Thesis Project Portfolio

POISONING ATTACKS AND SUBPOPULATION SUSCEPTIBILITY

(Technical Report)

ALGORITHMIC BIAS AND DISCRIMINATION: AN INTERDISCIPLINARY PERSPECTIVE

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Evan Rose

Spring, 2023

Department of Computer Science

TABLE OF CONTENTS

SOCIOTECHNICAL SYNTHESIS

Machine learning has achieved remarkable successes over the past decade, solving problems once thought to be impossible. Autonomous vehicles are able to navigate busy roads with the confidence of experienced drivers, AI agents can defeat the world's best players in complex strategy-based games, and complicated protein structures can be understood with unprecedented speed and accuracy. Recently, machine learning-powered AI systems have even surpassed such impressive technical milestones that their existence has reignited a conversation about the very meaning of intelligence.

As exciting as these accomplishments are, they come with drawbacks. The development of these tools often relies on the large-scale collection of unfiltered, untrustworthy data used for training, and these data may contain various biases resulting from historical and societal prejudices. If these biases are not properly addressed, they can be amplified and propagated by the machine learning tools which learn from them. As developers of tools which can have major impacts on individuals' life opportunities, we must reflect carefully on what it means for AI systems to exhibit bias, as well as what responsibility, if any, we have in accounting for it. In the same vein, the uncareful circumstances of machine learning development makes machine learning systems a prime target for malicious actors who may leverage their control over the system to influence its behavior. For example, collection of training data from public-facing sources presents an attack surface for malicious actors to influence the resulting tool by publishing their own malicious data on which a system may be trained. With so much at stake, it is more important now than ever to consider deeply the full set of ramifications of using such powerful technologies. Thus, the broad purpose of this research is to study ways to make machine learning and AI technologies safer, fairer, and more reliable.

The technical project studies data poisoning attacks, a type of attack in which an adversary influences the behavior of a machine learning system by interfering with the data used to train it. In this work, we study poisoning attacks against subpopulations of a data distribution. Several poisoning attack experiments are designed against both synthetic and real datasets in order to understand what factors contribute to attack difficulty. We show that the susceptibility of a machine learning system to subpopulation poisoning attacks varies based on the properties of the dataset and the targeted subpopulation. By leveraging visualization techniques, poisoning attacks against both the low-dimensional and high-dimensional classifiers can be also be visualized, creating a rich depiction of the attack process and elucidating the factors contributing to attack difficulty.

The STS research studies algorithmic bias and discrimination, especially as it arises in the context of algorithmic decision-making. Social and technical interactions between algorithmic systems and the social groups the affect gives rise to a cohesive network of sociotechnical relationships by which the development of algorithmic technologies can be understood. The Social Construction of Technology (SCOT) framework is used to formulate the relationships between social groups and algorithmic decision-making technology. After describing each social group's common meaning with respect to the technology, we explore goals, requirements, and challenges regarding the technology.