of the model's knowledge (maximizing utility) but with less certainty that nothing remains of the target data. If unlearning tools become highly reliable, they could offer a way to honor erasure rights in a manner that is proportionately compliant with the request—avoiding the "nuclear option" of destroying a model trained on thousands of people's data just because one person opted out.[235] However, until such methods are proven effective, organizations run

[231] *Id.* at 6.

[232] M. Chen et al., *supra* note 3, at 896.

[233] *See, e.g.*, Neel et al., *supra* note 57, at 934; Golatkar et al., *supra* note 81, at 9307.

[234] See discussion *infra* Section 4: "Right to Object to or Withdraw Consent from Processing."

[235] *See* Cooper et al., *supra* note 1, at 14.

a risk: a model that has been subject to a deletion request but not fully purged could be a ticking time bomb of non-compliance if it ever divulges the supposedly erased information. In the EU, non-compliance with erasure rights can lead to severe penalties.[236] This puts pressure on AI developers to err on the side of caution. In borderline cases, some may choose to retrain models from scratch (true deletion) or heavily restrict what the model can do, rather than rely on unlearning alone.

In sum, unlearning is an imperfect but pragmatic tool for responding to data deletion demands. It seeks to balance the individual's right to be forgotten against the practical reality that one person's data is intertwined with a model built on many people's information. Whether it achieves an equivalent level of privacy protection is case-specific, and so far, it appears that pure unlearning rarely matches the completeness of full model deletion. Thus, from a legal perspective, unlearning helps effectuate the rights to correction and erasure, but it may need to be supplemented with other steps (like model edits, additional training, or usage constraints) to fully realize those rights in practice.

4. Right to Object to or Withdraw Consent from Processing

Data privacy laws often give individuals the right to object or withdraw their consent to certain processing of their personal data, after which the organization must stop using their data for those purposes.[237] In the AI training context, this creates a scenario similar to that of the right to erasure: if someone originally allowed their data to be used to train a model, or if the data was used under an assumption of lawful basis, and later the data subject objects or rescinds their consent, the data controller should cease processing of that person's data. For a deployed machine learning model, "ceasing processing" logically means the person's data should no longer affect the model's operations. Short of turning the model off entirely, the way to achieve this is to remove the person's data from the model (i.e., to unlearn it). Thus, unlearning is directly relevant as a mechanism to honor objections or withdrawn consent. It enables a model owner to prospectively exclude an individual's data from further influence, without having to discard the model in entirety that was built only in part on that data. In effect, unlearning is a form of "update" to the model when the legal basis for using certain data has evaporated.

While this sounds relatively straightforward, there are hurdles in practice. One issue is timing. Laws like the GDPR require that when consent is withdrawn or an objection is lodged, the controller must stop the processing within a reasonable timeframe.[238] If a user opts out of a dataset today, a company cannot wait a year before effectuating that removal (at least, not in the traditional personal data

[236] *See* Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317, ¶ 93 (May 13, 2014).

[237] GDPR, art. 7(3).

[238] GDPR, art. 17(1).

context).[239] Yet retraining or unlearning in a massive model is not instantaneous. It might be infeasible to do a fresh training run for each individual withdrawal request in close to real-time.[240] Production ecosystems compound timing challenges: transformer-based services stitched from multiple components and federated settings where data never leave devices mean there is no single warehouse to purge.[241] Unlearning may require coordinated updates across dispersed controllers/processors, elongating timelines and complicating proof of completion. As such, some scholars have suggested batching unlearning requests and updating models periodically (say, retraining every few months or on an annual cycle).[242] For example, a service could accumulate all deletion/withdrawal requests and incorporate them in a scheduled model update, thereby efficiently handling many removals at once.[243]

Regulators might tolerate a brief delay if it's reasonable under the circumstances (considering technical difficulty), but there is an overarching legal grey area around how quickly a model must forget someone who has revoked permission (GDPR Art. 7, notably, gives not even a general timeframe for compliance with withdrawn consent compliance).[244] If compliance without "undue delay" is expected, as it is with the right to erasure,[245] and if this is interpreted strictly, companies may need to develop faster unlearning pipelines or use architectures that allow quicker updates. Otherwise, firms risk non-compliance by virtue of technical slowness. Policymakers may eventually need to clarify expectations here, possibly by explicitly allowing batched or periodic compliance updates for AI models so long as they occur within a certain timeframe.[246]

Another challenge is providing proof to the individual (or regulator) that the data truly no longer influences the model. Essentially, it must be demonstrated that the objection or withdrawal consent request has been honored. In a simple database, proof is shown by evidence that the record has been deleted. In a complex model, one might need a certificate of unlearning or other similar, validated proof. Architectures that enable deterministic, shard-level retraining can improve verifiability. For example, sharded-isolation (SISA)-style designs produce predictable before/after weight snapshots for a given deletion, enabling auditors who hold both versions to confirm that the same deletion yields the same retrained

[239] *See* GDPR, art. 12(3) ("The controller shall provide information on action taken on a request under Article [] . . . 22 [automated decision-making] without undue delay and in any event within one month of receipt of the request. The period may be extended by two months where necessary, taking into account the complexity and number of requests."); *cf.* Liu, *supra* note 179.

[240] Cooper et al., *supra* note 1, at 5.

[241] Nerella et al., *supra* note 64, at 30 (discussing shared training model which leverages data from fragmented sources without divulging sensitive patient information because of how federated learning communicates between various data sources).

[242] Liu, *supra* note 179, "Section 1: A bit of history & motivations for unlearning."

[243] *Id.*

[244] GDPR, art. 7.

[245] GDPR, art. 17(1) ("without undue delay").

[246] *See* Liu, *supra* note 179, "Section 1: A bit of history & motivations for unlearning."

weights—a practical audit hook when certifying that a record's influence has been removed.[247]

However, researchers and commentators assert that generating a verifiable "proof of unlearning" is not always possible with current techniques.[248] Unlearning processes might be heuristic and not leave a clear audit trail that can be externally validated. This complicates the right to object to or withdraw consent. The individual might reasonably ask, "How do I know my data isn't still somewhere in that model?" while the state-of-the-art contemporary techniques still do not yet offer easy answers. One could imagine tools in the future, though, that output an unlearning report or quantitative measure of data influence removed. But for now, trust is required. In regulatory terms, this issue of proof could be handled via oversight provisions (e.g., requiring companies to submit models for third-party testing if challenged, or to use reliably provable unlearning methods once they do mature). Until then, the exercise of objection/withdrawal rights in AI contexts will rely on the controller's representations and the general robustness of their unlearning compliance program.

