ISSUE BRIEF

ARTIFICIAL INTELLIGENCE

Technologist Roundtable: Diving Deeper into "Unlearning" and Technical Guardrails

Post-Event Summary and Takeaways

August 2025

Stacey Gray, Marlene Smith

FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

About FPF

The Future of Privacy Forum (FPF) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. For more about FPF, please visit us at fpf.org/about and learn more about the FPF Center for Artificial Intelligence at fpf.org/issue/ai-ml/.

Event Summary

On **July 17, 2025**, the Future of Privacy Forum (FPF) hosted the second in a series of Technologist Roundtables with the goal of convening an open dialogue on complex technical questions that impact law and policy, and assisting global data protection and privacy policymakers in understanding the relevant technical basics of large language models (LLMs). We invited a range of academic technical experts to convene with each other and data protection regulators from around the world.

In emerging research literature and policy, the topic of "machine unlearning" and its related technical guardrails concerns the extent to which information can be "removed" or "forgotten" from an LLM or similar generative AI model or from an overall generative AI system. The topic is relevant to a range of policy goals, including complying with individual data subject deletion requests, respecting copyrighted information, building safety and related content protections, and overall performance. Depending on the goal at hand, different technical guardrails and means of operationalizing "unlearning" have different means of effectiveness.

This post-event summary contains highlights and key takeaways from the Roundtable on July 17, 2025. The takeaways build upon earlier discussions of the nature of LLMs and transformer architecture, and do not take a position on the legal question of whether personal information "exists within a model." As legal interpretations emerge and evolve, navigating these complex emerging issues increasingly requires understanding the technical literature and nature of a range of technical guardrails.

About the Authors

This event summary was drafted by Marlene Smith, FPF AI Research Assistant, and edited by Stacey Gray, FPF Senior Director for Artificial Intelligence.

Experts Acknowledgements

Thank you to Dr. Rob van Eijk, Elisabeth Gory, and to the academic researchers who attended this Roundtable: A. Feder Cooper (Stanford University, Microsoft), Ken Ziyu Liu (Stanford University), Weijia Shi (University of Washington), and Pratyush Maini (Carnegie Mellon University).

Table of Contents

1. Overview of Machine Unlearning

The roundtable began with a brief presentation that defined "machine unlearning" and gave an overview of its application to the components of a generative AI system, its different types (often categorized as "exact" and "approximate"), methods of data suppression, and an overview of "policy mismatches," or areas in which unlearning may not align with the policy goal or aim being sought.

a. Generative AI Systems and Memorization

Machine unlearning has arisen in response to a number of policy concerns, including generative AI systems that have generated, or leaked, sensitive information to end users. Early examples of this can be found in the NYT v. OpenAI lawsuit over ChatGPT's output of New York Times articles, and in research from Google DeepMind, which demonstrated the ability to extract verbatim training data from an early (2023) ChatGPT model through targeted prompting. Cases like these, in which models are able to output information from training datasets, are instances of ***model memorization.*** While "memorization" may have multiple definitions in machine learning literature, it is used here to refer to a model (nearly)-exactly learning specific training examples (i.e., verbatim memorization of training data).

While there are clear-cut concerns over this kind of model behavior (with respect to copyright, privacy, etc.), solutions are not so simple. The difficulty of removing information from a generative AI system hinges on the way machine learning models learn about and store information in the first place. Unlike databases, or systems that search and access stored information, machine learning models are trained to make predictions based on patterns deduced from vast amounts of training data. While generative AI models are often discussed as stand-alone objects, the models themselves are also better understood as a component in a larger generative AI system. Generative AI systems typically include additional components like a user interface, developer API, and input and output filters. Broadly defined, machine unlearning solutions may act on any element of the overall system.

b. Types of Unlearning

The process of removing information from a machine learning model is not well defined. While data can be deleted from a training set, this will not impact a model that has already been trained on this data. Broadly defined, machine unlearning methods aim to remove the *effect* of particular training data from a trained model in a targeted way.

