Machine Unlearning for AI Regulations

Anonymous Author(s)
Affiliation
Address
email

Abstract

The "right to be forgotten" and the data privacy laws that encode it have motivated machine unlearning since its earliest days. Now, some argue that an inbound wave of artificial intelligence regulations — like the European Union's Artificial Intelligence Act (AIA) — may offer important new use cases for machine unlearning. However, we argue this opportunity will only be realized if researchers proactively bridge the (sometimes sizable) gaps between machine unlearning's state of the art and its potential applications to AI regulation. To demonstrate this point, we use the AIA as an example. Specifically, we deliver a "state of the union" as regards machine unlearning's current potential (or, in many cases, lack thereof) for aiding compliance with the AIA. This starts with a precise cataloging of the potential applications of machine unlearning to AIA compliance. For each, we flag the technical gaps that exist between the potential application and the state of the art of machine unlearning. Finally, we end with a call to action: for machine learning researchers to solve the open technical questions that could unlock machine unlearning's potential to assist compliance with the AIA — and other AI regulation like it.

1 Introduction

Since its inception, Machine Unlearning (MU) has been motivated by the so-called "right to be forgotten" (RTBF) [11], which is encoded in data privacy laws like the European Union's General Data Protection Regulation (GDPR) [33, Art. 17]. Worldwide, multiple AI regulation efforts are working their way through the legislative process [7, 6, 128] or have graduated it and gone into effect [35, 22]. As these frameworks take shape, researchers have begun to explore whether MU can play a role in supporting compliance with these new AI regulations [75, 87, 43, 90, 85, 56, 60]. However, recent works call into question whether this motivation really holds water [24].

In this paper, we argue that MU's potential to assist compliance with AI regulation will only be realized if researchers close the technical gaps between MU's state of the art and this prospective new application.

We use the European Union's Artificial Intelligence Act (AIA) [35] as an example to support our argument. This starts with a thorough cataloging of the AIA requirements that MU can hypothetically assist compliance with. For each potential use case, we flag any legal ambiguities that lawmakers, in order to clarify MU's potential as an AIA compliance tool, should address when amending, updating, or translating the AIA into codes of practice or technical specifications. What is more, we scrutinize, from a technical perspective, whether the state of the art (SOTA) of MU can really support the hypothesized application. In many cases, we identify considerable gaps between the two. Finally, we conclude with a pointed call for the AI research community to take action and fill these gaps in order to help make MU a viable tool for assisting AI regulation compliance.

2 Machine Unlearning

To set the stage for our analysis, we set forth, in this section, we define and provide an overview of MU and its key concepts:

2.1 Formal Definition of Unlearning

Let $M = A(D)$ denote a model trained on dataset D using algorithm A. An **unlearning query** specifies a **forget-set** $D_f \subset D$, with the **retain-set** defined as $D_r = D \setminus D_f$. The goal of an unlearning algorithm U is to remove the influence of D_f from M, yielding an unlearned model $M_u = U(M; D_f, D_r)$. Depending on the approach, U may not require access to D_r [130].

Definition 2.1. Following [44], U is an (ϵ, δ)-**unlearner** if the distribution of $U(M; D_f, D_r)$ is (ϵ, δ)-close to that of $A(D_r)$. Specifically, two distributions μ and ν are (ϵ, δ)-close if for all measurable events B: $\mu(B) \leq e^{\epsilon}\nu(B) + \delta \quad \text{and} \quad \nu(B) \leq e^{\epsilon}\mu(B) + \delta$.

This definition provides a natural taxonomy for MU algorithms. When $\epsilon = \delta = 0$, U is termed **exact unlearning**; otherwise, it is referred to as **approximate unlearning**.

2.2 Informal Definitions

While Def. 2.1 provides a rigorous formulation of MU, researchers commonly use informal interpretations, typically phrased as `removing` x `from` M. However, deriving informal definitions directly from Def. 2.1 can be challenging, as the entity to remove may not be explicitly identifiable. For example, in generative models, x often corresponds to a fact or concept without explicit representation in M or D.

Additionally, MU is broadly applied to various methods, but overly general definitions introduce unnecessary complexity, potentially obstructing clear scientific discourse. Therefore, we restrict MU in this paper to ML techniques that explicitly modify the model's parameter-set (e.g., deletion and retraining, fine-tuning, parameter addition or removal). This scoped definition allows MU to remain a practical tool for applications such as safeguarding and alignment, while methods like guardrailing (or "output suppression" as per [24]) remain distinct, meriting separate evaluation in regulatory and other contexts.

2.3 Evaluation metrics

While Def. 2.1 is widely accepted in the MU community, it presents several challenges in practical settings. First, some works question whether this definition is necessary or sufficient to achieve true MU [108]. Second, in large-scale applications, it is computationally infeasible to directly measure the closeness between the distributions $A(D_r)$ and $U(M; D_f, D_r)$. As a result, researchers often resort to alternative proxies to verify MU. These proxies include performance metrics (e.g., classification accuracy [47] or generative performance using metrics like ROUGE for large language models [84]) and privacy attacks, such as membership inference attacks [57, 110].

2.4 Trade-offs and risks

MU algorithms strive balance three key objectives: **Model Utility**, **Forgetting Quality**, and **Efficiency**. In certain privacy-centric applications, forgetting can be synonymous with achieving privacy [79]. Hyperparameters and regularizers impact these trade-offs. For example, in MU via `Fine-tune`, the number of steps and learning rate dictate the balance between forgetting quality and efficiency. Similarly, in `Gradient Ascent`, the number of steps determines the trade-off between effective MU and preserving model's utility.

Additionally, forgetting may sometimes conflict with privacy due to two phenomena. First, unlearning specific data points can inadvertently expose information about others in the retained set due to the "onion effect" of privacy [13]. Second, over-forgetting [70] a data point may reveal its membership in the original training set—a phenomenon known as

the "Streisand Effect" [47]. Addressing these challenges requires careful calibration of MU methods to ensure a delicate equilibrium across these competing objectives.

Beyond these trade-offs, MU introduces risks associated with *untrusted parties* [73] and *malicious unlearning* [97]. Malicious entities could exploit MU to make fake deletion queries, or introduce computation overhead to systems [60].

3 The EU's Artificial Intelligence Act

The AIA sets forth requirements for AI systems and models placed on the market or put into service in the EU [35, Art. 1]. These requirements largely target two categories of AI: AI systems and general-purpose AI ("GPAI") models. Here, we define these categories and, for each, review some the AIA requirements that relate to the discussion at hand.

3.1 AI Systems

The AIA broadly defines AI systems to include any "machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" [35, Art. 3.1]. An example of a system that might meet this criteria is ChatGPT [38].

In laying out its rules for these AI systems, the AIA relies on a "risk-based" approach [83], by which an AI system's perceived degree of risk determines the exact rules that apply to it. Here, the strictest requirements — and the ones most relevant to our discussion — are reserved for those AI systems deemed to be *high-risk* [35, Art. 6]. Such high-risk AI ("HRAI") systems are subject to a bevy of requirements [35, Chap. III]. Among them, the following are the most relevant to us:

Risk management: HRAI systems must implement risk management systems that identify known and reasonably foreseeable risks that the system may pose to health and safety or to fundamental rights [35, 67, Art. 9.2(a)]. Here, risks to *health and safety* includes risks to mental and social well-being as well as physical safety. [2, 31]. Meanwhile, risks to *fundamental rights* includes, among other things, the right to non-discrimination [32], including from biased results [3].

Importantly, wherever these risks are identified, they should be "reasonably mitigated or eliminated through the development or design" of the AI system [35, Art. 9.2-3].

Accuracy and cybersecurity: HRAI systems must be designed and developed so as to achieve an "appropriate level" of accuracy and cybersecurity [35, Art. 15.1]. In its Recitals, the AIA stresses that these appropriate levels are a function of the system's intended purpose as well as the SOTA [35, Rec. 74]. When it comes to cybersecurity, the AIA specifically requires that HRAI systems be "resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities" [35, Art. 15(5)] and take technical measures to "prevent, detect, respond to, resolve and control for ... data poisoning" as well as "confidentiality attacks" [35, Art. 15.5].

3.2 GPAI models

In contrast to an AI system, a GPAI model is defined as any AI model that is "trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market" [35, Art. 3.63]. Some see this as being synonymous with "foundation model" [1]. An example of a GPAI model that might meet this criteria is GPT 3.5, the model powering ChatGPT [38].

In laying out its requirements for GPAI models, the AIA again uses a risk-based approach, with the strictest requirements reserved for GPAI models deemed to carry *systemic risk* [35,

Art. 55]. This is defined as the risk of "having a significant impact on the [EU] market due to [its] reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain" [35, Art. 2.65; Annex III]. This status can be established through proxies, including performance on benchmarks and the amount of compute used during training [35, Art. 51]. While the AIA itself does not further elaborate on what constitutes systemic risk, a companion piece to the AIA posits that it covers risks related to: (1) cyber offense; (2) chemical, biological, radiological, and nuclear (CBRN); (3) loss of control; (4) automated use of models for AI research and development; (5) persuasion and manipulation; and (6) large-scale discrimination [36].

Among the AIA's requirements for GPAI models that do display systemic risk — and those that don't — the following are the most relevant to our analysis:

Copyright: All GPAI model providers must "put in place a policy to comply with Union law on copyright and related rights" [35, Art. 53.c]. Among other things, this policy must respect rightsholders' requests, per [34, Art. 4(3)], to opt out of text and data mining (TDM) on their copyrighted works [35, Rec. 105, Art. 53.c].

Systemic risk: GPAI models that display systemic risk must "mitigate" it [35, Art. 55(a-b)]

Cybersecurity: GPAI models with systemic risk are additionally required to "ensure an adequate level of cybersecurity" [35, Art. 55(d)].

4 MU for AIA compliance: a catalog

This Section comprehensively catalogs the potential applications of MU to assist AIA compliance. For each, we analyze the SOTA and its ability to support the potential application, then identify any open questions the research community must resolve in order to bridge the gap between the two. In sum, we find that the potential applications of MU to assist AIA compliance ultimately roll up into just six separate applications (Fig. 1):

- **Accuracy:** Improve accuracy per EU [35, Arts. 9, 15];
- **Bias:** Mitigate bias per EU [35, Arts. 9, 55];
- **Privacy Attack:** Mitigate confidentiality attacks per EU [35, Arts. 9, 15, 55]);
- **Data Poisoning:** Mitigate data poisoning per EU [35, Arts. 15]);
- **GenAI risk:** Mitigate other risks of generative outputs per EU [35, Arts. 9, 55]);
- **Copyright:** Aid compliance with copyright laws, per EU [35, Art. 53])

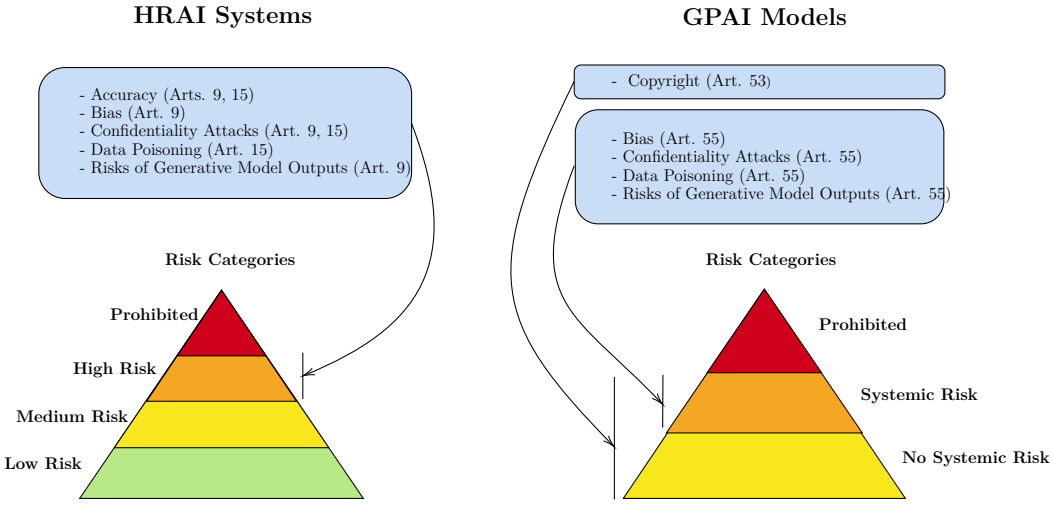

Figure 1: AIA Uses Cases for Machine Unlearning.

4.1 Accuracy

Two AIA provisions may compel HRAI systems providers towards higher levels of accuracy. a), HRAI systems must achieve a level of accuracy appropriate to their intended use and the SOTA [35, Art. 15.1; Rec. 74]. b) , HRAI systems' risk management must include measures to mitigate/eliminate risks to health and safety [35, Art. 9], which could stem from low accuracy in domains like medicine [65, 64, 58]. MU can boost accuracy by removing the problematic data from the model.

The accuracy use case should not require privacy guarantees on the unlearned data [45]. Because the goal is strictly to boost accuracy to the level deemed appropriate [35, Art. 15] or until the overall residual risk to health and safety posed by the inaccuracy is judged to be "acceptable" [35, Art. 9]. In measuring that, AI providers will presumably account for any inadvertent, counteracting degradation in accuracy caused by the MU itself [72, 10, 121].

Current SOTA MU theoretically offer paths towards improving model accuracy by forgetting mislabeled [45, 107, 12, 15, 52], out-of-date and outlier training data points [69, 122, 121, 71], or, potentially, removing noise from medical data [96, 28]. The largest hurdle for this use case might be identifying all of the data points that are leading to inaccuracy (e.g., the mislabeled examples), which can be difficult [46]. It may be good enough to identify only a subset of these examples — so long as accuracy is boosted to levels deemed "appropriate" in light of the intended purpose as well as the SOTA [35, Art. 15.1; Rec. 74]. MU based on subset forget sets have shown success in boosting accuracy. However, other studies have suggested that you need all of polluted data, not just some of this, or it might backfire [46]. It is also important to note that evaluating unlearning success is application dependent. And, that, approximate unlearning should not be expected to yield higher accuracy than exact retraining without the low-quality data.

Key Points: (i) Multiple AIA requirements may benefit from MU. (ii) Theoretical guarantees may not be needed. (iii) Evaluation measure is application-dependent.

