Learning to Unlearn, Failing to Forget? Assessing Machine Unlearning Through Ethics and Epistemology

Iqra Aslam[1], Donal Khosrowi[1], Rahul Nagshi[2]

[1]Leibniz Universität Hannover

[2]Independent Researcher

iqra.aslam@stud.uni-hannover.de, donal.khosrowi@philos.uni-hannover.de, rahul.nt@outlook.com

Abstract

Machine Unlearning (MU) aims to remove the influence of unwanted data from trained AI models, driven by ethical/legal concerns like privacy (e.g., the Right to be Forgotten), bias mitigation, security, and copyright protection. This paper critically examines MU, arguing that it is currently unclear whether its technical methods and ethical goals are suitably aligned. Currently, important questions around what MU does, what it should do, and how its efforts align with stakeholder needs remain unaddressed. Drawing on insights from social epistemology and the ethics of forgetting, the paper makes progress in clarifying what MU is and whether it aligns with relevant goals. It does so by distinguishing three different senses of unlearning that vary in regard to what stakeholder needs they can adequately cater to. Drawing on a series of stylized cases, the paper highlights potential alignment gaps between MU's methods and its wider goals, and emphasizes the need for more concrete guidelines to assess MU's effectiveness, clearer ethical foundations, and improved stakeholder engagement.

1 Introduction

Artificial intelligence (AI) systems, particularly large language models (LLMs) like ChatGPT or LLaMA, and text-to-image (TTI) models like Stable Diffusion, continue proliferating across a wide range of domains such as education, medicine, creative industries, finance, and law (Chen et al. 2024). However, the very scale and power that make these models so versatile simultaneously introduce significant ethical, legal, and regulatory challenges. Some of the most urgent challenges surround the web-scale scraping of data needed to train these models. With models trained on extremely large amounts of uncurated data, numerous ethical and legal concerns arise, including regarding the unauthorized ingestion, retention and disclosure of personal data (e.g. people's medical records); algorithmic bias introduced by unrepresentative or inherently discriminatory training data; and copyright and other rights infringements by models that enable the reproduction of protected works (see Cooper et al. 2024). Jointly, these and other concerns raise important questions about how to remove unwanted data, information or knowledge from trained AI systems[1].

Concerns around sensitive information and knowledge stored and disseminated by digital technologies are not new or unique to AI. A well-known example is the landmark 2014 case involving Mario Costeja González, who successfully sued Google Spain to have outdated information about his past financial difficulties delisted from search results (Floridi 2015). This victory was pivotal in establishing the Right to be Forgotten (RTBF) as a recognized entitlement under the European Union's General Data Protection Regulation (GDPR) (Oesterling et al. 2023). Yet, while removing a hyperlink from a search index may be technically possible (albeit not entirely straightforward), applying the GDPR's Article 17 (the "right to erasure") to complex AI systems like LLMs presents formidable practical and conceptual hurdles (Zhang et al. 2024; Liu et al. 2025). Unlike databases or search indexes, neural networks do not store information in discrete, easily localizable units; their "memory" is distributed, statistical, and subsymbolic, encoded in intricate patterns over a model's parameters learned from vast datasets. As models increasingly store unwanted information in ways that are difficult for humans to detect and understand, they move us closer to a world where *forgetting*, as the erasure of information, may become increasingly elusive. This inherent complexity raises particularly acute challenges as AI systems are embedded in domains where the ethical and legal consequences of encoding and retaining unwanted information are particularly significant.

[1] While AI systems based on neural networks do not literally 'store' data, they may (and indeed perhaps should) encode information or knowledge contained in their training data. Throughout the paper, we use the terms 'information' and 'knowledge' conjunctively to denote epistemically relevant things a trained ML system encodes (alongside other things). Notably, while information denotes the kinds of things encoded by a model that may change the epistemic state of a user (e.g. their credences) regardless of their factual accuracy, knowledge requires a truth or accuracy condition.

In response to these challenges, the ML research community has embarked on a novel research program, called Machine Unlearning (MU). Broadly, MU is the project of engineering techniques to selectively remove unwanted information or knowledge from trained ML models, without incurring the cost of retraining models from scratch while also maintaining model performance (Wang et al. 2024).

However, while significant technical progress has been made in the MU literature in recent years, this paper argues that there are major unresolved tensions within the MU research program: First, there is significant ambiguity around what 'unlearning' and 'forgetting' mean. Second, it often remains unclear whether the methods proposed under the banner of MU align robustly with the ethical and legal goals they purport to serve, and with varied stakeholder demands arising across a complex topology of cases. In particular, the practical realities of implementing MU often involve approximations and trade-offs, raising urgent questions about whether the resulting modes of 'forgetting' truly satisfy pertinent moral requirements.

This paper centrally aims to draw out tensions and gaps between what kinds of unlearning and forgetting MU methods may enable, and the relevant unlearning needs of stakeholders. To be sure, we are not the first to explore such tensions, nor the first to scrutinize how MU fits with its stated objectives. Yet, while recent work (Cooper et al. 2024; Hine et al. 2024) pursues similar aims in regard to better understanding the goals of MU, our paper is the first to look specifically at MU's ability to cater to *ethical* rather than only legal or policy desiderata. What is more, our discussion also provides novel conceptual resources that aid in fine-graining our understanding of how MU's abilities map onto a range of plausible ethical goals. Finally, in doing so, we establish so far unrealized connections between MU and the social epistemology and ethics of forgetting literature (Basu 2020; Michaelian 2011), applying epistemic concepts such as epistemic innocence, epistemic virtue and duty, and epistemic responsibility to assess MU's functioning relative to stakeholder needs and pertinent moral values.

Throughout the paper, we refer to 'stakeholders' primarily in the ethically significant sense of those whose data, identity, or well-being are affected by how ML systems store, disseminate, or unlearn information, especially data subjects. While stakeholder interests can be varied and sometimes conflicting, our use of the term focuses on those whose claims can be grounded in moral and legal entitlements, such as rights to privacy, dignity, and freedom from discrimination. Importantly, we do not suggest that what stakeholders demand and what ethical principles require always overlap - after all, stakeholders may also make unsubstantiated claims. Instead, we focus on stakeholder claims that embody, and can be potentially grounded in, deeper normative commitments (e.g. to moral rights and other goods), thus requiring serious ethical consideration. Our inquiry hence treats stakeholder needs not as mere policy inputs or pragmatic constraints, but as expressions of morally relevant demands that MU systems ought to address.

The discussion proceeds as follows. In *Section 2*, we provide a short overview of the MU research program, reviewing its core motivations, key approaches, as well as its technical limitations. We then draw out gaps between the proposed goals of MU and their realization, pointing out that insights from the social epistemology and ethics of forgetting literature can help us understand and critically appraise MU and related practices surrounding the needs of data deletion. In *Section 3*, we clarify the notions of unlearning and forgetting. We argue that distinguishing between three modes of forgetting helps better understand the capabilities and limitations of current MU techniques and assess their adequacy more rigorously. *Section 4* further explores the three modes of forgetting, using a series of stylized cases to discuss how MU may be implemented and measure it against what stakeholders (e.g. people harmed by unfair and unsafe ML systems) want from the MU project. This demonstrates how our distinctions can help assess the effectiveness of MU methods at catering to different stakeholder demands. Based on these insights, we also offer recommendations for how to improve the ethical appraisal of MU approaches going forward and outline a set of question clusters that require increased interdisciplinary attention. *Section 5* concludes the paper.

