A Marauder's Map of Security and Privacy in Machine Learning

An overview of current and future research directions for making machine learning secure and private

Nicolas Papernot
Google Brain
papernot@google.com

ABSTRACT

There is growing recognition that machine learning (ML) exposes new security and privacy vulnerabilities in software systems, yet the technical community's understanding of the nature and extent of these vulnerabilities remains limited but expanding. In this talk, we explore the threat model space of ML algorithms through the lens of Saltzer and Schroeder's principles for the design of secure computer systems. This characterization of the threat space prompts an investigation of current and future research directions. We structure our discussion around three of these directions, which we believe are likely to lead to significant progress. The first seeks to design mechanisms for assembling reliable records of compromise that would help understand the degree to which vulnerabilities are exploited by adversaries, as well as favor psychological acceptability of machine learning applications. The second encompasses a spectrum of approaches to input verification and mediation, which is a prerequisite to enable fail-safe defaults in machine learning systems. The third pursues formal frameworks for security and privacy in machine learning, which we argue should strive to align machine learning goals such as generalization with security and privacy desirata like robustness or privacy. Key insights resulting from these three directions pursued both in the ML and security communities are identified and the effectiveness of approaches are related to structural elements of ML algorithms and the data used to train them. We conclude by systematizing best practices in our growing community.

CCS CONCEPTS

• **Security and privacy → Systems security**;

KEYWORDS

machine learning, security, privacy

ACM Reference Format:
Nicolas Papernot. 2018. A Marauder's Map of, Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private. In *11th ACM Workshop on Artificial Intelligence and Security (AISec '18), October 19, 2018, Toronto, ON, Canada.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3270101.3270102

AISec '18, October 19, 2018, Toronto, ON, Canada

ACM ISBN 978-1-4503-6004-3/18/10.
https://doi.org/10.1145/3270101.3270102