SUPPORT POOL
OF EXPERTS PROGRAMME

AI-Complex Algorithms and effective Data Protection Supervision

Effective implementation of data subjects’ rights

by Dr. Kris SHRISHAK

As part of the SPE programme, the EDPB may commission contractors to provide reports and tools on specific topics.

The views expressed in the deliverables are those of their authors and they do not necessarily reflect the official position of the EDPB. The EDPB does not guarantee the accuracy of the information included in the deliverables. Neither the EDPB nor any person acting on the EDPB's behalf may be held responsible for any use that may be made of the information contained in the deliverables.

Some excerpts may be redacted or removed from the deliverables as their publication would undermine the protection of legitimate interests, including, inter alia, the privacy and integrity of an individual regarding the protection of personal data in accordance with Regulation (EU) 2018/1725 and/or the commercial interests of a natural or legal person.

TABLE OF CONTENTS

Document submitted in March 2024

INTRODUCTION

The General data Protection Regulation (GDPR) empowers data subjects through a range of rights. A data subject has the right to information (Articles 12-14), the right of access (Article 15), the right to rectification (Article 16), the right to erasure (Article 17), the right to restrict processing (Article 18), the right to data portability (Article 20), the right to object (Article 21) and the right not to be subject to a decision based solely on automated processing (Article 22).

This report covers techniques and methods that can be used for effective implementation of data subject rights, specifically, the rights to rectification and the right to erasure when AI systems have been developed with personal data. This report addresses these rights together because rectification involves erasure followed by the inclusion of new data. These techniques and methods are the result of early-stage research by the academic community. Improvements and alternative approaches are expected to be developed in the coming years.

1 CHALLENGES

AI systems are trained on data that is often memorised by the models (Carlini et al., 2021). Machine learning models behave like lossy compressors of training data and the performance of these models based on deep learning is further attributed to this behaviour (Schelter, 2020; Tishby & Zaslavsky, 2015). In other words, machine learning models are compressed versions of the training data. Additionally, AI models are also susceptible to membership inference attacks that help to assess whether data about a person is in the training dataset (Shokri et al., 2017). Thus, implementing the right to erasure and rectification requires reversing the memorisation of personal data by the model. This involves deletion of (1) the personal data used as input for training, and (2) the influence of the specific data points in the trained model.

There are several challenges to effectively implement these rights (Bourtoule et al., 2021):

1. **Limited understanding of how each data point impacts the model**: This challenge is particularly prevalent with the use of deep neural networks. It is not known how specific input data points impact the parameters of a model. The best known methods rely on "influence functions" involving expensive estimations (by computing second-order derivatives of the training algorithm) (Cook & Weisberg, 1980; Koh & Liang, 2017).

2. **Stochasticity of training**: Training AI models is usually performed by random sampling of batches of data from the dataset, random ordering of the batches in how and when they are processed, and parallelisation without time-synchronisation. All these make the training process probabilistic. As a result, a model trained with the same algorithm and dataset could result in different trained models (Jagielski et al., 2023).

3. **Incremental training process**: Models are trained incrementally such that an update relying on specific training data point will affect all subsequent updates. In other words, updates in the training process depend on all previous updates. In the distributed training setting of federated learning, multiple clients keep their data and train a model locally before sending the updates to a central server. In such a setting, even when a client only once sends its update and contributes to the global model at the central server, the data and the contribution of this client influences all future updates to the global model.

4. **Stochasticity of learning**: In addition to the training process, the learning algorithm is also probabilistic. The choice of the optimiser, for example, for neural networks can result in many different local minima (result of the optimisation). This makes it difficult to correlate how a specific data point contributed to the "learning" in the model.

2 HOW TO DELETE AND UNLEARN

1. **Data Curation and Provenance**: Essential elements to implement the rights in Articles 15-17 of GDPR are data curation and provenance. However, these are necessary but not sufficient for implementing these rights completely as they do not include information related to how the data influenced the trained model. These are prerequisites for the other approaches in this report.