2 Machine Unlearning: Definition, Goals, Methods, and Evaluation

Here, we offer a brief overview of MU, its central goals, techniques and evaluation methods. More in-depth discussions can be found in adjacent literature (e.g. Cooper et al. 2024). For our purposes, and glossing over many nuances and a rich history, we consider MU to be a research program with a key focus on selectively removing the influence of specific data (the "forget set" D_f) from a trained machine learning model, θ_o, ideally without the need for complete retraining on the remaining data (D_r), and without compromising the overall performance of the resultant model. The goal is to produce an updated model θ_r that behaves as if it were never exposed to D_f. Measuring success involves metrics like performance drops on D_f, ensuring stable performance on D_r, resisting membership inference attacks (MIAs) (Zhavoronkin et al. 2024; Sula et al. 2022), attempting to detect D_f's original presence (Di 2025; Liu et al. 2024), and ideally achieving statistical indistinguishability from a fully retrained model. An adjacent approach to MU focuses not on intervening with a trained model as such, but only its behaviors at inference time or beyond, e.g. trying to suppress the generation of harmful or toxic text or image outputs even as a model principally retains the ability to produce such

outputs. While these efforts, which draw on a variety of techniques such as using system prompts or output filters, may not be uncontroversially regarded as approaches to *unlearning*, they are nevertheless often considered part of the larger MU research program (see e.g. Cooper et al. 2024).

2.1 The Stated Goals of Machine Unlearning

MU is often presented as driven by ethical, legal, and practical goals (Cooper et al. 2024), such as privacy, bias mitigation, security, and safeguarding of rights (such as copyrights). The MU literature reflects these multifaceted goals, often highlighting distinct but complementary motivations. For instance, Wang et al. (2024) underscore the foundational legal impetus by focusing on GDPR and RTBF as the legal basis and motivation behind MU, highlighting a specific, legally mandated aspect of privacy. In contrast, Cooper et al. (2024, p. 2) point to a broader and evolving scope of motivations, observing that "machine unlearning's presumptive mandate has expanded significantly. More and more, research papers, policy briefs, and media reports suggest MU as an approach for meeting a broad range of objectives for both open and closed models and systems, spanning privacy, copyright, safety, and more." Complementing these perspectives, Nguyen et al. (2022) cite security, privacy, usability, and fidelity as the motivating goals behind MU techniques. Together, these perspectives illustrate a field motivated by both specific regulatory requirements and a broader call for responsible and adaptable AI systems.

2.2 Core MU Strategies: Exact vs. Approximate Unlearning

MU's central strategies are often motivated and explained by reference to *exact unlearning*. This 'gold standard' aims at producing a model that is identical in behavior to one trained from scratch without the forget-set D_f. While such an approach would offer the strongest erasure guarantees, best aligning with ethical and legal goals, complex models like deep neural networks often necessitate full retraining on D_r, which is why exact unlearning is often presented as computationally prohibitive for many real-world scenarios due to cost and time constraints.

By contrast, *approximate unlearning*, which is the central family of strategies pursued in MU research, aims at offering a more practically feasible approach. Approximate unlearning methods directly modify the originally trained model θ_o to efficiently reduce or minimize D_f's influence on model behaviors and outputs, using techniques like perturbed gradient descent or scrubbing (Golatkar, Achille, and Soatto 2020; Neel, Roth, and Sharifi-Malvajerdi 2021; Suriyakumar and Wilson 2022), to make it behave like θ_r. While significantly faster and more scalable, approximate methods lack strong erasure guarantees. They risk leaving residual information traces within the model, potentially extractable via sophisticated attacks, thus providing only an empirical, non-verifiable mode of forgetting that may fall short of meeting strict legal or ethical requirements.

The choice between exact and approximate approaches hence encodes a key tradeoff between balancing erasure guarantees and the ethical and legal goals they serve, against practical resource limitations that stand in the way of exact unlearning.

2.3 Benchmarking MU Effectiveness: Tasks and Limitations

Evaluating unlearning efficacy relies on benchmarking tasks designed to simulate real-world scenarios. Examples include TOFU (testing forgetting of specific factual associations) (Maini, 2024) or using membership inference attacks (MIAs) adversarially to probe for remaining data signatures.

Despite these efforts, evaluation remains a critical weakness. Current benchmarks lack standardization, making comparisons across different MU techniques difficult and unreliable (Zhou et al. 2024). They also often use synthetic data or narrow task definitions that may not generalize to complex, real-world unlearning needs (e.g., removing subtle, distributed biases). Furthermore, they typically test against known vulnerabilities, offering little assurance against future, more advanced methods for extracting supposedly forgotten information (Thaker et al. 2024). Consequently, existing benchmarks do not provide robust, trustworthy verification of unlearning success (Zhou et al. 2024; Cooper et al. 2024).

2.4 Persistent Technical Challenges

Beyond evaluation hurdles, MU faces a series of further technical obstacles (Xu et al. 2024a; Waerebeke et al. 2025; Hsu et al. 2024; Zhou et al. 2024). In particular, 1) *scalability issues* arise because scaling unlearning techniques to models with billions or trillions of parameters can raise distinct technical challenges (Xu et al. 2024a). 2) MU often faces so-called *utility-forgetting tradeoffs*, where aggressively removing information often degrades a model's overall performance, even on unrelated tasks, and balancing effective erasure with retrained model utility remains a largely unsolved problem (Waerebeke et al. 2025). 3) *Residual data issues* arise because approximate methods inherently risk incomplete erasure, leaving vulnerabilities undermining unlearning objectives, especially for sensitive data. (Hsu et al. 2024) 4) *Verification difficulties* arise because of the lack of reliable, standardized benchmarks and metrics, which makes it difficult to confidently assess or guarantee that unlearning has been achieved to some specified degree (Zhou et al. 2024). These challenges highlight that while MU seeks to address a range of important needs, significant technical and methodological maturation is needed before it reliably

delivers on its promise to enable controlled forgetting in AI systems.

2.5 Beyond Technical Challenges: What is MU Aiming for, Ethically?

Addressing the technical challenges outlined earlier is crucial for advancing the MU project. Yet, while it is somewhat clear what MU aims for epistemically, i.e. some kind of unlearning/forgetting, we contend that progress also demands a more fundamental clarification of MU's underlying ethical rationales and how they hang together with its epistemic aims. Currently, the stated goals of MU, often framed around privacy, security, bias mitigation, and copyright, remain vague and thus risk conflicting with the pertinent values and needs of stakeholders, casting doubt on MU's effectiveness.

Three specific concerns emerge. First, *practical gaps* arise because MU research frequently focuses on technical mechanisms for data influence removal, often evaluated in controlled settings, but so far often lacks deeper engagement with the complexities of diverse, real-world unlearning scenarios, leaving its real-world effectiveness in complex settings under-evidenced.

