ARE WE REALLY UNLEARNING? THE PRESENCE OF RESIDUAL KNOWLEDGE IN MACHINE UNLEARNING

Hsiang Hsu, Pradeep Niroula, Zichang He, Chun-Fu Chen
JPMorganChase Global Technology Applied Research
{first.last, richard.cf.chen}@jpmchase.com

ABSTRACT

Machine unlearning seeks to remove a set of forget samples from a pre-trained model to comply with emerging privacy regulations. While existing machine unlearning algorithms focus on effectiveness by either achieving indistinguishability from a re-trained model or closely matching its accuracy, they often overlook the vulnerability of unlearned models to slight perturbations of forget samples. In this paper, we identify a novel privacy vulnerability in unlearning, which we term residual knowledge. We find that even when an unlearned model no longer recognizes a forget sample—effectively removing direct knowledge of the sample—residual knowledge often persists in its vicinity, which a re-trained model does not recognize at all. Addressing residual knowledge should become a key consideration in the design of future unlearning algorithms.

1 INTRODUCTION

The extensive use of user data in training machine learning (ML) models has raised significant privacy concerns, especially when users invoke their "Right to be Forgotten," as highlighted in recent regulations such as the EU's General Data Protection Regulation (GDPR) (Voigt & Von dem Bussche, 2017). This right mandates that a ML-driven system must completely erase user data not only from databases but also from the models themselves upon a user's request (Shastri et al., 2019). Consequently, simply deleting user data from databases is often inadequate, as the data can still be extracted (Carlini et al., 2023) or reconstructed (Li et al., 2024) from the models themselves, particularly from Deep Neural Networks (DNNs). Ideally, a machine learning model may "exactly" eliminate the influence of user data by re-training from scratch with a new dataset that excludes the specified data for each removal request. However, this re-training approach is computationally expensive and time-consuming, making it impractical for real-time or large-scale implementation.

To address this challenge, a more scalable approach known as *machine unlearning* has been proposed (Cao & Yang, 2015). This technique attempts to selectively forget specific data (referred to as forget samples) from a trained model without requiring complete re-training. Although machine unlearning methods cannot achieve exact unlearning, they are capable of achieving "approximate" unlearning. The criteria for approximate unlearning stem from the concept of Differential Privacy (DP) (Dwork et al., 2014), which requires that the model, after unlearning, remains statistically indistinguishable from a model that has been fully re-trained without the forget samples. Here, the model obtained through re-training serves as the gold standard for evaluating the effectiveness of machine unlearning algorithms.

Although the concept of approximate unlearning provides a theoretical foundation for the effectiveness of unlearning, its evaluation remains challenging since the exact distributions of model weights are generally infeasible to determine[1]. In practice, unlearning algorithms are evaluated by comparing an unlearned model's accuracy on forget samples to that of a re-trained model. However, even if the accuracy of the unlearned model matches that of the re-trained model, there is no guarantee that this agreement in accuracy is stable under slight, imperceptible perturbations to the forget samples. If a forget sample, when modified by a barely perceptible perturbation, can be re-identified by the unlearned model, it introduces an additional layer of privacy risk.

[1]Approximate unlearning and its variants have primarily been evaluated on linear models; see, e.g., Guo et al. (2019); Chourasia & Shah (2023); Chien et al. (2024).

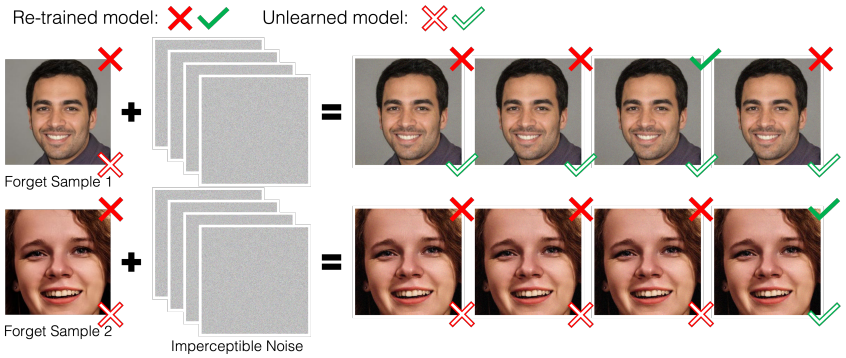

Figure 1: Residual knowledge in machine unlearning is illustrated by comparing the prediction consistency of the re-trained and unlearned models. Solid check or cross marks indicate the correctness of predictions from the re-trained model, while hollow ones represent the unlearned model. For forget sample 1, both models exhibit the same predictions (left); however, the unlearned model often correctly predicts more of the perturbed samples (four images on the right, which are perturbed by the imperceptible noise in the middle). Ideally, as seen with forget sample 2, both the re-trained and unlearned models should consistently predict correctly across the original and all perturbed samples.

