ARTIFICIAL INTELLIGENCE
IN IT SECURITY

MARTIN STEINEBACH AND MICHAEL WAIDNER
TECHNISCHE UNIVERSITÄT DARMSTADT

Expected developments of AI in IT Security

In IT security today, the usage of AI is already established in multiple domains. SPAM detection is a well-known example where support vector machines try to distinguish wanted from unwanted emails. Author attribution combines natural language forensics and machine learning. Deep learning helps in identifying illicit images and has improved malware detection as well as network intrusion detection.

Besides these applications, we can observe that the role of AI shifts from a tool used by IT security to a technology protected by it. This additional perspective is a common development: IT security is most often added to existing software that was developed without taking security into account. Many business solutions have embraced AI recently, often under pressure as users wanted to harvest the benefits of it as quickly as possible. This commonly leads to applications with high security risks due to bugs and error-prone routines. This includes the systems where AI training is executed and those where decisions are made based on the trained nets. Besides the hardening of code and thereby closing gaps for attacks, they require common methods of access control and rights management; AI systems do not differ from other Big Data or cloud systems here.

Not only the systems though, but also the AI algorithms and trained nets become the target of attacks. Various approaches try to mislead or influence AI -based decisions, requiring countermeasures of IT security protecting the core assets of AI.

As a third aspect, AI will also become a tool for attackers. IT security needs to be prepared for attacks that are capable of adapting more quickly to complex security measures, just like intrusion detection systems today aim to identify complex attacks with the help of AI.

Adversarial machine learning will become more common in IT security. Whenever a security challenge can be described by a relatively simple concept on the one hand, but can also be addressed by machine learning on the other hand, the other side, be it defender or attacker, will use adversarial machine learning to efficiently identify weaknesses in the strategy of the other party and deploy specialised attacks or defences against it.

The detection of SPAM, spear phishing or fake news as well as image attribution in security scenarios are examples of where this can be expected or is already common. For example, if an attacker wants to learn about the SPAM filtering capabilities of an online provider or a company, he can probe the solution with the help of a stolen or otherwise acquired email account within a certain perimeter of the target. AI can now send SPAM to the address and verify its receipt. The AI automatically modifies a blocked message until it passes through the SPAM filter. At this point, the attacker AI has beaten the SPAM filter and can now deliver the SPAM to its targets.

In the case of identity documents, an AI can take a passport photo of a person and modify it until it is recognised by a biometric system or any other trained AI. The attacker AI will eventually find the level of modification that is sufficient enough to fool the verification system with minimal visual changes to the photo.

It is most likely that the trend of these two examples will occur in most scenarios where AI has the role of a security checkpoint: the defender AI will become an oracle for attacker AIs trying to circumvent the defender. A race between defender and attacker AI will take place, similar to the obfuscation and recognition of malware or the hacking of systems and fixing of security issues.

Impact on Society

A prominent example of AI with a potential impact on society is predictive policing. While the actual performance and benefits are still being discussed and evaluated, it can be seen as a herald for future developments. The availability of data and the capability of complex analytics will lead to a wave of prediction approaches in a variety of domains.

In predictive policing, aspects like the influence of prejudice and biased data have already been discussed. Similar discussions have taken place on the topic of customer “scoring”, even before the rise of AI. The lesson learned from these discussions is the need for transparency if these technologies are to be widely accepted. Transparency not only addresses equations or algorithms as in scoring, but also the data used for training the AI. Data transparency is closely linked to the concept of privacy by design, which also includes the demand to inform data subjects and for methods to correct false data.

Interpreting the results of AI will be an even more important challenge in the future than it is already today. For example, while it is relatively simple to describe the architecture of deep learning, a common tool in AI, the trained net resulting from combining the architecture with tagged training data is often beyond human interpretation. Given that deep learning is still subjected to relatively high error rates depending on the training quality, an important decision based on an AI result should only be made with a second opinion coming from a human expert. In other words, it is important that AI is not seen as an incomprehensible decision engine, but as an assistant for human experts as long as AI results cannot be interpreted comprehensibly by the subject of the decision.

But at some point in the future the fundamental choice must be made regarding whether we allow algorithms to quickly decide upon important questions on their own based on the collected data. When the AI technology becomes ubiquitous, the sheer number of decisions made by it will render it impossible to execute effective human control. We see preliminary discussions about AI in cars, where its behaviour is questioned from an ethical perspective, for example in the case of an accident. Similar discussions will be necessary in multiple domains, since only a set of accepted rules will allow AI usage that is trusted by society. The role of IT security will then be to verify if the regulations have been implemented successfully, comparable with today's software and hardware penetration tests.

Need for action

Data collected today may (and most likely will) be used in unexpected ways tomorrow. As we see in the advancement of AI-assisted data analysis, it will be possible in the future to combine different data to breach the privacy of individuals up to a level beyond anything currently known and already criticised. Social media accounts, fitness tracker data, consumer behaviour and smart home data will be linked to advanced user profiles if the data is not sufficiently protected. