Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice

A. Feder Cooper[*‡1,2,3] **Christopher A. Choquette-Choo**[*4] **Miranda Bogen**[*5,6]
Matthew Jagielski[*4] **Katja Filippova**[*4] **Ken Ziyu Liu**[*3]
Alexandra Chouldechova[2] **Jamie Hayes**[4] **Yangsibo Huang**[7] **Niloofar Mireshghallah**[8]
Ilia Shumailov[4] **Eleni Triantafillou**[4] **Peter Kairouz**[7] **Nicole Mitchell**[7]
Percy Liang[3] **Daniel E. Ho**[9] **Yejin Choi**[8] **Sanmi Koyejo**[3] **Fernando Delgado**[10]
James Grimmelmann[1,11,12] **Vitaly Shmatikov**[11] **Christopher De Sa**[13] **Solon Barocas**[2]
Amy Cyphert[14] **Mark Lemley**[9] **danah boyd**[2] **Jennifer Wortman Vaughan**[2]
Miles Brundage **David Bau**[15] **Seth Neel**[16] **Abigail Z. Jacobs**[17] **Andreas Terzis**[4]
Hanna Wallach[2] **Nicolas Papernot**[4] **Katherine Lee**[‡1,4]

[*]First author [‡]Lead, correspondence: {afedercooper, kate.lee168}@gmail.com

[1]The GenLaw Center [2]Microsoft Research [3]Stanford University [4]Google DeepMind
[5]Center for Democracy & Technology [6]Princeton CITP [7]Google Research
[8]University of Washington [9]Stanford Law School [10]Lighthouse [11]Cornell Tech
[12]Cornell Law School [13]Cornell University [14]West Virginia University, College of Law
[15]Northeastern University [16]Harvard Business School [17]University of Michigan

Abstract

We articulate fundamental mismatches between technical methods for machine unlearning in Generative AI, and documented aspirations for broader impact that these methods could have for law and policy. These aspirations are both numerous and varied, motivated by issues that pertain to privacy, copyright, safety, and more. For example, unlearning is often invoked as a solution for removing the effects of targeted information from a generative-AI model's parameters, e.g., a particular individual's personal data or in-copyright expression of Spiderman that was included in the model's training data. Unlearning is also proposed as a way to prevent a model from generating targeted types of information in its outputs, e.g., generations that closely resemble a particular individual's data or reflect the concept of "Spiderman." Both of these goals—the targeted *removal* of information from a model and the targeted *suppression* of information from a model's outputs—present various technical and substantive challenges. We provide a framework for thinking rigorously about these challenges, which enables us to be clear about why unlearning is not a general-purpose solution for circumscribing generative-AI model behavior in service of broader positive impact. We aim for conceptual clarity and to encourage more thoughtful communication among machine learning (ML), law, and policy experts who seek to develop and apply technical methods for compliance with policy objectives.

Authors' expertise and intended audience. Our contributions require expertise in ML, law, and policy. We intend for our audience to be members of all of those communities. We organized a team of experts in each discipline. The resulting paper reflects the efforts of a large-scale collaboration across academic institutions, civil society, and industry labs. We intend for this paper to be a standalone document: one with the necessary (and sometimes elementary) background to make our contributions legible to our diverse intended audience, and at the appropriate level of abstraction to encourage effective cross-disciplinary communication about machine unlearning.

1 Introduction

"Machine unlearning" has recently captured public attention as a potential general-purpose approach for purging unwanted information from machine-learning (ML) models. It raises problems of technical interest, but perhaps more significantly, machine unlearning also finds broader appeal outside of technical circles for its perceived ability to support law and policy aims (Section 2). Since around 2016, technical experts and policymakers have invoked unlearning as a way to operationalize compliance with an individual's "right to be forgotten," with respect to removing personal data from deployed models, as granted in the E.U.'s General Data Protection Regulation (GDPR) [102].[1] Now, with the emergence of Generative AI, machine unlearning's presumptive mandate has expanded significantly. More and more, research papers, policy briefs, and media reports suggest machine unlearning as an approach for meeting a broad range of objectives for both open and closed models and systems,[2] spanning privacy, copyright, safety, and more [e.g., 51, 63, 71, 75, 96, 118, 134, 140].

Unfortunately, the fit between unlearning and policy is not so straightforward in practice. Machine unlearning is a set of technical methods and here, as always, there are critical gaps—gaps that are too often overlooked—between what technical methods do and what policy aims to achieve [31]. Our goal is to provide conceptual clarity that elicits these gaps, and to encourage more thoughtful communication among ML, law, and policy experts who seek to develop and apply technical methods for compliance with law and policy objectives (Sections 6 & 7). In summary:

Deleting information from an ML model is not well-defined. First, information cannot be deleted from an ML model in the same way that it can from a database. During training, data are transformed into patterns that get encoded in the model's parameters—patterns that are not directly or easily interpretable (Section 2). There is no way to cleanly identify, target, and delete specific, contained pieces of information from these parameters. Instead, it is possible, even if computationally expensive, to train a *new* model on a dataset that does not contain problematic data (Section 4)—for example, a specific scientific paper on designing novel flu viruses or a specific in-copyright image of Spiderman. This is typically what it means to "remove" data from a generative-AI model in machine unlearning, which deviates from intuitive understandings of the term. Removal applies to discrete pieces of data *in the training dataset before training occurs*; it cannot target the latent patterns that a trained model has learned across different data examples (Section 3). For example, there is no clear way to remove the more general concepts of "how to synthesize a toxic molecule" or "Spiderman" from a model; there is no single obvious or appropriate way to go about translating such open-ended aims to concrete tasks that can be implemented by an algorithm (Section 5).

Removing information from a model does not provide guarantees about model outputs. Second, removing information *from a model's parameters* does not guarantee that this model could never produce related information *at generation time*. Even if one removed all in-copyright images of Spiderman from a model's training data, this does not mean it would be impossible for the model to generate outputs that resemble Spiderman when put to use. Generative-AI models are impressive in part because they are able to generate novel outputs that transcend the information that is exactly contained in their training data. It is therefore a mistake to think that making a limited set of targeted changes to a model's parameters is sufficient to make promises about what types of outputs that model could or could not possibly produce (Section 5). This point is further complicated by the fact that users can introduce information at generation time through prompts. In the context of machine unlearning, user prompts can even reintroduce information whose effects were previously removed from the model's parameters. Combining such a prompt with a model's reasoning abilities, it may be possible to produce outputs that are effectively the same as those that would have been produced if an unlearning method had never been used in the first place (Section 5).

In general, removal on its own is often neither necessary nor sufficient to constrain model outputs in a controlled manner.[3] Instead, *suppressing* certain model outputs from being surfaced to users may be a more appropriate area of focus for technical methods [e.g., 39]. While it is now common to

[1] This paper focuses on generative-AI models, but unlearning methods originate from classification.

[2] Our observations address cross-cutting issues applicable to both open and closed technology.

[3] One important exception where removal may be necessary (but still not sufficient) concerns illegal content that may have been included in the model's training data, such as child sexual abuse material (CSAM) [101].

include output suppression under the umbrella of "machine unlearning," arguably, these methods have nothing to do with "unlearning" some information from a model; they serve as guardrails on model and system outputs that bear more resemblance to alignment techniques (Section 4).

Even seemingly innocuous model outputs can be put to undesirable uses. Lastly, even if technical research were to shift the focus to methods for suppressing undesirable model outputs, this would not immediately provide solutions for all law and policy ends. Generative-AI systems are dual-use, where the appropriateness of downstream use wholly depends on context. There remain familiar (but fundamentally irresolvable) tensions that are inherent to highly generative, dual-use technologies [151], like the PC and the Internet (Section 7). Just as a PC could be used as a tool to perpetrate fraud or to write the next great Broadway musical, a generative-AI system can similarly be put to malicious or beneficial uses. Further, on their own, generated outputs may be innocuous or have significant legitimate uses; yet these same outputs could be pressed into service for adversarial or malicious downstream uses. To greater or lesser extents, different unlearning methods can remove the influence of specific training data from a model's parameters or suppress undesirable model outputs (Section 4). But the type of control these functions provide is localized to the model or system. Additional controls on downstream use would require anticipating how a person or other agent might behave with outputs in a potentially infinite number of contexts—none of which is reasonably under the purview of machine unlearning (Section 5).

In this paper, we explore these observations in detail, and examine their various implications for privacy, copyright, and safety—three areas for which machine unlearning has been suggested as a viable approach for operationalizing compliance with law or policy requirements (Section 6). Rather than following the well-trod path of surveying [82, 83, 109, 114, 117, 143] or evaluating [81, 83, 89] existing unlearning methods, we take a step back and think conceptually about what, in principle, machine unlearning could reasonably accomplish.

We articulate fundamental mismatches between machine unlearning—as a technical problem of study in ML research—and aspirations for the broader impact that methods emerging from this research could have for law and policy. In contrast to common opinions in policy research [e.g., 9, 63], we show that machine unlearning—both the entire class of methods and specific techniques—should not be misunderstood as a general-purpose solution for circumscribing model behavior in service of broader positive impact. 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. However, given the fundamental mismatches we identify, it is also hard to imagine that, even with time, technical solutions for unlearning will ever wholly achieve desired law and policy objectives. In light of these limitations, we provide recommendations on how ML experts should focus their research and how policymakers can adjust their expectations and norms concerning reasonable best efforts when using an unlearning method in practice (Section 7).

We organize the remainder of the paper as follows:

Section 2. We begin with some necessary background on machine unlearning—both its technical motivations and evolving motivations for Generative AI from law and policy.

Section 3. We identify different targets—observed information, latent information, and higher-order concepts—that model developers or custodians may want to address with unlearning.

Section 4. Defining these targets (Section 3) helps us make clear which types of information a specific unlearning method may apply to and which it does not. Some methods can *remove* targeted pieces of observed information before training occurs. For Generative AI, most methods aim to *suppress* model outputs that contain undesirable content.

Section 5. Together, our discussion illuminates four important mismatches between unlearning motivations (Section 2), targets (Section 3), and methods (Section 4). These mismatches lay the groundwork for understanding how there are substantive aims that cannot, from first principles, be addressed with unlearning methods alone.

Section 6. We examine how these mismatches (Section 5) manifest differently and exhibit various implications for issues related to privacy, copyright, and safety contexts.

Section 7. We suggest takeaways and possible future directions for ML research and AI policy.

2 Background and Motivations for Machine Unlearning

The natural starting place for our discussion is to address first what machine *learning* attempts to accomplish. This will let us provide an intuition, grounded in an interpretation of the E.U.'s General Data Protection Regulation (GDPR) that is prevalent in ML research, for why one might want to do *un*learning to revert or change the results of this process (Section 2.1). From this intuition, we provide a loose definition for machine unlearning that originates from traditional AI settings (Section 2.2). We then discuss evolving motivations for machine unlearning in response to the ascendance of Generative AI (Section 2.3). These new motivations have encouraged an expanded definition for machine unlearning (Section 2.4), which we will rely on throughout the paper.

2.1 ML research and its interpretation of the GDPR

In brief, **machine learning** is an area of computer science and engineering that uses techniques from probability and statistics to develop algorithms that produce models that encode patterns learned from data. Model **architectures** range from simple linear models or decision trees to complicated, large-scale neural networks. In a bit more detail, we rely on the GenLaw glossary [31]:[4]

> Machine-learning neural-network **models** all contain **parameters**. . . . During an **algorithmic process** called **training**, these parameters are repeatedly updated based on the training data within the **training dataset** that the model has seen. Each update is designed to increase the chance that when a model is provided some input, it outputs a value close to the target value we would like it to output. By presenting the model with all of the **examples** in a dataset and updating the parameters after each presentation, the model can become quite good at doing the task we want it to do.

Each training-data **example** in the training dataset "is a self-contained piece of data, such as an image, a piece of text (e.g., content of a web page), a sound snippet, a video, or some combination of these" [31]. These examples can also include personal information—home addresses, sensitive demographic attributes, health information, personal photos, and more.

In some jurisdictions, individuals have rights associated with the control of their personal data. Notably, since its adoption in 2018, Article 17 of the E.U.'s GDPR provides the "Right to erasure" (more commonly called the "right to be forgotten") [45, 102], which gives individuals rights (with exceptions) to demand that companies delete their personal data.[5] ML researchers often interpret Article 17 to apply to both to the individual's data examples that have been used as training data and to the resulting trained models themselves [e.g., 10, 24, 73, 74, 84, 90, 106] (Appendix A).[6] This presents a problem because, in almost all cases, the model would not *just* be trained on a specific, right-exercising individual's data. It would also be trained on data associated with thousands of others, if not many more. Wholesale erasure of a trained model, in response to one individual's deletion request, would therefore likely be an extreme, over-broad interpretation of Article 17.[7]

This problem raises a natural question for ML research: rather than deleting a trained model altogether, is it possible to develop algorithms that can achieve more targeted removal of training data from the model? In the specific context of ML research's common interpretation of the GDPR: is it possible to remove the influence of the right-exercising individual's data from the model, without imposing on the model controller the undue burden of the cost of retraining a new model from scratch without that individual's data examples (Section 4.1)? Machine unlearning is the area of ML research that attempts to address this question.

[4] We reprint with permission excerpts from the GenLaw Glossary. This glossary provides definitions in machine learning, law, and policy at the same level of abstraction that is intended for our audience.

[5] There are several exceptions to this right, for instance, when keeping the data is in the public interest (e.g., the data are used to comply with a legal ruling) or for certain research purposes, etc. [102, Article 17(3)].

[6] While this interpretation is common in unlearning papers, it is still very much up for debate. Exactly how Article 17 may or may not apply to ML models is under active discussion. (See Section 6.1.)

[7] Also consider that many thousands of people might make such a request, each one requiring retraining a new model from scratch (Section 4.1). Retraining runs could perhaps be periodically batched—removing multiple individuals' personal data together in the same run to reduce the number of times a model is retrained. Even still, retraining could happen a large number of times for the same model.

2.2 A loose definition for machine unlearning

What is referred to as "machine unlearning" in the technical literature actually corresponds to a wide variety of different methods and techniques, which are loosely grouped together. For this reason, we will rely on a loose, intuitive (rather than rigorous) definition of machine unlearning that we construct in relation to this common underlying technical motivation. In relation to early research, **machine unlearning** is a subarea of machine learning that develops methods for the *targeted removal* of the effect of training data from the trained model. We will soon refine this definition (Section 2.4), in response to changing motivations in the field for when to use unlearning (Section 2.3) and how to implement it (Section 4).