Second, *ethical gaps* arise because MU largely operates without a robust, explicitly articulated ethical foundation. While there is frequent reference to legal rights such as RTBF, moral goods (e.g. rights, liberties) and values are rarely invoked in motivating specific MU undertakings. This makes it challenging to ascertain which moral goods or rights (e.g., rights to autonomy or privacy, or goods like dignity) MU aims to protect or promote, and how the technical interventions it affords serve these ends. Without this clarity, MU's motivations risk being perceived as mere technical compliance rather than genuine ethical commitment.

Third, MU is subject to *misalignment gaps*: given the practical and ethical ambiguities, it becomes uncertain whether MU's technically defined objectives truly align with the nuanced and varied concerns and complaints of those seeking erasure. Do current methods effectively meet epistemic-ethical rights, moral principles, or, more concretely, the specific and diverse needs of stakeholders? Or should different motivations guide MU development? And when, if ever, are MU approaches inappropriate, and developers should instead focus on building models that do not learn unwanted information in the first place?

Existing MU research does not provide sufficient clarity on these questions. In light of these gaps, a more critical examination is required of 1) what MU techniques are currently doing, 2) what MU is supposed to do (e.g., promote compliance with moral/legal rights, enact meaningful forgetting), and 3) whether these two aspects genuinely cohere and align. In line with the gaps identified here, our central claim is that significant uncertainty exists around whether MU's methods effectively cater to its stated aspirations. As we argue shortly, an integrated epistemic-ethical lens is needed to evaluate MU's effectiveness. Such a lens acknowledges that MU is not only a technical endeavour that focuses on the pursuit of concrete epistemic goals, but, at the same time, an irreducibly *ethical* endeavour that must be put on clear foundations regarding the moral goods and rights it seeks to promote. Drawing on literature from social epistemology and the ethics of forgetting, in the next section, we outline a threefold distinction of different modes of forgetting that helps better understand, appraise, and eventually improve the MU project.

3 Flavors of Unlearning

As outlined earlier, MU is often presented as a technical solution to ethical and legal challenges, including protecting user privacy, mitigating bias, addressing security threats, and complying with copyright law. However, these motivations rely on ambiguous notions of *unlearning* and *forgetting* that remain under-investigated. Such ambiguities raise several basic questions, such as: what exactly does it mean for a machine to unlearn/forget? What moral demands and complaints can and should unlearning/forgetting speak to? Can forgetting in machines mirror entrenched epistemic-ethical practices surrounding human (and institutional) forgetting? What kinds of unlearning can be morally demanded, given what is technically feasible? In order to make progress on addressing these questions and make it easier to identify and mitigate gaps between what MU does and what it is supposed to do, we propose a threefold epistemological distinction between different epistemic states, each encoding distinct modes of ignorance/unlearning/forgetting. In particular, we distinguish between:

1. Never knowing X, or knowing and then forgetting X *exactly*[2]
2. Knowing and then forgetting X, but *imperfectly*
3. Knowing X but choosing not to *act* on it.

Each of these modes reflects different epistemic and ethical considerations and maps onto different strategies and challenges within MU. Below, using insights from social epistemology and the ethics of forgetting literature, we highlight the utility of these distinctions, proceeding mode-by-mode and using stylized examples to illustrate. This will help

[2] We treat 'knowing' in AI systems in a minimalist way as a functional or operational property, where an AI system 'knows' X *iff* it has encoded X in such a way that it can retrieve/reproduce or act on X (e.g. by producing certain outputs) under appropriate conditions. Notably, this does not require that AI systems have epistemic or mental states like belief. See also He and Yang (2025), Paseri and Durante (2025), and Fierro et al. (2024)

highlight potential misalignments between the kind of forgetting stakeholders may desire or deserve and what MU does or can deliver, which we further elaborate in Section 4.

3.1 Never Knowing X: Epistemic Innocence and Moral Constraints on Knowledge

The first and strongest mode of ignorance is where an agent (machine or human) either a) never learns a piece of information X, or else b) forgets X *exactly* as if X had never been known to it. Applied to ML systems, a) reflects a case where a system is never trained on X in the first place, and b) mirrors *exact unlearning*: retraining a model from scratch on a dataset that excludes the forget set, yielding a model that is indistinguishable from one that never saw the undesired training data. As discussed earlier, exact unlearning is often framed as practically infeasible, but in the real world, stakeholders might desire, demand, and even deserve this mode of forgetting.

This first mode, irrespective of how it is achieved, is both demanding and potentially controversial: knowledge is often assigned a positive epistemic value, where knowing something is better than not knowing it (Basu 2020; Michaelian 2011). Developers of large AI systems like LLMs or TTI models are often motivated by, and draw on, purported scaling laws (Kaplan et al. 2020), both in terms of dataset and model size, and thereby seemingly commit to the assumption that learning more is better. However, this is far from an obvious assumption, and the epistemological literature has drawn out various ways in which knowing more is not always ideal for an epistemic agent and sometimes *not* knowing can be beneficial or desired by agents for both epistemic and ethical reasons (Basu 2020; Michaelian 2011; Singer et al. 2021). One epistemic concept discussed in philosophical literature is *epistemic innocence*. The idea behind epistemic innocence is that sometimes knowing less or having false beliefs can be tolerated and/or even needed to reach an overall epistemically beneficial state (Puddifoot and Bortolotti 2019). In a similar vein, it is not immediately clear that it is always strictly better for ML systems to be exposed to more data.

A real-world example touching on these issues is the removal of questionable explicit material from the training data set of TTI models like Stable Diffusion (SD). Specifically, earlier image generation models like SD 1.5 were trained on the extensive LAION-5B dataset (Schuhmann et al. 2022), which was largely compiled from unguided web crawling and thus included a significant amount of explicit material, even with attempts by LAION-5B developers to classify and filter out underage explicit content. In a significant shift, Stable Diffusion 2.0 was developed with a more stringent approach aiming to remove unsafe content from the training set. This resulted in a relative lack of explicit material in the training set for SD 2.0, making it more difficult for the model to generate such content (Thiel 2023).

This modification of the training data and from-scratch retraining for SD 2.0 can be seen as a notable real-world example of efforts to prevent a model from producing undesirable outputs. It also highlights the importance of the concept of epistemic innocence, where removing explicit material made it difficult for SD 2.0 to generate questionable outputs due to the lack of exposure to undesirable information. While in some (unwanted) respects the performance of the model was thereby degraded, the intervention achieved a more aligned model than SD 1.5. We will revisit the decisions of Stability AI in more detail in Section 4 to assess their effects on the use of these models at large, especially when taking into account the needs and expectations of different stakeholders.

3.2 Knowing and Forgetting X: An Epistemic Virtue or Duty

The second mode of forgetting we highlight here is when an agent or entity knows or learns a piece of information, X, but then forgets it, albeit in a potentially *imperfect* way. This is weaker than never knowing X because it does not guarantee that X will never resurface. In MU, this type of forgetting corresponds to approximate unlearning, where the goal is to remove the effects of unwanted data, while avoiding the need to retrain a model from scratch.