In this paper, we draw attention to this new privacy risk, which we term the *residual knowledge of machine unlearning*. We observe that even for a forget sample that the unlearned model no longer recognizes, it is often possible to apply an imperceptible perturbation on the sample such that the unlearned model correctly identifies while a re-trained model does not recognize it at all. See Figure 1 for an illustration. This phenomenon suggests that although unlearning algorithms remove direct knowledge of the forget sample itself, there may still be residual knowledge about the sample remains in its vicinity—identifiable information that would not have existed had the model been re-trained from scratch. We provide empirical evidence demonstrating the existence of this phenomenon across several existing unlearning algorithms.

2 Background and Related Work

We start by defining the notations and providing a brief overview of existing machine unlearning algorithms, along with the metrics commonly used to evaluate their effectiveness.

Let $\mathcal{S} \triangleq \{\mathbf{s}_i = (\mathbf{x}_i, y_i)\}_{i=1}^n$ be a training dataset with n sample points $\mathbf{s}_i$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbf{R}^d$ represents the feature vector and $y_i \in \mathcal{Y}$ denotes the target. The hypothesis space, denoted as $\mathcal{H}$, consists of functions parameterized[2] by $\mathbf{w} \in \mathcal{W}$ that map from $\mathcal{X}$ to $\mathcal{Y}$, i.e., $\mathcal{H} \triangleq \{h_\mathbf{w} : \mathcal{X} \to \mathcal{Y}; \mathbf{w} \in \mathcal{W}\}$. Additionally, let $\|\mathbf{z}\|_p$ denote the ℓ_p-norm of a vector $\mathbf{z}$ for a scalar $p \geq 1$.

Existing unlearning algorithms. Consider a randomized learning algorithm $A : \mathcal{S} \to \mathcal{H}$, such as Stochastic Gradient Descent (SGD). Machine unlearning is a mechanism $M(h_\mathbf{w}, \mathcal{S}, \mathcal{S}_f)$, applied to a trained model $h_\mathbf{w} = A(\mathcal{S})$, that aims to remove the influence of certain samples within a subset $\mathcal{S}_f \subseteq \mathcal{S}$ (Xu et al., 2023). Here, the subset $\mathcal{S}_f$ is commonly referred to as the forget set. The simplest mechanism for machine unlearning involves obfuscating model weights (Golatkar et al., 2020a), such as $M(A(\mathcal{S}), \mathcal{S}, \mathcal{S}_f) = \mathbf{w} + \sigma\mathbf{n}$, where $\mathbf{n} \sim N(0, \mathbf{I}_{|\mathbf{w}|})$ is the isotropic Gaussian noise. However, as σ increases, this approach can lead to poor overall model performance by making $\mathbf{w}$ independent of $\mathcal{S}$. It is therefore essential to define a retain set $\mathcal{S}_r \triangleq \mathcal{S} \backslash \mathcal{S}_f$ and design the unlearning mechanism to preserve the model's performance (e.g., accuracy) on $\mathcal{S}_r$.

Besides re-training, in which we have $M(A(\mathcal{S}), \mathcal{S}, \mathcal{S}_f) = A(\mathcal{S}_r)$, numerous algorithms[3] have been developed to achieve the goal of unlearning. For instance, Guo et al. (2019) leverages influence functions to assess the impact of forget samples and directly update model weights using a one-step Newton method. This Newton-based approach was later extended by Golatkar et al. (2020b), which linearizes neural networks using the Neural Tangent Kernel (NTK). Kodge et al. (2024) pro-

[2]We consider a parameterized $\mathcal{H}$ since machine unlearning for non-parametric methods, such as a nearest-neighbor classifier, can be trivially achieved by simply removing the training sample in constant time.