2. **Retraining of models**: Deleting the model, removing the personal data requested to be erased, and then retraining the model with the rest of the data is the method that implements the rights in Articles 16-17 of the GDPR effectively. For small models, this method works well. However, for larger models, the training cost is very expensive and often alternative approaches might be required, especially when numerous deletion requests are expected. Furthermore, this approach, and many of the other approaches, assumes that the model developer is in possession of the training datasets when the requirement to delete and retrain arises.

3. **Exact unlearning**: To avoid retraining the entire model, approaches to unlearn the data have been proposed. Despite the growing literature, there are very few unlearning methods that are currently most likely to be effective.

   a. **Model agnostic unlearning**: This method is not dependent on the specific machine learning technique. It is the only approach which has been shown to work for deep neural networks. This approach either (1) relies on storing model gradients (Wu et al., 2020)**,** or (2) relies on the measurement of sensitivity of model parameters to changes in datasets used in federated learning (Tao et al., 2024), or (3) modifies the learning process to be more conducive to unlearning (Bourtoule et al., 2021)**.**

      The latter, known as SISA **(Sharded, Isolated, Sliced, and Aggregated), is currently the best-known approach. It involves modifying the training process, but is independent of specific learning algorithms** (Bourtoule et al., 2021)**. This approach presets the order in which the learning algorithm is queried to ease the unlearning process.** The approach can be described as follows:

      i. The training dataset is divided into multiple "shards" such that each training data point is present in only one "shard". This allows for a non-overlapping partition of the dataset. It is also possible to further "slice" the "shards" so that the training is more modular and deletion is eased further.

      ii. The model is then trained on each of these shards or slices. This limits the influence of the data points to these specific shards or slices.

      iii. When a request for erasure or rectification arrives, unlearning is performed, not by retraining the entire model, but by retraining only the shard or slice that had included the "delete requested" data.

This method is flexible. For instance, the shards can be chosen such that the most likely "delete request" data are in one shard. Then, fewer shards will need to be retrained, assuming that personal data and non-personal data are separated as part of data curation.

b. **Model intrinsic unlearning**: These methods are developed for specific AI techniques. For instance, the methods that are suitable for decision trees and random forests have been shown to be effective (Brophy & Lowd, 2021) by using a new approach to develop decision trees and then relying on strategic thresholding at decision nodes for continuous attributes, and at high-level random nodes. Then the necessary statistics are cached at all the nodes to facilitate removal of specific training instances, without having to retrain the entire decision tree.

c. **Application specific unlearning**: While exact unlearning is generally expensive in terms of computation and storage, some applications and their algorithms are more suitable to exact unlearning. Specifically, recommender systems based on k-nearest neighbour models are well suited due to their use of sparse interaction data. Such models are widely used in many techniques including collaborative filtering and recent recommender system approaches such as next-basket recommendation. Using efficient data structures, sparse data and parallel updates, personal data can be removed from recommendation systems (Schelter et al., 2023).

4. **Approximate Unlearning:** Significant amount of technical literature on machine unlearning focuses on approximate unlearning, where the data is not deleted, but instead, the model is adjusted such that the probability of the influence of the data, estimated based on proxy signals, on the model is reduced. Approximate unlearning is less expensive in terms of computation and storage requirements.

   a. **Finetuning:** Once a model is trained, it can be finetuned for many purposes including the approximate removal of the effect of the data that has been requested to be deleted (Golatkar et al., 2020; Warnecke et al., 2023). When a deletion request along with the "removal dataset" (the data to be removed) is received, the model is trained again for a few epochs on this "removal dataset" such that the model "forgets" it.

   b. **Influence unlearning:** Approximate unlearning approaches have been proposed that rely on estimating the influence of specific data on the model (Izzo et al., 2021; Koh & Liang, 2017). This estimation is then used to update the model for unlearning, which is akin to finetuning. Usually, these approaches also require additional model training. However, to reduce the computation, it is also possible to prune the model (or reduce the size) before the unlearning process (Jia et al., 2023).

   c. **Intentional misclassification:** When a request to delete specific data about a person is received, the model owner intentionally misclassifies these data points. This can be achieved with access to the pre-trained model and the data points provided by the data subject with the deletion request but does not require access to the rest of the training dataset (Cha et al., 2024). Another approach, saliency unlearning, tackles the problem of unlearning at the level of weights rather than data or model. It relies on estimating the weights that are most relevant (salient) for unlearning before deploying random labels for the data to be deleted (Fan et al., 2024)**.** This approach has been proposed for image classification and generation.

   d. **Parameter deletion**: Another approach to unlearn without deleting the data from the model but removing its influence involves storing a list of data and parameter updates during the training process. When a deletion request arrives, the parameter updates are undone (Graves et al., 2021). Due to the need to store the parameter updates, this approach has a high storage requirement, especially for large models, although less than that for exact unlearning.

