Penetration Testing Artificial Intelligence

by Simon Tjoa (St. Pölten UAS, Austria), Christina Buttinger (Austrian Armed Forces), Katharina Holzinger (Austrian Armed Forces) and Peter Kieseberg (St. Pölten UAS, Austria)

Securing complex systems is an important challenge, especially in critical systems. Artificial intelligence (AI), which is increasingly used in critical domains such as medical diagnosis, requires special treatment owing to the difficulties associated with explaining AI decisions. Currently, to perform an intensive security evaluation of systems that rely on AI, testers need to resort to black-box (penetration) testing.

In recent years, artificial intelligence (AI) has significantly changed the way we do business and research. Applications that previously seemed possible only in science fiction (e.g. personal assistants like Siri and Alexa) are now a reality. AI components are also becoming increasingly important in the automated decision-making routines behind many systems that are used in areas such as cancer research, open-source intelligence (OSINT) and intrusion detection.

However, there is one huge drawback that limits the use of this technology. Often it remains unclear what exactly deep neural networks or similar approaches have learned and whether the software can be trusted. For some applications, either substantial research is conducted to gain a deeper understanding of their inner workings, or a human is involved in the process to ensure valid operation (i.e. 'human-in-the-loop'). While this approach is feasible in many cases, e.g. the doctor-in-the-loop, many applications, especially those that concern decision-making in critical infrastructures, do not scale with a human in the loop, often due to their time-critical nature. Furthermore, many of these decision-making processes need to be based on large amounts of inferenced datasets, thus making manual analysis practically impossible. This greatly reduces the trust in the results derived from such systems. In addition, in some applications, such as self-driving vehicles, it is not possible to use explainable AI or human intervention. Therefore, it is crucial to use an attacker's mindset to test the robustness and trustworthiness of the artificial system – especially considering the large attack surface posed by these systems and the massive developments in adversarial machine learning [1]. Combined with the inability to explain results, a lot of damage could be caused by attackers manipulating intelligent systems for their own gain.

We propose a high-level concept of a systematic process to test AI systems within the data science lifecycle. This is mainly done by combining techniques from risk management (i.e. assessing the business risk, existing controls and the business case for an attacker), adversarial machine learning (i.e. evaluating the trustworthiness and robustness of the algorithm and trained abilities) and traditional penetration testing (i.e. evaluating the security of the implemented system, e.g. manipulation of sensor data).

Figure 1 gives an overview of the generic approach, focussing on the AI components. While standard penetration testing of the underlying platform is required to mitigate threats and security gaps on this level (the greyed-out part labelled 'platform' in Figure 1), this method extends the standard approaches to achieve certain tasks required for the AI components. The main problem with AI components is explainability; it is usually not possible to gain a detailed understanding of why a certain decision was made [2]. Thus, testers resort to black-box security testing, trying to generate unwanted results either by using completely random (fuzzied) input material or by using nearby or extreme values. When using algorithms that learn from past decisions, it is vitally important to attack the underlying knowledge. We must assume that an attacker might be in possession of parts of the underlying (training) data or even have a (black-box) environment running the algorithms in question. The latter would enable the attacker to run many executions using arbitrary data or fuzzied information, trying differential attacks and feeding specially structured information into

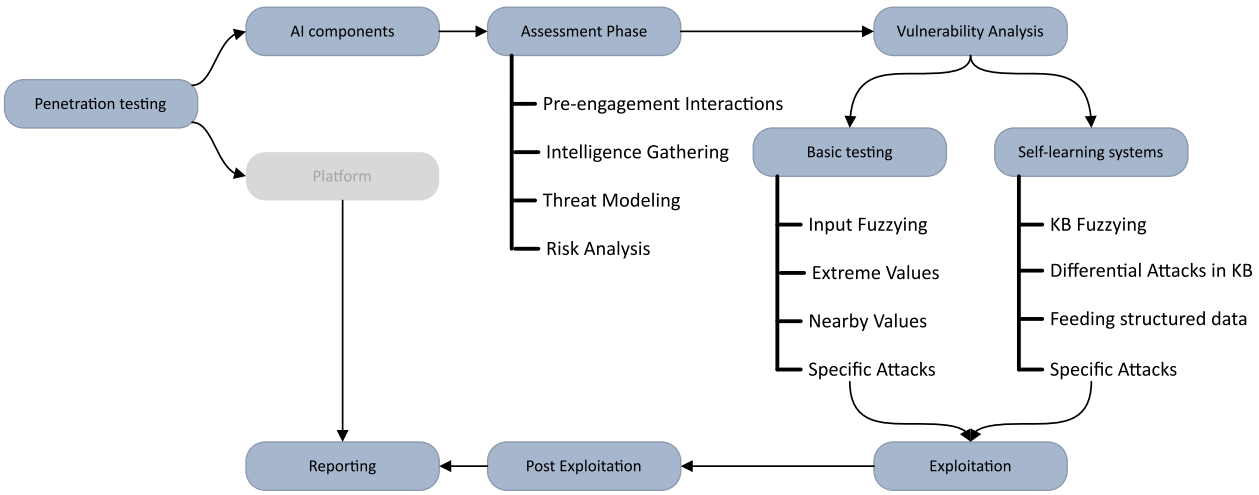

Figure 1: A high-level approach to penetration testing AI systems.