"***Exact unlearning***" is the most straightforward version, and involves retraining a model on a refined version of the original training dataset, from which the undesirable data has been removed. One issue with exact unlearning is the cost: retraining a model from the ground up is expensive, and doing so in response to many unlearning requests is impractical.

Another issue with exact unlearning is that it can be difficult to determine the scope of information to remove from a training dataset in order to sufficiently alter model behavior. By definition, one can only remove data from a dataset that actually appears there. This information, which is explicitly presented to the model during training, is called "***observable information.***" Sometimes, however, the target information may be "***latent information***," or information that does not explicitly appear in a dataset but can be deduced from combining one or more pieces of observable information. For example, a dataset that includes two observable text elements "Lucy is going to John's house on Saturday" and "John's house is in Washington, DC" would include the latent information that Lucy will be in Washington, DC on Saturday. Latent information may rely on much more complicated patterns or networks of observable information; other examples of latent information could include abstract concepts like "truth" or "personhood."

Given these challenges, "***approximate unlearning***" methods aim to *approximate* the results of exact unlearning without retraining. These methods focus on removing information from the underlying model by adjusting its weights. Additionally, "***suppression methods***" work by suppressing information from surfacing to the end user by aligning the underlying model or adding system guardrails like suppression filters. These methods are designed to be more efficient than exact unlearning, and may aim to address issues related to information scope.

A specific definition of machine unlearning is constantly evolving, and different experts have different opinions on what should be included. The more stringent definition of unlearning includes 1) ***exact unlearning (retraining)*** and 2) ***approximate unlearning (altering model weights)***, but excludes 3) ***suppression methods that suppress system outputs without adjusting the underlying model***. A more expansive definition of unlearning includes all three of these approaches. While different researchers disagree on exact definitional boundaries, they agree that each of these methods has a role to play and that methods should be carefully disambiguated based on their specific approach.

c. Policy Mismatches

The current understanding of unlearning suffers from mismatches between policy goals and technical solutions. Specifically:

- Output suppression is not a replacement for removing training data in cases for which it matters what a model was trained on (even if it cannot generate that content);
- Conversely, removing training data and retraining (exact unlearning) does not guarantee meaningful output suppression. This is because new information can always be introduced by users, and generative AI is inherently capable of generating novel outputs;
- Models are not equivalent to their outputs; and
- Models are not equivalent to how their outputs are put to use (users can put the same information to various uses).

Presentation Highlights:

- Defined broadly, machine unlearning solutions can include exact unlearning (retraining), approximate unlearning (altering model weights), and suppression filters (preventing model input or output).
- Different solutions act on different components of an overall generative AI system.
- Observable information, or data that appears explicitly in a training dataset, is easier to remove and "unlearn" than latent information, or data that does not appear explicitly in the training dataset but may be learned through secondary connections.
- Policy goals determine the effectiveness of a given solution, and mismatches can occur. For example: Output suppression is not a replacement for removing training data in cases for which it matters what a model was trained on (even if it cannot generate that content); and conversely, removing training data does not guarantee meaningful output suppression.

2. Core Unlearning Methods (Exact and Approximate)

- ❖ *What are the differences between exact and approximate unlearning, and in what situations would each technique be deployed?*
- ❖ *How are both methods implemented and evaluated in practice?*
- ❖ *What are the limitations of exact and/or approximate unlearning; specifically, what kinds of information might they fail to remove?*

In this discussion, experts began by exploring the differences between exact and approximate unlearning. In general, exact unlearning can be thought of as the standard that approximate unlearning seeks to recreate more efficiently. For technologists, exact unlearning has a kind of mathematical guarantee that the removed data has never been seen by the model, while approximate unlearning (which involves altering model weights directly, without re-training) lacks this guarantee. Because of this, approximate unlearning solutions require probabilistic or empirical verification that the model no longer exhibits knowledge of the unlearned information.