5. **Differential privacy and model retiring policy**: Differential privacy gives a mathematical guarantee that there is a bound on the contribution of individual data point to the model and that this contribution is small. However, the contribution is not zero,[1] thus necessitating "unlearning" (Chandrasekaran et al., 2021). One approach is to combine differential privacy with a policy to periodically retire or delete the model and retrain a differentially private model, instead of retraining for every deletion request.

   When a deletion request is received, if the relevant personal data is in the possession of the data controller, then the data should be deleted. The model deletion is not performed for every request because it is unclear how individual personal data points impact the differentially private model. However, once there is a sufficiently large number of requests, then, put together, these data points would affect the model (still unknown how exactly), and thus there is reason enough to delete the model and retrain the model with differential privacy.

3 WHAT TO UNLEARN

1. **Samples**: A deletion request for a specific piece of information or sample about a person. Methods described in the previous section have been developed for this setting.

2. **Features**: In some applications, features and labels may hold certain personal characteristics that are to be deleted. An approximate unlearning method has been proposed for this purpose by estimating the influence of specific features on the model parameters (Warnecke et al., 2023). This method can be used to unlearn features in a trained model for thousands of data subjects. Another approach involves estimating the correlation between features that could represent the personal characteristics and then to progressively unlearn these features (Guo et al., 2022). This method is most applicable for deep neural networks in the image domain, for example, facial recognition systems, where the deeper layers of the neural networks are smaller (Nguyen et al., 2022).

3. **Class**: AI systems can be designed to classify outputs into one, two or many different classes. In certain applications, the data to be deleted is represented as a class in the trained model. In some facial recognition applications, all data points about a person in the form of facial images belong to a particular class and if a person requests for their personal data to be deleted, then the classification should not work for this person's class. A couple of approximate unlearning methods introduce noise such that the classification error for the deletion class is maximised and then the model is "repaired" to maintain the performance for the rest of the data (Chundawat et al., 2023; Tarun et al., 2024). These methods do not delete all the samples associated with the class, but instead manipulate the trained model for this class directly.

[1] It would be impossible for a model to learn from the training data if the contribution is zero (Bourtoule et al., 2021).

When image classification or facial recognition technology is developed by training Convolutional Neural Network (CNN) models with federated learning, the class is selectively pruned based on extracting features in the images that contribute to different classes (Wang et al., 2022). The person making the deletion request locally extracts these features for their images and sends it to the central server, who then prunes the class from the global model.

4. **Client:** When AI systems are developed with federated learning that includes contribution from multiple clients, a client (or a person) might request that their entire contribution to the global model due to their local dataset be deleted. Due to the incremental training process, only deleting the updates to the global model made by this client is insufficient to remove the influence of this client's data. An approach known as FedEraser stores historical parameter updates at a central server to sanitise all updates that followed the updates of this client (Liu et al., 2021). The sanitisation process involves collaborative updates from the remaining clients whose contributions are still part of the global model.

4 APPROXIMATE UNLEARNING VERIFICATION

Approximate unlearning methods have been proposed with the claim that they are indistinguishable from retraining the model from scratch without the deleted data. The claims are usually based on metrics such as indistinguishability to a hypothetically model retrained from scratch, unlearning accuracy, remaining accuracy and membership inference attacks.