Open Problems: (i) Lack of reliable methods for identifying problematic data to unlearn. (ii) Lack of controllability over trade-offs

4.2 Bias

Providers of both HRAI systems and GPAI models with systemic risk must mitigate certain types of model bias. The former must take measures to mitigate or eliminate risks to fundamental rights, which includes the right to non-discrimination [35, Arts. 9]). The latter must take steps to mitigate their models' systemic risk [35, Art. 55], which includes the risk of large-scale discrimination [36]. Bias can occur because unrepresentative or incomplete data prevent the model from perform fairly on different groups or, in the case of generative models, cause it to produce stereotyped or otherwise discriminatory outputs [39]. In all cases, MU can ostensibly help forget the data points or training data patterns causing the bias [95, 71, 101, 68, 15, 132]. An important limiting factor on this use case is that training data that is not there to begin with cannot be forgotten; if the bias is due to a data *deficit*, MU will not help. Because the goal here is to reduce or eradicate bias, success should ultimately be measured using traditional bias metrics like the difference in performance on various subgroups [25] or, in the case of generative models, the propensity for biased outputs as measured with benchmarks [93].

Current SOTA Model debiasing has a longer history than MU. [125], . Recently, both exact [54]) and approximate [19, 91, 54]) MU solutions have been offered to mitigate model biases. In the debiasing literature, solutions include *pre-processing*, *in-processing*, and *post-processing* methods [86]. MU, can mainly be considered as a post-processing method. However, it is difficult to draw a separating line between debiasing and MU methods. MU works usually re-use some of the evaluation metrics in the debiasing literature, however, how to evaluate bias is, generally, considered an "open problem" [99]. In order to preserve accuracy by not forgetting data points holistically, [120] use MU to forget only those the features that lead to bias.

Key Points: (i) MU may aid compliance with multiple bias-related AIA requirements. (ii) MU is only subtractive and never additive, limiting its application to this use case. (iii) De-biasing solutions are not limited to MU

Open Problems (i) Lack of methods for identifying bias counterfactuals. (i) Lack of controllability over trade-offs. (iii) Difficulty of guaranteeing full unlearning of biases, due to generalization.

4.3 Confidentiality attacks

The AIA requires providers of both HRAI systems and GPAI models with systemic risk to resolve and control for confidentiality attacks. Providers of HRAI systems must ensure their systems achieve an "appropriate level" of cybersecurity given the intended use and the SOTA, including by taking technical measures to prevent, detect, respond to, resolve and control for confidentiality attacks [35, Art. 15.5; Rec. 74]. Meanwhile, providers of GPAI models with systemic risk must ensure those models reflect "an adequate level of cybersecurity" [35, Art. 55.d], which presumably also includes defending against confidentiality attacks. While the AIA does not define confidentiality attacks, we take them to include any attacks, including data reconstruction and membership inference attacks, that cause a model to reveal confidential details about its training such as data points or membership [112, 21]. This may include confidential training data memorized by generative models Cooper et al. [24], Gu et al. [50], Łucki et al. [132]. Where such attacks occur — or where there is reason to think they might — MU can ostensibly help defend against them by forgetting the confidential information vulnerable to attack [60, 71, 13, 17, 122, 99, 5].For this use case, the measure of success should be whether confidentiality attacks succeed in the wake of the MU [49], though use case-specific metrics have been developed Maini et al. [84]. When it comes to this use case, there are, importantly, other viable options for protecting against confidentiality attacks, including training with DP [114, 66].

Current SOTA. Multiple MU techniques have been proposed to mitigate confidentiality attacks (or the related problem of inadvertent model leakage of personal data) [26, 4, 81, 16, 115, 11, 10, 9]. As is, applying MU to this use case can carry sizable trade-offs. For example, unlearning some data points for the sake of protecting them from recovery by attackers can jeopardize the privacy of other data points that neighbor the unlearned ones Carlini et al. [13] or even increase the risk of membership inference attacks that recover the unlearned data points Chen et al. [18], Barez et al. [5], Kurmanji et al. [70]. Differently, approximate unlearning, when used to delete particular data points, can carry a bias trade-off [126, 92] and an accuracy trade-off that rises as more data is forgotten [48, 84]. It is also important to note that current MU methods usually fail on new emergent attacks that are devised with new assumptions [129, 62].

Key Points (i) MU may aid compliance with multiple confidentiality attack-related AIA requirements. (ii) Due to attack diversity, success should be measured on case-by-case basis. (iii) DP is a strong alternative to MU for this use case.

Open Problems (i) Difficulty of providing formal guarantees of attack susceptibility. (ii) Difficulty of applying MU to new, emergent attacks. (iii) Identifying, localizing, and measuring memorization of confidential data.

4.4 Data poisoning

In data poisoning, specially-crafted data points are injected into a training set to alter (e.g., degrade or bias) model behavior to the attacker's benefit [8]. Backdoor attacks are a type of data poisoning where the injected data points create "triggers" the attacker can exploit during inference [76]. The AIA obligates the providers of both HRAI systems and GPAI model with systemic risk to address such attacks. HRAI system providers must ensure their systems achieve an "appropriate level" of cybersecurity, including via technical measures to "prevent, detect, respond to, resolve and control for" data poisoning attacks [35, Art. 15(5)]. Providers of GPAI models with systemic risk, meanwhile, must "ensure an adequate level

of cybersecurity" in their models [35, Art. 55.d], which presumably also includes defenses against data poisoning. Where it is known that data poisoning has (or could) occur, MU may help remove the effects of the poisoned data points on the model and, thus, help satisfy these requirements [121, 80, 11, 12, 104]. When it comes to measuring success for this use case, because the "primary goal is to unlearn the adverse effect due to the manipulated data," the ideal benchmark would seem to be whether those adverse effects — be they vulnerability to backdoor triggers, bias, or lower accuracy — are eliminated or reduced [46]. For example, Goel et al. [46] measure MU efficacy based on whether proper accuracy on backdoor triggers is restored.

Current SOTA Though some work has demonstrated MU can succeed for this use case [116, 102] other works question the effectiveness of using MU to address data poisoning or backdoor attacks specifically [51, 109, 94, 122]). As always, identifying the full forget set (here, the poisoned samples) remains challenging [46]. Some methods, moreover, can have a significant accuracy trade-off on this use case [94]. Such trade-offs can be particularly difficult as poisoned data overlaps with the clean data and, in most cases, they are even visually indistinguishable from each other.

Key Points (i) MU may aid compliance with several data poisoning-related AIA requirements. (ii) A proper benchmark should measure the elimination of adverse effects.

Open Problems (i) Finding contaminated data at scale is challenging. (ii) Unlearning the backdoor pattern without hurting unaffected data is challenging. (iii) Current MU methods mostly fail on data poisoning use case.

4.5 Other risks of generative outputs

Generative outputs may pose risks to health, safety, and human rights or pose systemic risk that providers of HRAI systems and GPAI models, respectively, must mitigate. For example, HRAI systems' risk management systems must strive to mitigate or eliminate risks the system poses to health, safety, and fundamental rights [35, Art. 9]. Generative outputs may pose risks to health and safety, e.g., by issuing bad medical advice [118, 55]), and may pose risks to the fundamental right of non-discrimination, e.g., by producing stereotyping outputs [88]. For GPAI models with systemic risk, providers of such models must mitigate that risk [35, Art. 55], which could be brought on by generative model outputs that display offensive cyber capabilities, knowledge of CBRN, and more [36, 89]. In all these cases, MU may help mitigate the non-compliant outputs by unlearning the data points or even the concepts in the training set that are causing them [132, 24]. Computationally, it may offer advantages even as compared to other popular alignment techniques like reinforcement learning [123]. Measuring success for this use case should arguably be "context dependent" [123]. That is, the best way to measure the MU's efficacy is to benchmark the exact behavior that we desire to repair [123]. This could utilize existing benchmarks unrelated to MU [5]. Differently, [74] propose a benchmark for measuring MU of CBRN knowledge and approaches that examine the model parameters for remnants of the unlearned concepts have also been proposed [61].