However, experts agreed that the word "exact" can be misleading. While exact unlearning, at least in principle, aims for a provable guarantee, this guarantee is limited. Exact unlearning can *only* remove observable data from the training dataset, and thus makes no guarantees about what kind of latent data

may remain in the model. For example, a model trained completely on non-copyrighted images may still be able to reproduce content that depicts copyrighted characters (thus, exactly unlearning copyrighted material may not meet the intended goal of suppressing copyrighted outputs).

a. Unlearning verification

Different thresholds and methods for verifying or measuring "unlearning" may apply in different scenarios. For example, empirical measures of effectiveness could look at instances of verbatim regurgitation, semantic similarity, close summaries/paraphrasing, or factual overlap. Any of these might be appropriate choices in a given situation, so it is important to carefully outline the goals of any particular unlearning method and ensure that it is verified appropriately. For example, a method that seeks to limit exact regurgitation for copyright reasons should be verified to suppress exact regurgitation rather than summarization; conversely, a method that seeks to unlearn facts about individuals for privacy purposes should be verified to suppress factual information even if it is not directly summarized from an existing source. Additionally, the goals of unlearning should be contextually motivated; different use cases will raise different kinds of concerns.

One expert added that while suppression and removal are often seen as having binary measures (yes/no, is the data suppressed or removed), empirical measures are inherently not binary. Empirical measures test how a system behaves in practice in order to make a claim about *how likely* it is to consistently behave in a desired way (i.e., a percent chance).

b. Limitations of Unlearning in Practice

In practice, the group agreed that exact and approximate unlearning methods are primarily a research area of study, and that neither is being routinely deployed (yet) by developers. Instead, most developers are currently opting for suppression filters to prevent generative AI systems from surfacing undesirable information, such as copyrighted material.

Experts explained that weight editing unlearning solutions, which in theory are more efficient than re-training for developers, remain challenging to implement because of how information in a model is "tangled up." At times, unlearning that is targeted at specific information can have an undesirable impact on the system as a whole. For example, unlearning dangerous information about bioweapons may decrease a system's performance on science-related topics, or even affect seemingly unrelated capabilities, like essay writing.

Given costs of implementing unlearning, a potential option for developers might be to "batch" multiple unlearning requests and respond to them all at once during regularly scheduled retraining. One expert highlighted that while companies may release new versions of models or post-train relatively frequently, they may be hesitant to implement unlearning in this way. As discussed, removing large amounts of

information from a training dataset may affect more than just the knowledge/capabilities being targeted and may degrade model utility. One expert also highlighted that as model families become more mature, new releases may happen less frequently, and developers may opt to update a base model rather than completely retrain new models from scratch. Both of these aspects would make a batching approach impractical.

As previously mentioned, unlearning methods often struggle to remove latent information, or information that is implicitly inferrable from, rather than explicitly contained in, the training dataset. However, even observable data can be difficult to remove in its entirety. First, observable data can appear in variable ways, which may make it difficult to identify all instances for removal. For example, there are many ways to write out someone's address, and removing this information would require identifying each particular instance for removal.

Finally, data in a training set may not be structured or identified in any particular way, meaning that finding every piece of information linked to any particular person or thing may require searching it out. Finding effective ways to search data requires training good classifiers, which is not necessarily difficult to do, but can be costly and impractical to implement at scale. In relation to copyright, it is possible to train a classifier to comb through a training set in order to identify every chunk of text (of some length or longer) that appears in the first book of Harry Potter. However, doing this for every book, or every piece of copyrighted text, may be less feasible. Issues like this, which relate to the scale of the training data used to train these kinds of models, introduce important questions about operationalizing unlearning.

Discussion Takeaways (Part 1):

- Practical limitations of exact and approximate unlearning include: scoping challenges (observable vs. latent or inferred information); potential impact on seemingly unrelated model performance; and effective structuring of datasets at scale.
- Verification methods should be constructed with use, context, and policy goals in mind.
- In practice, most developers today are deploying suppression solutions, rather than exact or approximate unlearning, due to perceived challenges with the scope of data observability, the potential impact on model performance, and costs.

