Situating and Understanding Machine Unlearning, Ethically

IQRA ASLAM, Leibniz Universität Hannover, Germany

Machine Unlearning (MU) aims to remove unwanted data and its effects from machine learning models while preserving performance. Driven by ethical and legal concerns such as privacy (Right to be Forgotten), security, bias mitigation, and copyright protection, MU faces challenges, including technical limitations, ethical ambiguities, and conflicting stakeholder expectations. This paper critically examines MU's motivations and effectiveness, arguing that it remains unclear 1) what MU does, 2) what it should do, and 3) how efforts and goals fit together. To clarify, I introduce a tripartite epistemological distinction: 1) never knowing X, 2) learning X and forgetting it, and 3) acting as if one doesn't know X. Analyzing cases of copyright, data privacy, and intellectual property, the paper shows inconsistencies between MU's goals and outcomes, stressing a need for clearer ethics, stakeholder engagement, and transparency. Refining MU is crucial to ensuring that it effectively serves its intended purposes.

Keywords: Machine Unlearning; Right to Be Forgotten; Ethics of AI; Ethics of forgetting; Epistemology

Reference Format:
Aslam. 2025. Example EWAF'25 Submission. In Proceedings of the Fourth European Workshop on Algorithmic Fairness (EWAF'25). Proceedings of Machine Learning Research, 5 pages.

1 Introduction

Recent disruptive generative AI (GenAI) technologies like Large Language Models (LLMs) have found their way into education, science, medicine, finance, and other fields. This development gives rise to a range of ethical concerns, such as violations of privacy, security breaches, and unfair practices where *forgetting* some information/knowledge is desirable. For example, the Right to be Forgotten (RTBF) in the EU General Data Protection Regulation (GDPR) gives people the right that certain information about them be removed or made inaccessible to the public. However, this is not straightforward to implement in the case of large models like LLMs. Models trained on large-scale datasets encode and 'remember' a lot of information, including sensitive information, such as patients' private medical records, or misleading and biased information, such as racial or sexist stereotypes. Yet, due to the massively distributed nature of neural networks, it is difficult to know how and where models encode specific tokens of information and how to suppress or remove them. As a result, the proliferation of sensitive data or unwanted information through LLMs continues to threaten a range of moral and legal rights like privacy and the rights not to be discriminated against (e.g., racial prejudice) and create tensions with laws like RTBF [1]. This happens through the well-documented tendency of LLMs to produce biased outputs [2,3] and security threats through jailbreaking attempts to retrieve sensitive private information [4,5].

Authors' Contact Information: Iqra Aslam, Leibniz Universität Hannover, Germany, iqra.aslam@stud.uni-hannover.de

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

2 Machine Unlearning

Various moral and legal norms regulate who can access, store, and retrieve information, but enforcing them is difficult given the complicated nature of LLMs. Hence, norms alone are insufficient to ensure privacy, security, and bias mitigation: we need technological progress to make machines 'forget' what they have memorized when retaining that particular information is no longer desirable. In response to these and similar concerns, data scientists and machine learning researchers have launched research into a new area called Machine Unlearning (MU). The motivation behind MU is to remove the unwanted data and its effects from a model while retaining its ability to perform a range of desired tasks. An ideal, gold-standard approach to achieve this is removing undesirable data from a training set and retraining a model from scratch. This technique is called *exact unlearning* [6]. Yet, while exact unlearning is ideal for achieving a model that retains only the desirable data and performs its task, it is 1) computationally taxing to retrain models from scratch and 2) assumes that removing certain data from a training set will not result in unwanted model behaviors. Addressing mainly the first concern, the key alternative approach pursued in the MU literature is *approximate unlearning* [7], where the goal is to remove the effects of the unwanted data (ideally to the point that it resembles exact unlearning) without retraining from scratch. While various MU techniques have been proposed and are promising next-best solutions, the complexity of MU undertakings also hampers MU's initial appeal, as various studies indicate that MU techniques do not always mitigate the problems they aim to solve [8,9]. Key technical challenges include: 1) scalability: large-scale models with billions of parameters make efficient unlearning difficult [10]; 2) utility vs. forgetting trade-offs: balancing model performance with effective data removal remains challenging [11]; 3) residual information: ensuring complete removal of sensitive data without traces is complex [12]; and 4) lack of standardized benchmarks: the absence of unified evaluation metrics complicates assessment [13].