In summary, machine unlearning is poised to become a key method by which companies attempt to effectuate individuals' rights to stop certain data uses, and it is a useful mechanism that translates a legal right ("don't use my data anymore") into a technical change in an AI model ("the model no longer uses your data"). It certainly facilitates the exercise of these rights by offering alternatives to shutting down an entire model that was partly trained on objectors' data. However, unlearning also complicates these rights because it introduces uncertainty into what it means to stop "using" data in an AI setting. If the model cannot be perfectly purged or if verifying data deletion is impossible, then asserting the right to withdraw consent may not guarantee the outcome that an individual expects. This is another illustration of how existing privacy norms strain under the weight of AI's complexities: the rights remain the same on paper but fulfilling them requires new technical and possibly legal innovations.

D. Unlearning as Complement or Enhancement to Existing Remedies

Given the above tensions, one might ask if machine unlearning can be a replacement for traditional privacy remedies or if it should simply be a complement to them. The scholarly and practical consensus is that unlearning is best viewed as a useful augment to existing data protection measures and not a panacea on its own.[249] Put otherwise, unlearning can and should be used to enhance privacy compliance and accountability in the AI context, but it optimally works in tandem with other strategies and with its limitations acknowledged. Operationally,

[247] *See* Bourtoule et al., *supra* note 7, at 142; Thudi et al., *supra* note 5, at 4009 ("Reproducing the alleged computation is synonymous to showing its plausibility."); *See also* Nguyen et al., *supra* note 40, at 3 ("[A] verification (or audit) is needed to prove that the model actually forgot the requested data and that there are no information leaks.").

[248] Liu, *supra* note 179, "Section 3: Evaluating unlearning."

[249] Cooper et al., *supra* note 1, at 11.

unlearning could be implemented as a layered stack: (i) exact/certified forgetting for legally significant takedowns; (ii) approximate edits (targeted fine-tuning, amnesiac pruning, Fisher-style perturbations) to localize repair; (iii) front-end output suppression to prevent resurfacing; and (iv) DP-aware training to cap per-record influence ex ante.[250] This stack offers granularity and proportionality, preserving lawful utility while addressing specific defects.

On the remedial front, compare unlearning to stricter remedies such as algorithmic disgorgement and full model deletion, which are blunt but decisive remedies used to address data misuse. Algorithmic disgorgement (as ordered by the FTC *In the Matter of Everalbum, Inc.*[251]) requires a company to entirely delete models derived from unlawful data, ensuring that ill-gotten gains or benefit is not retained.[252] By destroying the learned knowledge entirely, the intended privacy protection-through-remedy is achieved. Unlearning, by contrast, attempts a more surgical strike: remove just the pieces of learned knowledge that are problematic (e.g., derived from a specific person's data) while keeping the rest of the model intact.[253]

The strength of unlearning lies in this precision. It is far more targeted than retraining from scratch or deleting whole models. Studies have shown that, as such, certain unlearning techniques or model designs can forget a data point at a tiny fraction of the computational cost of total retraining.[254] For example, one method partitioned a model into components such that a data deletion affected only a small subset, achieving forgetting with about 0.3% of the training cost of rebuilding the entire model.[255] This efficiency makes unlearning a practical tool for ongoing compliance: a company can respond to removal requests or rectify its training set without incurring the massive expense (and downtime) of full model redevelopment. In that sense, unlearning complements other legal enforcement methods by providing a means to carry out corrective orders that would otherwise be onerous. Unlearning procedures give companies a way to comply with regulators' demands without losing the benefit of other (lawfully retained) data. Unlearning also offers a measure of proportionality: it can be scaled to the scope of the violation (forget one user, not punish all users).

Unlearning's targeted nature can also enhance privacy-by-design. If developers anticipate the need to remove data, they can design models in modular ways or keep track of data influence, making future unlearning easier and more exact. This could be seen as an improvement on traditional data governance. Instead of treating trained models as 'black boxes' that are forever and mysteriously influenced by whatever is put in, developers could better maintain the ability to edit and purify

[250] *See* Guo et al., *supra* note 30; Thudi et al., *supra* note 5; Bourtoule et al., *supra* note 7; Dwork & Roth, *supra* note 103; Ginart et al., *supra* note 104.

[251] Decision and Order, Everalbum, Inc., *supra* note 204, at 5.

[252] Slaughter, *supra* note 162, at 39.

[253] Cooper et al., *supra* note 1, at 4.

[254] Achille et al., *supra* note 21, at 4.

[255] *Id.*

models more precisely and as needed for compliance. Additionally, unlearning methods can be combined with other privacy-enhancing techniques. For instance, some research suggests using differential privacy during training to limit each data point's influence, which in turn makes any single-point removal less disruptive and more provably effective.[256] In that way, unlearning (removing data) plus differential privacy (adding noise to mask a data's impact) together could yield AI models where regulators and users can be more confident that no individual's data is embedded in an irremovable way. Unlearning can also work alongside contractual and organizational measures, e.g., a company might promise in its privacy policy to use unlearning if a user exercises their rights, thus adding an extra layer of privacy accountability (failure to do so could be deemed a breach of contract or deceptive trade practice).

A related privacy-preserving strategy is output suppression, which constrains what the model says rather than altering what it knows. Techniques such as reinforcement learning from human feedback (RLHF) train a reward model to favor compliant or refusal-style answers (for instance, declining to reveal personal data) without changing the underlying parameters that still contain that data.[257] While attractive because it needs modest retraining and no access to the original corpus, RLHF provides only a behavioral guarantee: it curbs disclosure so long as the refusal policy is not circumvented. Lawyers should understand that the model's weights (and thus the information they encode) remain intact; RLHF merely teaches the system to respond, "I'm sorry, but I can't help with that request," rather than truly to forget. Hence, RLHF mitigates exposure risk but does not satisfy deletion or erasure rights in the strict legal sense.

Beyond in-model alignment, firms often deploy external filtering layers, stand-alone classifiers that screen prompts or outputs for prohibited content before delivery to end users.[258] OpenAI's Moderation API exemplifies this architecture: the filter flags responses with high probabilities of containing personal identifiers or other sensitive categories and then blocks or edits them. These modular "watchdogs" can be improved or domain-tuned (e.g., medical-record filters) without retraining the base model, a practical advantage for sector-specific compliance.[259] Yet the approach remains whack-a-mole: filters can miss novel disclosure patterns or over-block lawful speech, and they leave the underlying data untouched. As Shumailov et al. (2024) demonstrate, a determined attacker with weight-level access can still extract the hidden information.[260] Accordingly, external filters and moderation APIs are risk-mitigation—not deletion—tools; they

[256] M. Chen et al., *supra* note 3, at 907.

[257] Niloofar Mireshghallah et al., *Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity* Theory, ARXIV 1, 7 (June 28, 2024), https://arxiv.org/pdf/2310.17884 [https://perma.cc/2NJY-UT8A]; Ouyang et al., *supra* note 74, at 27732.