Current SOTA Multiple works use MU to curb undesirable generative model outputs [27, 124, 117, 41]. However, the task is difficult, without agreed-upon best practices [24]. Broad concepts like non-discrimination tend to go beyond individual data points, to latent information which is not easily embodied as a discrete forget set [24, 77]. Even if data points that are intrinsically harmful (e.g., the molecular structure of a bioweapon) are removed, models may still assemble dangerous outputs from latent information in the rest of the dataset [24, 111]. Trying to remove that latent knowledge can risk model utility [24]. As a separate but related issue, AI systems in these scenarios may be dual-use, where the appropriateness of outputs depends on downstream context; this, too, makes identifying the forget set difficult and increases the likelihood of a utility trade-off as the model forgets desirable knowledge alongside undesirable knowledge [24, 105, 99]. All of these issues, in turn, make it difficult if not impossible to specify formal guarantees on the MU [77].

Key Points (i) MU may aid compliance with several AIA requirements related to generative outputs.

Open Problems (i) Defining the forget set when what we seek to forget is conceptual. (ii) Difficulty of guaranteeing full unlearning of unwanted behaviors, due to generalization. (iii) Mitigating the forgetting of useful knowledge alongside undesirable knowledge.

4.6 Copyright

All GPAI model providers must have a policy for complying with EU copyright law [35, Art. 53.c]. Among other things, this policy must honor the TDM opt-outs of rightsholders [35, Art. 53.c; Rec. 105], which is often a feature of AI training [100, 53]. When it comes to AI and copyright law, a distinction is sometimes made between the "input" (training) phase and the "output" (inference) phase of the AI life cycle [100, 98]. At this point in time, the primary compliance risk during the input phase seems to be that an AI training set could include data points that violate TDM opt-outs. When this happens, we assume that using MU to remove the opt-out data points from the trained model does not cure the violation, since the violation occurred at the moment the opt-out data was used for training. That said, MU may still represent a valuable component of a copyright-compliance policy by helping prevent, at the "output" phase, further violations of copyright law when the opt-out data points — or any other copyright-protected data points in the training set — are reproduced to some degree in model outputs [100]. This is a real risk with generative models, which often memorize training data [23, 14]. When MU is applied to this use case, we may measure success by tracking how likely the model is to generate works that are sufficiently similar to the copyrighted works. For example, we might rely on existing benchmarks that measure the tendency of models to produce copyrighted materials [78, 20]. Differently, Ma et al. [82] produce a benchmark for the success of MU in the copyright context.

Current SOTA Wu et al. [119] unlearn copyrighted works from diffusion models. At first glance, exact MU would seem to provide a guarantee that copyrighted works in the training set will not be reproduced in outputs [77]. But the fact is that retraining from scratch without the copyrighted data may not be a bulletproof solution for preventing copyright infringement in outputs because substantially similar representations of copyrighted "expressions" (e.g., images of characters like Spiderman) could still appear in outputs based on how the model generalizes from the latent information extracted from the rest of the training set Cooper et al. [24]. For the same reason, approximate unlearning aimed at removing the influence of the copyright data points on the model, on top of being hard to prove [77], also cannot ensure that copyrights are not infringed by outputs. In general, the SOTA of approximate unlearning has been deemed "insufficient" for the copyright use case, which may be why practitioners currently lean towards pre- and post-processing tools like prompting and moderation to bring AI into compliance with these laws [77, 106]. [29] "unlearn" copyrighted materials in LLM pre-training datasets by identifying and removing specific weight updates in the model's parameters that correspond to copyrighted content, evaluating their method by measuring the similarities between the model's outputs and the original content. The task of measuring whether substantially similar outputs are being produced is quite challenging [24].

Key Points (i) MU does not help with TDM opt-out violations; the damage is already done. (ii) MU may, however, help with downstream copyright violations in outputs. (iii) To avoid malicious unlearning, TDM opt-outs will have to be verified.

Open Problems (i) Difficulty in identifying copyright-infringing works in a dataset. (ii) Difficulty of verifying whether model output owes to copyrighted data or generalization. (iii) Localizing and measuring memorization of copyrighted data is itself an open problem.

5 Discussion

MU might offer potential solutions for some of AIA compliance requirements, but it is not a silver bullet. Throughout this work, we have balanced enthusiasm for MU's capabilities

with a clear-eyed view of its limitations. A recurring challenge across use cases—such as accuracy, bias, and confidentiality—is the difficulty of identifying and isolating harmful or low-quality data. In modern AI models, such information is often encoded in distributed representations, making precise removal difficult and risking forgetting useful knowledge.

In many cases, the target of unlearning (e.g., a fact or concept) lacks a discrete representation. Still, recent work in generative models shows promise: concept editing in diffusion models [59], data attribution [113], and inversion-based techniques [42] all offer ways to trace and remove implicit or emergent representations. Another important challenge is verification [108]. Approximate MU methods currently offer limited guarantees, complicating auditing. Going forward, we advocate for the development of formal forgetting guarantees that can underpin regulator-endorsed standards.

While some applications—like correcting mislabeled data to improve accuracy [69]—are feasible with today's methods, others (e.g., bias mitigation or copyright control) face steeper barriers. In some cases, MU may be an unnecessarily complex solution relative to alternatives. However, overlaps between applications (e.g., boosting both fairness and accuracy) suggest that well-designed MU interventions could serve multiple regulatory goals simultaneously.

6 Conclusion

There are still sizable challenges that must be cleared before MU will be a viable tool for assisting compliance with the AIA (and, by extension, since AI regulations tend to feature recurring principles [37, 30], other AI regulations). To realize MU's potential for these use cases, AI researchers should help solve the open technical problems logged by this paper. Among other things, this includes work on identifying forget set data points, on resolving the privacy and performance trade-offs of MU, and on resolving the particular challenges to these use cases that generative model outputs present. Working collaboratively, we can all help unlock MU's potential to assist compliance with AI regulation and, by extension, help safeguard the important social values these regulations encode.

7 Related works

The arguments against using MU as a tool for compliance with the AIA or other AI regulation would likely point to its shortcomings, trade-offs, and risks as well as the viable substitutes for MU in these scenarios. Some recent works, for example, broadly question whether MU can really achieve its goals, especially in the generative domain [24, 5, 131, 106]. Other works scrutinize MU's trade-offs around performance, privacy, security, and cost [121, 13, 127, 68, 131, 40]. These factors could reasonably make alternative methods, training with DP, or post-training alignment tuning more appealing for the AI regulation use cases highlighted in this paper Łucki et al. [132], Cooper et al. [24].

These alternative approaches come with their own limitations. For instance, while some may consider DP [63] as a strong alternative to MU, several caveats deserve attention. First, DP mechanisms often struggle to balance tight privacy guarantees with acceptable model utility [103]. This trade-off becomes especially pronounced in high-utility applications. Second, unlike traditional privacy settings where protection is applied uniformly across all data points, MU typically targets a specific subset of data—the so-called "forget set." In large-scale training corpora that combine individually identifiable data with more publicly available content, applying DP globally may offer overly broad protections that are both inefficient and unnecessary [47]. Third, there are use cases where DP is not sufficient or optimal. For instance, if the objective is to remove a harmful or undesired behavior from a generative model (e.g., misinformation, bias, or offensive content), a DP-trained model may still require explicit MU interventions to mitigate such behaviors.