the system. This is very similar to the cryptoanalytic counterparts of partially known and chosen plaintext attacks. Furthermore, depending on the algorithms in use, specific attacks might exist that need to be considered during the penetration test.

While penetration testing is extremely valuable to evaluate such systems, proper risk analyses are often overlooked. These are important to: (i) carve out the attack surface, and (ii) help determine mitigation strategies and possible attack scenarios. Further research into possible attack scenarios is particularly important as the potential damage caused by manipulation of intelligent systems is often not clear even for the system's designers. Possible outcomes range from the introduction of broad bias into decision-making processes through to an attacker being able to launch fine-tuned attacks. Thus, together with identifying the (information) assets, the security analyst will also need to determine possible attack and damage scenarios in order to develop a feasible mitigation strategy.

The proposed workflow is at first draft stage and requires additional methods to tailor it to specific systems and the technologies. Nevertheless, it can be used as a template to provide a basic level of security in AI-based systems. Importantly, penetration testing can never give a security guarantee; at best, the testers will find all bugs that could have been found by attackers, but as history has shown, even very prominent software stacks can be susceptible to newfound or newly introduced errors [3]. We are investigating these issues in two academic projects, the COIN-project "Big-Data Analytics" [L1] and the FORTE-project "exploreAI" [L2]. In these projects we are conducting in-depth research into efficient penetration testing against intelligent systems and future implications for critical infrastructure security.

Links:

[L1] https://research.fhstp.ac.at/en/projects/big-data-analytics
[L2] https://kwz.me/h1Q

References:

[1] A. Chakraborty et al.: "Adversarial attacks and defences: A survey", arXiv preprint arXiv:1810.00069, 2018.
[2] K. Holzinger et al.: "Can we trust machine learning results? artificial intelligence in safety-critical decision support", ERCIM NEWS, (112), pp.42-43, 2018.
[3] Z. Durumeric et al.: "The matter of heartbleed", in proc. of the 2014 conference on internet measurement (pp. 475-488), 2014.

Please contact:

Simon Tjoa
University of Applied Sciences St. Pölten, Austria
simon.tjoa@fhstp.ac.at

Future Cyber-security Demands in Modern Agriculture

by Erwin Kristen, Reinhard Kloibhofer (AIT Austrian Institute of Technology, Vienna) and Vicente Hernández Díaz (Universidad Politécnica de Madrid)

The European agricultural sector is transforming from traditional, human labour-intensive work to data-oriented digital agriculture that has great potential for semi- or fully autonomous operation. This digital transformation offers many advantages, such as more precise fact-based decision making, optimised use of resources and big changes in organisation – but it also requires improved cyber-security and privacy data protection.

To feed the world's growing population and compensate for the loss of arable soil, the agricultural sector needs to increase efficiency, productivity and food quality, while simultaneously reducing labour costs and environmental impacts. The current approaches aim to use more powerful machines in the field, make these machines semi- or fully autonomous, and to plan precise fertilisation, irrigation, pest control and harvesting regimes based on detailed environmental data. The race for solutions has started: fields, crops and livestock are supplied with numerous sensors that monitor the environment. Machines are equipped with intelligent algorithms to perform their daily work with high precision and provide extensive operation status information, enabling a 24/7 availability. The agricultural system infrastructure, composed of numerous networked digital devices, is called Agriculture Internet of Things (AIoT).

The resulting bulk data can guide precise decision making on the farm and inform product development by machinery manufacturers. However, the colossal data gathering activities are very attractive for cyber-attacks, including theft, manipulation and misuse of data.

In the "AFarCloud" project, a group of European partners are working to implement the AIoT concept. We are currently developing an abstraction layer for AIoT-based architectures. This is the middleware that defines the software components and procedures, acting as an interface between the field layer and the cloud-based data processing layer where the farm management services are located (Figure 1).

The field layer includes the sensors, actuators, outdoor devices, vehicles and livestock. One important part of the middleware is the cross-layer cyber-security management (CSM) service, which handles the security maintenance process, providing a security process definition for periodic security assessment and security improvement recommendations. It facilitates a trouble-free, secure operation.

Cyber-security measures protect the production plant against attacks. In the early days of automation, only the information technology (IT) sector (farm management and middleware) was affected by cyber-security threats. The operational tech-