[258] Pisano et al., *supra* note 122, at 2.

[259] *Id.*

[260] Shumailov et al., *supra* note 19, at 5.

should accompany, not substitute for, genuine unlearning when legal erasure is required.

Together, RLHF and external filtering illustrate the behavioral flank of privacy protection: they manage outputs but not memory, underscoring why subsequent enforcement frameworks must distinguish suppression from unlearning.

As much as unlearning improves flexibility, it has notable weaknesses and should not be solely relied upon for compliant privacy protection. A key weakness of unlearning is the incompleteness of data removal as latent traces can persist. In scenarios where absolute assurance is required (for example, removal of illicit contraband data like child abuse images), unlearning alone could be deemed too uncertain, and a combination of model inspection, output monitoring, and perhaps partial architecture changes could be needed to really eliminate the influence.[261] Another weakness is its technical complexity and potential model degradation. Repeatedly unlearning data points from a model could lead to accumulated error or reduced performance, especially if the unlearning methods are not exact. A model might become less accurate or exhibit anomalies after many deletion operations, which might conflict with other legal obligations like fairness or accuracy in automated decision-making. In extreme cases, if a huge number of removals are required, it may actually be simpler and more reliable to retrain from scratch as unlearning is not infinitely scalable. Thus, unlearning complements, but does not fully replace, the fallback option of full retraining when needed.

Realistically, some large AI models, like today's giant GPT-style models, are already so complex that current unlearning techniques struggle to handle them at scale. The feasibility of unlearning in models with billions of intertwined data points is still being tested. One piece of scholarship on this topic bluntly noted that these models contain "an arbitrarily high number of data dimensions and statistical correlations," making it very difficult to determine the specific effect of any given training example on the model.[262] In such cases, the only sure way to remove a data point's effect might be to rebuild the model without it included—precisely the costly process unlearning is meant to avoid. Future research may invent either new methods or architectures that are more traceable, but until then, unlearning for the largest models may be more theoretical than practical. In short, unlearning is a promising technique to enhance existing privacy remedies. It can make compliance more attainable and less damaging to useful AI functionality, but it is not a magic wand. Policymakers and practitioners increasingly recognize that unlearning methods are imperfect and may serve as only one approach of many in a privacy protection toolkit.[263] It is best deployed as part of a layered strategy: for instance, use privacy-by-design to minimize data usage up front, employ unlearning to handle individual deletions and corrections, and retain the option of stronger

[261] Cooper et al., *supra* note 1, at 2.

[262] Achille et al., *supra* note 21, at 1.

[263] Cooper et al., *supra* note 1, at 3.

measures (like model deletion or output gating) if sensitive residuals still pose a risk.

Machine unlearning sits at the intriguing intersection of cutting-edge AI engineering and fundamental principles of privacy law. This section has explored the central tension: while unlearning techniques aim to satisfy core legal requirements on paper—allowing AI models to forget unlawful data, limit processing to intended purposes, and honor individual rights like erasure and objection—in practice, these techniques often fall short of the law's expectations for complete privacy protection. Residual traces of "forgotten" data in model parameters and the possibility of reconstructive outputs mean that a model may continue to leak personal information even after unlearning has occurred.[264] Thus, an organization might technically comply with an erasure request by running an unlearning algorithm, yet the individual's privacy could still be compromised if the model retains an algorithmic shadow of the data. This gap between formal compliance and actual risk reveals a need for additional safeguards. Moving forward, both policy and technical developments will be crucial to bridge this gap. Regulators may need to set clearer standards for what counts as "adequate" unlearning and develop oversight mechanisms to ensure genuine data removal, such as requiring proof or certifications of the unlearning process.[265] On the technical side, researchers are exploring hybrid approaches—from differential privacy guarantees to more transparent model architectures—to reinforce unlearning and prevent the emergence of new privacy vulnerabilities when models are updated.[266] Ultimately, machine unlearning should be seen not as a silver-bullet solution but as one emerging tool that, combined with robust privacy governance and possibly new legal norms, can help uphold individuals' rights in the age of AI.[267]

The next part of this Article will build on these insights. It considers how law and policy might adapt to support the effective deployment of unlearning techniques and what alternative or complementary measures might be necessary to truly "forget" personal data in machine learning contexts.

IV. OPERATIONALIZING MACHINE UNLEARNING FOR THE ENFORCEMENT OF PRIVACY AND DATA PROTECTION LAW

Part IV introduces a framework for operationalizing machine unlearning in privacy law's enforcement, with unlearning proposed as one component of a broader privacy intervention spectrum. It considers a range of measures specifically aligned with the enforcement and rulemaking authority of the Federal Trade Commission under Article 5 of the FTC Act, Department of Justice Consumer Protection Division's policy and litigation stratagem, enumerated powers of states Attorneys General, and emerging global standards. For example, it proposes

[264] Cooper et al., *supra* note 1, at 2; M. Chen et al., *supra* note 3, at 896.

[265] Liu, *supra* note 179.

[266] M. Chen et al., *supra* note 3; Liu, *supra* note 179.

[267] Cooper et al., *supra* note 1, at 23.

preventive measures which may be included in agencies' published industry guidelines, such as incorporating privacy-preserving techniques into training models (e.g., differential privacy), as well as reactive measures like output suppression model deletion and algorithmic disgorgement to be used in settlements and litigation remedies. The framework explores when, where and how such approaches could be operationalized despite mismatches between technicalities in unlearning and broader privacy concepts. It does so by addressing practical concerns such as computational cost, scalability, and tensions between efficiency and accuracy in its recommendations. This Part emphasizes that privacy governance for generative AI must be multifaceted and dynamic, reflect the diverse harms and risks posed by these systems, and avoid reinforcing "compliance-by-design" approaches that prioritizes technical fixes over substantive accountability. We treat unlearning not as a single tool but as a graduated stack—heavyweight certified forgetting for legally dispositive removals and lightweight suppression for everyday safety[268]—so agencies can calibrate remedies to risk and feasibility.

U.S. privacy enforcement involves multiple actors, namely the Federal Trade Commission (FTC) under Section 5 of the FTC Act, the Department of Justice (DOJ) Consumer Protection Branch, and state Attorneys General enforcing state consumer protection and privacy statutes.[269] Each can pursue companies for privacy-invasive practices, including misuse and deceptive use of personal data in AI training.[270] For example, the FTC has ordered deletion of algorithms developed with ill-gotten data (so-called "algorithmic disgorgement") in settlements like *In the Matter of Everalbum*, where a photo app had to delete facial recognition models built on users' images.[271] The DOJ, often in tandem with the FTC, has enforced data privacy laws like COPPA in court, as in *United States v. Kurbo Inc.* (requiring deletion of models trained on children's data).[272] State AGs likewise bring actions under state law for data misuse, sometimes leading to injunctions or fines. Globally, data protection regulators enforce analogues to these rights under regimes such as

[268] Guo et al., *supra* note 30; Thudi et al., *supra* note 5; Bourtoule et al, *supra* note 7; Dwork & Roth, *supra* note 103; Ginart et al., *supra* note 104.