This definition is deliberately broad, rather than prescriptive. We intentionally do not include specific requirements for *how* certain information is "targeted" or "removed."[8] For now, we also are not prescriptive about what the exact "effects" are of "learned information" on the trained model's behavior. We will address this in more detail in Section 3, where we discuss different types of learned information that could reasonably be targeted for unlearning.

This definition covers a variety of methods in the technical literature. It encompasses prior work from the last 10 years that has studied unlearning in clustering [53], classification and regression [11, 42, 100, 122], federated learning [67, 85], and more [17, 146]. It also applies to the classic paper by Cauwenberghs and Poggio [19], which studies the problem of unlearning in support vector machines (SVMs) under the name "decremental learning" over two decades ago.

2.3 Generative AI and evolving motivations for machine unlearning

Given the particularly high cost of (re)training large-scale generative-AI models, there is a developing interest to apply efficient unlearning methods in this area. However, translating unlearning methods to generative-AI models exhibits some important technical challenges, since these models differ from the more traditional ML models to which much prior work in unlearning has applied [e.g., 11, 53]. Traditional AI settings tend to involve models that produce concise outputs from a bounded and typically fixed set, for example, classification labels like `dog` or `cat`. After using an unlearning method on such a model, its outputs for a given input may change (e.g., its classification may flip from `cat` to `dog`), but the set of possible outputs (e.g., `cat` and `dog`) generally remains the same. In contrast, generative-AI models produce "information-rich" [26] outputs of the same modality as their training data. The set of possible outputs is significantly more expansive. For example, text-to-text models like Llama 3 [87] and those embedded in systems like Claude [4], ChatGPT [103], and Gemini [127] produce long-form text outputs.

With this key difference, the desired goals for what machine unlearning could achieve have also shifted. They have begun to expand beyond the scope of our loose definition for unlearning—beyond *removal* of the influence of *training-data inputs* on the trained *model's parameters*—to also encompass desired effects on the model's possible *generated outputs* when the model is *put to use*.

2.4 An expanded, loose definition for machine unlearning

In the context of more recent research on Generative AI, the loose definition for unlearning in Section 2.2 has widened in scope. In relation to these developments, we offer an expanded loose definition: **machine unlearning** is now a subarea of machine learning that both develops methods for (1) the *targeted removal* of the effect of training data from the trained model and (2) the *targeted suppression* of content in a generative-AI model's outputs. Later, we will organize our discussion of concrete unlearning methods in relation to this split (Section 4).

In other words, the scope for unlearning no longer just concerns what we, following Cooper and Grimmelmann [26], will refer to as **back-end** considerations: "characteristics and capabilities of the *model itself* that directly result from its training." Unlearning also concerns **front-end** considerations: "how the *model behaves* in generating outputs in response to … specific prompt[s]"(emphasis

[8] There is arguably a spectrum between *targeted* machine unlearning and unintentional (and undesired) loss of representation of information in the model, which prior work has studied in various settings, for example, **catastrophic forgetting** [56, 116]. Our focus is the former. See also Section 5, Mismatch 2.

(a) Back-end (b) Front-end

Figure 1: Both the (**a**) back-end and (**b**) front-end involve processes that have their own inputs and produce their own outputs (simplified here). This is why we use this additional terminology for clarifying which inputs and outputs are under discussion. There is nothing complicated here; it is just shorthand to signal different aspects of the trained model at different points in time.

added) [26].[9] Both the back-end and front-end involve processes that have their own inputs and produce their own outputs (Figure 1).[10] On the back-end, the training dataset is an input and the trained model is an output; the back-end involves making choices for which training data to include, which training algorithm to run, etc. On the front-end, the prompt is an input and the generation is an output; the front-end includes the process of **inference** (i.e., producing generations), system-level filters that may prevent the processing of certain undesirable user prompts or the user-facing output of certain undesirable generations (Section 4.2), etc. On the back-end, the trained model *is an output*; on the front-end, the trained model *is used to produce outputs*. Throughout this paper, we will use this back-end/front-end terminology as a shorthand for distinguishing the different points in time where unlearning is of interest, and which artifacts a particular unlearning method is intended to affect—the *model parameters* or the *model's possible generations* (Section 4).

Extending the example at the beginning of this section of an individual exercising their "right to be forgotten" under the GDPR: for a generative-AI model, the goal for unlearning would no longer simply be to remove the influence of the individual's personal data from the model's parameters on the back-end, but also to ensure that the resulting model could not produce outputs that reflect that individual's personal data on the front-end. It is an appealing idea that machine unlearning could serve both of these ends. Indeed, it would be remarkably convenient if machine unlearning could be both an approach for mitigating the influence of problematic training data on a trained model's parameters *and* for effectively moderating a generative-AI model's possible outputs. If so, it would also perhaps be reasonable to assume, as many researchers and organizations have, that machine unlearning could, on its own, be used to solve issues related to problematic model outputs in a variety of policy-relevant domains: novel privacy challenges [12, 23, 73, 94, 148], copyright [26, 43, 77, 137, 150], safety [80, 81, 86], and more.

However, as we discuss below, these two back-end and front-end goals are very different in kind (Sections 4 & 5). Tying them together ultimately muddles what concrete unlearning methods could reasonably achieve on their own for desired policy ends (Section 6). To arrive at this conclusion, we first need some additional language that will enable us to refine our discussion of what types of information machine unlearning could address.

3 Targets for Machine Unlearning

Our loose definition for unlearning is abstract (Section 2.4); in fact, it is so abstract that it allows for an enormous number of reasonable interpretations and possible techniques that satisfy it. In this section, our aim is to provide some language that can help us be more precise. Building on our loose definition, we now pin down useful ways to think about what a "piece of information" could mean—what types of information one might want to target with unlearning. In the sections that follow, we show that the targets we define are places where concrete unlearning methods could potentially apply in practice (Section 4). We will also note that some articulated goals for unlearning escape these target definitions altogether, highlighting instances where unlearning methods could not be applied rigorously or reliably for certain desired ends (Section 5).

[9] Data can enter a generative-AI system on the back-end as training data (i.e., for pre-training, fine-tuning, alignment) and the front-end via prompts, generation-time plug-ins, and retrieval-augmented generation (RAG). We refer to Lee et al. [77, Part I] for more information on data and the generative-AI supply chain.

[10] The utility of this framing becomes clear in Section 4, where we discuss concrete inputs and outputs of unlearning methods applied at these different stages. Using the words "input" and "output" would be unclear, as they are overloaded with different meanings at different stages. Also note that this is a different usage of "back-end" and "front-end" from Internet software, where "back-end" refers to server-side components like storage and "front-end" refers to client-side components like a user interface .

We define three overarching (and, as we will see, overlapping) **targets**: **observed information** (Definition 1), **latent information** (Definition 2), and **higher-order concepts** (Definition 3). These definitions build upon each other and become more abstract and indeterminate. After presenting the definitions, we discuss how this indeterminacy surfaces challenges for designing and implementing unlearning methods in practice.

Definition 1 Observed information. *Data that are explicitly presented to the model during training. These data serve as inputs to computations that update the model's parameters.*

Observed information includes training-data examples: the contiguous pieces of data that are the base-level unit of input to model training (Section 2.1). For example, consider that the text "Susan's phone number is 555-123-4567" is included as an example in an LLM's training data. Since this text is used directly to train the LLM, it is observed information. Observed information also captures sets of training examples, such as all examples in the overall training dataset that mention Susan. It also includes data contained within examples, such as just the phone number "555-123-4567" in "Susan's number is 555-123-4567."

Effective trained models **generalize**: the learning process instills models with complex patterns that are derived from the observed information in the training data—patterns that models can apply to previously unseen information when they are put to use for inference or generation (Section 2.1). This learned information is latent in the training data.[11]

Definition 2 Latent information. *Data that are not explicitly presented to the model during training, but that can be derived or otherwise elicited from a trained model based on the patterns that the model has learned during training.*

Unlike observed information (Definition 1), latent information is not *literally* observed in the training data. However, there are ML-based methods that claim to identify latent information and make it observable in the trained model's parameters[12] or indirectly through a model's outputs when the model is put to use.[13] Latent information can include simple deductions [68, 111]. For example, given the observed information "Carlos is going to Susan's house for a birthday party this Thursday" and "Susan lives in Philadelphia," a possible piece of latent information is that Carlos is going to be in Philadelphia on Thursday.[14] This information is not literally contained in the training data; it is derived from relationships learned from observed information. Of course, latent information can also be significantly more complex than such simple deductions. The power of large-scale models trained on enormous datasets [76] comes from their flexibility to capture all sorts of latent information—across observed information, across latent information, or across some combination of the two. Indeed, information can interact to produce sophisticated, higher-order information that ML research often refers to as "knowledge" or "capabilities."

Definition 3 Higher-order concepts. *Combinations of latent and observed information that manifest in the model as complex and coherent abstractions, knowledge, capabilities, or skills.*

Before giving some examples of higher-order concepts, some disclaimers are in order. Definition 3 is not intended to suggest something particularly deep about how models organize information or exhibit complex behaviors. (This is not, after all, a paper about ontology or metaphysics.) Instead, we give a definition of higher-order concepts for convenience: to align with how the ML technical literature tends to refer to conceptual learned representations. But it is nevertheless reasonable to think of higher-order concepts as complex combinations of latent information—that

[11] For useful models, most learning is generalization; however, models also memorize (near) exactly a portion of their training data [e.g., 18, 26, 48, 77, 99, 108, 121]. Understanding the relationship between memorization and generalization is an active area of research.

[12] For example, some methods identify "concept neurons:" model parameters that relate to human-interpretable concepts [8, 40, 52].

[13] In some cases, it may be possible to identify non-exhaustively the data examples that contributed to latent information. However, currently, this is not true in general (Section 5). Eliciting information from the model's parameters in useful ways is one of the goals of **mechanistic interpretability** research [e.g., 25, 59, 97, 98].

[14] Of course, just as with observed information, there is no guarantee that latent information is factually correct. (In this example, perhaps Carlos attends the party remotely over a video call.)

there is, loosely speaking, a spectrum of complexity for latent information (Definition 2), with simple deductions drawn directly from observed information on one end, and significantly more complex patterns (often called capabilities or emergent abilities [112, 123, 126, 138]) at the other.

This spectrum reveals that Definition 3 is somewhat arbitrary, since it is not clear how to distinguish when a piece of latent information is sufficiently complex to be considered a higher-order concept. We do not attempt to draw these lines. Nevertheless, we still find it useful in the discussion that follows to have a target definition that lets us to refer to the unlearning of higher-order concepts, since this is a type of information that could reasonably be—and, in some cases, is claimed to be—a target for machine unlearning (Sections 4, 5 & 6).

Given these disclaimers, we enumerate a few examples that satisfy Definition 3. A model's representation of a "person" [107] is a higher-order concept. "People" is also a higher-order concept (perhaps generalized from latent information about relationships between different "person"s). So, too, is the knowledge that composes concrete subjects like "Spiderman," "Marie Curie," and "basketball;" knowledge of abstract ideas like "justice" and "toxicity;" and notions of "artistic style" and "scientific phenomenon" (as well as instances of particular artistic styles and phenomena, like "Cubism" and "gravity"); and the ability to reason about the relationships between different concepts, including "mathematical reasoning."

4 Unlearning Methods and Evaluating Evidence for Their Success

With an understanding of the different types of information one may want to target with machine unlearning (Section 3), we next discuss concrete unlearning methods that aim to address them. By focusing on targets, we show how unlearning in generative-AI contexts attempts to have targeted effects in two overarching ways. First, there are methods that, in line with the original loose definition of unlearning (Section 2.2), address the targeted **removal of observed information** (Section 4.1) on the back-end (Figure 1a). Second, in response to the shifting motivations for unlearning in generative-AI contexts (Sections 2.3 & 2.4), there are methods for **output suppression** (Section 4.2) of targeted information in a model's outputs on front-end (Figure 1b).

Our treatment of specific unlearning methods for removal and suppression is fairly brief.[15] This is because our purpose is to provide sufficient framing that will enable us to elicit important conceptual gaps and limitations—fundamental mismatches between unlearning motivations, targets, and methods (Section 5). It is these mismatches that are the heart of our paper, and are relevant for understanding misalignment with law and policy aims (Section 6).

4.1 Methods for removal (of observed information)

As discussed above, one of the articulated goals for machine unlearning is to purge unwanted information from models (Sections 1 & 2.2). This is a fundamentally challenging technical problem because an ML model is not like a database. For a database, it is typically the case that specific pieces of information can be identified, targeted, and deleted; but there is no direct analogue for deleting targeted information from a generative-AI model. While each of a model's training examples "is a self-contained piece of data" [31], this is not the case for how information learned from these examples is arranged in a trained model's parameters. The training process encodes patterns learned from training data in the model's parameters in ways that are not directly or easily interpretable (Sections 2.1 & 3).

As a result, "removal" of information from a generative-AI model deviates from intuitive understandings of the term "removal." Instead, the most straightforward way one might "remove"[16] information is to replace the original model with a *new* model that is trained on a dataset that does not contain problematic examples—for example, a specific scientific paper on designing novel flu viruses or a specific in-copyright image of Spiderman. This process removes specific observed information (Definition 1) from a model's *training dataset*, instead of literally removing it from a

[15] As stated previously, we deliberately do not provide an in-depth survey or taxonomy of state-of-the-art techniques that are branded as machine-unlearning methods. Several groups of authors have already done so from different perspectives [e.g., 83, 109, 117].

[16] We drop the quotation marks going forward; this is the sense of "removal" that we use in this paper.

model's parameters. This is, in some—though, as we will see, limited—senses an easier technical problem to solve. Even though we cannot treat a model like a database, we can treat a training dataset like one: observed information can be relatively easily and directly identified, targeted, and removed from the dataset *before training* transforms this information and makes its effects difficult to locate in the model's parameters.

The "gold standard" for machine unlearning. The method above is often referred to as **retraining from scratch** or the **"gold standard"** [e.g., 54, 83, 91]. It is the "gold standard" because the targeted information was literally never observed by the training process; by definition, it is guaranteed that this specific, targeted information could not have influenced the model's parameters.