The second mode draws on two ideas: 1) a disposition to forget harmful or epistemically disadvantageous information can be an *epistemic virtue*, and 2) there may also be *epistemic-ethical duties* to forget such information. Let us explain these two ideas in turn.

Epistemic virtue is a concept that originated in virtue epistemology (VE), a normative approach to epistemology that shifts the focus from *what* constitutes a justified belief that is a candidate for knowledge to *who* forms such beliefs and through which *traits*. These traits are commonly called intellectual or epistemic virtues (Turri et al. 2021). In virtue reliabilism, associated with philosophers such as Sosa (2007), Greco (2010), and Goldman (1992; 2002), such virtues are understood as cognitive faculties or abilities (such as vision, memory, and introspection) that reliably lead to the formation of true beliefs.

Meanwhile, more pertinent to our analysis are virtue responsibilist accounts, such as Linda Zagzebski's (1996), which emphasize intellectual character traits like open-mindedness, intellectual courage, and inquisitiveness as central to epistemic agency. Informed by such accounts, we propose that the disposition to intentionally forget information when that information is no longer epistemically valuable and/or is recognized as false or harmful, can be considered an epistemic virtue, insofar as it contributes to responsible belief management and intellectual flourishing.

A second route to put forgetting on more principled epistemic-ethical grounds can be motivated in terms of epistemic-ethical duties. Here, we particularly consider the work by Rima Basu (2022) who argues that forgetting can be a moral duty, especially when remembering X threatens to constrain a person's identity or social status unjustly. Practices like expunging juvenile arrest records or choosing to blur memories around a friend's past problematic behaviors rest on the idea that outdated or unrepresentative facts should not indefinitely define individuals.

In the context of MU, these virtue- and duty-based ideas apply when people, or groups, may not want the world to know or remember them in a certain way, and thus require machines to forget some information about them, catering to individuals' needs and freedoms to change and grow. Since active forgetting is likely easier to engineer in machines than in humans, there may also be heightened responsibilities to do so, e.g., when confronting demands for debiasing ML systems in the interest of promoting fair and safe AI. However, as outlined earlier, it remains unclear whether MU techniques can promote this form of forgetting to an extent that is satisfactory for stakeholders, that caters to pertinent epistemic-ethical duties or embodies relevant virtues.

To see this, consider a hypothetical case where a series of false allegations of fraud were posted on anonymous online forums about an individual named Dorian. The allegations never spread beyond the series of posts that were later removed from the web, but the posts were captured in the dataset used to train an LLM. Due to the uniqueness of Dorian's name, the model neatly encoded the link between his name and the false allegations. When prompted with Dorian's full name, the model now reproduces the accusations in various degrees of explicitness, even in response to prompts about Dorian that are unrelated to his alleged criminal conduct. Aware of this, Dorian invokes RTBF or similar rights to have this association removed from the model and an approximate unlearning method is employed to modify the model's parameters to minimize the influence of the posts in question, while maintaining the model's ability to provide other information about Dorian, e.g. that he is an innovative entrepreneur and a skilled gardener.

Our example represents the kind of targeted removal that approximate unlearning strives for and that could embody meaningful forgetting as an epistemic-ethical virtue or duty. However, as we argue shortly in Section 4, technical limitations inherent in approximate unlearning methods (e.g. concerning residual traces or verification challenges) leave a critical gap: aspirations to fulfil morally grounded duties or exhibit relevant virtues, e.g. in forgetting inaccurate, misleading, and harmful information about others, may often not be adequately met by current MU capabilities.

3.3 Knowing X but Choosing Not to Act on It: Epistemic Responsibility

The third mode of forgetting, knowing X but choosing not to *act* on it, is even weaker than knowing and forgetting X, *imperfectly*. It is, in fact, not *true forgetting* in a strict sense, but *suppression* (see also Cooper et al. 2024): an agent (or a machine) still retains information on X but does not *use or reveal* it through actions or behaviors, such as responses to a question or prompt. Technically, this is implemented through guardrails, moderation layers, and safety filters (Hacker, Engel, and Mauer 2023). In the context of LLMs, guardrails are algorithms that monitor and filter the inputs and outputs of trained LLMs (Dong et al. 2024).

This mode of forgetting relates to the concept of *epistemic responsibility*, the idea that epistemic agents may not only have obligations to know or not know X, but also to be *reflective* about how they *use* their knowledge (Elgin 2013). One way to exercise epistemic responsibility is to recognize when to *withhold* or *suspend* a specific piece of information, especially in contexts involving sensitive or controversial information. The concept of epistemic responsibility thus applies to most human epistemic agents, given they have some epistemic autonomy and control over their beliefs and how these figure in making inferences and informing actions (Rudy-Hiller 2018). Machines may not have this level of agency-related responsibility, yet this does not imply that their behaviors should not be governed by norms that are designed to withhold, suspend, or suppress information from being used or disseminated, when doing so is appropriate (MirzaeiGhazi and Stenseke 2024; Robbins 2025).

Importantly, however, knowing X but choosing not to act on it is an especially weak mode of forgetting, and may not always adequately cater to stakeholder demands. For instance, it might not be enough for a data subject to know that guardrails have been put in place to stop an LLM from divulging personal information about them, as the mere fact that the model encodes such information without the subject's consent may already provide grounds for legitimate complaints. To use an analogy, if a state, say, collects detailed personal information about the whereabouts of its citizens, it may not be enough to receive (even the most compelling) assurances that this information will never be used or shared - the point of a data subject's complaint in such a setting is rather that the state simply should not have that information to begin with. Returning to AI systems, retention and suppression invariably allow for the possibility of models regurgitating sensitive information they continue to encode, especially when extracted through adversarial attacks.

An important domain where this approach is used is in the use of various post-training techniques, system prompts and content moderation filters used to steer and regulate the out-

puts of LLMs like ChatGPT (see e.g. Liu et al. 2023; Mahomed et al. 2024), for instance, aiming to block harmful outputs like hate speech or the dissemination of private information. Despite advances in such techniques, the general worry persists that the underlying model may still possess the capacity to generate questionable outputs through various 'jailbreaking' techniques (e.g., DAN prompts, see Valentino 2023). The presence of such vulnerabilities illustrates the limitations of suppression as an effective form of epistemic responsibility engineered into AI systems. Suppression techniques strike specific tradeoffs between attending to stakeholder expectations and the costs of implementing unlearning. We explore these and other tensions further in the following section.

4 Towards an Ethical MU Project

Based on our distinction of three modes of forgetting, we now showcase how it can be useful to assess potential misalignments between stakeholder demands and MU approaches. Such assessments, we insist, are crucial for evaluating MU's effectiveness in addressing a range of plausible goals and needs.