[269] See, e.g., *Protecting Consumer Privacy and Security: Privacy & Security Enforcement*, FED. TRADE COMM'N, https://www.ftc.gov/news-events/topics/protecting-consumer-privacy-security/privacy-security-enforcement [https://perma.cc/22NU-GBB9] (last visited Nov. 8, 2025); *Consumer Protection Branch – Privacy Practice Areas*, U.S. DEP'T OF JUST., https://www.justice.gov/civil/consumer-protection-branch-practice-areas#Privacy [https://perma.cc/77B8-HW7J] (last visited Nov. 8, 2025); *Privacy: Consumer Protection 101*, NAT'L ASS'N OF ATT'YS GEN., https://www.naag.org/issues/consumer-protection/consumer-protection-101/privacy/ [https://perma.cc/7ZBJ-J859] (last visited Nov. 8, 2025).

[270] *See, e.g.*, FED. TRADE COMM'N, FTC POLICY STATEMENT ON UNFAIRNESS (Dec. 17, 1980), https://www.ftc.gov/legal-library/browse/ftc-policy-statement-unfairness [https://perma.cc/DB8Y-HYA4]; Children's Online Privacy Protection Rule (COPPA), 16 C.F.R. pt. 312 (2024), https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa [https://perma.cc/QK2Y-JT67]; Decision and Order, Everalbum, Inc., *supra* note 204.

[271] Decision and Order, Everalbum, Inc., *supra* note 204, at 5.

[272] Proposed Stipulated Order, United States v. Kurbo, Inc., No. 3:22-cv-04477 (N.D. Cal. Mar. 30, 2022), at 8, https://www.ftc.gov/system/files/ftc_gov/pdf/wwkurbostipulatedorder.pdf [https://perma.cc/SC4C-9JET].

the EU's GDPR, which empowers authorities to order deletion of unlawfully processed personal data and even restrict model use derived from such data. Indeed, the GDPR's right to erasure[273] and similar laws (California's CCPA, Canada's PIPEDA, etc.) have legally solidified this right to be forgotten in the machine learning context.[274]

Because privacy harms from AI are multifaceted, regulators require a range of interventions beyond machine unlearning alone. Technical unlearning—i.e., removing personal data from a model—is just one remedial tool. Enforcement bodies also emphasize broader measures: penalties to deter misconduct, mandates for better data governance, and ongoing oversight of AI systems. Purely technical fixes cannot address all privacy risks. For instance, deleting a piece of training data may not fully erase its "algorithmic shadow" —the persistent imprint that data leaves on a model's behavior.[275] Privacy regulators therefore blend ex ante guidance with ex post remedies. They recognize that robust privacy protection requires preventive steps (to avoid problematic data use in the first place) and multiple remedial options (to fully redress harms), rather than over-reliance on any single technique like unlearning. The next sections examine how agencies operationalize this mix, situating machine unlearning within a broader enforcement toolkit alongside comparative insights from the GDPR.

A. Preventive Measures in Published Guidelines

Regulators increasingly use published guidelines and promulgated rules to urge organizations to adopt privacy-preserving practices before problems arise.[276] This can reduce the burden on later unlearning or deletions in AI contexts. The FTC's business guidance and consent orders often require the implementation of comprehensive data governance programs, which implicitly further these goals by mandating internal review and deletion of data that should not be retained (thereby preventing it from ever influencing a model).[277] Global regulators echo these expectations; for instance, the GDPR obligates controllers to implement data protection by design and by default, which includes data minimization and the ability to purge personal data on request—effectively front-loading the capacity to

[273] GDPR, art. 17.

[274] M. Chen et al., *supra* note 3, at 906.

[275] Li, *supra* note 158.

[276] *See, e.g.*, FED. TRADE COMM'N, ARTIFICIAL INTELLIGENCE COMPLIANCE PLAN (September 2025), https://www.ftc.gov/system/files/ftc_gov/pdf/FTC-AI-Use-Policy.pdf [https://perma.cc/BS2S-G5BW]; FED. TRADE COMM'N, *AI (and Other) Companies: Quietly Changing Your Terms of Service Could Be Unfair or Deceptive*, OFFICE OF TECHNOLOGY BLOG (Feb. 13, 2024), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive [https://perma.cc/5ALE-NFHC]; U.S. DEP'T. OF JUST., OVERVIEW OF THE PRIVACY ACT OF 1974 (2020 EDITION): AGENCY REQUIREMENTS (Oct. 22, 2022), https://www.justice.gov/opcl/overview-privacy-act-1974-2020-edition/agency-requirements [https://perma.cc/88NK-994X].

[277] Decision and Order, Everalbum, Inc., *supra* note 204.

remove data and its downstream effects.[278] In general, key preventive measures include data minimization (collecting and retaining only what is necessary) and privacy by design principles baked into model development. For example, agencies may recommend rigorous dataset vetting to exclude data that cannot be lawfully used or that might trigger removal requests down the line. In practice, this means robust dataset audits—reviewing training data for compliance (consent, legality, relevance) and documenting provenance—as a standard compliance step. But by catching problematic data early, companies can avoid having to "unlearn" it later under regulatory pressure. Guidance should also encourage "ready-to-forget" architectures. Expert-mixture designs (ARCANE-style) segment learning into narrow specialists that can be retrained surgically when a deletion request targets a confined topic, while SISA-like sharding produces deterministic retrains that double as verification mechanisms.[279] Both approaches shrink blast radius and increase auditability, aligning ex ante design with foreseeable erasure duties.

Technical literature reinforces that prevention can ease remedial burdens. Researchers have proposed designing ML models from the start to accommodate future deletion. One approach is compartmentalization of training data: splitting data into disjoint shards with separate sub-models, so that if a particular subset must be removed, only that shard's sub-model needs retraining.[280] This proactive "sharding" strategy, recommended in computer science research on machine unlearning, can dramatically cut the cost of later deletions—one study showed forgetting a data shard could take as little as 0.3% of the time of full retraining, albeit with a slight accuracy trade-off.[281] Agencies like NIST have highlighted such techniques in their AI risk management frameworks, and we can imagine guidelines encouraging firms to adopt "ready-to-forget" architectures. Other preventive tools include embedding differential privacy during training, which injects noise to statistically obscure individual data contributions. By limiting each data point's influence on the model, differential privacy can ensure that removing any single individual's data has negligible impact—effectively aligning with the principle of minimization and making compliance with deletion requests more manageable.[282] Regulators may cite these techniques in guidance documents or future rules as best practices (especially for high-risk AI uses) because they reduce reliance on after-the-fact unlearning. In sum, through guidelines and soft law, enforcement bodies promote upstream privacy safeguards—minimizing sensitive data use, securing opt-in consent for training data, using privacy-preserving model methods—to lessen the need for drastic remedies later. Prevention, in the form of sound data hygiene and privacy engineering, serves as the first line of defense.