At first glance, this seems like a reasonable (albeit expensive) solution to the unlearning problem. However, the "gold standard" exhibits important limitations: it casts machine unlearning as problem to be solved with respect to back-end inputs (i.e., training data) and, as a result, it does not directly apply to all of the types of targets one might want to address with unlearning. In particular, the "gold standard" does not directly apply to latent information (Definition 2) or higher-order concepts (Definition 3), as these are types of information that emerge and get encoded in a model's parameters during training. In general, it is often not clear exactly which observed information contributes to latent information and higher-order concepts (Section 3). Targeted removal of observed information from the training dataset can affect both the latent information and higher-order concepts that the model learning during training; however, in general, this relationship is not well understood. As such, the "gold standard" may have an indirect effect on these targets; however, it may not be effective with respect to ensuring unwanted information is not latent in the model's parameters, nor with respect to preventing unwanted information from manifesting on the front-end—in the model's outputs at generation time (Sections 4.2 & 5).[17]

Further, in practice, implementing the "gold standard" is expensive—often prohibitively expensive for today's enormous models trained on enormous datasets by expending immense computing resources. This has motivated the development of lower-cost methods for removal of *structured* information in the training dataset to produce models that have similar properties to those that have been retrained from scratch.

Structural removal. Methods for structural removal make the "gold standard" of retraining from scratch more computationally efficient. To do so, instead of requiring the whole model be retrained, these methods design custom model-training procedures that limit the amount of retraining that needs to be conducted to exclude targeted observed information [e.g., 11, 144].[18] There also exist methods that attempt to approximate structural removal, often by changing the original model's parameters rather than retraining from scratch. These methods often involve the development of algorithms that rely on mathematical theory to prove (under specific theoretical assumptions) that the modified model is (by some mathematical definition) "similar" to a model that has been retrained from scratch [58, 74]. Of course, such approximations are not literally equivalent to retraining from scratch; they often involve a probabilistic guarantee—not absolute certainty—that the targeted information has been successfully removed.[19]

Most methods for (approximate) structural removal have been developed for traditional AI settings, not Generative AI. There are a few methods for generative-AI contexts that have drawn inspiration from this work [e.g., 22, 73]. However, for two overarching reasons, traditional AI methods do not naturally translate to this newer setting. First, since structural-removal methods

[17] This is why we put the term "gold standard" in quotation marks, which are typically absent in the technical literature. These limitations also have broader implications, which we examine in Section 5, Mismatch 2.

[18] For this reason, structural removal is commonly referred to as **exact unlearning** in the ML literature [e.g., 143]. Even though these methods are different from the "gold standard," they retain the *exact* same guarantees of the "gold standard," with respect to removing the effect of targeted observed information. We avoid the term "exact unlearning" because it can be reasonably misunderstood to mean that such methods are able to "exactly unlearn" all types of targets. However, these methods only apply to observed information, not to latent information that is encoded in a perhaps unidentifiable (i.e., unstructured) place in the model. Further, these methods do not guarantee effective output suppression of targets. (See Section 5, Mismatch 2.)

[19] Practical implementations do not always align with theoretical mathematical assumptions. In such settings, methods may still work reasonably well empirically, but they may lose their theoretical guarantees.

typically require specific training processes for the original model, they cannot be applied to trained models that did not use those processes. This means that existing models that were not trained with structural removal in mind, such as Llama 3 450B [87], cannot post hoc be made compatible with these methods. Second, both structural-removal methods and methods that approximate them are very computationally expensive at generative-AI scale [83]. For both of these reasons, removal algorithms are challenging to implement for Generative AI in practice. Later, we will discuss how these practical challenges have important implications for legislative requirements around data deletion for production generative-AI systems (Section 6.1).

4.2 Methods for output suppression

The majority of unlearning methods in Generative AI focus on output suppression (Sections 2.3 & 2.4). In this setting, potentially problematic training data is observed during the training process, and there is no attempt to guarantee (with certainty or probabilistically) that this is not the case. Instead, output-suppression methods aim to prevent undesirable content from appearing in generations on the front-end, rather than attempting to remove the effects of targeted observed information on the back-end. These methods tend to be more computationally feasible than retraining from scratch (Section 4.1). Since they focus on model outputs, they are not limited to observed information; they also apply (to varying degrees of success) to latent information (Definition 2) and higher-order concepts (Definition 3).

We organize our discussion around two overarching approaches to output suppression: (1) methods that make modifications to the trained generative-AI model, and (2) methods that leave the model unchanged, but implement guardrails in the system in which model is embedded, in order to constrain the outputs that are presented to end users. Both of these approaches include a wide range of techniques that operate very differently from the removal methods discussed above. (See Section 5, Mismatch 1.) While it is now common to include output suppression under the umbrella of "machine unlearning," arguably, these methods have nothing to do with "unlearning" some information from a model; they bear more resemblance to alignment techniques.

Methods that modify the generative-AI *model*. Some output-suppression methods *modify the original model* to attempt to direct the model away from being able to produce outputs that reflect undesirable content. These methods cover a variety of different alignment-inspired techniques (e.g., different types of additional training, reinforcement learning) [66, 88, 91, 145, 149] and model editing [93, 95]. They all use back-end modifications to the trained model to try to alter the model's outputs at generation time on the front-end. As we have noted throughout, this is challenging to do in a *targeted* way because the relationship between model parameters and model outputs is not straightforward or, in some cases, possible to determine (Sections 3 & 4.1). As a result, while model-based methods for output suppression can make the generation of undesirable content less likely, they do not provide guarantees that the model could never produce such content.

Methods that implement guardrails in the generative-AI *system*. Some output-suppression methods leave the trained generative-AI model unchanged and instead take effect in the generative-AI system in which the model is embedded. For example, **output filters** may be wrapped around model outputs in order to prevent generations that contain certain undesirable content from being surfaced to end users [129]. This requires no change to the generative-AI model: output filters operate entirely on the front-end. For example, a user may prompt the system to generate the molecular formula for the smallpox virus and, in response, the model may generate that formula; but, the output filter may identify the formula as problematic, and not surface it to the user. Similarly, a system developer could implement **input filters** that filter problematic user prompts [104]—e.g., a filter that flags the user's prompt to generate the smallpox formula, and prevents the prompt from ever being supplied as an input to the model. These filters may themselves be implemented with ML models (e.g., more traditional ML classifiers), which are imperfect; they may attempt to target certain types of information, but may do so with greater or lesser degrees of precision and accuracy. Other proposed methods utilize the **system prompt** for output suppression. A system prompt is a piece of developer-chosen text that the system adds internally to the context of all user-supplied prompts, often to coax the model away from producing generations that contain undesirable content [106, 129]. Such in-context mechanisms may (or may not) work in practice; they are generally imprecise.

Since all of these methods focus on suppressing outputs, their success is most often evaluated by examining how they affect the types of generations that are produced in some downstream task. This often involves prompting the model or system with respect to content that the method intended to suppress, and observing if the resulting generations do not reflect that information [e.g., 20, 43, 91].[20] For example, in safety contexts, such evaluations often rely on the WMDP benchmark [81], which is a multiple-choice question dataset that focuses on biological, chemical, and cyber-security risks. One might test the original model on this question dataset as a baseline, and then apply an output-suppression method, re-test, and quantify changes in the answers as a proxy for determining if "unsafe" knowledge is no longer reflected in the model's answers [e.g., 125].[21] They might also perform a similar test for a model trained using the "gold standard" as another point of comparison [131] (i.e., evaluate the front-end behavior of a model that has had information removed on the back-end). Beyond evaluations like these, it is also common to test if the application of an unlearning method has effects on information that was *not* intentionally targeted—to evaluate if metrics for overall model utility are preserved [11, 69, 74, 83].

5 Mismatches between Unlearning Motivations, Targets, and Methods

Four important problems emerge directly from our discussion of removal of observed information (Section 4.1) and output suppression (Section 4.2) above. Output suppression is not a replacement for removal of observed information (Mismatch 1). Conversely, removal of observed information does not guarantee meaningful output suppression (Mismatch 2). More generally, models are not equivalent to their outputs (Mismatch 3) or, relatedly, to how their outputs are put to use (Mismatch 4). We address each of these points in turn.

Mismatch 1 *Output suppression is not a replacement for removal of observed information.*

Methods that aim to suppress certain model outputs on the front-end are intrinsically different from back-end removal of observed information from the model's training dataset (Section 4). With output-suppression methods, it is possible that a target could still be represented in the model, and it is possible that this target could manifest in the model's outputs.[22] These details could have important consequences for law and policy. For example, if a piece of legislation were to call for the explicit removal of a piece of training data from a model's training data set—to guarantee that a particular piece of information was never observed during training—unlearning methods that fall short of a guaranteeing structural removal (Section 4.1) would likely not suffice [50]. In other cases, modifications to the model or system to suppress certain types of observed information may be sufficient for some compliance requirements (Section 4.2). In general, the appropriateness of unlearning methods for removal or suppression to operationalize compliance with legislation in practice will depend on the exact details. These details include the particular legal domain in question, and perhaps also the circumstances of the use that potentially exposes information that was meant to be addressed (e.g., if some atypical, adversarial usage pattern is necessary for exposure of problematic observed information).

Mismatch 2 *Removal of observed information does not guarantee meaningful output suppression.*

Structural removal (Section 4.1) is insufficient to suppress model outputs that bear some resemblance to the removed data. Exactly removing a piece of observed information on the back-end, like a particular phone number, does not guarantee that it would be impossible on the front-end for a model to generate that phone number. Given latent information the model may contain about other phone numbers (and about numbers in general), it may be possible for the model

[20] There are various other types of evaluations, for example, probing latent information in the model. We defer to Lynch et al. [89] for further discussion on evaluation strategies.

[21] In practice, given the open-ended "information-rich" outputs of generative-AI models, it is very challenging (and an open research area) to come up with methods that reliably measure properties of model and system outputs [136]. Evaluation benchmarks like WMDP attempt to mitigate this complexity by setting up tasks (in this case, multiple-choice questions) that constrain the open-endedness of generated outputs.

[22] As in our discussion of shifting goals for machine unlearning (Section 2) and unlearning methods (Section 4), we continue to see a slippage between what model *is* (i.e., what is stored in its parameters) and the outputs that a model *could produce*. We attend to this in more detail below in Mismatch 3.

to generate a specific phone number for which all associated observed information has been removed (Section 6.1). Similarly, on the back-end, one could remove all in-copyright images of Spiderman from an image-generation model's training dataset and retrain from scratch with the hope this suffices to remove the higher-order concept of "Spiderman." However, this does not guarantee that, on the front-end, the new model could not possibly produce an output that might be "substantially similar" to copyrighted expression of Spiderman, based on how the model generalizes from latent information derived from the remaining training examples (Section 6.2). In both cases, removal could perhaps make the generation of similar outputs less likely; however, this cannot be assured in general (Sections 4 & 7.2).

From these examples, our main point is that there is a meaningful slippage that occurs when employing a removal technique (Section 4.1) in service of output suppression (Section 4.2): it is unclear which set of information should be targeted for removal in order to prevent the generation of certain outputs at generation-time. Removal of a narrow set of observed information (e.g., examples that contain phone numbers) from a model's training data can easily be *under-inclusive*, with respect to effectively suppressing the contents of that information at generation time. Being *over-inclusive* with how to instantiate targets for removal is also a potential problem, especially for cases that attempt to handle indeterminate higher-order concepts (Definition 3). One could remove all information related to comic books, spiders, the colors blue and red, the humanoid form, etc. But this is arguably too broad: it may be more effective at preventing generations that reflect "Spiderman," but it also removes significantly more information that one did not originally intend to target [e.g., 64, 89].[23]

Both sides of these examples—of over-inclusiveness and under-inclusiveness—further clarify how the "gold standard" can be challenging to implement and interpret as a baseline for unlearning (Section 4.2). As discussed in Section 4.1, the "gold standard" involves retraining a model from scratch with a set of examples removed from the training dataset; it applies directly to observed information. Using the "gold standard" can indirectly affect latent information or higher-order concepts, but it cannot ensure removal of or prevent generations that reflect these types of information. To try to capture some amount of these types of information in practice, implementations of this approach require navigating difficult, if not arbitrary, trade-offs to draw boundaries around what exactly to include for removal. For example, one could choose to retrain without all in-copyright training-data images of Spiderman that they manage to identify, but this would not necessarily include pictures of people in Spiderman Halloween costumes (Section 6.2). How to make these choices is clearly not a straightforward task, and yet it is essential when evaluating a particular unlearning method against the "gold standard" as a baseline, in order to make judgments about its efficacy.

Mismatch 3 *Models are not equivalent to their outputs.*

The slippage discussed above—between back-end, targeted removal from a model's training dataset and effective, front-end suppression of certain information at generation time—runs deep in unlearning research. Notably, it is typical to evaluate the success of an unlearning method not by examining changes in the model's *parameters*, but by prompting the model and measuring the extent to which certain types of *outputs* are no longer generated (Section 4.2).

This slippage has consequences for how we should think about gauging the success of an unlearning method. For example, consider that an individual p_0 has associated data examples and that a model trainer retrains a model from scratch without those examples, i.e., p_0's examples have been removed on the back-end. But now consider that there are also training examples related to individuals $p_1, \ldots, p_n$ that are by some quantitative measure similar to p_0's. On the front-end, a user prompts the retrained model with some (perhaps public) information about p_0 (e.g., demographics, address), with the goal of revealing information about p_0's health status. Combining the latent information in the model (from training on examples concerning $p_1, \ldots, p_n$) with the additional information provided in the user's prompt about p_0, the model generalizes to produce an output that reveals sensitive information about p_0's health status. This problem is related to what Shumailov et al. [115] calls "ununlearning": "unlearned knowledge gets reintroduced in-context, effectively rendering the model capable of behaving as if it knows the forgotten knowledge."

[23] It is in a sense possible to make unlearning effective by being so over-inclusive—by removing or suppressing so much information—that the model loses its ability to produce *anything* useful. However, it is also arguable that this is not a successful application of unlearning, since it is not "targeted" in a meaningful way.