3. Technical Guardrails and Risk Mitigation

- *What is output suppression or scrubbing? At what stage does it intervene?*
- *What are the strengths of technical guardrails and how do they address limitations of unlearning? Where do both kinds of solutions fall short?*
- *How do different techniques (unlearning and other technical guardrails) relate to different policy or legal concerns?*
- *What novel empirical evaluation frameworks might help us better understand the residual information leakage after unlearning attempts?*
- *Given the inherent distributed nature of concepts in foundation models, how might we develop more sophisticated techniques to identify and bind the "information footprint" of specific data?*

a. Additional Technical Guardrails

In contrast to exact and approximate unlearning, there are other kinds of unlearning solutions and related “technical guardrails” that can be used at various stages of the model lifecycle. For example, experts discussed using **differential privacy** techniques during an initial round of training to limit the amount of information a model can learn from any single piece of data.

Some routine practices, like **deduplicating** training data, may decrease the likelihood that a model memorizes certain elements of the training set, although one panelist pointed out that the research on model memorization is evolving and may be more nuanced. In other words, deduplication may *not* be a sure way to discourage memorization or regurgitation, as models have been shown to memorize data based on factors other than repetition, such as novelty. There may also be strategic ways to **scaffold the training process** to make eventual unlearning easier. For instance, by saving model “checkpoints” before introducing new data during the training process, exact unlearning can be performed more easily, by retraining the entire model from the version stored immediately before it was trained on unwanted data.

Finally, experts discussed various **suppression filter** solutions, and agreed that these are the kinds of unlearning solutions being implemented today. Some suppression solutions function simply to prevent the system from outputting unwanted information by adding input and/or output filters, or prompts. These kinds of system-level features ‘check’ user inputs before prompting a model and/or ‘check’ model outputs before surfacing them to a user. More sophisticated solutions may also **fine-tune** a model, in order to teach it not to output certain information it learned in the initial round of training.

At least one expert emphasized that suppression filters can be jailbroken, and advocated for the idea that we should care about what information a model stores, regardless of what it may be able to routinely output.