[278] GDPR, art. 25.

[279] H. Yan et al., *supra* note 7, at 4011 ("When large data unlearning, . . . the accuracy of ARCANE would not degrade too much [and] the training and unlearning of ARCANE is much faster than SISA."); Bourtoule et al., *supra* note 7, at 154; Thudi et al., *supra* note 5, at 4009; Nguyen et al., *supra* note 40, at 3.

[280] Achille et al., *supra* note 21, at 2.

[281] *Id.* at 4.

[282] *Id.* at 5.

B. Reactive Remedies and Model-Based Obligations

When preventive measures fail or violations occur, regulators turn to reactive remedies. Increasingly, these remedies focus not just on deleting raw data, but on the models and algorithms that have absorbed that data. In the context of AI enforcement, this has given rise to novel obligations at the model level—effectively compelling organizations to unlearn illicit data and its effects. We outline three categories of such reactive interventions: (1) model deletion or algorithmic disgorgement; (2) targeted output suppression and selective unlearning; and (3) operationalizing these remedies in enforcement actions.

1. Model Deletion & Algorithmic Disgorgement

Model deletion, also termed algorithmic disgorgement or algorithmic destruction, is the strongest form of machine unlearning remedy, requiring a company to eliminate not only the offending data but also any AI models or algorithms trained on that data.[283] Recent scholarship frames this as an essential new enforcement tool to address the "algorithmic shadow" problem—the idea that once personal data has been ingested by an AI, simply deleting the source data is insufficient, because the model retains a latent imprint of that data.[284] Tiffany Li (2022) introduced algorithmic destruction as a privacy remedy precisely to tackle these residual harms, arguing that regulators must sometimes force the deletion or retraining of models built on misused personal information.[285] This remedy has already started to appear in enforcement: the FTC's orders in *Everalbum* (2021) and related cases required the company to delete "any models or algorithms" developed with improperly obtained data.[286] Scholars Jevan Hutson and Ben Winters (2024) hail such model deletion as a way to meaningfully sanction companies whose AI products are tainted by unlawful data collection, ensuring that businesses causing algorithmic harm face deletion—effectively destruction—of significant portions of their AI/ML work products including trained models and datasets.[287] Proponents note this remedy can improve privacy and deter misconduct by stripping wrongdoers of any unfair advantage gained from misused data.[288]

However, scholars also urge caution and nuance. Daniel Wilf-Townsend (2024) observes that in its current form, algorithmic disgorgement can be a grossly disproportionate penalty if applied inflexibly.[289] For instance, deleting an entire large-scale model (representing millions in research and development) because a tiny fraction of training data was problematic might overshoot, causing undue

[283] Li, *supra* note 158, at 491; Hutson & Winters, *supra* note 158, at 127.

[284] Li, *supra* note 158, at 479.

[285] *Id.*

[286] Decision and Order, Everalbum, Inc., *supra* note 204.

[287] Hutson and Winters, *supra* note 158, at 131.

[288] Wilf-Townsend, *supra* note 158, at 1829.

[289] *Id.* at 1809.

collateral damage.[290] Wilf-Townsend argues for a balanced doctrine that considers factors like the degree of a defendant's culpability, how much the tainted data contributed to the model, and the availability of less drastic alternatives.[291] In other words, model deletion should be applied proportionately. Current scholarship thus frames algorithmic disgorgement as a powerful yet blunt tool, one that can be refined by "surgical" unlearning techniques. Machine unlearning research offers ways to selectively remove a data point's influence without scrapping the entire model. By incorporating such techniques, regulators and courts could require partial model purging in lieu of total deletion, achieving compliance while mitigating the remedy's severity. Indeed, Achille et al. (2024) highlight methods to pinpoint and eliminate a particular dataset's influence on a model (sometimes termed *selective forgetting*).[292] These innovations suggest that model-based remedies need not be all-or-nothing—unlearning can complement disgorgement by allowing more granular deletions (e.g. re-training only certain layers or components). In sum, model deletion is emerging as a key legal tool for AI accountability, and machine unlearning techniques stand to both bolster its effectiveness and temper its overbreadth.

2. Targeted Output Suppression & "Selective Unlearning"

Not all post-harm interventions require retraining or destroying a model; regulators might also seek output-focused remedies. These involve directing an AI system to stop producing certain content derived from tainted data—essentially a form of selective unlearning at the output level. If a generative model was improperly trained on a person's personal information or a copyrighted work, for example, authorities could mandate measures to prevent the model from ever generating that specific personal data or work again. This can be achieved through targeted output suppression techniques. One approach is deploying filters or wrappers around the model. As recent research notes, developers can implement system-level filters that intercept and block disallowed content either in the input (user prompt) or output stage.[293] For instance, a regulator might require a large language model to have an integrated filter that recognizes and omits a particular data subject's name or other personal facts in its responses. Such filters are already used to enforce content policies (blocking hate speech, certain biometrics, etc.), and they could be repurposed as privacy remedies—ensuring that "certain undesirable generations" never reach end-users.[294]

Where full retraining is unwarranted, agencies can require targeted fine-tuning to down-weight disallowed content and, where training data are unavailable,

[290] *Id.* at 1855.

[291] *Id.* at 1809.

[292] Achille et al., *supra* note 21, at 2.

[293] Cooper et al., *supra* note 1, at 6.

[294] Wilf-Townsend, *supra* note 158, at 1820.

Fisher-guided perturbations to attenuate residual signal.[295] These selective edits operationalize proportionality—narrow fixes for narrow harms—while consent orders can pair them with output filters to prevent resurfacing.

Another tactic is partial model adjustment without full retraining. This could involve fine-tuning the model on a clean dataset (with the problematic data removed) just enough to blunt its capacity to reproduce the offending information. Sometimes called selective retraining, this method aims to forget a narrow piece of knowledge. In computer vision and natural language processing (NLP) research, algorithms for selective forgetting can remove or alter a model's memory of specific classes or entries.[296] Regulators could, for example, require a company to run a targeted unlearning procedure so that a facial recognition AI "unlearns" a particular person's face embeddings, rather than deleting the entire model. This kind of remedy falls between pure model deletion and doing nothing—it surgically remedies the issue by suppressing the illicit output. We might analogize it to a recall or patch: the model remains mostly intact but is patched not to output or utilize the specific data at issue.