Perhaps this is an obvious possible outcome: the model has not unlearned *the ability to generalize* about p_0 from removing p_0's data alone. (Indeed, generalization is arguably the main goal of machine learning; see Section 3.) When taking front-end outputs into consideration, not just back-end removal of information, it is arguably the case that, for some contexts (Section 6), information about p_0 has not been successfully unlearned in a meaningful way. That is, removing sensitive information about p_0 from the model on the back-end does not mean that the model could not be prompted to produce sensitive information about p_0 on the front-end (e.g., it may still be possible to use the model to make sensitive inferences about p_0). Structural removal of only p_0's examples is perhaps under-inclusive, when taking into consideration how the retrained model might respond to prompts. But removing examples related to $p_1, \ldots, p_n$ would be over-inclusive (Mismatch 2).

Mismatch 4 *Models are not equivalent to how their outputs are put to use.*

A corollary follows from Mismatch 3, which involves another slippage. Among those who mistakenly believe that unlearning is a standalone solution for effectively moderating possible model outputs on the front-end, some reason further that this could help curtail further downstream undesirable or malicious model uses in practice.[24] It is obvious, but nevertheless important, to emphasize that seemingly innocuous outputs could be put to undesirable downstream uses. To greater or lesser extents, different unlearning methods can remove the effect of observed information from models or suppress certain types of model outputs; but the type of control this provides is localized to the model's parameters and outputs. Additional control would require anticipating how a person or other agent might behave with generative-AI outputs in an unbounded number of contexts—-none of which is reasonably under the purview of machine unlearning.

6 Machine Unlearning in Policy and Practice

The mismatches (Section 5) between unlearning targets (Section 3), methods (Section 4), and goals (Section 2) present clear technical and substantive challenges. We next consider how these mismatches manifest in specific ways and introduce complications for three law and policy areas where researchers and organizations have suggested that unlearning could help achieve certain desired ends for broader impact: privacy (Section 6.1), copyright (Section 6.2), and safety (Section 6.3).

A common theme for these areas is the underlying assumption that using unlearning methods to constrain model outputs could potentially act in the service of more general ends for content moderation—to prevent users from generating potentially private, copyright-infringing, or unsafe outputs. For each, bringing in domain-specific details amplifies the mismatches that we describe in Section 5, revealing an even deeper disconnect between the use of unlearning methods in practice, actual policy considerations, and regulatory compliance. To address this disconnect, judges and policymakers will need to set reasonable expectations concerning the imperfect outcomes of best-effort implementations of unlearning methods to support specific policy goals (Section 7.2).

6.1 Privacy

Given the breadth of data generative-AI models ingest for training, many experts worry about models revealing private information that they were trained on through their generations [e.g., 9, 12, 31, 94, 99, 120]. These concerns relate to privacy rights in different jurisdictions and associated remedies to preserve those rights. As discussed above (Section 2.1), in a number of jurisdictions, individuals have the right to request that organizations delete their personal data, also referred to as the "right to be forgotten," following Article 17 of the GDPR [102]. Regulators may seek remedies that require the removal of a set of data examples used for model training (Section 4.1) that they assess to have been unlawfully or improperly collected. In other cases, remedies may be more far-reaching; regulators may seek to delete a trained model in its entirety, which is often referred to as **algorithmic disgorgement** [1, 77, 141].

[24] Similar observations have been made in algorithmic-fairness contexts: a model that produces risk scores for criminal recidivism is distinct from the distribution of scores that model produces over a given population, which is again distinct from how the (distribution of) scores gets used for decision-making, e.g., a magistrate using those risk scores to inform their judgments about whether or not to grant a defendant bail upon rearrest [e.g., 7, 32]. Nevertheless, this slippage takes on an expanded meaning for generative-AI contexts.

In the context of data protection and privacy of personal information, deletion requirements[25] also often demand removal of data within a certain time limit. For example, the California Consumer Privacy Act (CCPA) requires businesses to reply to data-deletion requests within 45 business days, extendable to 90 business days [14]. While deleting specific records from a traditional database or dataset in this time frame is often technically feasible, some laws recognize that, in other cases, deletion may be less feasible or require "disproportionate effort." In such circumstances, some jurisdictions may provide exceptions to deletion requirements, for example, in cases where the information is otherwise publicly available, or is necessary to complete a transaction, achieve purposes in the public interest, or comply with other legal obligations [e.g., Article 17(3), 102]. Some jurisdictions have also ruled that information should be suppressed from being presented to users, even if the underlying data could not be deleted [39].

Such requirements and possible remedies have motivated attention to methods in machine unlearning as an approach for achieving compliance with privacy legislation. For example, at least in principle, unlearning perhaps seems like a direct match for satisfying data-deletion requests in a more efficient and targeted way: unlearning methods could be a finer-grained alternative to complete model disgorgement, less expensive and more efficient than retraining models from scratch on new datasets, and, due to improved efficiency, more suitable for satisfying deletion time frames generally required by privacy and data protection laws. More generally, unlearning methods have appeal because they seem to strike a balance between desires to enable large-scale training of AI models and to retain a toolkit of interventions that advance privacy. But of course, as with assessing any potential remedy, it would likely be necessary to consider the feasibility or reasonableness of using a particular unlearning method in practice, with respect to desired targets, costs, and overall effectiveness.

We address some of these considerations below, organized around three broad goals that we observe for unlearning-related efforts for Generative AI that pertain to privacy concerns grounded in regulatory frameworks: (1) data deletion (i.e., removing observed information from a model's training dataset), as well as suppression at generation time of (2) outputs that resemble personal information and (3) latent information.

Data deletion requests (i.e., removal of observed information). Data deletion involves entirely removing observed information (Definition 1) from training datasets. It is often motivated by (1) legal rights of individuals (often called **data subjects**) to request the deletion of personal data coupled with (2) implicit expectations that removing observed information will help to mitigate against models outputting verbatim pieces of potentially private training data.

In some cases, for Generative AI, data deletion is most straightforwardly implemented by retraining a model from scratch or, if applicable or feasible,[26] some structural-removal method (Section 4.1) with the right-exercising data subject's examples removed from the training dataset.[27] Depending on the reason for deletion, this might on its own suffice for certain deletion requests. For example, if the main issue being remedied is lack of consent for the use of a data subject's personal data, output suppression might not be a relevant remedy. It also may not be an issue if inferences can still be made about the data subject, so long as those inferences are based on training and prompt data that have been processed with proper consent.

However, even though removal methods seem like a direct match for implementing data deletion, they are not a straightforward solution in practice. In general, there are significant technical challenges in identifying all instances of observed information that meet certain privacy-relevant

[25] Other legislation, e.g., the Virginia Consumer Data Protection Act [135], also has requirements for data correction, not just data deletion. Correction could be operationalized as removal of the incorrect data and replacing them with (i.e., retraining with) the corrected data.

[26] Recall that structural-removal methods are not currently widely usable for generative-AI contexts. Methods that approximate structural removal do not guarantee that the targeted observed information is actually removed. (See Section 4.1 & Section 5, Mismatch 1.)

[27] Separately, some argue that models trained with methods like **differential privacy** are sufficient to preserve a data subject's privacy; in such cases, some believe that the data subject's privacy is retained even though their examples are included in the training data, so it would not be necessary to use machine-unlearning methods to remove them. See Brown et al. [12] for additional discussion.

criteria within large-scale datasets.[28] Consider a deletion request to remove images of a particular data subject from the training data used to produce an image-generation model. This may be computationally expensive at scale and require the use of ML tools that are themselves imperfect at identifying all such examples. Even if we had tools to guarantee perfect identification of a given set of examples, in privacy contexts, there are fundamental challenges for drawing boundaries around what this set ought to include (Mismatch 2). Should the set to remove be conservative and include images that only feature the right-exercising data subject? Should it be more (and perhaps overly) inclusive, and also cover family photos where other data subjects are also present? Photos where the right-exercising data subject is in the background?[29]

Further, even if perfect removal of observed information were to satisfy deletion requests in name, this would not guarantee that a model could not produce outputs that reflect these data, which could matter in contexts where model outputs are also of concern. Even in the best-case scenario, an unlearning method that satisfies privacy requirements on the back-end (i.e., with respect to how models are trained) would be insufficient to guarantee privacy is preserved on the front-end (i.e., at generation time) (Mismatches 2 & 3). In general, on their own, methods for removal are insufficient to guarantee privacy is preserved on the front-end.

Output suppression of observed information (and information that resembles it). Stakeholders may also focus efforts on suppressing certain pieces of observed information from outputs, either through modifying the generative-AI model (e.g., with RLHF) or system-level filters (Section 4.2). This would cover cases in which a particular piece of observed information was for some reason not included in a deletion request (e.g., because there was a failure to identify or deem it appropriate for deletion), as well as cases in which latent information in the model (Definition 2) enables the generation of outputs that resemble the right-exercising individual's personal information.

Output suppression would prevent surfacing observed information to end users, but would not actively remove it from models or training datasets. This approach most closely resembles the notion of the "right to be forgotten" that follows from the Court of Justice of the European Union (CJEU) ruling in 2019. The CJEU ruled that Google should act on requests from data subjects by suppressing information from a viewable index in relevant jurisdictions, but did not necessarily require deleting that information from underlying data storage [39].[30] Approaches that support suppression of certain types of outputs are imperfect (Section 4.2); it would be likely that efforts to suppress observed information would be subject to a test of reasonable or proportionate effort, with effectiveness determined by an evaluation of how difficult it would be to extract the suppressed observed information from the model following the application of technical or procedural interventions, for example, through red teaming and related procedures.[31]

Output suppression of latent information. Additionally, many privacy practitioners have come to recognize that simply restricting the collection or processing of certain observed information may not mitigate privacy concerns. This can be the case if technology enables an actor to infer information about a particular data subject based on latent information derived from similar data

[28] Current legislative deletion provisions tend to have concrete scopes for deletion criteria, such as deletion of data associated with a particular user account [e.g., 14]. Such boundaries are less clear for training datasets, for which the underlying training examples tend not to be organized in relation to their provenance [76], e.g., according to the user to which the data relate.

[29] Methods that approximate structural removal (Section 4.1) inherit these challenges. Unlike structural-removal methods, they do not guarantee with certainty that the chosen set of examples is removed from the model (Mismatch 1). The extent to which more efficient, but less accurate, approximate-structural methods could be sufficient to stand in for structural ones is a question for policymakers and regulators [28].

[30] In brief, the court found that the request in question fell under the law [39, paragraph 52], that removal of the information from all domains (not just those that reflect the Member States of the European Union) was an over-broad interpretation of the authority and scope for the relevant laws [39, paragraphs 59-65], and that it would suffice to de-reference the information "on the versions of that search engine corresponding to all the Member States, using, where necessary, measures which, while meeting the legal requirements, effectively prevent or, at the very least, seriously discourage an internet user conducting a search from one of the Member States on the basis of a data subject's name from gaining access, via the list of results displayed following that search, to the links which are the subject of that request" [39, paragraph 74].

[31] For discussions of red teaming, we refer to Feffer et al. [46] and Chouldechova et al. [21].

subjects who have consented to (or not objected to) data processing (e.g., inferring p_0's health status in Mismatch 3), or to infer sensitive characteristics from benign data that may be subject to fewer restrictions. For instance, California privacy regulators view such inferred information to still be personal information about consumers over which they can exercise their rights, when such information is used to make a profile about them [14].

Stakeholders may be concerned about two distinct elements of inferences like these that are derived from latent information. First, if a generative-AI model or system has explicitly generated and then stored inferred information about an individual, such that a new data point has been explicitly created (e.g., a data point about p_0's inferred health status), deleting that new data point from storage may be expected in order to meaningfully preserve that individual's privacy. However, if a model has the *capability* to draw connections about a data subject via latent information in its parameters and additional information provided in the user's prompt, output-suppression approaches may be more appropriate to prevent generations that compromise that data subject's privacy.[32]

Importantly, all three areas discussed above are neither mutually exclusive nor independent. They could each be implemented in the service of satisfying privacy aims. But this is also not straightforward, as sometimes different privacy goals and the relevant technical approaches to attempt to accomplish them may be in tension. For instance, implementing an output-suppression intervention may require a system operator to retain the information that should be prevented from being surfaced, in order to filter out this information from an output or filter out prompts that aim to solicit it.[33] Removal of observed information, meanwhile, may create the (as we have seen, false) impression that a model will not be able to produce a specific piece of information. Failing to implement efforts to prevent the generation of that information may lead to similar concerning impacts of the collection or retention of that data. Lastly, just as it is challenging to draw clear boundaries around which data to remove to satisfy a deletion request, it is similarly a difficult and open-ended problem to draw boundaries around what to suppress from model and system outputs.

6.2 U.S. Copyright

At first glance, as part of a response to claims in the U.S. that allege copyright infringement in connection with generative-AI models and systems, it may seem appealing to attempt to use machine-unlearning methods to target higher-order concepts (Definition 3) that relate to creative expression, as perhaps a way to operationalize notice-and-takedown requests [38].[34] However, U.S. copyright is not a straightforward problem, and unlearning is not a straightforward solution.[35]

We begin with some brief background on U.S. copyright law. Copyright law protects "original works of authorship fixed in any tangible medium of expression" [33]. This means that copyright protection extends to a particular image or a particular paragraph of writing, but not to any ideas or facts contained in it. Because copyright law gives creators the exclusive right to prepare reproductions (copies) and derivative works, courts examine whether potential copies are "substantially similar" to the original work, and whether those copies thus infringe on the rights of the copyright holder. Substantial similarity is a challenging concept with a varied and complicated history in copyright caselaw. Common tests to determine substantial similarity are subjective [110]; judgments for substantial similarity cannot "be reduced to a simple formula that can easily be applied across different works and genres" [77, p. 72].

[32] As in Section 3, the factuality of a piece of latent information is not relevant for our purposes. For example, making an incorrect inference about p_0's health status may still violate p_0's privacy. More generally, such false or "hallucinated" outputs can still cause harm.

[33] This tension—of needing to retain information in order to facilitate suppression—is also relevant for copyright (Section 6.2) and safety (Section 6.3). More generally, this tension pre-dates interest in machine unlearning for Generative AI. For instance, in the past, Facebook attempted to address the spread of NCII on its social-media platform by requesting users to upload the images in question to another Facebook-hosted tool, so that Facebook could identify and remove the images from the platform [62].