References

[1] Ada Lovelace Institute. Foundation models and general purpose AI systems: Understanding impacts and implications. Project report, Ada Lovelace Institute, 2024. URL `https://www.adalovelaceinstitute.org/project/foundation-models-gpai/`.

[2] A. Armstrong, R. Butler, and K. Gambrell. Ai and product safety standards under the EU AI Act. Research paper, Carnegie Endowment for International Peace, March 2024. URL https://carnegieendowment.org/research/2024/03/ai-and-product-safety-standards-under-the-eu-ai-act.

[3] L. Arnold. How the European Union's AI Act provides Insufficient Protection Against Police Discrimination. *University of Pennsylvania Carey Law School News*, May 2024. URL https://www.law.upenn.edu/live/news/16742-how-the-european-unions-ai-act-provides.

[4] T. Ashuach, M. Tutek, and Y. Belinkov. Revs: Unlearning sensitive information in language models via rank editing in the vocabulary space, 2024. URL https://arxiv.org/abs/2406.09325.

[5] F. Barez, T. Fu, A. Prabhu, S. Casper, A. Sanyal, A. Bibi, A. O'Gara, R. Kirk, B. Bucknall, T. Fist, L. Ong, P. Torr, K.-Y. Lam, R. Trager, D. Krueger, S. Mindermann, J. Hernandez-Orallo, M. Geva, and Y. Gal. Open problems in machine unlearning for ai safety, 2025. URL https://arxiv.org/abs/2501.04952.

[6] J. Beardwood. The Canadian Artificial Intelligence and Data Act and the EU AI Act: Will sanity prevail as they more closely align? – Part 2 — Changes to both Acts bring them closer together... but not too close. *Computer Law Review International*, 25(5):129–137, 2024. doi: doi:10.9785/cri-2024-250501. URL https://doi.org/10.9785/cri-2024-250501.

[7] L. Belli, Y. Curzi, and W. B. Gaspar. AI regulation in Brazil: Advancements, flows, and need to learn from the data protection experience. *Computer Law & Security Review*, 48: 105767, 2023. ISSN 0267-3649. doi: https://doi.org/10.1016/j.clsr.2022.105767. URL https://www.sciencedirect.com/science/article/pii/S0267364922001108.

[8] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL http://icml.cc/2012/papers/880.pdf.

[9] J. Borkar. What can we learn from data leakage and unlearning for law? *CoRR*, abs/2307.10476, 2023. doi: 10.48550/ARXIV.2307.10476. URL https://doi.org/10.48550/arXiv.2307.10476.

[10] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019. URL https://doi.org/10.1109/SP40001.2021.00019.

[11] Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015. doi: 10.1109/SP.2015.35.

[12] Y. Cao, A. F. Yu, A. Aday, E. Stahl, J. Merwine, and J. Yang. Efficient repair of polluted machine learning systems via causal unlearning. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, page 735–747, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355766. doi: 10.1145/3196494.3196517. URL https://doi.org/10.1145/3196494.3196517.

[13] N. Carlini, M. Jagielski, C. Zhang, N. Papernot, A. Terzis, and F. Tramer. The privacy onion effect: Memorization is relative, 2022. URL https://arxiv.org/abs/2206.10469.

[14] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models, 2023. URL https://arxiv.org/abs/2301.13188.

[15] J. Chen and D. Yang. Unlearn what you want to forget: Efficient unlearning for llms, 2023. URL https://arxiv.org/abs/2310.20150.

[16] K. Chen, Y. Wang, L. Zhao, C. Jiang, H. Mai, Y. Wu, H. Hong, Y. Shen, J. Mo, L.-L. Huang, J. Peng, X. Wang, and Q. Yang. Private data protection with machine unlearning for next-generation networks. *IEEE Open Journal of the Communications Society*, PP:1–1, 01 2024. doi: 10.1109/OJCOMS.2024.3518503.

[17] K. Chen, Z. Wang, and B. Mi. Private data protection with machine unlearning in contrastive learning networks. *Mathematics*, 12(24), 2024. ISSN 2227-7390. doi: 10.3390/math12244001. URL https://www.mdpi.com/2227-7390/12/24/4001.

[18] M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 896–911. ACM, Nov. 2021. doi: 10.1145/3460120.3484756. URL http://dx.doi.org/10.1145/3460120.3484756.

[19] R. Chen, J. Yang, H. Xiong, J. Bai, T. Hu, J. Hao, Y. Feng, J. T. Zhou, J. Wu, and Z. Liu. Fast model debias with machine unlearning, 2023. URL https://arxiv.org/abs/2310.12560.

[20] T. Chen, A. Asai, N. Mireshghallah, S. Min, J. Grimmelmann, Y. Choi, H. Hajishirzi, L. Zettlemoyer, and P. W. Koh. Copybench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation, 2024. URL https://arxiv.org/abs/2407.07087.

[21] CLTC. Adversarial machine learning, 2024. URL https://cltc.berkeley.edu/aml/. CLTC Research Guide.

[22] Colorado GA. Artificial intelligence regulation and disclosure act, May 2024. URL https://leg.colorado.gov/bills/sb24-205. Senate Bill 24-205.

[23] A. F. Cooper and J. Grimmelmann. The files are in the computer: On copyright, memorization, and generative AI. *Chicago-Kent Law Review*, 2024. forthcoming.

[24] A. F. Cooper, C. A. Choquette-Choo, M. Bogen, M. Jagielski, K. Filippova, K. Z. Liu, A. Chouldechova, J. Hayes, Y. Huang, N. Mireshghallah, I. Shumailov, E. Triantafillou, P. Kairouz, N. Mitchell, P. Liang, D. E. Ho, Y. Choi, S. Koyejo, F. Delgado, J. Grimmelmann, V. Shmatikov, C. D. Sa, S. Barocas, A. Cyphert, M. Lemley, danah boyd, J. W. Vaughan, M. Brundage, D. Bau, S. Neel, A. Z. Jacobs, A. Terzis, H. Wallach, N. Papernot, and K. Lee. Machine unlearning doesn't do what you think: Lessons for generative ai policy, research, and practice, 2024. URL https://arxiv.org/abs/2412.06966.

[25] D. DeAlcala, I. Serna, A. Morales, J. Fierrez, and J. Ortega-Garcia. Measuring bias in ai models: An statistical approach introducing n-sigma, 2023. URL https://arxiv.org/abs/2304.13680.

[26] G. Dhingra, S. Sood, Z. M. Wase, A. Bahga, and V. K. Madisetti. Protecting LLMs against privacy attacks while preserving utility. *Journal of Information Security*, 15:448–473, 2024. doi: 10.4236/jis.2024.154026.

[27] O. Dige, D. Arneja, T. Yau, Q. Zhang, M. Bolandraftar, X. Zhu, and F. Khattak. Can machine unlearning reduce social bias in language models? pages 954–969, 01 2024. doi: 10.18653/v1/2024.emnlp-industry.71.

[28] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete. Unlearning scanner bias for mri harmonisation in medical image segmentation. In *Medical Image Understanding and Analysis: 24th Annual Conference, MIUA 2020, Oxford, UK, July 15-17, 2020, Proceedings 24*, pages 15–25. Springer, 2020.