Underlying our assessments is a key moral principle that helps us tell which ethically grounded unlearning demands stakeholders may justifiably place on developers and deployers: the *ought-implies-can* principle (see Vranas 2007 for a discussion). In our everyday epistemic and ethical practices, we often place varying moral demands on each other with regard to what we may, should or shouldn't know about each other. For instance, demanding that another human being, say a friend, exactly forgets information about something embarrassing we did is over-demanding: human memory is not amenable to interventions that exactly remove unwanted information. Here, the ought-implies-can principle regulates what counts as a legitimate moral demand. It maintains that only those things that an agent can, at least in principle, do can be morally demanded of them; otherwise, moral complaints and demands have no normative force.

Crucially, ML systems offer a different scope of forgetting than humans. While we cannot retrain human minds from scratch to forget embarrassing episodes or traumatic events, ML systems can, at least in principle, be retrained in such a way. This shifts the boundary of moral demands we can legitimately place on ML systems' makeup and their behaviors (or at least those who build, control and deploy these systems); since more advanced forms of forgetting are possible, they may now also be obligatory.

With this principle in mind, let us turn to examine different types of cases corresponding to the different modes of forgetting, to make progress on understanding how MU approaches might be appraised ethically.

The first case-type concerns never knowing. In the MU literature, this mode is largely neglected. MU is typically presented as a post-training technique of righting a possible wrong without much discussion of how this and other related wrongs may have been prevented in the first place. In essence, MU builds various sorts of sophisticated fire extinguishers, yet the literature has little to say on how to prevent fires. More data is often taken to imply better models, and once desired performance is achieved, then difficult tradeoffs between performance and other ethical or legal desiderata are negotiated. In particular, exact unlearning is mainly referenced as a vignette of an ideal gold standard and not typically considered as a viable method, citing computational and financial feasibility constraints. But this ignores cases where individuals and groups might have legitimate demands for complete erasure.

To elaborate this point further, let's revisit the Stable Diffusion example outlined in Section 3.1, where, before releasing SD 2.0, a form of exact unlearning was achieved by modifying the training set and retraining from scratch to mitigate the problems in SD 1.5. In this process, explicit content was removed from the training dataset to avoid unwanted outputs. While it may be commendable that Stability AI undertook this effort towards exact unlearning, other concerns persist. For example, given the open-source nature of Stable Diffusion, exact unlearning of unsafe content for a new model does not resolve the problem that earlier versions still circulate via unofficial channels. So, even if attempts are made to achieve 'never knowing' through retraining, the practicalities of information control in open ecosystems mean ethical issues can linger, suggesting that even the strongest MU techniques alone might not fully remedy past (and ongoing) infringements and safety concerns. This raises the question of whether models like SD 1.5 should have been released at all, or if they should not have been trained on a largely uncurated dataset like LAION-5B in the first place, requiring informed decisions taken before training.

Being trained on unsafe materials contained in LAION-5B, however, is not the only controversy surrounding Stable Diffusion. Another: the prominent dispute concerns the way SD 1.5 had learnt artists' styles and was thereby able to closely reproduce these styles as well as more specific copyrighted content (Yuan et al. 2023). As SD 1.5 gained prominence, artists raised objections, arguing the model had effectively assimilated their unique styles without consent, which led to a variety of ongoing legal challenges and significant pressure on Stability AI (Goetze 2024). The artists demanded that their works should be removed entirely from the training data, requiring data disgorgement akin to exact unlearning (Salkowitz 2024; Heikkilä 2022).

As mentioned earlier, when Stability AI subsequently retrained the model to produce version 2.0, it removed certain categories of problematic content, such as pornographic and

nude imagery. However, contesting artists' works were not excised from the training set of this version; a decision explicitly confirmed by company founder Emad Mostaque (Vincent 2022). These decisions demonstrate that exact unlearning is a possibility, despite frequent suggestions in the MU literature that it is largely infeasible. In particular, Stability AI did not employ exact unlearning with regard to artists' work. This may have been to ensure optimal or even improved model performance, or to appease the demands of users benefitting from generating compelling images through prompts. However, they were clearly able and willing to employ exact unlearning with regard to explicit material. This case highlights the gap between the demands of the stakeholders whose data is being used (artists) and the actions of companies that, even when they *could* employ stronger unlearning techniques to satisfy pertinent stakeholder demands, choose not to do so.

More recently, Stability AI has offered artists an opt-out way to tag their images for removal from the training set for Stable Diffusion 3.0 (Heikkilä 2022). Although this indicates that unlearning demands are being taken more seriously, one might argue that it shifts the burden of responsibility from developers and deployers to data subjects and stakeholders (e.g., artists), raising important questions about AI literacy and informed consent, both of which require further attention and research (e.g., debates around what constitutes meaningful consent, fair use, and creators' control over their works and styles) (Lovato et al. 2024; Kyi et al. 2025).

Opting to retrain and exclude artist styles, despite the associated epistemic cost to stylistic breadth, would have signaled a clear ethical prioritization: valuing artistic ownership and consent over model performance. This reinforces that the 'never knowing' standard, while technically challenging and often deemed computationally prohibitive, may directly satisfy potential moral obligations for certain information to remain unacquired or precisely expunged by a machine.

The worry, hence, remains that despite being at the center of MU's stated motivations, ethical concerns are only secondary in the real-world application of MU. What is more, as emphasized earlier, even if stronger unlearning efforts had been undertaken, earlier SD model checkpoints, preserving artists' styles, continue to be available on the web, thus further highlighting the weakness of even the strongest mode of unlearning.

The second case-type concerns knowing X and forgetting it, albeit imperfectly. As elaborated earlier, this mode is mirrored by approximate unlearning techniques, which most of the technical MU literature is concerned with. In Section 3.2, we illustrated this case with the hypothetical example of Dorian, who has an unlearning request to remove false allegations of fraud against him from the training dataset of an LLM. Plausibly, Dorian's expectation here would be a complete removal so as to ensure his identity is safeguarded. However, it is unclear to what extent current MU techniques can address such needs.

Much of the MU literature reports approximate unlearning successes for an increasing range of approaches and methods (Vatter et al. 2023; Qu et al. 2023; Nguyen et al. 2022). For example, concerning the case of removing artistic styles like those of Van Gogh, different approximate unlearning methods are used to achieve different degrees of removal (Biswas et al. 2025; Schioppa et al. 2024; Gandikota et al. 2023). While these techniques can make it difficult for people to generate images in a particular artist's style, they do not make it entirely impossible, as some traces may remain within the model. For instance, Biswas et al. (2025, p. 9) acknowledge that: "…highly adversarial prompts may still trigger partial leakage." Similarly, Schioppa et al. (2024, p. 10) note that their devised unlearning metric, "while effective in quantifying the probability of generating the undesired concept and rooted in previous work, does not account for more nuanced qualitative features", advocating for further research to develop the techniques more comprehensively. These evaluations suggest that the technical reality of MU often clashes with stakeholders' expectations, such as artists', who may desire and expect their intellectual property or style to be entirely and irreversibly removed from ML systems, or data subjects like Dorian, who need questionable training data items removed to prevent harm.