[34] See Lee et al. [77, Part II.G] for more detailed discussion on Section 512 and Generative AI.

[35] We limit our specific discussion to U.S. copyright. Other jurisdictions exhibit differences in copyright doctrine and caselaw, for example, with respect to exceptions to copyright-holders' exclusive rights. While we draw from U.S. doctrine and caselaw, the overarching points that we make in this section about unlearning, substantial similarity, and causation have broader relevance.

When one looks at the processes that go into the training, deployment, and use of generative-AI systems, there are several places where a "copy" could be made, e.g. copying the data examples in a training dataset within a program, training a model, and generating substantially similar copies of the data examples at generation time. Not all such copying is copyright infringement. Depending on the circumstances, the defense of "fair use" may protect uses of a work for a different purpose (potentially including training a model) and some uses that modify the work in "transformative" ways. Both "fair use" and "transformative use" are also technical terms in copyright [34, 79]. Whether a particular use is fair depends on the facts of each case, including the effect on the market for the copyrighted work in question. But courts in the most analogous situations have held intermediate copies made for purposes of generating new output (e.g., for model training) to be fair use [6, 78]. However, they are less likely to hold the output of a generative-AI model or system is fair use if it is substantially similar to the original—unless it parodies or comments on the original [60, 77].

Following from this brief background, we focus our discussion of copyright and machine unlearning with respect to training-data inputs on the back-end and generated outputs on the front-end. We do not address potential implications for intermediate artifacts, e.g., a trained model's parameters [26].

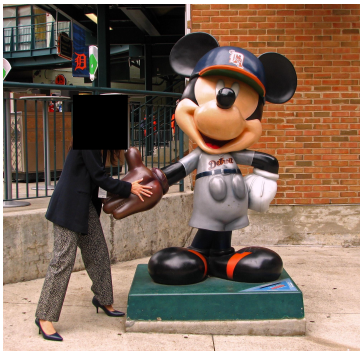
Suppression of substantially similar outputs. If a model generates an output that is substantially similar to a copyrighted work, in response, it may be tempting to use machine-unlearning methods to remove the ability to do so. For the reasons discussed above concerning output-suppression methods (Section 4.2), this is challenging for each unlearning target because there is no notion of similarity that can be used to programmatically and comprehensively determine which works are substantially similar in the interest of copyright law [77, 110]. A feature, like a color scheme, may be problematic if copied from one work but not from another work (so long as it is copyrightable subject matter). The techniques used to suppress generations similar to a particular in-copyright image of Mickey Mouse may not generalize to suppressing generations that are similar to another image of Mickey Mouse—or of any other Disney character [77, Part II.F]. That is, while such techniques could prevent potentially problematic outputs in some cases, they cannot generalize to all cases.

Further, in order to suppress certain outputs, for example, those that resemble in-copyright expression of "Spiderman," the overall generative-AI system likely needs to have learned information about this expression in order to *not* present it to the end user. For example, an output filter would need to be able to identify "Spiderman" (likely, from being trained on data that contain "Spiderman"-related expression) in order to filter it out. So, even if one were to remove all instances of "Spiderman" from the generative-AI *model*, more generally, it might be infeasible to remove all information about "Spiderman" from the generative-AI *system*; such information might be required to effectively implement output suppression at the system level.

Removal of specific training examples. If one were to remove a particular data example from the model, by contrast, this is a direct application of removal of a piece of observed information (Section 4.1). However, more generally, removal still might not always be the appropriate approach; it could constitute over-reach, since copyright does not forbid all forms of copying (e.g., fair uses, internal copies [57]). And removal could also likely be overbroad, as discussed in Mismatch 2, but with particular implications for copyright. Removal of a data example could prevent transformative, non-infringing uses of the data example in addition to potentially infringing ones. None of the unlearning methods we have described can or do distinguish between transformative fair uses and non-transformative superseding uses,[36] and transformativeness is not the only relevant factor for fair use. It is unreasonable to expect unlearning methods to capture these nuances. As evident from caselaw, courts themselves struggle to draw the line of fair use.[37]

[36] A superseding use is when, in the market for an original work, a new work replaces an original work (e.g., purchases of a fourth edition of a textbook replace the third edition). A non-transformative superseding use is a superseding use in which the new, replacing work does not change the character or purpose of the original work (e.g., a freely circulated digital PDF of a textbook dramatically changes the market for for-purchase hard copies of the textbook) [16].

[37] This challenge extends beyond unlearning methods. Distinguishing between transformative fair uses and non-transformative superseding uses requires context (e.g., how will the generation be used?) that is typically not currently available in generative-AI systems.



(a) Image from the training dataset

(b) Generation for the prompt `"Mickey Mouse"`

Figure 2: CommonCanvas is a research tool and text-to-image model [55], trained only using images with Creative Commons licenses. One can think of this model as a "gold-standard" baseline that does not contain in-copyright images of Mickey Mouse: the only examples in the training data that reflect the higher-order concept of "Mickey Mouse" are from personal photographs, e.g., (**a**) (redacted for privacy). Even without unlicensed, in-copyright training examples of Mickey Mouse, the model can generate outputs that resemble "Mickey Mouse," e.g., (**b**).

On the front-end at generation time, unlearning may also be ineffective in a variety of cases. As we have discussed throughout the piece, it is possible to run the "gold standard" for machine unlearning—to retrain a model from scratch without a specific piece of observed information (Section 4.1)—and still generate an output that is similar to that information or is otherwise similar to some higher-order concept reflected in that information (Mismatch 2). In such cases, this can be due to the presence of elements of the original work in other works in the training data, for example, related works, duplicates, or otherwise similar works that themselves may or may not be deemed infringing copies of the original work. For a concrete example, consider CommonCanvas, a text-to-image generation model, for which the training dataset's images all have Creative Commons licenses [55]. The training dataset does not contain reproductions of unlicensed, in-copyright images of Mickey Mouse; and yet, based on inclusion in the training dataset of licensed personal photographs (e.g., from Disney World), it is still possible for CommonCanvas to generate images that could be judged substantially similar to "Mickey Mouse" (Figure 2).[38]

Unlearning methods as tools for causation. We next consider the use of a machine-unlearning method as a tool for determining causation in a copyright infringement suit. If a generated output is alleged to be too similar to a particular plaintiff's creative work, the plaintiff (e.g., an artist) will have to prove copying to establish copyright infringement. Defendants (e.g., a generative-AI company) may attempt to use a counterfactual argument to challenge causation. (Indeed, this is the same type of counterfactual that the "gold standard" for machine unlearning attempts to establish. See Section 4.1.) For instance, consider that a model generates an output that is substantially similar to the plaintiff's work, which was included in the model's training dataset; if a model had been trained without the inclusion of the plaintiff's work, would the model's generated output still be the same or very similar to the substantially-similar, potentially infringing output in question? If so, that may seem to suggest that the presence of the plaintiff's particular work in the training dataset for the original model did not cause the output. This is significant in copyright law because independent creation [47] is a defense to copyright infringement claims [113]. In this case, for example, the "gold standard" of retraining the model from scratch could be used to produce the counterfactual model without the plaintiff's work.

This reasoning is tempting but incorrect. Whether the counterfactual model was trained without the plaintiff's work is remarkably hard to assure. This, again, is because there may exist other derivative works of the plaintiff's work in the training data (e.g., see Figure 2). Those derivative works may have reasonable fair-use arguments. The converse, if a plaintiff can show that the

[38] Here, the training process still has **access** to images that contain information related to "Mickey Mouse," even if those images are not exact copies of unlicensed, in-copyright images of Mickey Mouse. Access to a copyrighted work is one type of evidence for proving copying in a copyright infringement suit [5].

output would have been significantly different without the plaintiff's work, is perhaps more convincing, but also flawed in practice. This is because the model training process (Section 2) is inherently non-deterministic.[39] Two models *trained on the same dataset* (let alone different ones) may generate significantly different sets of outputs for the same prompt. Judging whether these sets of outputs are meaningfully (and perhaps subtly) different is not a straightforward task to evaluate—neither with technical tools in machine learning [136] nor with respect to making judgments about similarity for copyright [26].

In all, these difficulties show that, while unlearning may seem appealing for copyright remedies, judges and practitioners must be careful to consider their current capabilities and limitations. Our discussion in this section shows that unlearning with current techniques will not map perfectly to the contours of copyright law. Output-suppression methods could be deemed acceptable if courts accept the empirical evaluations of these methods; but judges who consider unlearning as a remedy to copyright infringement will have to weigh the practical limitations of unlearning methods, as well as the potential unexpected consequences of unlearning on unrelated content. That is particularly true because copyright law imposes significant penalties for noncompliance, including statutory damages [37], destruction of infringing artifacts [36], and even criminal sanctions [35].

6.3 Safety

Last, we address concerns about AI safety, which span a wide range of issues and communities [e.g., 2, 3, 9, 13, 15, 27, 29, 44, 61, 105, 124, 130, 132, 139, 142, 147]. Among this variety, there is one recurring theme that is especially important to address in relation to machine unlearning: the concern that "dual-use," large-scale generative-AI models exhibit

> high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by … substantially lowering the barrier of entry for non-experts to design, synthesize, acquire, or use chemical, biological, radiological, or nuclear (CBRN) weapons [130].

We draw the quote above from the U.S. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence; however, similar concerns and language can be found in a variety of legislative and policy documents, including the E.U. AI Act [44, Recital 110], the International Scientific Report on the Safety of Advanced AI produced by the AI Seoul Summit [9, Chapter 4], and OpenAI's Preparedness Framework [105]. Generative-AI models and systems "are sometimes called 'dual-use' because of their potential for both benefit and harm" [133].

Some researchers and policymakers claim that, to limit potential harmful uses, machine-unlearning methods could be used to remove "unsafe," "hazardous," or otherwise "undesirable behaviors" from generative-AI models [e.g., 9, 81, 83, 86, 89, 149]. For one notable example, in the cross-stakeholder AI Seoul Summit report, Bengio et al. claim that "'Machine unlearning' can help to remove certain undesirable capabilities," e.g., those "that could aid malicious users in making explosives, bioweapons, chemical weapons, and cyberattacks" [9, p. 75].

Unclear boundaries for removal. For now, we set aside questions of output suppression, and note that there are particular challenges for safety contexts with respect to drawing lines around what to target for removal from a model (Mismatch 2). For example, some proponents of unlearning as an approach for improving safety assume that specific topics with dual-use potential, such as synthetic biology or molecular generation, can be successfully targeted for removal. But topics like these are broad and under-specified, and relate to all sorts of observed information, latent information, and higher-order concepts (Section 3). How can such topics be adequately translated into specific observed-information targets to remove? How should one go about determining reasonable boundaries for which observed information should be kept and which should be targeted for removal?[40]

[39] One source of non-determinism is randomness in the order in which examples are surfaced to the model during training. Example ordering has an effect on the ultimate trained model's parameters and the models outputs [30, 41].

[40] For example, even after unlearning unsafe information related to biohazards to reduce unsafe question-answering capabilities, researchers have shown it is possible to recover such capabilities by further training the model on unrelated benign information. The boundary around which observed information (regardless of whether it is considered "safe" or "unsafe") contributes to these unsafe capabilities is unclear [152].

In certain cases, it may be possible to remove observed information that is intrinsically harmful or has the high potential to be put to harmful uses—for example, respectively, observed information for non-consensual intimate imagery (NCII) [101] or the molecular structure of the smallpox virus.[41] However, many types of observed information (Definition 1) do not fit into these categories. Instead, potential safety issues come about from latent information (Definition 2): the fact that many potentially dangerous items can be assembled using observed information that is itself innocuous or has significant legitimate uses. For instance, from all of the information in a high-school chemistry curriculum, it is possible to derive formulas for toxic molecules. But removal of all knowledge of high-school chemistry from a model to foreclose the possibility of such latent information is likely overbroad.

So far, we have only considered information on the back-end that pertains to the trained model; safety challenges are also difficult when we consider the front-end. As discussed in Section 2.3 and with respect to Mismatch 3, the open-ended format of user inputs to generative-AI models means that, via their prompts, end-users can introduce additional information into the model's context at generation time. This information could be otherwise absent from the model's training data and not reflected in the model's parameters. It could also overlap with or reflect observed information that was removed from the model using an unlearning method. By bringing this information back into the generative-AI system via the prompt, the model can still be used on the front-end to reason about the information it has unlearned; its output might even be the same as if an unlearning method had not been applied in the first place (Section 5, Shumailov et al. [115]).[42]

Inherent tensions for unlearning in dual-use systems. Separate from the practical difficulties discussed above, there is an even more fundamental challenge originating from the inherent nature of dual-use systems. By definition, dual-use systems can be put to potentially beneficial or potentially harmful uses [133]. It is not just the case that innocuous observed information could lead to potentially unsafe latent information in the trained model; it is also possible for generated outputs that are innocuous in isolation to be put to unsafe or otherwise undesirable downstream uses.

Consider the example of unlearning all information for "how to synthesize a toxic molecule," first introduced in Section 1. Setting aside the tractability of translating this into concrete targets to unlearn, knowledge of how to actually produce such a molecule is not a property of the model in isolation. The ability to actually synthesize it also depends on the knowledge of the user [119].[43] The particular user is clearly an important factor to consider with respect to downstream use. What if the user already has a recipe for making such a molecule (obtained from another source), and the generative-AI model lowers the barrier for creation of the molecule for the user by explaining, in detail, how to understand the details and nuances of the recipe that they do not understand on their own? What if the model provides a single "missing piece" of information that is innocuous on its own (e.g., details of a single chemical reaction) that, in combination with everything else this user knows, enables them to create the molecule?[44]

[41] Of course, a formula for a molecular structure is not sufficient to produce a molecule (Mismatch 4); but for sufficiently dangerous molecules, the formula itself might be considered a safety risk.

[42] Evaluations for the success of unlearning methods in safety contexts often do not explicitly test for this issue (Section 4.2). Many such evaluations rely on the WMDP benchmark [81], which is a multiple-choice question dataset that focuses on biological, chemical, and cyber-security risks. Setting aside the observation that multiple-choice questions may not in general be the most effective way to measure such risks, this evaluation setup does not allow for the type of more open-ended reasoning that this scenario presents.