[29] G. Dou, Z. Liu, Q. Lyu, K. Ding, and E. Wong. Avoiding copyright infringement via large language model unlearning, 2024. URL https://arxiv.org/abs/2406.10952.

[30] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101896. URL https://www.sciencedirect.com/science/article/pii/S1566253523002129.

[31] EC. Questions and answers: Coordinated plan on artificial intelligence 2021 review. Press Release QANDA/21/1683, European Commission, April 2021. URL https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683.

[32] EU. Charter of fundamental rights of the European Union, 2000. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT. Official Journal of the European Communities, C 364/1.

[33] EU. General data protection regulation (gdpr), April 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679. Official Journal of the European Union, L 119/1.

[34] EU. Directive (EU) 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, April 2019. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790. Official Journal of the European Union, L 130/92.

[35] EU. Artificial intelligence act, March 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. Official Journal of the European Union.

[36] EU AI Office. First draft of the general-purpose AI code of practice. Policy document, European Union, November 2024. Independent expert draft for stakeholder consultation.

[37] S. Feldstein. Evaluating Europe's push to enact AI regulations: How will this influence global norms? *Democratization*, 31(5):1049–1066, 2024. doi: 10.1080/13510347.2023.2196068. URL https://doi.org/10.1080/13510347.2023.2196068.

[38] D. Fernández-Llorca, E. Gómez, I. Sánchez, et al. An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI. *Artificial Intelligence and Law*, 2024. doi: 10.1007/s10506-024-09412-y. URL https://doi.org/10.1007/s10506-024-09412-y.

[39] E. Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1), 2024. ISSN 2413-4155. doi: 10.3390/sci6010003. URL https://www.mdpi.com/2413-4155/6/1/3.

[40] L. Floridi. Machine unlearning: Its nature, scope, and importance for a "delete culture". *Philosophy & Technology*, 36(42), 2023. doi: 10.1007/s13347-023-00644-5.

[41] M. Fore, S. Singh, C. Lee, A. Pandey, A. Anastasopoulos, and D. Stamoulis. Unlearning climate misinformation in large language models, 2024. URL https://arxiv.org/abs/2405.19563.

[42] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[43] J. Geng, Q. Li, H. Woisetschlaeger, Z. Chen, Y. Wang, P. Nakov, H.-A. Jacobsen, and F. Karray. A comprehensive survey of machine unlearning techniques for large language models, 2025. URL https://arxiv.org/abs/2503.01854.

[44] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

[45] S. Goel, A. Prabhu, A. Sanyal, S.-N. Lim, P. Torr, and P. Kumaraguru. Towards adversarial evaluations for inexact machine unlearning, 2023. URL https://arxiv.org/abs/2201.06640.

[46] S. Goel, A. Prabhu, P. Torr, P. Kumaraguru, and A. Sanyal. Corrective machine unlearning, 2024. URL https://arxiv.org/abs/2402.14015.

[47] A. Golatkar, A. Achille, and S. Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020.

[48] L. Graves, V. Nagisetty, and V. Ganesh. Amnesiac machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11516–11524, May 2021. doi: 10.1609/aaai.v35i13.17371. URL https://ojs.aaai.org/index.php/AAAI/article/view/17371.

[49] K. Grimes, C. Abidi, C. Frank, and S. Gallagher. Gone but not forgotten: Improved benchmarks for machine unlearning, 2024. URL https://arxiv.org/abs/2405.19211.

[50] K. Gu, M. R. U. Rashid, N. Sultana, and S. Mehnaz. Second-order information matters: Revisiting machine unlearning for large language models, 2024. URL https://arxiv.org/abs/2403.10557.

[51] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. doi: 10.1109/ACCESS.2019.2909068.

[52] E. Gündogdu, A. Unal, and G. Unal. A study regarding machine unlearning on facial attribute data. In *18th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2024, Istanbul, Turkey, May 27-31, 2024*, pages 1–5. IEEE, 2024. doi: 10.1109/FG59268.2024.10581972. URL https://doi.org/10.1109/FG59268.2024.10581972.

[53] Hamburg Regional Court. Hamburg regional court, germany [2024]: Robert kneschke v. laion e.v., case no. 310 o 227/23, September 2024. URL https://www.wipo.int/wipolex/en/judgments/details/2381. Judgment concerning copyright and text and data mining exceptions under the DSM Directive and German law. Part of the 2024 WIPO Intellectual Property Judges Forum collection.

[54] L. Han, H. Huang, D. Scheinost, M. Hartley, and M. R. Martínez. Unlearning information bottleneck: Machine unlearning of systematic patterns and biases. *CoRR*, abs/2405.14020, 2024. doi: 10.48550/ARXIV.2405.14020. URL https://doi.org/10.48550/arXiv.2405.14020.

[55] T. Han, S. Nebelung, F. Khader, et al. Medical large language models are susceptible to targeted misinformation attacks. *npj Digital Medicine*, 7:288, 2024. doi: 10.1038/s41746-024-01282-7.

[56] A. Hatua, T. T. Nguyen, F. Cano, and A. H. Sung. Machine unlearning using forgetting neural networks, 2024. URL https://arxiv.org/abs/2410.22374.

[57] J. Hayes, I. Shumailov, E. Triantafillou, A. Khalifa, and N. Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. *arXiv preprint arXiv:2403.01218*, 2024.

[58] W. D. Heaven. Hundreds of ai tools have been built to catch covid. none of them helped. *MIT Technology Review*, July 2021. URL https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.

[59] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv.org/abs/2208.01626*, 1, 2022.

[60] E. Hine, C. Novelli, M. Taddeo, et al. Supporting trustworthy ai through machine unlearning. *Science and Engineering Ethics*, 30:43, 2024. doi: 10.1007/s11948-024-00500-5. URL https://doi.org/10.1007/s11948-024-00500-5.

[61] Y. Hong, L. Yu, H. Yang, S. Ravfogel, and M. Geva. Intrinsic evaluation of unlearning using parametric knowledge traces, 2024. URL https://arxiv.org/abs/2406.11614.

[62] S. Hu, Y. Fu, S. Wu, and V. Smith. Jogging the memory of unlearned models through targeted relearning attacks. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[63] Y. Huang and C. L. Canonne. Tight bounds for machine unlearning via differential privacy. *arXiv preprint arXiv:2309.00886*, 2023.

[64] C. James, J. Ranson, R. Everson, and D. Llewellyn. Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients. *JAMA Network Open*, 4:e2136553, 12 2021. doi: 10.1001/jamanetworkopen.2021.36553.

[65] K. R. Jongsma, M. Sand, and M. Milota. Why we should not mistake accuracy of medical AI for efficiency. *NPJ Digital Medicine*, 7(1):57, mar 2024. doi: 10.1038/s41746-024-01047-2.

[66] G. Kaissis, J. Hayes, A. Ziller, and D. Rueckert. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy, 2023. URL https://arxiv.org/abs/2307.03928.

[67] M. E. Kaminski. Legal fictions in the AI Act. *Boston University Law Review*, 103, November 2023. URL https://www.bu.edu/bulawreview/files/2023/11/KAMINSKI.pdf.

[68] S. Keskpaik. Machine unlearning. Techsonar report, European Data Protection Supervisor, jan 2024. URL https://edps.europa.eu/techsonar/machine-unlearning. TechSonar Series.

[69] M. Kurmanji, E. Triantafillou, and P. Triantafillou. Machine unlearning in learned databases: An experimental analysis, 2023. URL https://arxiv.org/abs/2311.17276.