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

However, beyond making progress on technical challenges, I argue that MU must first clarify its ethic-epistemological foundations. Currently, MU's goals are predominantly articulated by ML developers within academic research, often driven by efforts to meet legal mandates like the Right to be Forgotten (RTBF), and sometimes shaped by ethical narratives surrounding fairness and safety [14,15]. These goals may, however, conflict with the diverse values of various stakeholders, thereby casting doubts on MU's practical effectiveness.

Broadly, stakeholders in MU encompass any individuals or groups impacted by or invested in how ML systems acquire, encode, and disseminate knowledge or information. This may include: **1) Data providers and users/subjects:** Individuals whose data is used to train models (e.g., artists, patients) and those affected by ML systems' outputs, particularly concerning bias and safety (e.g., marginalized communities), **2**) **Technology developers, distributors, and regulators:** Entities involved in the research, development, deployment, and governance of ML and MU (e.g., developers, tech companies, policymakers, etc.). This paper will primarily focus on stakeholders whose data and creative work are used to train models, as well as users/subjects, along with communities vulnerable to algorithmic bias. The rationale for this narrower focus is the heightened personal risks these stakeholders face, as opposed to other stakeholders whose primary concerns might be monetary or professional.

When the MU goals, as articulated by MU researchers, are juxtaposed with the interests of these particularly vulnerable stakeholders, three levels of concern emerge:

- **Practical gaps:** MU aims to make machines forget sensitive data, citing goals like privacy, security, copyright, and bias mitigation. However, it lacks a thorough analysis of real-world cases beyond RTBF-related research (see p. 35 [16]).
- **Ethical gaps:** MU lacks a clear ethical foundation, leaving it uncertain which moral goods and rights are at stake and how MU's efforts promote them. Without deeper analysis, its motivations remain unclear.
- **Misalignment:** Given these gaps, whether MU's stated goals align with stakeholder concerns is uncertain. If they do, are they met effectively? If not, what other motivations should drive MU? Current research does not provide sufficient clarity.

In response to this situation, I argue that it is crucial to assess more critically what MU is, i.e. what it: a) is currently doing, b) is supposed to do, and c) whether a) and b) "fit well together." My central claim is that MU's stated goals often diverge from its real-world effects. An integrated ethic-epistemological lens can help better understand whether MU serves the right goals and does so successfully. Examining MU from an epistemological perspective is essential because, at its core, it is a knowledge and/or information management project. It concerns what information about the world and the people is encoded within, and potentially removed from, ML systems. Therefore, to fully grasp MU and its objectives, we must consider the fundamental concepts of knowing, not knowing, learning, and forgetting specific pieces of information. Crucially, this inquiry concerns knowledge *about people*, which carries the power to shape individual identities and reinforce or challenge societal marginalization by disseminating certain personal information or particular stereotypes. The central question in this investigation of MU goals shifts from merely 'Does system A know X?' to 'What is A *permitted* to know about subject B?'. This is also relevant in the current social epistemological debates surrounding ethics of forgetting, where philosophers claim that, in some cases, forgetting/not knowing a piece of information X about another individual is our moral duty [17]. In the case of MU, this technology plays a pivotal role in this dynamic by dictating what information concerning B is retained, removed, or modified within a model and to what degree. Consequently, MU is a matter of social epistemology, as it engages the social dimensions of knowledge and information acquired, encoded, and disseminated by ML systems. Furthermore, because the knowledge (or lack thereof) disseminated by ML models can have profound ethical and, subsequently, legal ramifications for individuals and society at large, the ethics and epistemology are intertwined in motivating the MU project.