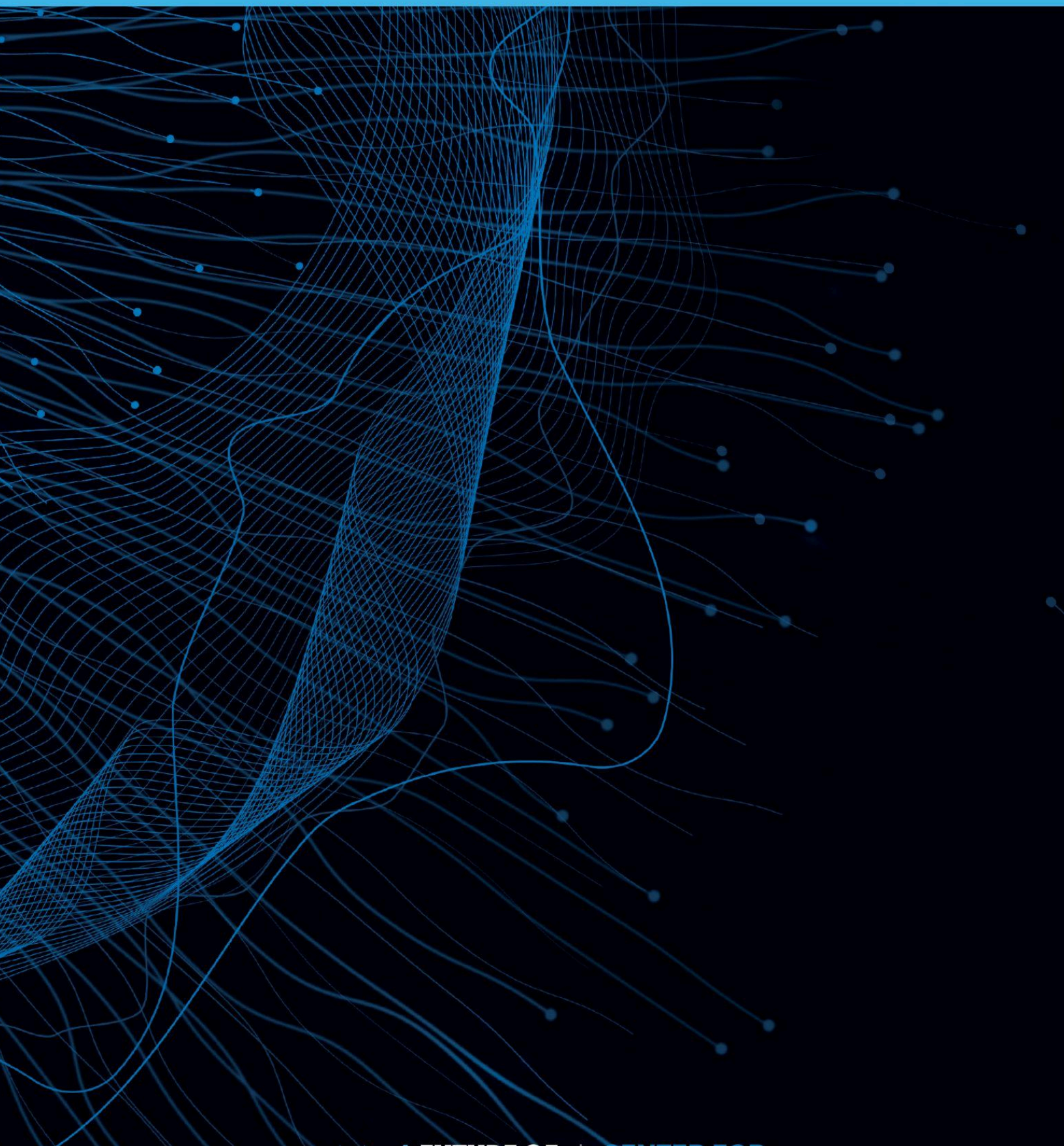
b. Looking Forward

Many expressed the need to create more strategic model architectures or training methodologies to facilitate unlearning. This could be achieved by training a base model on safe information before fine-tuning it with data that is more likely to need to be unlearned. Experts also emphasized the importance of carefully defining the policy goals of unlearning, as well as carefully constructing methods to verify that these goals have been met. Relatedly, one expert introduced the idea that unlearning is just one possible tool to address generative AI related harms to consumers. The expert encouraged exploring outside solutions like compensating and/or crediting people for their data.

Lastly, one expert framed the novel issues with generative AI as just one piece of an ongoing debate about the exciting and potentially harmful capabilities of generative technologies. Like the PC and the internet, generative AI systems are defined by their ability to produce generative outputs. This panelist warned that overly constraining generative AI systems would have important effects on the central capabilities of these systems. Additionally, constraining these models will not necessarily constrain end users from putting outputs to malicious use. The expert urged audience members to consider the tradeoffs of restraining these kinds of generative AI systems and to understand how different interventions may serve to better meet underlying issues.

Discussion Takeaways (Part 2):

- Other technical guardrails exist at different stages of the AI lifecycle to prevent memorization or output of undesired information. During training, these include differential privacy, deduplication of training data, or model checkpointing to reduce the need for unlearning or make unlearning easier later.
- Suppression techniques, which are currently widely in use, include alignment methods as well as system filters and prompts.
- Moving forward, we should design and train models in ways that facilitate unlearning and/or any elements of LLM design that align with policy goals, while keeping in mind the unique traits of generative technology and design solutions that thoughtfully interact with these systems, rather than targeting their defining (generative) capabilities.



FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

Washington, DC | Brussels | Singapore | Tel Aviv

info@FPF.org | FPF.org