Unlearning accuracy is the accuracy of the unlearned model on the data expected to be forgotten. Remaining accuracy is the accuracy of the unlearned model on the remaining data. Membership inference attacks (MIAs) are used in an attempt to extract "deleted" data from the updated model. If the probability of such extraction is around 50%, then the "deletion" is treated as a success. However, MIA is a privacy attack and relying on it for testing is unreliable. A well-developed model will not be susceptible to MIA, in which case, MIA cannot be used as a proxy signal to test unlearning.

Furthermore, approximate learning lacks strong guarantees. These metrics do not address a very basic concern: it is possible to obtain two models with similar weights and parameters with non-overlapping training data (Thudi et al., 2022). That is, removing an influence of a particular parameter is not sufficient to have "deleted" the data as the influence could have been from a different data. Moreover, the assumption of having to unlearn a model that is indistinguishable from retraining from scratch itself may not be the right approach. This is because a model retrained from scratch could have different model distributions due to the stochasticity of training (Goel et al., 2022; Yang & Shami, 2020).

5 CONCERNS WITH MACHINE UNLEARNING

1. **Privacy**: Just like machine learning, machine unlearning also introduces privacy concerns. Membership inference attacks (Shokri et al., 2017) that have been shown to attack machine learning can also be used against machine unlearning (Chen et al., 2021). The concern here is that when it is possible to query a model twice, once before unlearning and once after unlearning, the person querying could deduce which data was deleted.

2. **Bias**: When deletion requests are made, minority classes are more adversely affected because the datasets in the real world are not balanced. When it comes to data deletion requests, not everyone is equally likely to make such requests. It has been shown that there is a correlation

between the unlearning probability and class labels (Koch & Soll, 2023). Thus, it is imperative that accuracy of models for sub-categories are assessed after unlearning to assess for bias.

6 LIMITING PERSONAL DATA OUTPUT FROM GENERATIVE AI

The approaches discussed thus far address applications including facial recognition technology where personal data processing is concerned. AI systems are susceptible to privacy leakages and to adversarial attacks such as MIA. This is also true of generative AI systems, which could generate personal data as part of its output. Text generation AI based on large language models have been shown to be more susceptible to MIA than small models (Carlini et al., 2021).

In generative AI systems, personal data is output when explicitly prompted (E.g., Give me the birth date of [person name]). The same can take place with image and video generation tools as well. Personal data is also output when not explicitly prompted. These generative AI tools make things up or "hallucinate" (Maynez et al., 2020) and generate factually incorrect content that could reveal personal data about people. E.g., when information about one person is asked and a large language model outputs information about another person (with their name) (D. Zhang et al., 2023).

The area of research to limit generation of personal data from generative AI is new, and much less mature than the field of machine unlearning, which by itself is quite young.

1. **Model finetuning**: In the case of diffusion models (e.g., Stable Diffusion), a method has been proposed to finetune the model such that specific concepts are not output in the images (Gandikota et al., 2023). This method eliminates visual concepts such as specific artistic styles, nudity and certain objects. A similar approach can be used to prevent generation of images with specific personal characteristics (E. J. Zhang et al., 2023). Another approach known as "selective amnesia" applies continuous learning to forget concepts from generative models based on variational autoencoders and diffusion models (Heng & Soh, 2024).
2. **Data redaction**: A variant of model finetuning uses data and class redaction techniques to limit generation of specific outputs in generative adversarial networks (GANs). A set of data that should not be generated is selected as a redaction set, which is then used to generate a "fake distribution" such that outputs falling within the redaction set are penalized (Kong & Chaudhuri, 2023). This approach is based on similar approaches that re-train models to limit generation of specific outputs (Asokan & Seelamantula, 2020; Hanneke et al., 2018; Sinha et al., 2021).
3. **Output modification**: The output of image generators can be modified to not generate specific kinds of images. This can be achieved by training a machine learning classifier to modify outputs before they are revealed to the end users (Rando et al., 2022) or by incorporating additional information and guiding the inference process (Schramowski et al., 2023). Alternatively, reinforcement learning with human feedback can be used (Bai et al., 2022; Ouyang et al., 2022) to prevent generation of personal data. However, such methods have many shortcomings (Casper et al., 2023) and are shown to be easy to circumvent, especially when the end user has access to the parameters, as is the case with fully open-source models.[2]