The third and final mode of forgetting, knowing X but choosing not to act on it, is achieved through applying guardrails/content moderation filters (Wang and Singh 2024). The case of GitHub Copilot offers a pertinent example. Upon its release, users found that Copilot, trained on vast amounts of public code including GPL-licensed repositories, could generate code nearly identical to existing public code snippets, triggering legal concerns about license violations and improper attribution (Xu et al. 2024b; DOE 1 v. GitHub, Inc. 2025). Rather than retraining the model (Codex, from OpenAI) to 'forget' this code entirely, Microsoft/GitHub implemented filters designed to block the output of exact matches to known public code (Salva 2023). The model, however, preserves the underlying patterns learned from that code (Risk Assessment of GitHub Copilot, 2023).

Being the weakest, easiest, and cheapest way of addressing unlearning needs, suppression of information is also more heavily used by companies like OpenAI than other methods and sometimes presented as a one-size-fits-all solution (Coglianese and Crum 2025). However, suppression means the information is not removed at all, and there may be a risk of unwanted data resurfacing at any time, as in the case of GitHub Copilot outlined earlier.

One may argue that this is more a problem of guardrails that are not (yet) good enough. We could imagine that with

advancements in methods, it could be *guaranteed* that suppressed information will never resurface. However, we maintain that even then, stakeholders may legitimately insist that suppression is insufficient and their specific information be removed from a system, as pointed out by the analogy of a state acquiring sensitive information about its citizens in Section 3.3. No matter how strongly the state guarantees that the information will never be used, the individual's rights may remain violated.

Differentiating these three ways of forgetting helps address the gaps identified earlier. It allows us to assess what kind of forgetting a specific MU technique actually delivers (modes 2 or 3) versus what kind might be ethically required in a given context (e.g. mode 1 for highly sensitive data). It operationalizes philosophical concepts where 'never knowing' implies moral constraints on knowledge; 'knowing and forgetting' aligns with epistemic duties; and 'knowing but not acting' relates to engineered epistemic responsibility. This integrated epistemic-ethical perspective casts MU as more than a technical enterprise, revealing that method choice matters ethically for whether MU caters to core moral values like privacy, justice, dignity, or accountability.

4.1 Recommendations

Our distinctions between different modes of unlearning/forgetting enable a more nuanced approach to MU development and governance going forward. MU techniques can benefit by tailoring strategies and policies that display sensitivity to which mode of forgetting is desired by data subjects and other stakeholders, and whether their demands are relevant and justified. In particular, we put forward three recommendations and insights that build on our analysis.

First, developers and deployers should explicitly state which mode of forgetting (1, 2, or 3) MU interventions aim to achieve and which specific normative goals justify this choice. Especially in industry settings, there is sometimes significant ambiguity around exactly what modes of unlearning, proprietary MU approaches pursue (see e.g. Luria 2025 for an example). Moreover, whenever mode 2 or 3 are chosen as appropriate, developers and deployers should be able to justify this choice over pursuing mode 1, explaining why, exactly, the costs associated with stronger forms of unlearning are not justified by the urgency of relevant stakeholder demands. This approach, which puts focus on publicly justifying model-building and deployment choices in regard to unlearning needs, would help MU efforts stand on more explicit and interrogable ethical foundations, speaking more directly to stakeholders' needs and enabling wider debate around which MU efforts are appropriate in what contexts.

Second, the threefold distinction of different modes of unlearning/forgetting offers policymakers a vocabulary to create more nuanced regulations. Regulatory instruments aimed at protecting stakeholder rights should accordingly specify the required mode of forgetting based on data sensitivity, potential harm, and the legal rights at stake (e.g., demanding approaches approximating mode 1 for RTBF requests concerning sensitive personal data, while perhaps permitting mode 3 for certain types of public information or content moderation). This would help align legal obligations with technical possibilities more realistically.

Third, achieving stronger modes of forgetting, such as mode 1 and mode 2, often requires architectural choices made early in the ML pipeline. Ethics-readiness, informed by the desired modes of forgetting, should be integrated from the start, including data provenance tracking, provisions for metadata tagging for deletable content, modular model design facilitating easier retraining or targeted modification, and transparency in training practices.

These recommendations and insights, derived from our earlier analysis, aim to shift MU from a reactive, often technically limited practice towards a proactive, ethically grounded field capable of better aligning ML models with human values. That said, effectively implementing such recommendations also requires further interdisciplinary research and societal dialogue to engage a wide range of open questions that lie beyond the scope of the current discussion. In particular, we see six major question clusters that should be engaged by the research community as well as the wider stakeholder landscape, including developers, deployers, policy-makers and regulators, as well as data subjects and activists:

1. What constitutes legally and morally 'sufficient' thresholds for forgetting, particularly for approximate methods (mode 2) or suppression (mode 3)? How, exactly, do judgments of sufficiency vary across contexts and between stakeholder groups? How should disagreement between such assessments be negotiated?
2. Who ultimately bears responsibility for ensuring information dissemination by AI systems is ethically sound and legally compliant, and how should accountability for unlearning failures be distributed?
3. How should MU governance be approached across different international jurisdictions with varying legal frameworks and epistemic-ethical norms regarding memory, privacy, and justice?
4. Could MU frameworks, informed by our analysis, be designed to support epistemic pluralism, allowing diverse communities to negotiate and implement their own epistemic-ethical norms for machine unlearning?
5. How can MU techniques be safeguarded against misuse, e.g. allowing powerful actors to effect forms of digital amnesia or narrative manipulation and curate AI-mediated knowledge landscapes based on political

or commercial interests, rather than ethical and epistemic principles like harm reduction or truthfulness? (see Chen 2025)

6. How can genuine efforts at putting MU on more solid ethical foundations resist the danger of ethics proceduralization, where implementing a technical unlearning method becomes a substitute for genuine engagement with moral rights (see Ratner and Moeslund 2025)?

Exploring these and other related questions, we insist, is central to further progress in navigating the complex ethical topology of MU.

5 Conclusions

Machine Unlearning (MU) emerges as a promising research program to manage the persistent, often problematic, ways in which AI systems acquire, encode, retain and disseminate unwanted information. This paper traced the outlines of MU, highlighting significant gaps between its technical implementations, often limited by feasibility constraints and evaluation uncertainties, and the ethical goals of privacy, fairness, safety, and accountability it purports to serve. We argued that bridging these gaps requires moving beyond purely technical considerations to embrace a view of MU as an irreducibly ethical project. To this end, we introduced a philosophically grounded distinction between 'never knowing X or knowing X but exactly unlearning it,' 'knowing X and forgetting it, but imperfectly, ' and 'knowing X but deciding not to act on it'. Drawing out pertinent connections to the social epistemology and ethics of forgetting literature, we showed how these distinctions allow for a more nuanced assessment of how different MU techniques map onto concrete ethical demands arising in diverse contexts.

By highlighting the distinct ethical weight carried by different modes of forgetting, our analysis underscores the need for more nuanced thinking about how MU is developed, evaluated, and governed. It calls for explicitly aligning technical methods with clearly defined ethical objectives, establishing context-sensitive evaluation standards through stakeholder participation, designing legal and policy frameworks sensitive to varying modes of unlearning, and embedding 'forgettability' into system design and engineering pipelines from the outset. Simultaneously, we urge that there is a range of important but yet unresolved questions for the community to engage with. Ultimately, fostering trustworthy AI requires treating MU not just as a technical challenge but as a crucial moral frontier demanding ongoing interdisciplinary engagement, ethical reflection, and a steadfast commitment to aligning automated systems with human moral values and rights.