3.1 Ethics of Forgetting

MU is often motivated by ethical grounds such as privacy (specifically the RTBF), bias mitigation, fairness, security, and copyright. At the centre of achieving these moral goods is the idea of making machines forget or unlearn certain information or knowledge. Yet, these broad categories primarily justify technical research without understanding, clarifying, and addressing what it means to unlearn or forget.

To ensure that MU addresses real stakeholder concerns instead of enlisting overgeneralized motivations, I propose to examine social epistemology and the ethics of forgetting to determine what forms of unlearning are necessary based on stakeholder rights and ethical considerations and what MU methods are currently targeting and achieving. Therefore, using the social-epistemological lens on the ethics of forgetting, I propose a tripartite distinction to clarify different possible forms of unlearning. It differentiates between: **1) Never knowing X** (Where X can be any data that stakeholders want the system to unlearn), **2) learning X and then forgetting X,** and **3) knowing X but not acting in accordance with the knowledge of X**.

This distinction helps assess MU's effectiveness in safeguarding stakeholder rights and ethical interests. To illustrate its applicability, let me offer three brief cases where MU could play a crucial role.

3.1.1 Case: 1) Stable Diffusion's Removal of Artist's Style - Never Knowing X

Ensuring a piece of information is never known to anyone is the ideal scenario when someone doesn't want the world to know X about them. This amounts to a state of complete ignorance with respect to X, and in the context of machine learning, it can be achieved by ensuring that X, or other related information through which X can be inferred, is not included in the training data. This, of course, requires that developers and other responsible parties exercise suitable precautions, which, for various reasons, is often not the case.

The next-best possible action, as per MU, is exact unlearning, also corresponding to the "never knowing" type in the tripartite distinction. Exact unlearning means completely deleting the unwanted data from the training set and then training the ML model again. Unlike mere forgetting, this form of unlearning aims for total ignorance. However, achieving this requires retraining from scratch, which is costly [18]. In some cases, a full retraining may be necessary, such as removing racial bias in a facial recognition model trained predominantly on one racial group, and so retraining with a diverse dataset may be the only way to ensure fairness if the model performs poorly on other groups. Similarly, exact unlearning has been applied in response to ethical and legal challenges, as seen in the case of Stable Diffusion. Early versions of Stable Diffusion (before SD 2.0) were trained on public datasets containing copyrighted, celebrity images, and other questionable material such as nude and pornographic material As text-to-image systems gained traction, artists protested that the model had effectively "memorized" their styles, leading to lawsuits and pressure on Stability AI [19]. While artists demanded that their works be removed entirely from the training data, requiring data disgorgement akin to exact unlearning [20, 21], Stability AI did not follow these requests. However, they did remove other images, such as nude and pornographic material, from the training data before retraining Stable Diffusion 2.0 from scratch [22]. This clearly shows that even when Stability AI was conducting retraining from scratch, mainly for performance improvement, they only cherry-picked, removing one type of content from training data (pornographic material) and ignoring the demands of other stakeholders (artists). One may then ask what this gap between the MU goals of data providers and users/subjects and those of companies/developers entails for the ethical and legal concerns surrounding ML and MU.

3.1.2 Case: 2) Facial Recognition and Data Privacy - Knowing X but Forgetting It

To understand unlearning in terms of knowing X but forgetting it, consider a facial recognition model trained on thousands of images, including those of a specific individual. Later, that person requests that their data be erased. An MU method (approximate unlearning) [23] attempts to remove their identifying information while ensuring the model retains its overall performance. This is done by adjusting the model's parameters to minimise the influence of the forget set, ideally making it as if that data was never part of the training process. While this represents progress in MU, it remains mainly at the academic stage. Yet, even at this level, challenges persist, such as ensuring that unlearning is sufficient to safeguard the rights of the stakeholders. For example, while the data may be largely inaccessible or erased, different users may have different expectations about what it means for their data to be truly forgotten. This creates tension between technical feasibility and the stringent requirements of privacy concerns, including but not limited to RTBF, leaving open questions about whether current approaches can fully satisfy such ethical rights.

