Project Summary: CAREER: Securing Large-Scale Algorithmic Decision-Making Systems

Thanks to the availability of massive training data and the abrupt advances in machine learning techniques, algorithmic decision-making systems (ADMSes) are playing increasingly vital roles in our everyday lives. However, we still lack understanding about their security properties, which is highly concerning given their increasing use in security-critical domains (e.g., financial services, driverless vehicles, healthcare diagnosis). The goal of this proposal is to explore the potential security breaches of ADMSes and to develop rigorous yet practical solutions to mitigate such vulnerabilities. Specifically, the PI plans to investigate ADMSes' vulnerabilities to two general types of attacks: (i) adversarial inputs, which are malicious samples crafted by adversaries to trigger the systems to misbehave, and (ii) adversarial modules, which are malicious components incorporated as building blocks of ADMSes. In both cases, the PI plans to develop rigorous yet practical mitigation solutions. For adversarial input attacks, the PI aims at universal, attack-agnostic defense mechanisms that work effectively against unseen attack strategies, complement existing defense solutions, and provide comprehensive diagnosis information about the potential risks in classification outputs. For adversarial module attacks, the PI aims at scalable mechanisms that detect and repair infected modules *in situ*. Especially, the PI plans to explore modularized remedies that can be applied as patches to other ADMSes affected by similar attacks.

Intellectual Merit

The PI proposes to empirically study a range of real ADMSes deployed in security-critical domains to deepen our understanding of the security vulnerabilities inherent in ADMSes. The PI further proposes to develop a set of technologies and engineering systems which defend ADMSes against adversarial attacks without compromising their desirable predictive power. The proposed designs, at a high level, follow two general principles: (i) developing a rigorous analytical framework to formulate the invariant properties of attacks, and (ii) designing systems that leverage such analytical results to develop universal, attack-agnostic defense solutions effective for unseen attack variants and complement existing defense solutions. The application of this general approach leads to the following specific intellectual merits of the proposed research: (i) Insightful understanding of the security breaches of ADMSes to adversarial attacks; (ii) Significant theoretical advances in the invariant properties of adversarial attacks; (iii) Technical innovations of enforcing universal, attack-agnostic protection for ADMSes against malicious attacks following varied attack strategies; (iv) Development of a system testbed which enables automated investigations of attacks and defense strategies within heterogeneous contexts.

Broader Impacts

ADMSes hold great promise to transform the way people live, work, and travel. Yet, to fully realize such promise, their potential security vulnerabilities must be carefully shielded. The transformative nature of the proposed research is to completely rethink the way we defend ADMSes against adversarial attacks, so that (i) the defense is effective against unseen attack variants, (ii) the predictive power of ADMSes is intact, and (iii) the system operator is fully informed of the potential risks in the decision-making. The proposed project will help enable this vision through a unified research program consisting of the analytical formulations of vulnerabilities, attacks, and mitigations as well as the exploration and development of practical system prototypes. The success of the proposed research will not only significantly deepen our understanding of the security properties of ADMSes, but also greatly deepen our understanding on designing and implementing robust ADMSes for a wide range of domains. The results of this project will be applicable to areas including security, machine learning, and systems. The PI's education plan will employ the research results to promote engineering education. The PI will also continue to recruit and mentor female and under-represented minority students for this research. The PI plans to enhance the research experiences for undergraduates and K-12 students by actively involving them into the educational and outreach programs at Lehigh University. Finally, with the widespread use of machine learning (ML) in many sectors of our society, the security of ML-powered systems is becoming a societal issue. Using the research results from this project, the PI aims at making the general public more aware of the intricacies of ML security.