[2] https://www.reddit.com/r/StableDiffusion/comments/wv2nw0/tutorial_how_to_remove_the_safety_filter_in_5/

CONCLUSION

The GDPR offers data subjects with many rights. This report covers techniques and methods to implement the right to rectification and the right to erasure when AI systems process personal data. Implementing these rights is challenging but many technical approaches have been proposed. Data curation and provenance are prerequisites for these approaches. Some of the challenges such as stochasticity of training AI models can be modified to make compliance with data erasure requests easier (Bourtoule et al., 2021). Such design choices might have performance trade-off but are an aspect of data protection by design. Other important rights offered by the GDPR to data subjects are left to future projects.

As a strong recommendation regarding data protection, only the use of completely anonymised data for the development and deployment of AI models would avoid obligations related to the correction and deletion of personal data in AI models. If it is necessary to use personal data, including pseudonymised data, to develop an AI model then the legal obligations to implement data subject rights apply. The updates and changes made to the AI model should be adequately logged and documented such that subsequent request for rectification and erasure of personal data can be fulfilled.

BIBLIOGRAPHY

Asokan, S., & Seelamantula, C. (2020). Teaching a GAN what not to learn. *Advances in Neural Information Processing Systems*, *33*, 3964–3975.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. *CoRR*, *abs/2212.08073*. https://doi.org/10.48550/ARXIV.2212.08073

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine Unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, 141–159. https://doi.org/10.1109/SP40001.2021.00019

Brophy, J., & Lowd, D. (2021). Machine Unlearning for Random Forests. *ICML*, *139*, 1092–1104.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting Training Data from Large Language Models. *USENIX Security Symposium*, 2633–2650.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., … Hadfield-Menell, D. (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. https://doi.org/10.48550/ARXIV.2307.15217

Cha, S., Cho, S., Hwang, D., Lee, H., Moon, T., & Lee, M. (2024). *Learning to Unlearn: Instance-wise Unlearning for Pre-trained Classifiers* (arXiv:2301.11578). arXiv. http://arxiv.org/abs/2301.11578

Chandrasekaran, V., Jia, H., Thudi, A., Travers, A., Yaghini, M., & Papernot, N. (2021). *SoK: Machine Learning Governance* (arXiv:2109.10870). arXiv. http://arxiv.org/abs/2109.10870

Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., & Zhang, Y. (2021). When Machine Unlearning Jeopardizes Privacy. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 896–911. https://doi.org/10.1145/3460120.3484756

Chundawat, V. S., Tarun, A. K., Mandal, M., & Kankanhalli, M. (2023). Zero-Shot Machine Unlearning. *IEEE Transactions on Information Forensics and Security*, *18*, 2345–2354. https://doi.org/10.1109/TIFS.2023.3265506

Cook, R. D., & Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, *22*(4), 495–508.

Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., & Liu, S. (2024). SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=gn0mIhQGNM

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., & Bau, D. (2023). *Erasing Concepts from Diffusion Models* (arXiv:2303.07345). arXiv. http://arxiv.org/abs/2303.07345

Goel, S., Prabhu, A., & Kumaraguru, P. (2022). Evaluating inexact unlearning requires revisiting forgetting. *arXiv Preprint arXiv:2201.06640*.

Golatkar, A., Achille, A., & Soatto, S. (2020). Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. *CVPR*, 9301–9309.

Graves, L., Nagisetty, V., & Ganesh, V. (2021). Amnesiac machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(13), 11516–11524.

Guo, T., Guo, S., Zhang, J., Xu, W., & Wang, J. (2022). Efficient Attribute Unlearning: Towards Selective Removal of Input Attributes from Feature Representations. *CoRR*, *abs/2202.13295*.

Hanneke, S., Kalai, A. T., Kamath, G., & Tzamos, C. (2018). Actively avoiding nonsense in generative models. *Conference On Learning Theory*, 209–227.

Heng, A., & Soh, H. (2024). Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, *36*.

Izzo, Z., Smart, M. A., Chaudhuri, K., & Zou, J. (2021). Approximate Data Deletion from Machine Learning Models. *AISTATS*, *130*, 2008–2016.