[3]Due to space limitations, we cannot cover all existing unlearning methods. For a more comprehensive survey, please refer to Nguyen et al. (2022); Xu et al. (2023).

pose removing the influence of forget samples from activation spaces by applying singular value decomposition to layer-wise activations. Fine-tuning is also a widely adopted approach. For example, Gradient Descent (GD) fine-tunes the model exclusively on retain samples (Neel et al., 2021), whereas Gradient Ascent (GA) reverses the gradient updates associated with forget samples (Graves et al., 2021; Jang et al., 2022). More recently, Kurmanji et al. (2024) propose NegGrad+, an approach that fine-tunes a model by simultaneously applying GD on the retain set and GA on the forget set, effectively balancing learning and unlearning in a single optimization step.

Approximate unlearning and its variants. The certification of approximate unlearning was first introduced in Guo et al. (2019), where an unlearning algorithm M is said to satisfy (ϵ, δ)-certified removal if it is (ϵ, δ)-DP (Dwork et al., 2014). The concept of statistical indistinguishability in DP has further inspired Rényi unlearning (Chourasia & Shah, 2023; Chien et al., 2024), an alternative certification of approximate unlearning that extends Rényi DP (Mironov, 2017). An unlearning algorithm M is said to satisfy (α, ϵ)-Rényi unlearning if for all $\mathcal{S}$ and $\mathcal{S}_r$ that differ by only a single sample (i.e., there is exactly one forget sample), the α-Rényi divergence $D_\alpha(\cdot\|\cdot)$ (Rényi, 1961) satisfies $D_\alpha(\nu\|\nu') \leq \epsilon$, where $M(A(\mathcal{S}), \mathcal{S}, \mathcal{S}_f) \sim \nu$ and $A(\mathcal{S}_r) \sim \nu'$. Both (ϵ, δ)-certified removal and (α, ϵ)-Rényi unlearning rely on comparing the distributions of the unlearned and re-trained models. However, when $\mathcal{H}$ is complex, these distributions are generally infeasible to compute. A more practical approach to certified unlearning leverages a readout function, $r : \mathcal{H} \times \mathcal{S} \to \mathbb{R}$, which extracts data-specific information from a model h, such as accuracy on the retain or forget samples. This approach enables an information-theoretic perspective on unlearning, as proposed by Nguyen et al. (2020) and Golatkar et al. (2020a), where unlearning can be quantified via the Kullback-Leibler divergence $D_{\mathsf{KL}}(\cdot\|\cdot)$ (Kullback & Leibler, 1951): $D_{\mathsf{KL}}(\Pr[r(M(h, \mathcal{S}, \mathcal{S}_f), \mathcal{T})]\|\Pr[r(A(\mathcal{S}_r), \mathcal{T})] \leq \epsilon$, where $\mathcal{T} \subseteq \mathcal{S}$ can be any dataset such as the retain or forget sets. By shifting the focus from model distributions to measurable performance indicators, this perspective makes certified unlearning more practical and extends its applicability to non-convex settings (Zhang et al., 2024).

3 The Residual Knowledge after Unlearning

Although approximate unlearning certifications, such as Rényi unlearning, ensure that the weight distributions of the unlearned and re-trained models are similar, they do not guarantee robustness against adversarial perturbations. In fact, several prior studies have explored the impact of adversarial attacks on unlearning. For example, Marchant et al. (2022) propose an adversarial attack designed to increase the computational cost of data removal. Pawelczyk et al. (2024) demonstrate that existing unlearning methods fail to completely remove forget samples after a carefully crafted poisoning attack. Additionally, Zhao et al. (2024) reveal that a small fraction of malicious unlearning requests can significantly reduce the adversarial robustness of the unlearned model.

The objective of this paper differs fundamentally from those previous works. The proposed phenomenon of residual knowledge explores yet another dimension of how adversarial examples impact unlearning. Specifically, a forget sample can be easily modified so that the unlearned model correctly classifies it even when a re-trained model fails to do so. This suggests that the unlearning process is vulnerable, is susceptible to manipulation, and introduces new privacy risks. In other words, residual knowledge is closely related to the transferability of adversarial examples (Tramèr et al., 2017) between the unlearned and re-trained models.