[70] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou. Towards unbounded machine unlearning, 2023. URL https://arxiv.org/abs/2302.09880.

[71] R. Layne. How to make ai 'forget' all the private data it shouldn't have, February 2024. URL https://www.library.hbs.edu/working-knowledge/qa-seth-neel-on-machine-unlearning-and-the-right-to-be-forgotten. Featuring Seth Neel, discussing machine unlearning and data privacy.

[72] C. Li, H. Jiang, J. Chen, Y. Zhao, S. Fu, F. Jing, and Y. Guo. An overview of machine unlearning. *High-Confidence Computing*, page 100254, 2024. ISSN 2667-2952. doi: https://doi.org/10.1016/j.hcc.2024.100254. URL https://www.sciencedirect.com/science/article/pii/S2667295224000576.

[73] L. Li, X. Ren, H. Yan, X. Liu, and Z. Zhang. Pseudo unlearning via sample swapping with hash. *Information Sciences*, 662:120135, 2024.

[74] N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, L. Phan, G. Mukobi, N. Helm-Burger, R. Lababidi, L. Justen, A. B. Liu, M. Chen, I. Barrass, O. Zhang, X. Zhu, R. Tamirisa, B. Bharathi, A. Khoja, Z. Zhao, A. Herbert-Voss, C. B. Breuer, S. Marks, O. Patel, A. Zou, M. Mazeika, Z. Wang, P. Oswal, W. Lin, A. A. Hunt, J. Tienken-Harder, K. Y. Shih, K. Talley, J. Guan, R. Kaplan, I. Steneker, D. Campbell, B. Jokubaitis, A. Levinson, J. Wang, W. Qian, K. K. Karmakar, S. Basart, S. Fitz, M. Levine, P. Kumaraguru, U. Tupakula, V. Varadharajan, R. Wang, Y. Shoshitaishvili, J. Ba, K. M. Esvelt, A. Wang, and D. Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL https://arxiv.org/abs/2403.03218.

[75] N. Li, C. Zhou, Y. Gao, H. Chen, Z. Zhang, B. Kuang, and A. Fu. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2025. doi: 10.1109/TNNLS.2025.3530988.

[76] J. Lin, L. Dang, M. Rahouti, and K. Xiong. Ml attack models: Adversarial attacks and data poisoning attacks, 2021. URL https://arxiv.org/abs/2112.02797.

[77] K. Liu. Machine unlearning: What it is and why it matters, 2023. URL https://ai.stanford.edu/~kzliu/blog/unlearning.

[78] X. Liu, T. Sun, T. Xu, F. Wu, C. Wang, X. Wang, and J. Gao. Shield: Evaluation and defense strategies for copyright compliance in llm text generation, 2024. URL https://arxiv.org/abs/2406.12975.

[79] Z. Liu, G. Dou, E. Chien, C. Zhang, Y. Tian, and Z. Zhu. Breaking the trilemma of privacy, utility, and efficiency via controllable machine unlearning. In *Proceedings of the ACM on Web Conference 2024*, pages 1260–1271, 2024.

[80] Z. Liu, H. Ye, C. Chen, Y. Zheng, and K.-Y. Lam. Threats, attacks, and defenses in machine unlearning: A survey, 2024. URL https://arxiv.org/abs/2403.13682.

[81] T. Lizzo and L. Heck. Unlearn efficient removal of knowledge in large language models, 2024. URL https://arxiv.org/abs/2408.04140.

[82] R. Ma, Q. Zhou, Y. Jin, D. Zhou, B. Xiao, X. Li, Y. Qu, A. Singh, K. Keutzer, J. Hu, X. Xie, Z. Dong, S. Zhang, and S. Zhou. A dataset and benchmark for copyright infringement unlearning from text-to-image diffusion models, 2024. URL https://arxiv.org/abs/2403.12052.

[83] T. Mahler. Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act proposal. *The Swedish Law and Informatics Research Institute*, pages 247–270, Mar. 2022.

[84] P. Maini, Z. Feng, A. Schwarzschild, Z. C. Lipton, and J. Z. Kolter. Tofu: A task of fictitious unlearning for llms, 2024. URL https://arxiv.org/abs/2401.06121.

[85] M. A. Manab. Eternal sunshine of the mechanical mind: The irreconcilability of machine learning and the right to be forgotten, 2024. URL https://arxiv.org/abs/2403.05592.

[86] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[87] S. Mercuri, R. Khraishi, R. Okhrati, D. Batra, C. Hamill, T. Ghasempour, and A. Nowlan. An introduction to machine unlearning, 2022. URL https://arxiv.org/abs/2209.00939.

[88] L. Nicoletti and D. Bass. Humans are biased. generative AI is even worse. *Technology + Equality*, June 2023. URL https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

[89] NIST. NIST trustworthy and responsible AI: Artificial intelligence risk management framework – generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology (NIST), 2024. URL https://doi.org/10.6028/NIST.AI.600-1.

[90] A. Oesterling, U. Bhalla, S. Venkatasubramanian, and H. Lakkaraju. Operationalizing the blueprint for an ai bill of rights: Recommendations for practitioners, researchers, and policy makers, 2024. URL https://arxiv.org/abs/2407.08689.

[91] A. Oesterling, J. Ma, F. P. Calmon, and H. Lakkaraju. Fair machine unlearning: Data removal while mitigating disparities. In S. Dasgupta, S. Mandt, and Y. Li, editors, *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pages 3736–3744. PMLR, 2024. URL https://proceedings.mlr.press/v238/oesterling24a.html.

[92] A. Oesterling, J. Ma, F. P. Calmon, and H. Lakkaraju. Fair machine unlearning: Data removal while mitigating disparities, 2024. URL https://arxiv.org/abs/2307.14754.

[93] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. R. Bowman. Bbq: A hand-built bias benchmark for question answering, 2022. URL https://arxiv.org/abs/2110.08193.

[94] M. Pawelczyk, J. Z. Di, Y. Lu, G. Kamath, A. Sekhari, and S. Neel. Machine unlearning fails to remove data poisoning attacks, 2024. URL https://arxiv.org/abs/2406.17216.

[95] F. Pedregosa and E. Triantafillou. Announcing the first machine unlearning challenge, June 2023. URL https://research.google/blog/announcing-the-first-machine-unlearning-challenge/. Blog post.

[96] D. Preložnik and Ž. Špiclin. Improving brain mri segmentation with multi-stage deep domain unlearning. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 99–110. Springer, 2024.

[97] W. Qian, C. Zhao, W. Le, M. Ma, and M. Huai. Towards understanding and enhancing robustness of deep learning models against malicious unlearning attacks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1932–1942, 2023.

[98] J. P. Quintais. Generative AI, copyright and the AI act. November 2024. URL https://ssrn.com/abstract=4912701. Version 2.

[99] A. Reuel, B. Bucknall, S. Casper, T. Fist, L. Soder, O. Aarne, L. Hammond, L. Ibrahim, A. Chan, P. Wills, M. Anderljung, B. Garfinkel, L. Heim, A. Trask, G. Mukobi, R. Schaeffer, M. Baker, S. Hooker, I. Solaiman, A. S. Luccioni, N. Rajkumar, N. Moës, J. Ladish, N. Guha, J. Newman, Y. Bengio, T. South, A. Pentland, S. Koyejo, M. J. Kochenderfer, and R. Trager. Open problems in technical AI governance, 2024. URL https://arxiv.org/abs/2407.14981.