Jagielski, M., Thakkar, O., Tramèr, F., Ippolito, D., Lee, K., Carlini, N., Wallace, E., Song, S., Thakurta, A. G., Papernot, N., & Zhang, C. (2023). Measuring Forgetting of Memorized Training Examples. *ICLR*.

Jia, J., Liu, J., Ram, P., Yao, Y., Liu, G., Liu, Y., Sharma, P., & Liu, S. (2023). Model Sparsity Can Simplify Machine Unlearning. *NeurIPS*.

Koch, K., & Soll, M. (2023). No Matter How You Slice It: Machine Unlearning with SISA Comes at the Expense of Minority Classes. *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 622–637. https://doi.org/10.1109/SaTML54575.2023.00047

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. *International Conference on Machine Learning*, 1885–1894.

Kong, Z., & Chaudhuri, K. (2023). Data Redaction from Pre-trained GANs. *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 638–677. https://doi.org/10.1109/SaTML54575.2023.00048

Liu, G., Ma, X., Yang, Y., Wang, C., & Liu, J. (2021). FedEraser: Enabling Efficient Client-Level Data Removal from Federated Learning Models. *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, 1–10. https://doi.org/10.1109/IWQOS52092.2021.9521274

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W.-C., Yin, H., & Nguyen, Q. V. H. (2022). *A Survey of Machine Unlearning* (arXiv:2209.02299). arXiv. http://arxiv.org/abs/2209.02299

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *NeurIPS*.

Rando, J., Paleka, D., Lindner, D., Heim, L., & Tramèr, F. (2022). *Red-Teaming the Stable Diffusion Safety Filter* (arXiv:2210.04610). arXiv. http://arxiv.org/abs/2210.04610

Schelter, S. (2020). 'Amnesia'—Machine Learning Models That Can Forget User Data Very Fast. *CIDR*.

Schelter, S., Ariannezhad, M., & De Rijke, M. (2023). Forget Me Now: Fast and Exact Unlearning in Neighborhood-based Recommendation. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011–2015. https://doi.org/10.1145/3539618.3591989

Schramowski, P., Brack, M., Deiseroth, B., & Kersting, K. (2023). Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. *CVPR*, 22522–22531.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership Inference Attacks Against Machine Learning Models. *IEEE Symposium on Security and Privacy*, 3–18.

Sinha, A., Ayush, K., Song, J., Uzkent, B., Jin, H., & Ermon, S. (2021). Negative Data Augmentation. *ICLR*.

Tao, Y., Wang, C.-L., Pan, M., Yu, D., Cheng, X., & Wang, D. (2024). *Communication Efficient and Provable Federated Unlearning*. 1119–1131.

Tarun, A. K., Chundawat, V. S., Mandal, M., & Kankanhalli, M. (2024). Fast Yet Effective Machine Unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–10. https://doi.org/10.1109/TNNLS.2023.3266233

Thudi, A., Jia, H., Shumailov, I., & Papernot, N. (2022). On the Necessity of Auditable Algorithmic Definitions for Machine Unlearning. *USENIX Security Symposium*, 4007–4022.

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. *ITW*, 1–5.

Wang, J., Guo, S., Xie, X., & Qi, H. (2022). Federated Unlearning via Class-Discriminative Pruning. *WWW*, 622–632.

Warnecke, A., Pirch, L., Wressnegger, C., & Rieck, K. (2023). Machine Unlearning of Features and Labels. *Proceedings 2023 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium, San Diego, CA, USA. https://doi.org/10.14722/ndss.2023.23087

Wu, Y., Dobriban, E., & Davidson, S. (2020). Deltagrad: Rapid retraining of machine learning models. *International Conference on Machine Learning*, 10355–10366.

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316.

Zhang, D., Finckenberg-Broman, P., Hoang, T., Pan, S., Xing, Z., Staples, M., & Xu, X. (2023). *Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions* (arXiv:2307.03941). arXiv. http://arxiv.org/abs/2307.03941

Zhang, E. J., Wang, K., Xu, X., Wang, Z., & Shi, H. (2023). Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. *CoRR*, *abs/2303.17591*.

edpb

European Data Protection Board