

This paper presents a foundational advancement in the science of cybersecurity by addressing the critical problem of machine unlearning: the ability to efficiently and verifiably remove the influence of specific data from trained machine learning models. As AI systems are increasingly deployed in security-sensitive and mission-critical environments, the absence of principled mechanisms for data removal introduces significant risks related to privacy, data integrity, and adversarial manipulation.

The authors propose “Split Unlearning,” a novel framework that decomposes model training into separable components, enabling precise and efficient removal of targeted data contributions without requiring full model retraining. This approach transforms unlearning from an ad hoc and computationally expensive process into a tractable, measurable, and reproducible capability. The framework is supported by rigorous algorithmic design and comprehensive empirical evaluation across diverse datasets and model architectures.

Importantly, the framework enables measurable evaluation through metrics such as model accuracy degradation, quantification of data influence, and computational efficiency of unlearning operations. These properties establish machine unlearning as a scientifically grounded and reproducible discipline, advancing cybersecurity toward a more rigorous and quantifiable foundation.

The contribution is highly generalizable, applying across a wide range of machine learning systems and deployment contexts. This generality ensures relevance not only to current AI architectures but also to future systems, making the work broadly impactful across the cybersecurity landscape.

From a practical perspective, Split Unlearning addresses urgent real-world challenges, including compliance with data privacy regulations, mitigation of data poisoning attacks, and secure lifecycle management of AI systems. This capability is particularly critical for national security applications, where sensitive or compromised data must be rapidly and verifiably removed without degrading operational performance.

The paper bridging theoretical rigor with operational applicability, this work advances both the science and practice of cybersecurity. It introduces a new paradigm in which data influence within AI systems can be precisely controlled, audited, and reversed.

The idea is establishing machine unlearning as a fundamental security primitive that is on par with encryption and access control. This paper defines a new foundation for trustworthy AI systems. Its scientific rigor, generalizability, and immediate operational relevance make it an exceptionally strong candidate for the NSA Best Scientific Cybersecurity Paper Award.