[43] It is perhaps for this reason that OpenAI's Preparedness Framework categorizes CBRN risks in relation to both the model and the users. For example, this framework considers high risk to mean that the "Model enables an expert to develop a novel threat vector OR model provides meaningfully improved assistance that enables anyone with basic training in a relevant field (e.g., introductory undergraduate biology course) to be able to create a CBRN threat" [105, p. 9].

[44] One might critique this example for not qualifying to "substantially lower[] the barrier of entry for non-experts" to perform unsafe actions in the world. Perhaps the user could have found similar information through effective use of a non-generative-AI system, like a traditional search engine indexed over the public Internet, so it is questionable that using a generative-AI system was sufficiently "substantial" to make the task easier to execute. We set this question aside. It is not in scope for us to make claims about whether or not existing generative-AI models and systems meet the bar of what, for example, the U.S. executive order considers a meaningful safety risk [130]. Regardless, we can still address the extent to which unlearning can or cannot address cases like this one.

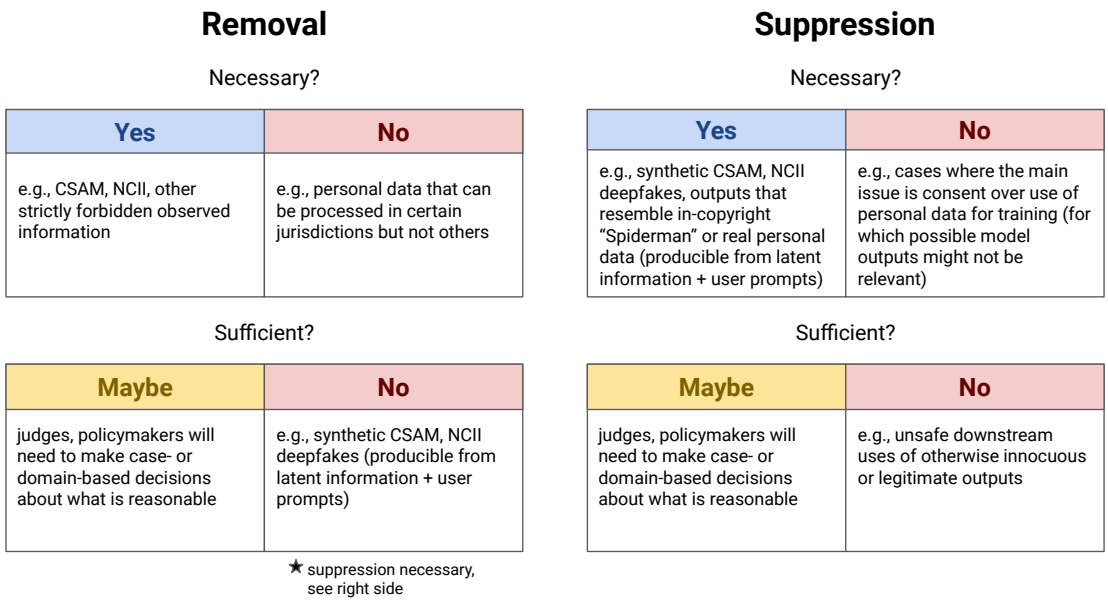


Figure 3: Following from the prior sections, four simple questions help clarify the usefulness of unlearning methods for removal and suppression to address policy aims for Generative AI. We consider if information removal of observed information is necessary and sufficient (**left**), and similarly if output suppression is necessary and sufficient (**right**). We provide examples of potential law and policy areas that could exhibit different answers to these questions. There are cases where removal may be necessary, but it is likely that removal is on its own insufficient. To moderate or constrain model outputs, suppression is likely always necessary, but suppression methods will also likely always be imperfect to catch all undesirable outputs.

For an additional example, consider a generative-AI system that includes a model trained for molecular generation—for suggesting formulas for new drugs and other molecules. Arguably, one of the purposes of such a system is to lower the barrier of expertise required for drug discovery. Even so, currently, a generative-AI system cannot on its own definitively determine that the molecules it produces are safe for human consumption; this is the point of lab experiments and drug trials [e.g., 128]. Once again, safety in this case is not an isolated property of the generative-AI model or system. Additional knowledge—in this case, derived through biochemical experimentation and human trials—is often needed to determine if the generative-AI outputs are beneficial or harmful. As discussed with respect to Mismatch 4, the types of control that methods for machine unlearning provide are incapable of preventing such downstream harmful outcomes. Unlearning can perhaps limit the information in the model or suppress model outputs, such that certain types of unsafe outputs are less likely, but it cannot guarantee that people will not put model outputs to unsafe uses [70].

7 Discussion and Conclusion

We have covered a lot of ground, but our main takeaway is fairly simple: there are significant mismatches between what technical methods for machine unlearning can achieve and aspirations for how these methods could make generative-AI models and systems operationalize law and policy aims in practice. To close, we briefly summarize key points about these mismatches (Section 7.1) and then we offer some concrete takeaways for ML research and AI policy (Section 7.2).

7.1 Recap: Machine unlearning doesn't do what you think

Now, having arrived at the end, we are able to revisit our main arguments through a set of relatively simple questions. With an understanding of the different goals (Section 2), targets (Section 3), methods (Section 4), and potential application domains (Section 6) for machine unlearning, we can ask whether (1) removing certain observed information or (2) suppressing certain outputs is (a) necessary or (b) sufficient to meaningfully comply with policy aims. (See Figure 3.)

Is observed-information removal necessary? Early goals for machine unlearning involved developing efficient methods to remove the effects of observed information (Section 3, Definition 1) from trained models (Section 2.2). This problem presents interesting and challenging technical problems for ML research, and also finds broader motivations in a particular (and contested) intepretation of data-deletion requirements in privacy legislation—namely, the "right to be forgotten" in the GDPR [102] (Section 2.1). Setting aside the case of the GDPR, removal from a model's training dataset (Section 4.1) may be necessary for some types of information, such as CSAM or NCII, where it is illegal or otherwise forbidden to observe this information in the model-training process (Section 6.3). In such cases, structural removal (Section 4.1) may be necessary and output-suppression methods (Section 4.2), which do not guarantee that information is removed from the model's parameters, may not suffice (Section 5, Mismatch 1).

In other cases, despite the intuitive alignment of the meanings of the words "removal" and "deletion," it is unclear if technical removal is indeed necessary to satisfy deletion requirements in law and policy. This lack of clarity is visible even in simpler cases that do not involve ML models. For instance, in some circumstances, the CJEU has ruled that suppressing a right-exercising individual's personal data in certain jurisdictions, rather than wholesale deletion across all jurisdictions, satisfies Article 17 of the GDPR (Section 6.1). While this example does not directly concern Generative AI, it is similarly unclear if removal is generally necessary to achieve desired ends in this setting.

Is observed-information removal sufficient? In limited cases, removal may be sufficient, for example, to satisfy data deletion requests (Section 6.1). In others, even if we could perfectly remove the effects of piece of targeted observed information from a model, this would likely not be enough to meet the prescribed goals of machine unlearning for Generative AI (Section 2.3). Instead, it may often be more important that a generative-AI model or system is unable to produce outputs that reflect certain observed information, certain latent information derived from it, or certain knowledge or abilities (Section 3). If the main goal is to moderate or constrain model outputs, removing a targeted piece of problematic observed information does not guarantee that a generative-AI model could not produce generations that reflect this information at generation time. For example, relying only on the "gold standard" approach to unlearning (Section 4.1) to remove real NCII from a model's training data cannot guarantee that the model could never be used to produce NCII deepfakes based on the combination of latent information in the model (Section 3, Definition 2) and the user's prompt (Section 5, Mismatch 3).

Is output suppression necessary? If preventing outputs like these is the main point (Sections 2.3 & 2.4), then output suppression is necessary. It is likely more important—both for technical methods (Section 4) and policy objectives (Section 6)—to devote attention to suppressing targeted types of model outputs. Output suppression, however, is a fundamentally different technical goal from removal of information from a model's training data (Section 5, Mismatches 2 & 3). To a certain extent, machine-unlearning research in Generative AI acknowledges these differences; one can see this in the shift to expand the family of methods for machine unlearning—to go beyond removal and to include technical approaches for output suppression (Section 4.2). (Indeed, if we attempt to draw lessons from cases like the CJEU example above, courts also clearly appreciate the conceptual differences between removal and suppression in rulings that find that there are cases in which suppression is a more appropriate operationalization of compliance with policy.)

Is output suppression sufficient? While there are cases where output suppression is important, it will not always on its own be sufficient. The appropriateness of some outputs will depend almost entirely on the end user and the context of use, not the model or system in isolation (Section 5, Mismatch 4). Seemingly innocuous or otherwise legitimate outputs could be put to all sorts of unsafe downstream uses (Section 6.3), and there is no feasible way for an output-suppression method like RLHF or an output filter (which is effectively a classifier) to anticipate this type of downstream outcome. More generally, output-suppression methods will likely always be imperfect (Section 4.2). There will likely be cases where generations violate a policy—e.g., contain information that resembles real personal data—but are not caught by an output filter and are surfaced to end users. This indicates that the sufficiency of a particular suppression method to prevent the generation of undesirable content will depend on the context of the particular system in which it is applied. For different contexts, policymakers and judges will need to identify appropriate guidance for what constitutes success for suppression, system developers will need to

figure out solutions for coming into compliance, and policymakers and judges will need to set reasonable expectations concerning whether a system developer has taken reasonable efforts to achieve compliance (Section 7.2).[45]

The main takeaway from asking these four questions is that the answers will depend on the specific context in law and policy (Figure 3). Removal may be necessary on occasion for specific types of data that models should never have seen during training. But removal often is not the main concern for the policy goals where machine unlearning gets invoked as a possible technical solution for Generative AI. Output suppression may generally be a more appropriate, but nevertheless imperfect, approach. In some cases, both removal and suppression may both be in order. In others still, there can be irreconcilable tensions between removal and suppression. That is, to effectively filter out some generations from being presented to end users, it may in fact be necessary for the overarching generative-AI system to retain and leverage related information (Section 4). Even if we were to remove "Spiderman" from a generative-AI model, the generative-AI system that contains this model would likely still need to have access to information about "Spiderman" in order to effectively suppress generations that resemble "Spiderman" (Section 6.2).

7.2 Takeaways for ML research and AI policy

Following from above, we offer five takeaways for ML researchers and AI policymakers.

Unlearning is just one approach in the ML and policy toolkit. There are clear gaps for what machine unlearning can do to achieve policy aims, both with respect to methods for removal of observed information and output suppression. Different methods may be useful to certain extents in specific contexts, but it is important to view unlearning as just one approach among may others (e.g., acceptable usage policies and responsible AI licenses [72, 92]) that could sometimes help achieve specific policy aims. Nevertheless, ML researchers should not claim—and policymakers should not misunderstand—that machine unlearning is generally on its own effective for making generative-AI models and their outputs compliant with any desired policy goals.

Evaluation of an unlearning method for a specific domain is a specific task. Further, such general claims about the broader impacts of unlearning are likely to be wrong from first principles because each legal and policy regime has its own specific expectations, which can be subtle and nuanced. To make rigorous claims about the broader usefulness of particular unlearning methods, as much as possible, ML experts need to evaluate specific unlearning techniques against specific regimes. This requires an understanding of these specific regimes, not just generalized ideas of how they might work—generalized ideas that may be so oversimplified that they are misleading or incorrect. To make claims about how an unlearning method might or might not be useful for operationalizing compliance with Article 17 of the GDPR, a layperson's reading of the text is not enough. It is important to be familiar with the complexity of different interpretations, rulings, and exceptions. To make claims about the relevance of an unlearning method for U.S. copyright compliance, it is important to make specific claims about specific areas of copyright law, rather than to treat copyright law as a monolith [77]. At a minimum, this requires understanding those specific areas of copyright.

The appropriateness of a particular technical mitigation hinges on these specifics. They cannot be overlooked or abstracted away. For one, as we have seen throughout, these specifics can illuminate whether removal or suppression is the right technical goal to pursue for a specific substantive end. In doing so, it becomes unclear if the original goal of removal of information from a model (Sections 2.1 & 2.2) is the most relevant technical end to pursue for law and policy impact. In many cases, it seems like output suppression is what interested parties really care about (Sections 2.3 & 2.4). Output suppression–which does not necessarily have anything to do with "unlearning" information from a model's parameters—is perhaps a more relevant area of focus for ML research that aspires to influence policy. For another, a clear understanding of the specific goals of specific pieces of law or policy is important for guiding the right set of solutions—technical or otherwise. In some regimes, perfect guarantees may be an unnecessary or undue burden for model developers and custodians. It may not be relevant to focus research efforts on producing methods that guarantee with certainty that a particular piece of information

[45] Both are non-trivial tasks, given the non-determinism of generative-AI system outputs. See Cooper and Grimmelmann [26, Part III.D], Wallach et al. [136], and Section 4.2.

is removed or suppressed. Reasonable efforts to remove or suppress may be sufficient in some legal contexts, even if their results are imperfect. Of course, this will depend on the needs judges and policymakers articulate for the particular domain.

Understanding unlearning as a generative-AI *systems* problem. From our discussion, it should be clear that machine unlearning for Generative AI does not only concern generative-AI *models*; it more generally concerns the generative-AI *systems* in which these models are embedded. Systems-level interventions (e.g., content filters) are an important tool for constraining outputs (Section 4.2); evaluating such interventions clearly requires systems-level analysis [28, 104]. Open-weight models, like Meta's family of Llama models [87], therefore present different challenges for unlearning. These models are released as their parameters; on their own, they cannot implement system-level guardrails—for unlearning or other purposes. In order to achieve this type of functionality, developers who use open-weight models for their own systems would need to implement their own mechanisms for output suppression, or to incorporate other available software that is intended for this purpose [e.g., 65].

Setting reasonable goals and expectations for unlearning. It is also important for judges and policymakers to realize that, in general, it is unlikely that technical solutions for unlearning will get significantly better anytime soon. It is unlikely that all that is needed is a few more years of research and development for unlearning methods to wholly achieve desired policy goals. Instead, it will be important to modify expectations for machine unlearning in policy norms. This necessarily includes thinking through specific policy goals and, when using technical methods to achieve those goals, what should constitute reasonable best efforts in different contexts with respect to removing or suppressing unwanted information from models and system outputs. For example, judges or regulators may expect best efforts to have observed information removed from a model and related information suppressed from its outputs; and, if a company meets this bar, they would not seek massive fines if somehow the model's outputs still approximate that information. We expect that the focus would then become on the remediation process—i.e., did a developer take reasonable steps—and not perfect results.