Acknowledgments

We would like to thank three anonymous reviewers for their insightful comments and constructive suggestions, which have helped improve this paper. Moreover, Iqra Aslam and Donal Khosrowi gratefully acknowledge financial support for this research from the Lower Saxony Ministry of Science and Culture (NMWK) and from the Gesellschaft für Wissenschaftsphilosophie (GWP).

References

Basu, R. 2022. The Importance of Forgetting. *Episteme* 19(4): 471-490.

Biswas, S. D.; Roy, A.; and Roy, K. 2025. CURE: Concept Unlearning via Orthogonal Representation Editing in Diffusion Models. *arXiv preprint arXiv:2505.12677*.

Chen, Z. Z.; Ma, J.; Zhang, X.; Hao, N.; Yan, A.; Nourbakhsh, A.; Yang, X.; McAuley, J.; Petzold, L.; and Wang, W. Y. 2024. A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law. *arXiv preprint*. arXiv:2405.01769 [cs.CL]. Ithaca, NY: Cornell University Library.

Chen, J. 2025. Digital amnesia: machine unlearning and the fragility of cultural memory. AI & Society. https://doi.org/10.1007/s00146-025-02308-8.

Coglianese, C., and Crum, C. R. 2025. Leashes, Not Guardrails: A Management-Based Approach to Artificial Intelligence Risk Regulation. *U of Penn Law School, Public Law Research Paper*, (25-04).

Cooper, A.F., Choquette-Choo, C.A., Bogen, M., Jagielski, M., Filippova, K., Liu, K.Z., Chouldechova, A., Hayes, J., Huang, Y., Mireshghallah, N., and Shumailov, I., 2024. Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice. arXiv:2412.06966.

Di, Z. 2025. Adversarial Machine Unlearning. *arXiv preprint arXiv:2406.07687v1*. https://arxiv.org/html/2406.07687v1

Ding, M.; Sharma, R.; Chen, C.; Xu, J.; and Ji, K. 2024. Understanding Fine-tuning in Approximate Unlearning: A Theoretical Perspective. arXiv preprint arXiv:2410.03833. https://arxiv.org/abs/2410.03833

DOE 1 v. GitHub, Inc. 2025. 4:22-cv-06823 - CourtListener.com. *CourtListener*. https://www.courtlistener.com/docket/65669506/doe-1-v-github-inc/

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., and Bensalem, S., 2024. Safeguarding large language models: A survey. arXiv preprint arXiv:2406.02622.

Elgin, C. Z. 2013. Epistemic Agency. *Theory and Research in Education* 11(2): 135-152.

Fierro, C.; Dhar, R.; Stamatiou, F.; Garneau, N.; and Søgaard, A. 2024. Defining Knowledge: Bridging Epistemology and Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 16096–16111. Miami, Florida, USA: Association for Computational Linguistics.

Floridi, L. 2015. "The Right to Be Forgotten: A Philosophical View". *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3853478

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing Concepts from Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2426-2436.

Goetze, T. S. 2024. AI Art is Theft: Labour, Extraction, and Exploitation: Or, On the Dangers of Stochastic Pollocks. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 186-196.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9301-9309. Seattle, WA, USA. doi: 10.1109/CVPR42600.2020.00932

Goldman, A. 1992. *Epistemology and Cognition*. Harvard University Press.

Goldman, A. 2002. The Unity of the Epistemic Virtues. *Philosophy and Phenomenological Research* 66, 2 (2002), 251–275.

Greco, J. 2010. *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge University Press.

He, J., and Yang, C. 2025. Testimony by LLMs. *AI & Society*. https://doi.org/10.1007/s00146-025-02366-y.

Hacker, P.; Engel, A.; and Mauer, M. 2023. Regulating ChatGPT and Other Large Generative AI Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112-1123.

Heikkilä, M. 2022. Artists Can Now Opt Out of the Next Version of Stable Diffusion. *MIT Technology Review*. https://www.technologyreview.com/2022/12/16/1065247/artists-can-now-opt-out-of-the-next-version-of-stable-diffusion/

Hine, E.; Novelli, C.; Taddeo, M.; and Floridi, L. 2024. Supporting Trustworthy AI Through Machine Unlearning. *Science and Engineering Ethics* 30(5).

Hsu, H.; Niroula, P.; He, Z.; and Chen, C. F. 2024. Are We Really Unlearning? The Presence of Residual Knowledge in Machine Unlearning. In *I Can't Believe It's Not Better: Challenges in Applied Deep Learning*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D., 2020. Scaling Laws for Neural Language Models. https://doi.org/10.48550/arXiv.2001.0836.

Kyi, L.; Mahuli, A.; Silberman, M. S.; Binns, R.; Zhao, J.; and Biega, A. J. 2025. Governance of Generative AI in Creative Work: Consent, Credit, Compensation, and Beyond. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–16. New York: Association for Computing Machinery.

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C.Y., Xu, X., Li, H., and Varshney, K.R., 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pp.1–14.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; and Liu, Y. 2023. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. *arXiv preprint arXiv:2305.13860*. https://arxiv.org/abs/2305.13860

Liu, Z.; Dou, G.; Tan, Z.; Tian, Y.; and Jiang, M. 2024. Machine Unlearning in Generative AI: A Survey. *arXiv preprint arXiv:2407.20516*. https://arxiv.org/abs/2407.20516.

Lovato, J.; Zimmerman, J. W.; Smith, I.; Dodds, P.; and Karson, J. L. 2024. Foregrounding Artist Opinions: A Survey Study on Transparency, Ownership, and Fairness in AI Generative Art. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7, 905–916. New York: Association for Computing Machinery.

Luria, B. 2025. Debiasing Llama 4 Scout: Pioneering Behavioral Unlearning for Safer AI. https://www.hirundo.io/blog/llama4-debiased, retrieved May 2025.

Mahomed, Y.; Crawford, C. M.; Gautam, S.; Friedler, S. A.; and Metaxa, D. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 660–686.

Maini, P. 2024. TOFU: A Task of Fictitious Unlearning for LLMs. *arXiv preprint arXiv:2401.06121v1*. https://arxiv.org/html/2401.06121v1

Michaelian, K. 2011. The Epistemology of Forgetting. *Erkenntnis* 74: 399-424.

MirzaeiGhazi, S.; and Stenseke, J. 2024. Responsibility Before Freedom: Closing the Responsibility Gaps for Autonomous Machines. *AI and Ethics*: 1-13.

Neel, S.; Roth, A.; and Sharifi-Malvajerdi, S. 2021. Descent-to-Delete: Gradient-Based Methods for Machine Unlearning. In *Algorithmic Learning Theory*, 931–962. PMLR.