Regulators can also require behavioral suppression measures such as RLHF or external moderation filters when full retraining is disproportionate.[297] RLHF aligns model behavior with policy norms (e.g., refusing to output personal data), while moderation APIs act as independent classifiers that block prohibited material.[298] Both improve consumer-facing safety but do not erase underlying representations; consequently, decrees relying on them should make clear that such measures satisfy output-control obligations, not formal deletion duties. A well-crafted order might pair RLHF or filtering with documentation of residual-risk testing to ensure that suppression does not masquerade as compliance.

Agencies and courts are beginning to contemplate such selective remedies. For instance, in settlements involving misuse of data for AI training, the FTC has mandated companies to abstain from using certain outputs or functionalities until compliance steps are taken (effectively shutting off parts of a system) in addition to deletion of associated models.[299] Looking ahead, an FTC consent decree or a state AG injunction could foreseeably specify that a generative AI service must filter out any output containing a complainant's personal information, or that a company must retrain portions of its model to eliminate a forbidden pattern (e.g. a specific copyrighted artwork style). These measures amount to "selective unlearning" obligations—more granular than wholesale model destruction, and

[295] Neel et al., *supra* note 57, at 933; Golatkar et al., *supra* note 81, at 9301; Eldan & Russinovich, *supra* note 71, at 1.

[296] Achille et al., *supra* note 158, at 9.

[297] Mattern et al., *supra* note 107, at 11330; Ouyang et al., *supra* note 74, at 27730; Pisano et al., *supra* note 122, at 1; OpenAI, *supra* note 74.

[298] OpenAI, *supra* note 74.

[299] Press Release, Rite Aid Banned from Using AI Facial Recognition After FTC Says Retailer Deployed Technology without Reasonable Safeguards, Fed. Trade Comm'n (Dec. 19, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without [https://perma.cc/RRC5-9RZV].

often more technically feasible in the short term. They do, however, require technical capabilities on the part of the company. A concern is that even this approach has limits: as researchers have pointed out, to reliably filter out a piece of information (say, all occurrences of "Spiderman" in a model's output), the system itself needs to know about "Spiderman" in order to catch it.[300] Thus, paradoxically, some information might remain embedded so the filter can function. Despite such nuances, output suppression and selective unlearning tools provide regulators with flexible options. These remedies focus on preventing harm (e.g., stopping a privacy-violating disclosure by the AI) without necessarily demanding the expensive step of rebuilding the model from scratch.

3. Operationalizing These Remedies

Translating the above remedies into enforceable orders requires careful crafting and monitoring. U.S. regulators have already included model-focused obligations in consent decrees, signaling how such remedies are operationalized. A prime example is the FTC's order in *Everalbum*, which not only compelled deletion of infringing algorithms but also imposed ongoing obligations to ensure compliance.[301] The company had to delete affected models and document that deletion, subject to FTC oversight. Similarly, in *United States v. Kurbo*, a DOJ action for a COPPA violation, the settlement mandated destruction of any models or training data derived from children's personal information and required regular reporting of compliance.[302] These cases illustrate a model for injunctive relief: regulators don't simply trust companies to unlearn on their own. Instead, they include provisions for verification, such as requiring companies to maintain records of what was deleted or to train staff in data removal protocols. Orders should specify verifiable procedures (e.g., shard-deterministic retrains (SISA-style)) that yield reproducible before/after weights, logged alongside unlearning manifests and output-filter rule sets.[303] This layered obligation (heavy model edits + light front-end suppression) provides both ex ante documentation and ex post monitoring suitable for third-party assessment.[304]

Future consent decrees may go further. We could see orders that require companies to build the capability for data unlearning on request. A settlement might stipulate, as illustration, that if any consumer exercises a deletion right, the company must not only erase the raw data but also update or retrain relevant models to remove the data's influence within thirty days. This essentially operationalizes

[300] Cooper et al., *supra* note 1 at 1.

[301] Decision and Order, Everalbum, Inc., *supra* note 204.

[302] Press Release, Weight Management Companies Kurbo Inc. and WW International Agree to $1.5 Million Civil Penalty and Injunction for Alleged Violations of Children's Privacy Laws, U.S. Dep't of Just. (March 4, 2022), https://www.justice.gov/archives/opa/pr/weight-management-companies-kurbo-inc-and-ww-international-inc-agree-15-million-civil-penalty [https://perma.cc/8YS4-Z2EF]; Proposed Stipulated Order, United States v. Kurbo, Inc., *supra* note 272.

[303] Guo et al., *supra* note 30; Thudi et al., *supra* note 5; Bourtoule et al, *supra* note 7.

[304] Bourtoule et al, *supra* note 7.

machine unlearning as an ongoing compliance duty. Regulators might also require dataset oversight committees or external audits to periodically review whether a model's training data includes any data that should have been purged (for legal or policy reasons). If such data is found, the order could trigger a mandatory partial retraining or unlearning process. In effect, these types of decrees function as live governance tools, keeping the company on a short leash regarding its AI training data.

There are also examples of tailored injunctive relief targeting outputs. A court or agency order could, for example, enjoin a company from deploying an AI model until it certifies that personal data of the complaining consumers has been scrubbed from the model, or that the model will no longer produce specified outputs (with penalties if prohibited outputs appear). One real-world parallel is in intellectual property: in the *Getty Images v. Stability AI* copyright dispute, observers have speculated that a settlement might require Stability AI to remove or disable the generation of certain copyrighted images from its Stable Diffusion model.[305] While largely hypothetical, it shows how selective unlearning has the opportunity to be compelled through legal agreement: the model might remain, but with enforced blind spots.

Finally, regulators must account for the afterlife of data once a model has been cloned, distilled, or embedded in downstream products. Deleting the source weights alone does not neutralize derivatives that inherited the tainted information. Recent enforcement confirms this expectation: in *Everalbum* the FTC required deletion not only of unlawfully retained photos but also of any face-recognition models trained on them.[306] Accordingly, any unlearning or deletion order should extend to progeny models and require firms to inventory, patch, or re-train all derivatives so that contested data are fully excised from the product ecosystem. Otherwise, perfect forgetting at the source leaves a compliance gap as wide as the downstream marketplace. These derivative-model obligations illustrate why enforcement frameworks must treat unlearning as a continuing duty, not a one-time event.

Overall, operationalizing unlearning remedies means writing enforcement orders with clear, measurable requirements (delete these files, retrain this model segment, filter these outputs) and follow-up mechanisms (compliance reports, third-party assessments). The trend in U.S. enforcement suggests an increasing comfort with these novel provisions. By embedding technical obligations into legal orders, regulators can ensure that machine unlearning moves from theory to practice in protecting consumers.