There are no general-purpose solutions to constrain generative technologies. Finally, and more generally, policymakers should resist the tendency to think that unlearning methods can lead to generative-AI models that can do "everything but X." One of the strongest appeals of many generative-AI systems is that they are general-purpose: they can be adapted to a wide range of uses and produce a wide range of useful outputs. A superficial understanding of machine unlearning is that it can surgically and completely remove specific capabilities from a model while leaving everything else about the model unchanged. As we have seen, this is not what unlearning methods actually accomplish. The same power to abstract and generalize that makes these models so useful also means that, with small targeted changes, they are often still capable of exhibiting similar behavior. To use a biological analogy, people who have forgotten a fact will often remember that same fact later when prompted differently to recall it; people who have suffered a stroke can sometimes regain main of the cognitive functions that were temporarily impaired. The brain is too complex and too capable for targeted unlearning to be workable; the *Men in Black* neuralyzer is science fiction.

This lesson is familiar from other generative technologies like the PC and the Internet [31, 151]. Ed Felten calls it the "Fallacy of the Almost-General-Purpose Computer" [49]. For example, the PC has the ability to be adapted to a wide range of computational tasks through suitable configuration and inputs; this means that there is no simple or reliable way to prevent a computer (let alone a generative-AI system) from ever being used to violate privacy, infringe copyright, or design a dangerous molecule—not without fundamentally compromising the flexibility and power that make it so useful. A toaster cannot design a bioweapon, but a toaster also cannot do much besides make toast. A computer can, and so can a generative-AI system. This is an inherent tension with all generative systems. We can try to tether or constrain different aspects of this generativity in different ways, but we will not be able block its capacity for harmful uses with one, comprehensive method—from either technology or policy—that will work in all possible contexts.

Acknowledgments

We thank the individual co-organizers of Evaluating Generative AI Systems: the Good, the Bad, and the Hype (GenLaw DC) for early conversations about machine unlearning that helped spark this collaboration. We also thank The K&L Gates Initiative in Ethics and Computational Technologies at CMU, The GenLaw Center, Georgetown Institute for Technology Law & Policy, and the Center for Democracy & Technology, who co-hosted the GenLaw DC workshop. We thank the reviewers and attendees of the *2nd Workshop on Generative AI + Law* at *ICML '24* for their feedback on earlier versions of this work. Lastly, we thank Jared Bomberg, Nicholas Carlini, Alexandra Givens, Milad Nasr, Adam Roberts, Pam Samuelson, and Jon Small for useful discussion and feedback on this piece.

References

[1] Alessandro Achille, Michael Kearns, Carson Klingenberg, and Stefano Soatto. AI Model Disgorgement: Methods and Choices, 2023. URL https://arxiv.org/abs/2304.03545.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, 2016. URL https://arxiv.org/abs/1606.06565.

[3] Anthropic. The case for targeted regulation, October 2024. URL https://www.anthropic.com/news/the-case-for-targeted-regulation.

[4] Anthropic. Meet Claude, 2024. URL https://www.anthropic.com/claude.

[5] Art Attacks Ink, LLC v. MGA Ent. Inc. 581 F.3d 1138, 1143 (9th Cir. 2009).

[6] Authors Guild v. Google, Inc. 804 F.3d 202 (2d Cir. 2015).

[7] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact, 2014.

[8] Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and Controlling Important Neurons in Neural Machine Translation, 2018. URL https://arxiv.org/abs/1811.01157.

[9] Yoshua Bengio et al. International Scientific Report on the Safety of Advanced AI: Interim Report, May 2024. URL https://assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf. AI Seoul Summit.

[10] Jaydeep Borkar. What can we learn from Data Leakage and Unlearning for Law?, 2023. URL https://arxiv.org/abs/2307.10476.

[11] Lucas Bourtoule, Varun Chandrasekaran, Christopher Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning. In *OAKLAND '21*, May 2021. URL https://arxiv.org/pdf/1912.03817.pdf.

[12] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What Does it Mean for a Language Model to Preserve Privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. URL https://doi.org/10.1145/3531146.3534642.

[13] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, 2018. URL https://arxiv.org/abs/1802.07228.

[14] California State Legislature. California consumer privacy act of 2018 (ccpa), 2018. URL https://cppa.ca.gov/regulations/pdf/cppa_act.pdf.

[15] California State Legislature. SB 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act. | Digital Democracy, 2024. URL https://digitaldemocracy.calmatters.org/bills/ca_202320240sb1047.

[16] Campbell v. Acuff-Rose Music, Inc. 510 U.S. 569 (1994).

[17] Yinzhi Cao and Junfeng 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.

[18] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying Memorization Across Neural Language Models. In *International Conference on Learning Representations*, 2023.

[19] Gert Cauwenberghs and Tomaso Poggio. Incremental and Decremental Support Vector Machine Learning. *Adv. Neural Inf. Process. Syst.*, 1, February 2001.

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

[21] Alexandra Chouldechova, A. Feder Cooper, Abhinav Palia, Dan Vann, Chad Atalla, Hannah Washington, Emily Sheng, and Hanna Wallach. Red Teaming: Everything Everywhere All at Once. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=KEggQCeDUA.

[22] Somnath Basu Roy Chowdhury, Krzysztof Choromanski, Arijit Sehanobish, Avinava Dubey, and Snigdha Chaturvedi. Towards scalable exact machine unlearning using parameter-efficient fine-tuning, 2024. URL https://arxiv.org/abs/2406.16257.

[23] Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-Shot Machine Unlearning. *Trans. Info. For. Sec.*, 18:2345–2354, January 2023. ISSN 1556-6013. URL https://doi.org/10.1109/TIFS.2023.3265506.

[24] Aloni Cohen, Adam Smith, Marika Swanberg, and Prashant Nalini Vasudevan. Control, Confidentiality, and the Right to be Forgotten, 2023. URL https://arxiv.org/abs/2210.07876.

[25] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability, 2023. URL https://arxiv.org/abs/2304.14997.

[26] A. Feder Cooper and James Grimmelmann. The Files are in the Computer: Copyright, Memorization, and Generative. *arXiv preprint arXiv:2404.12590*, 2024.

[27] A. Feder Cooper and Karen Levy. Fast or Accurate? Governing Conflicting Goals in Highly Autonomous Vehicles. *Colorado Technology Law Journal*, 20:249–277, 2022.

[28] A. Feder Cooper, Karen Levy, and Christopher De Sa. Accuracy-Efficiency Trade-Offs and Accountability in Distributed ML Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385534. doi: 10.1145/3465416.3483289.

[29] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 864–876, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533150.

[30] A. Feder Cooper, Wentao Guo, Khiem Pham, Tiancheng Yuan, Charlie F. Ruan, Yucheng Lu, and Christopher De Sa. Coordinating Distributed Example Orders for Provably Accelerated Training. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ISRyILhAyS.

[31] A. Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A. Choquette-Choo, Niloofar Mireshghallah, Miles Brundage, David Mimno, Madiha Zahrah Choksi, Jack M. Balkin, Nicholas Carlini, Christopher De Sa, Jonathan Frankle, Deep Ganguli, Bryant Gipson, Andres Guadamuz, Swee Leng Harris, Abigail Z. Jacobs, Elizabeth Joh, Gautam Kamath, Mark Lemley, Cass Matthews, Christine McLeavey, Corynne McSherry, Milad Nasr, Paul Ohm, Adam Roberts, Tom Rubin, Pamela Samuelson, Ludwig Schubert, Kristen Vaccaro, Luis Villa, Felix Wu, and Elana Zeide. Report of the 1st Workshop on Generative AI and Law. *arXiv preprint arXiv:2311.06477*, 2023.

[32] A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22004–22012, March 2024.

[33] Copyright Law of the United States. 17 U.S. Code § 102 - Subject matter of copyright: In general, December 1990. URL https://www.law.cornell.edu/uscode/text/17/102.

[34] Copyright Law of the United States. 17 U.S. Code § 107 - Limitations on exclusive rights: Fair use, October 1992. URL https://www.law.cornell.edu/uscode/text/17/107.

[35] Copyright Law of the United States. 17 U.S. Code § 506 - Criminal offenses, October 2008. URL https://www.law.cornell.edu/uscode/text/17/506.

[36] Copyright Law of the United States. 17 U.S. Code § 503 - Remedies for infringement: Impounding and disposition of infringing articles, December 2010. URL https://www.law.cornell.edu/uscode/text/17/503.

[37] Copyright Law of the United States. 17 U.S. Code § 504 - Remedies for infringement: Damages and profits, December 2010. URL https://www.law.cornell.edu/uscode/text/17/504.

[38] Copyright Law of the United States. 17 U.S. Code § 512 - Limitations on liability relating to material online, December 2010. URL https://www.law.cornell.edu/uscode/text/17/512.

[39] Court of Justice of the European Union. Judgment in Case C-507/17 Google LLC, successor in law to Google Inc. v Commission nationale de l'informatique et des libertés (CNIL), September 2019. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62017CJ0507. Press Release No. 112/19, Luxembourg.

[40] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge Neurons in Pretrained Transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581.

[41] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping, 2020. URL https://arxiv.org/abs/2002.06305.

[42] Thorsten Eisenhofer, Doreen Riepel, Varun Chandrasekaran, Esha Ghosh, Olga Ohrimenko, and Nicolas Papernot. Verifiable and Provably Secure Machine Unlearning, 2023.

[43] Ronen Eldan and Mark Russinovich. Who's Harry Potter? Approximate Unlearning in LLMs, 2023.

[44] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689. Official Journal of the European Union.

[45] Federico Fabbrini and Edoardo Celeste. The Right to Be Forgotten in the Digital Age: The Challenges of Data Protection Beyond Borders. *German Law Journal*, 21(S1):55–65, 2020. doi: 10.1017/glj.2020.14.

[46] Michael Feffer, Anusha Sinha, Wesley Hanwen Deng, Zachary C. Lipton, and Hoda Heidari. Red-Teaming for Generative AI: Silver Bullet or Security Theater?, 2024. URL https://arxiv.org/abs/2401.15897.

[47] Feist Publications v. Rural Telephone Service Company. 499 U.S. 340 (1991).

[48] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 954–959, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369794.

[49] Ed Felten. The Fallacy of the Almost-General-Purpose Computer, October 2002. URL https://freedom-to-tinker.com/2002/10/14/fallacy-almost-general-purpose-computer/.

[50] Luciano Floridi. Machine Unlearning: Its Nature, Scope, and Importance for a "Delete Culture". *Philosophy & Technology*, 36, 2023.

[51] Geneva Internet Platform. AI's right to forget – Machine unlearning. *digwatch*, August 2023. URL https://dig.watch/updates/ais-right-to-forget-machine-unlearning.

[52] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward Layers Are Key-Value Memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446.

[53] Antonio A. Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI forget you: data deletion in machine learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[54] Shashwat Goel, Ameya Prabhu, Philip Torr, Ponnurangam Kumaraguru, and Amartya Sanyal. Corrective Machine Unlearning, 2024. URL https://arxiv.org/abs/2402.14015.

[55] Aaron Gokaslan, A. Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. CommonCanvas: An Open Diffusion Model Trained with Creative-Commons Images. *arXiv preprint arXiv:2310.16825*, 2023.

[56] Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks, 2015.

[57] James Grimmelmann. Copyright for Literate Robots. *Iowa Law Review*, 101:657, 2016.

[58] Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. *CoRR*, abs/1911.03030, 2019. URL http://arxiv.org/abs/1911.03030.

[59] Sarah Hastings-Woodhouse. Introduction to Mechanistic Interpretability, 2024. URL https://aisafetyfundamentals.com/blog/introduction-to-mechanistic-interpretability/.

[60] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation models and fair use, 2023. URL https://arxiv.org/abs/2303.15715.

[61] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved Problems in ML Safety, 2022. URL https://arxiv.org/abs/2109.13916.

[62] Alex Hern. Facebook launching tools to tackle revenge porn. *The Guardian*, April 2017. URL https://www.theguardian.com/technology/2017/apr/05/facebook-tools-revenge-porn.

[63] Emmie Hine, Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. Supporting Trustworthy AI Through Machine Unlearning. *Science and Engineering Ethics*, 30, September 2024.

[64] Yangsibo Huang, Daogao Liu, Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Milad Nasr, Amer Sinha, and Chiyuan Zhang. Unlearn and Burn: Adversarial Machine Unlearning Requests Destroy Model Accuracy, 2024. URL https://arxiv.org/abs/2410.09591.

[65] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations, 2023. URL https://arxiv.org/abs/2312.06674.

[66] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge Unlearning for Mitigating Privacy Risks in Language Models, 2022. URL https://arxiv.org/abs/2210.01504.

[67] Hyejun Jeong, Shiqing Ma, and Amir Houmansadr. SoK: Challenges and Opportunities in Federated Unlearning, 2024.

[68] Susmit Jha and Sanjit A Seshia. A theory of formal synthesis via inductive learning. *Acta Informatica*, 54:693–726, 2017.

[69] Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. Soul: Unlocking the power of second-order optimization for llm unlearning, 2024. URL https://arxiv.org/abs/2404.18239.

[70] Erik Jones, Anca Dragan, and Jacob Steinhardt. Adversaries Can Misuse Combinations of Safe Models, 2024. URL https://arxiv.org/abs/2406.14595.

[71] Bjørn Aslak Juliussen, Jon Petter Rui, and Dag Johansen. Algorithms that forget: Machine unlearning and the right to erasure. *Computer Law & Security Review*, 51: 105885, 2023. ISSN 0267-3649. doi: https://doi.org/10.1016/j.clsr.2023.105885. URL https://www.sciencedirect.com/science/article/pii/S026736492300095X.

[72] Kevin Klyman. Acceptable Use Policies for Foundation Models. *arXiv preprint arXiv:2409.09041*, 2024.

[73] Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth. Privacy Adhering Machine Un-learning in NLP. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 268–277, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-ijcnlp.25. URL https://aclanthology.org/2023.findings-ijcnlp.25.

[74] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards Unbounded Machine Unlearning, 2023. URL https://arxiv.org/abs/2302.09880.