We formalize the mathematical definition of residual knowledge in the context of unlearning. For a forget sample $\mathbf{x}_f$, let $\mathcal{B}_p(\mathbf{x}_f, \tau) \triangleq \{\mathbf{x} \in \mathbf{R}^d; \|\mathbf{x} - \mathbf{x}_f\|_p \leq \tau\}$ represent the set of all possible perturbations of $\mathbf{x}_f$ within an ℓ_p-ball of radius $\tau \geq 0$. Let $m \sim M(h, \mathcal{S}, \mathcal{S}_f)$ denote an unlearned model and $a \sim A(\mathcal{S}_r)$ a re-trained model. We define the following non-negative ratio to quantify the residual knowledge around $\mathbf{x}_f$:

$$k((\mathbf{x}_f, y_f), m, a, \tau) \triangleq \frac{\Pr[m(\mathbf{x}'_f) = y_f; \mathbf{x}'_f \in \mathcal{B}_p(\mathbf{x}_f, \tau)]}{\Pr[a(\mathbf{x}'_f) = y_f; \mathbf{x}'_f \in \mathcal{B}_p(\mathbf{x}_f, \tau)]}. \tag{1}$$

If $k((\mathbf{x}_f, y_f), m, a, \tau) > 1$, we say that the unlearned model m suffers from residual knowledge of a forget sample $\mathbf{x}_f$, as it is more likely to correctly classify the vicinity $\mathcal{B}_p(\mathbf{x}_f, \tau)$ of the forget sample than a re-trained model. Conversely, if $k((\mathbf{x}_f, y_f), m, a, \tau) < 1$, it indicates that m excessively unlearn the forget sample, potentially leaking information about the training sample that was removed. We consider the unlearned model to have τ^*-robust unlearning if there

Table 1: The accuracy and ℓ_2-distance of unlearned models. The ℓ_2-distance between the re-trained model and each unlearning baseline implies that these unlearned models are similar to the re-trained model.

Metrics	Original	Re-train	NTK	Kodge	GA	GD	NegGrad+
Forget Accuracy	0.9800	0.8000	0.8400	0.4100	0.5800	0.9200	0.7700
Retain Accuracy	0.9989	0.9989	1.0000	0.9100	0.9244	1.0000	0.9344
Test Accuracy	0.9350	0.9350	0.9360	0.8550	0.9020	0.9510	0.9260
ℓ_2-distance	-	-	0.1628	0.1749	0.1612	0.1428	0.1608

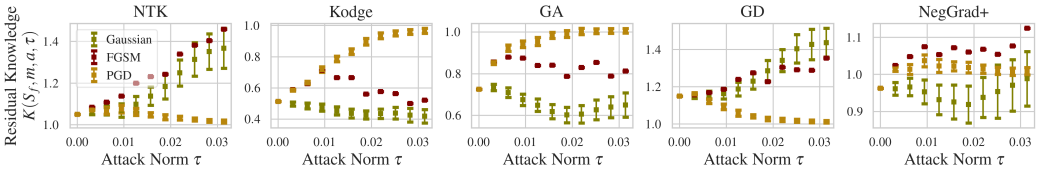

Figure 2: Residual knowledge in unlearned models varies significantly. NTK and GD exhibit substantial residual knowledge, while Kodge and GA excessively unlearn the forgotten samples. NegGrad+ is the most effective, consistently achieving $K(\mathcal{S}_f, m, a, \tau) \approx 1$ across different τ values under PGD.

exists $\tau^* = \inf_{\tau \geq 0} k((\mathbf{x}_f, y_f), m, a, \tau) = 1$, as this indicates that the unlearned and re-trained models perform similarly not only on the forget sample but also in its vicinity of τ^*. The definition in equation 1 can be readily generalized to a whole forget set $\mathcal{S}_f$ by $K(\mathcal{S}_f, m, a, \tau) \triangleq 1/|\mathcal{S}_f| \sum_{(\mathbf{x}_f, y_f) \in \mathcal{S}_f} k((\mathbf{x}_f, y_f), m, a, \tau)$.

We demonstrate the presence of residual knowledge using a smaller subset of the CIFAR-10 dataset (Krizhevsky et al., 2009), following a setup similar to that of Golatkar et al. (2020a). Specifically, we randomly sample 250 training images and 200 test images from each of the first five classes in CIFAR-10, with the validation set comprising 20% of the training data. We select 100 samples from class 0 to form the forget set $\mathcal{S}_f$. For our experiments, we use ResNet-18 (He et al., 2016) as the hypothesis space. As baselines, we adopt widely used unlearning methods that represent different unlearning strategies (cf. §2): NTK (Golatkar et al., 2020b), Kodge (Kodge et al., 2024), GA (Graves et al., 2021), GD (Neel et al., 2021), and NegGrad+ (Kurmanji et al., 2024).