[305] *Generative AI in the Courts: Getty Images v. Stability AI*, PENNINGTONS MANCHES COOPER (Feb. 16, 2024), https://www.penningtonslaw.com/news-publications/latest-news/2024/generative-ai-in-the-courts-getty-images-v-stability-ai [https://perma.cc/KM7U-PT9W]; Am. Ass'n of Indep. Music and Recording Indus. Ass'n of Am., Comments on A.I. and Copyright, Docket No. 2023–6 (2023), https://downloads.regulations.gov/COLC-2023-0006-8833/attachment_1.pdf [https://perma.cc/MF5H-C9NU] (discussing both *Getty Images v. Stability AI* and algorithmic disgorgement remedies).

[306] Decision and Order, Everalbum, Inc., *supra* note 204.

C. *Practical Tensions & Limitations*

Even as machine unlearning and related remedies become more common, several practical challenges temper their use. Policymakers and scholars have identified the following tensions in making these remedies effective and proportionate:

1. Computational Cost & Scalability

One major concern is the feasibility of repeatedly unlearning or retraining models at scale. Modern AI models are expensive and time-consuming to train; demanding frequent retraining or deletion in response to every data removal request can be technically onerous. Alessandro Achille et al. note that with today's massive models, any defect in the training corpus "cannot be trivially remedied by retraining the model from scratch"[307]—it is just too costly and slow. While research into more efficient unlearning methodologies (like compartmentalization) can reduce overhead, there is still a non-trivial burden.[308] Even "exact" sharded retrains can be hardware-intensive, with large-scale ImageNet-class experiments historically requiring multiple V100-class GPUs per shard; language-model baselines in the single-digit-billion parameter range can consume hundreds of thousands of GPU-hours for initial training, placing exact re-trains in the same order of magnitude.[309] Expert-mixture designs mitigate this by retraining only the affected expert, illustrating why agencies should permit architectural tailoring where consistent with the remedy's aims.

Regulators thus face a scalability issue in ordering remedies: it's one thing to require a small startup to rebuild a model, but ordering a tech giant to retrain a multi-billion-parameter model on demand might be impractical.[310] If compliance with deletion rights routinely forces companies to incur huge computational costs, they may resist or lobby against such requirements. This tension suggests unlearning will need to be targeted (used when truly necessary) and complemented by those preventive measures that minimize how often full retraining is needed.

2. Technical-Legal Mismatch

Gaps commonly exist between what legal directives envision and what technical remedies actually achieve. This is true also of unlearning. The law's requirement to delete personal data "in its entirety" is conceptually straightforward, but a trained model does not operate like a database; it generalizes from data clusters, causing traces of granular data to persist in multifaceted, indirect ways. Scholars describe how a model can retain a latent imprint of data even after an

[307] Achille et al., *supra* note 21, at 1.

[308] *Id.* at 3, 6-8.

[309] Bourtoule et al., *supra* note 7; Eldan & Russinovich, *supra* note 71; H. Yan et al., *supra* note 7.

[310] Wilf-Townsend, *supra* note 158, at 1816.

unlearning procedure. Tiffany Li's description of the "algorithmic shadow" captures this: the "persistent imprint of training data" is such that simply deleting data (or even performing the intended unlearning process) might not fully extinguish the data's influence from the model.[311] Indeed, a 2021 study demonstrated that machine unlearning can inadvertently create new privacy risks. By comparing an original model and an unlearned model, an attacker could infer whether a given data point had been in the initial training set, sometimes determining this more accurately than could have been figured from that first set alone—an ironic outcome.[312] Such findings where unlearning in letter still has left detectable knowledge in fact underscore that unlearning is not a magic wand. Regulators must understand the limitations: an order to "forget" data may need to be paired with validations that the model truly can no longer output or rely on that data. Technical experts may need to be involved in enforcement to bridge the understanding. Otherwise, a company might claim compliance by running a route unlearning algorithm, despite that the model could still indirectly reveal the "forgotten" information. This mismatch between legal expectation and technical reality is a call for careful oversight and possibly new standards (e.g., certification of unlearning efficacy).

3. Potential Overreach

Finally, there is a risk of overreliance on technical remedies in lieu of deeper accountability. If regulators lean too heavily on "compliance-by-design" mandates (e.g., requiring every AI system to have a built-in unlearning switch), firms might treat this as a box-checking exercise, focusing on the technical fix rather than the root cause of the privacy violation. Moreover, emphasizing post-hoc unlearning might divert attention from primary compliance obligations like obtaining valid consent, ensuring data quality, and preventing breaches in the first place. In other words, robust privacy governance might be overshadowed if organizations think "we can always unlearn later."

Similarly, mandating behavioral fixes like RLHF or external filtering without deeper data governance may create the illusion of compliance; these tools mask, but do not remove, the underlying data and therefore cannot alone discharge statutory erasure obligations.

Overzealous use of model deletion could also chill innovation: as Wilf-Townsend cautions, an uncalibrated deletion remedy can be disproportionate and unjust, especially if applied without regard to harm caused by the deletion.[313] Developers might fear that any minor data mistake could nuke an entire project.[314] Overreach in mandating unlearning could also raise practical enforcement dilemmas: for example, ordering deletion of a hugely popular algorithm (imagine

[311] Li, *supra* note 158, at 490-92.

[312] M. Chen et al., *supra* note 3, at 896-98.

[313] Wilf-Townsend, *supra* note 158, at 1841.

[314] *Id.*

forcing a public-facing AI service offline) could spark public backlash or have unintended societal costs.[315]

These limitations suggest a need for moderation and clear guidelines on when and which machine unlearning remedies are appropriate in a given remediation. Until then, regulators must guard against these complexities by using unlearning as part of a larger compliance toolkit, not a get-out-of-jail card for sloppy data practices. The goal at the privacy level is unchanged: to encourage accountability at every stage—from data collection to model deployment—rather than relying on after-the-fact purges.

D. Proposed Framework

To integrate machine unlearning into privacy enforcement effectively, we propose a flexible framework that spans the lifecycle of regulatory intervention. This framework treats unlearning as one tool among many to be deployed thoughtfully in investigations, settlements, and governance, and in coordination with global norms.