[75] Rachel Layne. How to Make AI 'Forget' All the Private Data It Shouldn't Have. *Harvard Business School*, February 2024. URL https://hbswk.hbs.edu/item/qa-seth-neel-on-machine-unlearning-and-the-right-to-be-forgotten.

[76] Katherine Lee, A. Feder Cooper, James Grimmelmann, and Daphne Ippolito. AI and Law: The Next Generation, 2023.

[77] Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain. *arXiv preprint arXiv:2309.08133*, 2023.

[78] Mark Lemley and Bryan Casey. Fair Learning. *Texas Law Review*, 99:743, 2021.

[79] Pierre N. Leval. Toward a Fair Use Standard. *Harvard Law Review*, 103(5):1105, 1990.

[80] Guihong Li, Hsiang Hsu, Chun-Fu Chen, and Radu Marculescu. Machine Unlearning for Image-to-Image Generative Models, 2024. URL https://arxiv.org/abs/2402.00351.

[81] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexander Wang, and Dan Hendrycks. The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning, 2024.

[82] Ken Ziyu Liu. Machine Unlearning in 2024, May 2024. URL https://ai.stanford.edu/~kzliu/blog/unlearning.

[83] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking Machine Unlearning for Large Language Models, 2024.

[84] Yang Liu, Zhuo Ma, Ximeng Liu, Jian Liu, Zhongyuan Jiang, Jianfeng Ma, Philip Yu, and Kui Ren. Learn to Forget: Machine Unlearning via Neuron Masking, 2021. URL https://arxiv.org/abs/2003.10933.

[85] Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The Right to be Forgotten in Federated Learning: An Efficient Realization with Rapid Retraining. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, page 1749–1758. IEEE Press, 2022. doi: 10.1109/INFOCOM48880.2022.9796721. URL https://doi.org/10.1109/INFOCOM48880.2022.9796721.

[86] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards Safer Large Language Models through Machine Unlearning, 2024. URL https://arxiv.org/abs/2402.10058.

[87] AI Meta Llama Team. The Llama 3 Herd of Models, 2024. URL https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.

[88] Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. Quark: Controllable text generation with reinforced unlearning, 2022. URL https://arxiv.org/abs/2205.13636.

[89] Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight Methods to Evaluate Robust Unlearning in LLMs, 2024.

[90] Ananth Mahadevan and Michael Mathioudakis. Certifiable Machine Unlearning for Linear Models, 2021. URL https://arxiv.org/abs/2106.15093.

[91] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A Task of Fictitious Unlearning for LLMs, 2024. URL https://arxiv.org/abs/2401.06121.

[92] Daniel McDuff, Tim Korjakow, Scott Cambo, Jesse Josua Benjamin, Jenny Lee, Yacine Jernite, Carlos Muñoz Ferrandis, Aaron Gokaslan, Alek Tarkowski, Joseph Lindley, A. Feder Cooper, and Danish Contractor. On the standardization of behavioral use clauses and their adoption for responsible licensing of ai. *arXiv preprint arXiv:2402.05979*, 2024.

[93] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

[94] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory, 2024. URL https://arxiv.org/abs/2310.17884.

[95] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-Based Model Editing at Scale. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mitchell22a.html.

[96] Christine Mui. AI's Next Challenge: how to forget. *Politico*, June 2024. URL https://www.politico.com/newsletters/digital-future-daily/2024/06/03/ai-machine-unlearning-forget-00161303.

[97] Neel Nanda. A Comprehensive Mechanistic Interpretability Explainer & Glossary, 2023. URL https://www.neelnanda.io/mechanistic-interpretability/glossary.

[98] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

[99] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable Extraction of Training Data from (Production) Language Models. *arXiv preprint arXiv:2311.17035*, 2023.

[100] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-Delete: Gradient-Based Methods for Machine Unlearning, 2020.

[101] Office of Science and Technology Policy. White House Announces New Private Sector Voluntary Commitments to Combat Image-Based Sexual Abuse, September 2024. URL https://www.whitehouse.gov/ostp/news-updates/2024/09/12/white-house-announces-new-private-sector-voluntary-commitments-to-combat-image-based-sexual-abuse/. The White House.

[102] Official Journal of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation), April 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[103] OpenAI. ChatGPT: Optimizing Language Models for Dialogue, 2022. URL https://web.archive.org/web/20221130180912/https://openai.com/blog/chatgpt/.

[104] OpenAI. GPT-4 System Card, March 2023. URL https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[105] OpenAI. Preparedness Framework (Beta), 2023. URL https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

[106] Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-Context Unlearning: Language Models as Few Shot Unlearners, 2024. URL https://arxiv.org/abs/2310.07579.

[107] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language Models as Knowledge Bases? *arXiv preprint arXiv:1909.01066*, 2019.

[108] USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V au2, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon, 2024. URL https://arxiv.org/abs/2406.17746.

[109] Youyang Qu, Xin Yuan, Ming Ding, Wei Ni, Thierry Rakotoarivelo, and David Smith. Learn to Unlearn: A Survey on Machine Unlearning, 2023.

[110] Rentmeester v. Nike, Inc. 883 F.3d 1111 (9th Cir. 2018).

[111] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035, 2019.

[112] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

[113] Selle v. Gibb. 741 F.2d 896 (7th Cir. 1984).

[114] Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. Exploring the Landscape of Machine Unlearning: A Comprehensive Survey and Taxonomy, 2024.

[115] Ilia Shumailov, Jamie Hayes, Eleni Triantafillou, Guillermo Ortiz-Jimenez, Nicolas Papernot, Matthew Jagielski, Itay Yona, Heidi Howard, and Eugene Bagdasaryan. UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI, 2024. URL https://arxiv.org/abs/2407.00106.

[116] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631 (8022):755–759, 2024.

[117] Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. Knowledge Unlearning for LLMs: Tasks, Methods, and Challenges, 2023.

[118] Alison Snyder. Machine forgetting: How difficult it is to get AI to forget. *Axios*, January 2024. URL https://www.axios.com/2024/01/12/ai-forget-unlearn-data-privacy.

[119] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. Can large language models democratize access to dual-use biotechnology?, 2023. URL https://arxiv.org/abs/2306.03809.

[120] Daniel J. Solove and Woodrow Hartzog. The Great Scrape: The Clash Between Scraping and Privacy, 2024. URL https://scholarship.law.bu.edu/faculty_scholarship/3917. Working Paper.

[121] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, 2022.

[122] David Marco Sommer, Liwei Song, Sameer Wagh, and Prateek Mittal. Towards Probabilistic Verification of Machine Unlearning, 2020.

[123] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, 2023. URL https://arxiv.org/abs/2206.04615.

[124] Elham Tabassi. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology (U.S.), Gaithersburg, MD, January 2023. URL http://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[125] Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-Resistant Safeguards for Open-Weight LLMs, 2024. URL https://arxiv.org/abs/2408.00761.

[126] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models, 2021. URL https://arxiv.org/abs/2102.02503.

[127] Gemini Team et al. Gemini: A Family of Highly Capable Multimodal Models, 2023.

[128] Thomas C. Terwilliger, Dorothee Liebschner, Tristan I. Croll, Christopher J. Williams, Airlie J. McCoy, Billy K. Poon, Pavel V. Afonine, Robert D. Oeffner, Jane S. Richardson, Randy J. Read, and Paul D. Adams. AlphaFold predictions are valuable hypotheses, and accelerate but do not replace experimental structure determination. *bioRxiv*, 2023. doi: 10.1101/2022.11.21.517405. URL https://www.biorxiv.org/content/early/2023/05/19/2022.11.21.517405.

[129] Pratiksha Thaker, Yash Maurya, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Guardrail baselines for unlearning in llms, 2024. URL https://arxiv.org/abs/2403.03329.

[130] The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, October 2023. URL https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. The White House.

[131] Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, Julio Jacques Junior, Ioannis Mitliagkas, Jun Wan, Lisheng Sun Hosoya, Sergio Escalera, Gintare Karolina Dziugaite, Peter Triantafillou, and Isabelle Guyon. Are we making progress in unlearning? Findings from the first NeurIPS unlearning competition, 2024. URL https://arxiv.org/abs/2406.09073.

[132] UK Government. Iconic Bletchley Park to host UK AI Safety Summit in early November, 2023. URL https://www.gov.uk/government/news/iconic-bletchley-park-to-host-uk-ai-safety-summit-in-early-november. Press release.

[133] U.S. Department of Commerce. Department of Commerce Announces New Guidance, Tools 270 Days Following President Biden's Executive Order on AI, July 2024. URL https://www.commerce.gov/news/press-releases/2024/07/department-commerce-announces-new-guidance-tools-270-days-following.

[134] UT News. Machine 'Unlearning' Helps Generative AI 'Forget' Copyright-Protected and Violent Content, 2024. URL https://news.utexas.edu/2024/03/21/machine-unlearning-helps-generative-ai-forget-copyright-protected-and-violent-content/.

[135] Virginia State Legislature. Virginia consumer data protection act of 2021 (vcdpa), 2021. URL https://law.lis.virginia.gov/vacodefull/title59.1/chapter53/.

[136] Hanna Wallach, Meera Desai, Nicholas Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Evaluating Generative AI Systems is a Social Science Measurement Challenge. *arXiv preprint arXiv:2411.10939*, 2024.

[137] Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A. Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. Evaluating Copyright Takedown Methods for Language Models, 2024. URL https://arxiv.org/abs/2406.18664.

[138] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models, 2022. URL https://arxiv.org/abs/2206.07682.

[139] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical Safety Evaluation of Generative AI Systems, 2023. URL https://arxiv.org/abs/2310.11986.

[140] Kyle Wiggers. Making AI Models Forget Undesirable Data Hurts Their Performance. *TechCrunch*, July 2024. URL https://techcrunch.com/2024/07/29/making-ai-models-forget-undesirable-data-hurts-their-performance/.

[141] Daniel Wilf-Townsend. The Deletion Remedy. *North Carolina Law Review*, 103, 2024. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4933011. Forthcoming 2025.

[142] Boming Xia, Qinghua Lu, Liming Zhu, and Zhenchang Xing. An AI System Evaluation Framework for Advancing AI Safety: Terminology, Taxonomy, Lifecycle Mapping, July 2024.

[143] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine Unlearning: A Survey, 2023.

[144] Haonan Yan, Xiaoguang Li, Ziyao Guo, Hui Li, Fenghua Li, and Xiaodong Lin. Arcane: An efficient architecture for exact machine unlearning. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4006–4013. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/556. URL https://doi.org/10.24963/ijcai.2022/556. Main Track.

[145] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning, 2024. URL https://arxiv.org/abs/2310.10683.

[146] Dayong Ye, Tianqing Zhu, Congcong Zhu, Derui Wang, Zewei Shi, Sheng Shen, Wanlei Zhou, and Minhui Xue. Reinforcement Unlearning, 2024.

[147] Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies, 2024. URL https://arxiv.org/abs/2406.17864.

[148] D. Zhang, P. Finckenberg-Broman, T. Hoang, et al. Right to be forgotten in the Era of large language models: implications, challenges, and solutions. *AI Ethics*, 2024. doi: 10.1007/s43681-024-00573-9. URL https://doi.org/10.1007/s43681-024-00573-9.

[149] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning, 2024. URL https://arxiv.org/abs/2404.05868.

[150] Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia Liu. UnlearnCanvas: A Stylized Image Dataset to Benchmark Machine Unlearning for Diffusion Models, 2024.

[151] Jonathan Zittrain. *The Future of the Internet–And How to Stop It*. Yale University Press, USA, 2008. ISBN 0300124872.

[152] Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An Adversarial Perspective on Machine Unlearning for AI Safety, 2024. URL https://arxiv.org/abs/2409.18025.

A Growth of unlearning papers over time

We provide some cursory evidence to support that there has been massive growth of machine unlearning papers in the last few years (Figure 4). Since we draw our results from arXiv, they do not include mentions of machine unlearning in technical reports (e.g., [9]) or other literature outside of computer science (e.g., [50]).

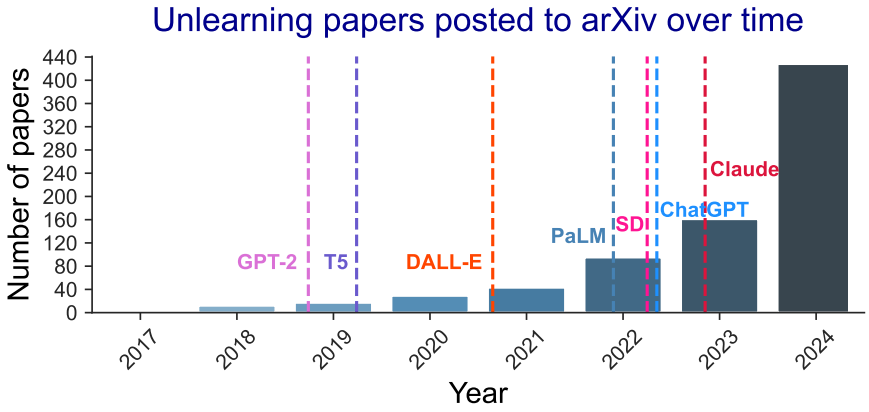


Figure 4: We scrape all papers that match `unlearn*` or `model forgetting` from arXiv and plot their counts over time, as of December 4, 2024. As of this date, there were a total of 810 papers starting from 1997 that matched out query. We indicate some important dates in the release of contemporary language and image generation models: **GPT-2**, **T5**, **DALL-E**, **PaLM**, **Stable Diffusion (SD)**, **ChatGPT**, and **Claude**.

GDPR was passed in 2016, and went into a effect in 2018. 790 of the papers have posting dates starting in 2016 (i.e., only 20 papers precede 2016). Of these 790 papers, 106 (i.e., 13.1%) mention "GDPR," "the right to be forgotten," or "RTBF" in the abstract. (These 106 papers are all from 2020-2024.) Given that we do not search the contents of all of the papers for these phrases, this serves as a lower bound of machine-unlearning papers that reference GDPR.

We also manually coded each paper with different categories, which we then used to assist with our literature review for the paper. Note that, as of December 4, there have been more unlearning papers (428) posted to arXiv in 2024 than there were in all prior years combined. While not easily visible in Figure 4, there were 3 papers in 2017.