Table 1 summarizes the performance of the original model (prior to unlearning), the re-trained model, and the unlearning baselines. We assess the accuracy of these models on the forget set, the retain set, and a hold-out test set not used during training. Most baselines demonstrate competitive accuracy on the retain and test sets compared to the re-trained model. However, some baselines, such as Kodge and GA, show reduced performance on the forget set.

Importantly, we find that some unlearning algorithms are highly susceptible to residual knowledge, where information in the vicinity of a forget sample can be exploited by adversarial perturbations (Kim, 2020). In Figure 2, we present $K(\mathcal{S}_f, m, a, \tau)$ of residual knowledge across varying attack norm τ. We empirically approximate the probability $\Pr[m(\mathbf{x}'_f) = y_f; \mathbf{x}'_f \in \mathcal{B}_p(\mathbf{x}_f, \tau)]$ using 100 adversarial examples $\mathbf{x}'_f$ generated by three different methods: injecting Gaussian noise ($p = 2$), FGSM ($p = \infty$) (Goodfellow et al., 2014), and PGD ($p = \infty$) (Madry et al., 2017). NTK achieves accuracy levels that closely match those of the re-trained model, as shown in Table 1. However, it retains a significant amount of residual knowledge, especially when $\mathbf{x}'_f$ is generated using Gaussian noise or FGSM. Similarly, GD preserves residual knowledge because it fails to effectively unlearn the forgotten sample initially. In contrast, Kodge and GA excessively unlearn the forgotten samples, achieving $K(\mathcal{S}_f, m, a, \tau^*) = 1$ only at $\tau^* = 0.03$ and $\tau^* = 0.02$, respectively. The most ideal case of unlearning is demonstrated by NegGrad+, which maintains a nearly stable $K(\mathcal{S}_f, m, a, \tau) \approx 1$ across varying attack strengths τ with PGD, effectively achieving $\tau^* \approx 0$.

Final remark. Residual knowledge in unlearning poses a potential privacy risk, warranting further research in several directions. First, the presence and extent of residual knowledge should be validated across a broader range of datasets and existing unlearning algorithms. Second, it would be valuable to define and measure residual knowledge in a manner analogous to differential privacy, along with conducting a theoretical analysis of this concept. Finally, developing an unlearning algorithm that ensures τ-robust unlearning remains an important challenge to address.

REFERENCES

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. *arXiv preprint arXiv:2401.10371*, 2024.

Rishav Chourasia and Neil Shah. Forget unlearning: Towards true data-deletion in machine learning. In *International Conference on Machine Learning*, pp. 6028–6073. PMLR, 2023.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020a.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pp. 383–398. Springer, 2020b.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

Sangamesh Kodge, Gobinda Saha, and Kaushik Roy. Deep unlearning: Fast and efficient gradient-free class forgetting. *Transactions on Machine Learning Research*, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images (technical report). *Universityof Toronto*, 2009.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in neural information processing systems*, 2024.

Guihong Li, Hsiang Hsu, Radu Marculescu, et al. Machine unlearning for image-to-image generative models. In *Internation Conference on Learning Representations (ICLR)*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050(9), 2017.

Neil G Marchant, Benjamin IP Rubinstein, and Scott Alfeld. Hard to forget: Poisoning attacks on certified machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7691–7700, 2022.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036, 2020.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Martin Pawelczyk, Jimmy Z Di, Yiwei Lu, Gautam Kamath, Ayush Sekhari, and Seth Neel. Machine unlearning fails to remove data poisoning attacks. *arXiv preprint arXiv:2406.17216*, 2024.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Supreeth Shastri, Melissa Wasserman, and Vijay Chidambaram. The seven sins of {Personal-Data} processing systems under {GDPR}. In *11th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 19)*, 2019.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), aug 2023.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards certified unlearning for deep neural networks. *arXiv preprint arXiv:2408.00920*, 2024.

Chenxu Zhao, Wei Qian, Yangyi Li, Wang Li, and Mengdi Huai. Rethinking adversarial robustness in the context of the right to be forgotten. In *International Conference on Machine Learning (ICML)*, 2024.