1. Investigations & Enforcement

During investigations or regulatory oversight (such as an FTC inquiry or State AG probe), agencies should leverage unlearning-related tools to assess compliance. This could include mandated data audits that require organizations to divulge the composition of training datasets and model inputs. By analyzing these, regulators can identify if protected personal data or unlawfully obtained information was used in training. If so, agencies may order interim relief such as a "data hold" (to prevent further training on suspect data) or demand disclosure of the model's unlearning capabilities. For instance, the FTC could ask: Do you have the technical ability to remove a consumer's data from your model upon request? If not, why not? Such questions not only signal expectations but also build a record of whether a company prepared for regulatory compliance. Investigators should request: (i) the provider's unlearning stack (heavy vs light methods, triggers, SLAs); (ii) architectural diagrams evidencing sharding/experts; (iii) determinism artifacts (e.g., SISA reproducible retrain hashes); and (iv) filter policies linked to specific takedowns. These materials allow agencies to test whether the system can actually forget and prove it.

In enforcement actions, agencies can also use machine unlearning as a detection tool. As an example, they could run their own experiments to see if a model produces personal data, which thus might indicate a failure to properly delete or "forget" it. U.S. regulators' coordination with GDPR authorities can be valuable here: EU data protection regulators might share helpful findings from audits (given that GDPR requires documentation of processing). A united front in investigations ensures that organizations operating internationally cannot present one face or

[315] *Id.*

argument to U.S. regulators and another to EU regulators; instead, they must meet a high standard of data and model hygiene in all jurisdictions.

2. Settlements & Remedial Agreements

When it comes to resolving a case, unlearning can be a 'surgical' remedy embedded in consent decrees or court orders. Rather than one-size-fits-all penalties, regulators should tailor settlements to require unlearning methodologies specific to the illegality of the data collection or use. For example, if a social media company's AI was trained on biometric data collected without consent, the settlement can obligate the company to retrain or adjust its models to purge that biometric influence rather than shutting down the AI entirely. Through its ability to pinpoint the violation, unlearning becomes a proportional remedy: it addresses the specific harm while allowing non-violative parts of the model to continue operating. Of course, if the taint of the unlawful activity is widespread or if the company lacks any mechanism to separate good from bad data, then broader model deletion may be warranted. The key is flexibility. Agreements might say "*to the extent feasible*, remove X's data from the model; if infeasible, then delete the model"—incentivizing companies to develop feasible unlearning methods preemptively. Consent orders should name permissible selective techniques (e.g., targeted fine-tuning, Fisher-guided perturbations, amnesiac pruning) and pair them with verification (e.g., pre/post benchmark suites; reproducible retrain manifests) and front-end gating.[316] Where a provider cannot meet these obligations within set timelines, fallback deletion (partial or full) should trigger, preserving proportionality while ensuring effectiveness.

Settlements can also impose forward-looking obligations, such as a requirement to honor any future deletion requests (perhaps under state privacy laws) by timely unlearning, subject to penalties for non-compliance. To enforce this, a consent order could last 20 years (as many FTC orders do) and include reporting provisions each time the company executes an unlearning action. Notably, Wilf-Townsend's suggestion to incorporate equitable factors can be operationalized here: settlements could explicitly take into account the company's intent (e.g., negligent vs. willful misuse of data) and adjust the stringency of the unlearning mandate accordingly. Global alignment is also crucial. A U.S. settlement should ideally require actions that satisfy GDPR expectations, too, so the company does not face inconsistent directives. We might see transatlantic cooperation where a U.S. order's remedial steps (like model retraining) are recognized by EU authorities as fulfilling an EU data erasure order, creating a more seamless compliance process for controllers.

3. Broader Governance

Beyond individual cases, integrating machine unlearning into the governance of AI involves policy coordination and standard-setting. Regulators should work

[316] *See* Neel et al., *supra* note 57; Golatkar et al., *supra* note 81; Eldan & Russinovich, *supra* note 71; Bourtoule et al., *supra* note 7; Thudi et al., *supra* note 5; Nguyen et al., *supra* note 40.

with organizations like the OECD and the Global Privacy Assembly to develop best practice guidelines on machine unlearning. This might yield certification frameworks for marking an AI system as "unlearning-compliant" (analogous to privacy seal programs). Also, U.S. agencies should coordinate with European Data Protection Authorities to issue joint guidance on how the GDPR's right to be forgotten can be fulfilled within an AI model context, providing necessary clarity that spans jurisdictions. Such coordination would help avoid a scenario where a company receives a deletion order in Europe but still faces uncertainty in the U.S. about whether to delete model data (potentially affecting U.S. consumers, too). A unified stance—or at least a mutual understanding—could be achieved through information-sharing agreements between the FTC and EU authorities, as well as through international fora. Additionally, unlearning should be part of industry standards and audits. For instance, the NIST AI Risk Management Framework could incorporate a guideline on data deletion and model update procedures. Sector-specific regulators (for example, HHS for health data) should also integrate unlearning principles into their promulgated rules for the AI systems under their purview, ideally via joint statements or in collaboration with other authorities for cross-sector consistency.

Ultimately, governance means setting the expectation that responsible AI development includes the ability to forget. By weaving this expectation into global policy documents and multi-agency strategies, regulators create a backdrop upon which machine unlearning is normalized as a component of accountability. This broader governance approach guards against regulatory arbitrage and helps foster technological solutions (perhaps new tools and services that specialize in efficient unlearning) that can serve many companies in complying with both U.S. and international privacy mandates.

V. CONCLUSION

Machine unlearning is poised to play a significant role in privacy law enforcement, but it must be understood as part of a broader privacy toolkit rather than a silver bullet. Implementing unlearning, whether via model deletion, selective retraining, or output suppression, can directly address the novel problem of AI systems' retaining personal data against individuals' wishes. It operationalizes the spirit of the "right to be forgotten" in the age of machine learning and offers regulators a tangible way to make controllers undo some of the harm from unlawful data collection and use. We have shown how it can complement traditional remedies by augmenting data deletion requirements so that models, too, are cleansed and how it can strengthen deterrence by preventing wrongdoers from easily profiting off of data misuse. At the same time, we have cautioned against over-reliance on these technical fixes. Unlearning should not become coverage for lax data practices, nor an automatic one-size-fits-all punishment without regard to its context. The goal of privacy-preserving machine unlearning should be to make use of these methodologies in a balanced manner by integrating them when it truly advances privacy interests and when the technical capability exists to do it effectively, or in tandem with other tools in the privacy law arsenal (fines, data use bans, etc.).

While crafting regulatory approaches to AI, it is important to continue refining the unlearning techniques themselves. Policymakers may need to invest in research and standards that improve our confidence in what unlearning can achieve so that legal mandates have the intended effect. Looking ahead, stronger collaboration between legal experts and computer scientists is vital in developing certifiable unlearning processes, much like data encryption has become a standard tool for security compliance. *Forget me not*—the question behind unlearning—is here to stay in privacy law. By operationalizing machine unlearning wisely within enforcement mechanisms, regulators can better hold organizations accountable for compliance with core privacy principles while ensuring that individuals' data truly fades from memory when required.