3.1.3 Case: 3) GitHub Copilot and Content Filtering – Knowing X but Choosing Not to Act on It

When GitHub Copilot was launched, developers discovered it could generate code snippets identical to the ones in GPL-licensed public repositories, raising legal concerns. Github Copilot, trained on open-source code, did not enforce license terms, and with specific prompts, it reproduced near-identical GPL code, proving memorization [24]. This led to lawsuits accusing Microsoft (which owns GitHub) and OpenAI (whose model, called Codex, was used for GitHub Copilot) of laundering open-source code without attribution. GitHub responded by adding filters to block exact GPL code snippets, but the model still retains their patterns [25]. Unlike Stable Diffusion, which modified its model to forget styles, Copilot "knows" GPL code but suppresses its output. This is akin to OpenAI's ChatGPT content moderation filters [26]. However, these content moderation filters of its earlier models were being bypassed by clever prompting techniques such as DAN (Do Anything Now) [27]. While OpenAI claims that the recent and current models have improved security [28], the attempts to develop jailbreaks persist [29]. Copilot and ChatGPT content filters are a weak form of unlearning, aligning with "Knowing X but Choosing Not to Act on It." They raise ethical concerns about intellectual property in the case of Copilot and about bias, privacy, and security in the case of ChatGPT. While the model avoids direct violations, it still internally "remembers" information, potentially leading to indirect infringements. This weak form of unlearning questions the sufficiency of content filtering in addressing the deeper issue of how models retain and reproduce learned information.

4 Conclusions and Recommendations

As evident from the cases above, the extent to which the MU efforts fit with stakeholder goals varies. Sometimes, MU efforts may help; e.g., artists might be satisfied if models don't regurgitate their style even if older models retain their style information. Sometimes, even comparatively lesser efforts, like OpenAI's content moderation filters, might suffice. However, such efforts may not always be enough - some stakeholders might demand more protection than current MU efforts. For example, in facial recognition systems trained on data sets that result in a persistent racial bias not removed with MU alone, stakeholder goals might legitimately require exact unlearning/never knowing. This shows that it is crucial to be more explicit and precise about values/goals and the fit relationship between concrete MU efforts and what they profess to achieve. This is necessary for the MU project to be a reasonable effort in securing moral goods and goals like privacy, bias mitigation, security, and copyright. Therefore, based on the arguments developed in this paper, further MU research and practical efforts should be directed towards:

1. aligning stakeholder goals and values with MU efforts. This requires developing frameworks to identify and reconcile diverse expectations with the technical realities of MU;
2. carefully creating policies around how to negotiate conflicts between stakeholders and AI developers/companies;
3. ensuring more transparency around how models are trained from the get-go and continually fashion future AI/ML projects after studying the moral claims of stakeholders to mitigate ethical issues before they arise and become difficult to handle.

Acknowledgements

I would like to thank Dr. Donal Khosrowi for supervising this project and helping shape and refine the ideas developed here. Moreover, I would like to thank Rahul Nagshi for explaining the technical aspects of Machine Learning and for his extensive moral support.

References

[1] Alexander X. Oesterling, Jiaqi Ma, Flavio P. Calmon, and Himabindu Lakkaraju. 2023. Fair Machine Unlearning: Data Removal while Mitigating Disparities. *arXiv.org* abs/2307.14754, (July 2023). https://doi.org/10.48550/arxiv.2307.14754

[2] Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in Large Language Models: Origin, Evaluation, and Mitigation. (November 2024). https://doi.org/10.48550/arxiv.2411.10915

[3] Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiao-Yan Zhao. 2023. Bias of AI-Generated Content: An Examination of News Produced by Large Language Models. *arXiv.org* abs/2309.09825, (September 2023). https://doi.org/10.48550/arxiv.2309.09825

[4] Das, B. C., Amini, M. H., and Wu, Y. 2025. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6), 1–39.

[5] Rui Zhang, Hongwei Li, Xiaoyan Qian, Weibin Jiang, and Hongxiang Chen. 2025. On large language models safety, security, and privacy: A survey. Journal of Electronic Science and Technology 100301.

