Trustworthy Autonomous Systems

Principal Investigator: Lalana Kagal, MIT

Brief statement of problem

Autonomous systems such as self-driving cars, automated surgical assistants, and unmanned aircrafts have begun making decisions previously entrusted to humans. As systems become more autonomous and complex, it not only increases the number of ways for them to fail, but also the number of ways they can be infiltrated. Cyberattacks against autonomous systems can have catastrophic consequences – they can destroy or disable critical infrastructure, disrupt service, and even steal sensitive classified information for military advantage. The potential for misusing these autonomous systems as weapons is also high. In 2011, researchers from the University of Washington and University of California at San Diego gained remote access to passenger vehicles by exploiting software vulnerabilities in General Motors' OnStar and Bluetooth systems, and were able to take physical control over the vehicle, such as controlling the display on the speedometer, shutting off the engine, and controlling the brakes. It is essential to ensure that these high stakes systems perform reliably and that their decisions are trustworthy even in the presence of anomalies and attacks. This proposal is focused on developing methodologies that will enable the deployment of robust autonomous systems, which are resilient as well as provably trustworthy.

Proposed approach to solve the problem

With the goal toward developing trustworthy intelligent autonomous systems, we will work on methodologies for anomaly detection, data provenance, explanability and protection of sensitive data.

(i) Anomaly detection: Complex systems work fairly well in practice, but when they fail, diagnosing the root-cause is difficult. Diagnostic systems have been successful in detecting errors for computers, spacecrafts, and cars, but these systems are static; they are not able to be augmented or easily re-configured to address and adapt to new errors and attacks and work on individual components. Diagnostic mechanisms for autonomous systems need to be able to absorb feedback, learn from their mistakes, and become better systems for future iterations. We will develop self-monitoring learning capability in order to improve robustness and diagnostics for identifying anomalous behavior of both individual components of an autonomous system as well as the system as a whole. Our work will combine commonsense reasoning [1] and machine learning to infer reasonableness of system decision or state.

(ii) Data provenance: Data provenance is an important part of being resilient as it helps systems understand whether the data is from a trusted source, whether it has been tampered with, whether it is accurate, how timely it is, etc. This requires that the data be tracked as it flows through the system and its changes be maintained in a log. Our focus will be on modeling data provenance using formal knowledge representation techniques [2] and designing protocols for provenance tracking and auditing.

(iii) Explanability: There is an appropriate lack of trust of autonomous systems, but this can be greatly reduced by giving these systems the ability to explain. Without explanations, how do we know if the systems are working in our best interest? Are the systems competent to do the jobs we assign them? When something goes wrong, as it inevitably will, how do we determine the reason for the problem? How do we assign blame, if that is necessary? How do we fix the problem so that the system will not make similar mistakes in the future? Will individual users and owners of these systems know enough about them to trust them? More so, developing an explanation of how and why they failed is even harder.

We will develop functionality to produce a detailed and verifiable explanation or proof. These explanations in conjunction with the data provenance and audit log can be used to verify the correctness of the system and ensure trustworthiness of the connected components. Cognitive autonomous systems are composed of several connected components including reasoners, machine learning, and programming blocks. In previous work we developed explanation mechanisms for reasoners [3, 4], in this project we will do so for machine learning by using rule extraction [5] and network dissection [6].

(iv) Protecting sensitive data: Information is the new weapon in cyberwars [7]. Even though a certain amount of openness and transparency is required for ensuring trust, sensitive data needs to be protected. In this project, we will study how to enable computation [8] and explanation over sensitive data without revealing the data itself. We will investigate how techniques such as zero knowledge proofs and differential privacy may be used in conjunction with detailed explanations to verify the decision/state of the system while maintaining privacy. An important aspect to make sure that this privacy does not inhibit cyber attribution. Dr. Bhargava is looking into cyber attribution, which provides insights into the source of the AI models and training sets and determining whether or not they can be trusted or are perhaps tainted by an adversary. We will collaborate with Dr. Bhargava's team to ensure that our privacy mechanisms do not hinder cyber attribution.

Benefit of the proposed approach over other possible and existing approaches to the problem

To the best of our knowledge, there is no other holistic aproach to trustworthiness of autonomous systems. The focus of existing approaches is identifying cyberattacks or providing diagnostic systems for anomalies and failures. Trust, however, is more complex than cybersecurity and includes proving that the system is behaving appropriately and that its data is not compromised. Self-monitoring constructs, like the one proposed in this proposal, along with data provenance and privacy are the way towards developing autonomous systems that are more trustworthy.

Approximate level of effort and budget, highlighting any additional resources expected above and beyond labor

The project will involve one PI and two Ph. D students. Budget will consist of $7,358 salary for faculty, and $78,443 salary for Ph. D students. The total budget including fringe benefits, tuition, and MIT overhead will be limited to approximately $200,000.

References

[1] L.H. Gilpin. Reasonableness Monitors. The 23rd AAAI/SIGAI Doctoral Consortium (DC) at AAAI- 18.

[2] O Seneviratne and L. Kagal, Enabling Privacy Through Transparency, IEEE Privacy Security and Trust 2012.

[3] A. Khandelwal, I. Jacobi, L. Kagal, Linked Rules: Principles for Rule Reuse on the Web, Fifth International Conference on Web Reasoning and Rule Systems (RR), August 2011

[4] L. Kagal, C. Hanson and G. Sussman, Explanations for Policy Decisions via Dependency Tracking, IEEE Workshop on Policies for Distributed Systems and Networks, June 2008

[5] J. R. Zilke, E. L. Mencía, and F. Janssen, "DeepRED–rule extraction from deep neural networks," in International Conference on Discovery Science. Springer, 2016, pp. 457–473.

[6] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in Computer Vision and Pattern Recognition, 2017.

[7] Kelly McSweeney, Cyberwarfare: The Most Stealthy Weapon Is Information, http://now.northropgrumman.com/cyberwarfare-stealthy-weapon-information/

[8] A. Heifetz, V. Mugunthan and L. Kagal, "Shade: A differentially-private wrapper for enterprise big data," *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, 2017, pp. 1033-1042.