Nguyen, T. T.; Huynh, T. T.; Ren, Z.; Nguyen, P. L.; Liew, A. W. C.; Yin, H.; and Nguyen, Q. V. H. 2022. A Survey of Machine Unlearning. *arXiv preprint arXiv:2209.02299*.

Oesterling, A.; Ma, J.; Calmon, F.; and Lakkaraju, H. 2024. Fair Machine Unlearning: Data Removal While Mitigating Disparities. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 3736-3744. PMLR.

Paseri, L., and Durante, M. 2025. Examining Epistemological Challenges of Large Language Models in Law. *Cambridge Forum on AI: Law and Governance* 1: e7. doi.org/10.1017/cfl.2024.7.

Puddifoot, Katherine, and Lisa Bortolotti. 2019. Epistemic Innocence and the Production of False Memory Beliefs. *Philosophical Studies* 176(3): 755–780.

Qu, Y.; Yuan, X.; Ding, M.; Ni, W.; Rakotoarivelo, T.; and Smith, D. 2023. Learn to Unlearn: A Survey on Machine Unlearning. *arXiv preprint arXiv:2305.07512*.

Ratner, H. F.; and Moeslund, T. 2025. Computational Implementations of Responsible AI: From the Right to Be Forgotten to Machine Unlearning. *AI & Society*. https://doi.org/10.1007/s00146-025-02261-6

Risk Assessment of GitHub Copilot. 2023. *Gist*. https://gist.github.com/0xabad1dea/be18e11beb2e12433d93475d72016902

Robbins, S. 2025. What Machines Shouldn't Do. *AI & SOCIETY*: 1-12.

Rudy-Hiller, F. 2022. The Epistemic Condition for Moral Responsibility. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, eds. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic/

Salkowitz, R. 2024. Artist And Activist Karla Ortiz On The Battle To Preserve Humanity In Art. *Forbes*. http://forbes.com/sites/robsalkowitz/2024/05/23/artist-and-activist-karla-ortiz-on-the-battle-to-preserve-humanity-in-art/

Salva, R. J. 2023. Introducing Code Referencing for GitHub Copilot. *The GitHub Blog*. https://github.blog/news-insights/product-news/introducing-code-referencing-for-github-copilot/

Schioppa, A.; Hoogeboom, E.; and Heek, J. 2024. Model Integrity when Unlearning with T2I Diffusion Models. *arXiv preprint arXiv:2411.02068*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schmidt, L.; Kaczmarczyk, R.; and Henning, J. 2022. LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. In *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*. Available at https://openreview.net/forum?id=M3Y74vmsMcY.

Singer, D. J.; Bramson, A.; Grim, P.; Holman, B.; Kovaka, K.; Jung, J.; and Berger, W. J. 2021. Don't Forget Forgetting: The Social Epistemic Importance of How We Forget. *Synthese* 198: 5373-5394.

Sosa, E. 2007. *A Virtue Epistemology: Apt Belief and Reflective Knowledge*. Oxford University Press.

Stable Diffusion Version 2. 2023. *GitHub*. https://github.com/Stability-AI/stablediffusion

Sula, N.; Kumar, A.; Hou, J.; Wang, H.; and Tourani, R. 2022. Silver Linings in the Shadows: Harnessing Membership Inference for Machine Unlearning. *arXiv preprint arXiv:2407.00866v1*. https://arxiv.org/html/2407.00866v1

Suriyakumar, V. M.; and Wilson, A. C. 2022. Algorithms that Approximate Data Removal: New Results and Limitations. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, Article 1372, 18892–18903. Curran Associates Inc., Red Hook, NY, USA.

Thiel, D. 2023. Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at https://purl.stanford.edu/kh752sm9123.
https://doi.org/10.25740/kh752sm9123.

Thaker, P.; Hu, S.; Kale, N.; Maurya, Y.; Wu, Z. S.; and Smith, V. 2024. Position: LLM Unlearning Benchmarks Are Weak Measures of Progress. *arXiv preprint arXiv:2410.02879*.

Turri, J.; Alfano, M.; and Greco, J. 2021. Virtue Epistemology. In *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), ed. E. N. Zalta. Stanford, CA: Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/epistemology-virtue/.

Van Waerebeke, M.; Lorenzi, M.; Neglia, G.; and Scaman, K. 2025. When to Forget? Complexity Trade-offs in Machine Unlearning. *arXiv preprint arXiv:2502.17323*. https://arxiv.org/abs/2502.17323

Valentino, Alexis. 2023. ChatGPT-DAN. GitHub repository. https://github.com/alexisvalentino/Chatgpt-DAN. Accessed May 23, 2025.

Vatter, J.; Mayer, R.; and Jacobsen, H. A. 2023. The Evolution of Distributed Systems for Graph Neural Networks and Their Origin in Graph Processing and Deep Learning: A Survey. *ACM Computing Surveys* 56(1): 1-37.

Vincent, J. 2022. Stable Diffusion Made Copying Artists and Generating Porn Harder and Users Are Mad. *The Verge*. https://www.theverge.com/2022/11/24/23476622/ai-image-generator-stable-diffusion-version-2-nsfw-artists-data-changes

Vranas, P. B. 2007. I Ought, Therefore I Can. *Philosophical Studies* 136: 167-216.

Wang, W.; Tian, Z.; Zhang, C.; and Yu, S. 2024. Machine Unlearning: A Comprehensive Survey. *arXiv preprint arXiv:2405.07406*. https://arxiv.org/abs/2405.07406

Wang, Y.; and Singh, L. 2024. Adding Guardrails to Advanced Chatbots. *arXiv preprint arXiv:2306.07500*. https://arxiv.org/abs/2306.07500

Xu, J.; Wu, Z.; Wang, C.; and Jia, X. 2024. Machine Unlearning: Solutions and Challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Xu, W.; Gao, K.; He, H.; and Zhou, M. 2024. A First Look at License Compliance Capability of LLMs in Code Generation. *arXiv e-prints, arXiv-2408*.

Yuan, G.; Cun, X.; Zhang, Y.; Li, M.; Qi, C.; Wang, X.; Shan, Y.; and Zheng, H. 2023. Inserting Anybody in Diffusion Models via Celeb Basis. arXiv:2306.00926.

Zagzebski, L. 1996. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge, UK: Cambridge University Press.

Zhang, D.; Finckenberg-Broman, P.; Hoang, T.; Pan, S.; Xing, Z.; Staples, M.; and Xu, X. 2024. Right to Be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions. AI and Ethics, 1–10.

Zhavoronkin, M.; Pautov, N.; Kalmykov, E.; Sevriugov, D.; Kovalev, O.; Rogov, I.; and Oseledets. 2024. Zapiski Nauchnykh Seminarov POMI Tom 540 [Notes of Scientific Seminars of POMI Vol. 540] (in Russian). *Записки научных семинаров ПОМИ Том 540*. http://ftp.pdmi.ras.ru/pub/publicat/znsl/v540/p046.pdf

Zhou, S.; Wang, L.; Ye, J.; Wu, Y.; and Chang, H. 2024). On the Limitations and Prospects of Machine Unlearning for Generative AI. *arXiv preprint arXiv:2408.00376*.