[6] Wei Wang, Zhi Tian, Chao Zhang, and Shui Yu. 2024. Machine unlearning: A comprehensive survey. arXiv:2405.07406.

[7] Jie Xu, Zhi Wu, Chen Wang, and Xiaohong Jia. 2024. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence.*

[8] Shuang Liu, Yao Yao, Jinyuan Jia, Steven Casper, Natalia Baracaldo, Peter Hase, et al. 2025. Rethinking machine unlearning for large language models. Nature Machine Intelligence, 1–14.

[9] Zhenyu Liu, Guangyu Dou, Zhiqiang Tan, Yiming Tian, and Ming Jiang. 2024. Machine unlearning in generative AI: A survey. arXiv:2407.20516.

[10] Xu, J., Wu, Z., Wang, C., and Jia, X. 2024. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

[11] 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.

[12] 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.

[13] Shiyu Zhou, Lei Wang, Jie Ye, Yuhong Wu, and Hongbin Chang. 2024. On the limitations and prospects of machine unlearning for generative AI. arXiv:2408.00376.

[14] W. Wang, Z. Tian, C. Zhang, and S. Yu. 2024. Machine Unlearning: A Comprehensive Survey. CoRR abs/2405.07406 (2024). https://doi.org/10.48550/arXiv.2405.07406

[15] A. F. Cooper, C. A. Choquette-Choo, M. Bogen, M. Jagielski, K. Filippova, K. Z. Liu, and K. Lee. 2024. Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice. arXiv preprint arXiv:2412.06966.

[16] Andrew F. Cooper, Catherine A. Choquette-Choo, Margaret Bogen, Matthew Jagielski, Kateryna Filippova, Kevin Z. Liu, et al. 2024. Machine unlearning doesn't do what you think: Lessons for generative AI policy, research, and practice. arXiv:2412.06966.

[17] R. Basu. 2022. The importance of forgetting. Episteme 19, 4 (2022), 471–490. https://doi.org/10.1017/

[18] Meyer, D. 2024. Why the cost of training AI could soon become too much to bear. *Fortune*, Apr. 4, 2024. Available at: https://fortune.com/2024/04/04/ai-training-costs-how-much-is-too-much-openai-gpt-anthropic-microsoft/.
[19] 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*, pp. 186-196.
[20] R. Salkowitz. 2024. Artist and Activist Karla Ortiz on the Battle to Preserve Humanity in Art. *Forbes*. Retrieved May 23, 2024 from https://forbes.com/sites/robsalkowitz/2024/05/23/artist-and-activist-karla-ortiz-on-the-battle-to-preserve-humanity-in-art/
[21] M. Heikkilä. 2022. Artists Can Now Opt Out of the Next Version of Stable Diffusion. *MIT Technology Review*. Retrieved December 16, 2022 from https://www.technologyreview.com/2022/12/16/1065247/artists-can-now-opt-out-of-the-next-version-of-stable-diffusion/
[22] Stability AI. 2023. Stable Diffusion Version 2. *GitHub*. Retrieved from https://github.com/Stability-AI/stablediffusion
[23] Choi, D., & Na, D. 2023. Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint arXiv:2311.02240*. https://arxiv.org/abs/2311.02240
[24] DOE 1 et al. n.d. *DOE 1 et al v. GitHub*. 4:22-cv-06823, (N.D. Cal.).
[25] James Vincent. 2022. Microsoft, OpenAI, and GitHub Copilot face class-action lawsuit for AI copyright violations. The Verge. Retrieved November 8, 2022 from https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data.
[26] Yusra Mahomed, Catherine M. Crawford, Sandeep Gautam, Sorelle A. Friedler, and Diana Metaxa. 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.
[27] Alexisvalentino. 2023. *Chatgpt-DAN*. GitHub. Retrieved May 19, 2025 from https://github.com/alexisvalentino/Chatgpt-DAN
[28]Metz, C., & Kang, C. 2023. A.I. regulation is coming. But what will it look like? *The New York Times*. https://www.nytimes.com/2023/03/03/technology/artificial-intelligence-regulation-congress.html.
[29] 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.