[100] E. Rosati. Infringing ai: Liability for ai-generated outputs under international, eu, and uk copyright law. *European Journal of Risk Regulation*, page 1–25, 2024. doi: 10.1017/err.2024.72.

[101] S. Sai, U. Mittal, V. Chamola, et al. Machine un-learning: An overview of techniques, applications, and future directions. *Cognitive Computation*, 16:482–506, 2024. doi: 10.1007/s12559-023-10219-3. URL https://doi.org/10.1007/s12559-023-10219-3.

[102] S. Schoepf, J. Foster, and A. Brintrup. Potion: Towards poison unlearning. *arXiv preprint arXiv:2406.09173*, 2024.

[103] J. Seeman and D. Susser. Between privacy and utility: On differential privacy in theory and practice. *ACM Journal on Responsible Computing*, 1(1):1–18, 2024.

[104] S. Shan, A. N. Bhagoji, H. Zheng, and B. Y. Zhao. Poison forensics: Traceback of data poisoning attacks in neural networks. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3575–3592, Boston, MA, Aug. 2022. USENIX Association. ISBN 978-1-939133-31-1. URL https://www.usenix.org/conference/usenixsecurity22/presentation/shan.

[105] W. Shi, J. Lee, Y. Huang, S. Malladi, J. Zhao, A. Holtzman, D. Liu, L. Zettlemoyer, N. A. Smith, and C. Zhang. Muse: Machine unlearning six-way evaluation for language models, 2024. URL https://arxiv.org/abs/2407.06460.

[106] I. Shumailov, J. Hayes, E. Triantafillou, G. Ortiz-Jimenez, N. Papernot, M. Jagielski, I. Yona, H. Howard, and E. Bagdasaryan. Ununlearning: Unlearning is not sufficient for content regulation in advanced generative ai, 2024. URL https://arxiv.org/abs/2407.00106.

[107] I. Sugiura, S. Okamura, and N. Yanai. Removing mislabeled data from trained models via machine unlearning. *IEICE Transactions on Information and Systems*, advpub:2024DAT0002, 2024. doi: 10.1587/transinf.2024DAT0002.

[108] A. Thudi, H. Jia, I. Shumailov, and N. Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4007–4022, 2022.

[109] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu. Data poisoning attacks against federated learning systems. In L. Chen, N. Li, K. Liang, and S. Schneider, editors, *Computer Security – ESORICS 2020*, pages 480–501, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58951-6.

[110] E. Triantafillou, P. Kairouz, F. Pedregosa, J. Hayes, M. Kurmanji, K. Zhao, V. Dumoulin, J. J. Junior, I. Mitliagkas, J. Wan, et al. Are we making progress in unlearning? findings from the first neurips unlearning competition. *arXiv preprint arXiv:2406.09073*, 2024.

[111] UK DSIT. International AI safety report: The international scientific report on the safety of advanced AI. Technical report, UK Government, January 2025. URL https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

[112] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. NIST AI 100-2e2023, National Institute of Standards and Technology, 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf.

[113] S.-Y. Wang, A. Hertzmann, A. Efros, J.-Y. Zhu, and R. Zhang. Data attribution for text-to-image models by unlearning synthesized images. *Advances in Neural Information Processing Systems*, 37:4235–4266, 2024.

[114] Y. Wang, Q. Wang, L. Zhao, and C. Wang. Differential privacy in deep learning: Privacy and beyond. *Future Generation Computer Systems*, 148:408–424, 2023. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2023.06.010. URL https://www.sciencedirect.com/science/article/pii/S0167739X23002315.

[115] Z. Wang, S. Chen, C. Li, L. Zhao, and Y. Liu. Applying machine unlearning techniques to mitigate privacy leakage in large language models: An empirical study. Sept. 2024. doi: 10.22541/au.172712647.70020033/v1. URL http://dx.doi.org/10.22541/au.172712647.70020033/v1.

[116] A. Warnecke, L. Pirch, C. Wressnegger, and K. Rieck. Machine unlearning of features and labels, 2023. URL https://arxiv.org/abs/2108.11577.

[117] X. Wei, N. Kumar, and H. Zhang. Addressing bias in generative ai: Challenges and research opportunities in information management. *Information & Management*, page 104103, 2025. ISSN 0378-7206. doi: https://doi.org/10.1016/j.im.2025.104103. URL https://www.sciencedirect.com/science/article/pii/S0378720625000060.

[118] K. Wu, E. Wu, D. E. Ho, and J. Zou. Generating medical errors: GenAI and erroneous medical references. feb 2024.

[119] Y. Wu, S. Zhou, M. Yang, L. Wang, H. Chang, W. Zhu, X. Hu, X. Zhou, and X. Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient, 2024. URL https://arxiv.org/abs/2405.15304.

[120] H. Xu, T. Zhu, W. Zhou, and W. Zhao. Don't forget too much: Towards machine unlearning on feature level, 2024. URL https://arxiv.org/abs/2406.10951.

[121] J. Xu, Z. Wu, C. Wang, and X. Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3):2150–2168, June 2024. ISSN 2471-285X. doi: 10.1109/tetci.2024.3379240. URL http://dx.doi.org/10.1109/TETCI.2024.3379240.

[122] Y. Xu. Machine unlearning for traditional models and large language models: A short survey, 2024. URL https://arxiv.org/abs/2404.01206.

[123] Y. Yao, X. Xu, and Y. Liu. Large language model unlearning, 2024. URL https://arxiv.org/abs/2310.10683.

[124] C. Yu, S. Jeoung, A. Kasi, P. Yu, and H. Ji. Unlearning bias in language models by partitioning gradients. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.375. URL https://aclanthology.org/2023.findings-acl.375.

[125] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[126] D. Zhang, S. Pan, T. Hoang, Z. Xing, M. Staples, X. Xu, L. Yao, Q. Lu, and L. Zhu. To be forgotten or to be fair: Unveiling fairness implications of machine unlearning methods. *arXiv preprint arXiv:2302.03350*, 2023.

[127] H. Zhang, T. Nakamura, T. Isohara, and K. Sakurai. A review on machine unlearning. *SN Computer Science*, 4(4), Apr. 2023. ISSN 2661-8907. doi: 10.1007/s42979-023-01767-4. URL http://dx.doi.org/10.1007/s42979-023-01767-4.

[128] J. Zhang. Australian mandatory ai guardrails proposed: — drawing from the canadian and european experience. *Computer Law Review International*, 25(6):162–165, 2024. doi: doi:10.9785/cri-2024-250602. URL https://doi.org/10.9785/cri-2024-250602.

[129] Z. Zhang, F. Wang, X. Li, Z. Wu, X. Tang, H. Liu, Q. He, W. Yin, and S. Wang. Does your llm truly unlearn? an embarrassingly simple approach to recover unlearned knowledge. *arXiv preprint arXiv:2410.16454*, 2024.

[130] K. Zhao, M. Kurmanji, G.-O. Bărbulescu, E. Triantafillou, and P. Triantafillou. What makes unlearning hard and what to do about it. *arXiv preprint arXiv:2406.01257*, 2024.

[131] S. Zhou, L. Wang, J. Ye, Y. Wu, and H. Chang. On the limitations and prospects of machine unlearning for generative ai, 2024. URL https://arxiv.org/abs/2408.00376.

[132] J. Łucki, B. Wei, Y. Huang, P. Henderson, F. Tramèr, and J. Rando. An adversarial perspective on machine unlearning for ai safety, 2024. URL https://arxiv.org/abs/2409.18025.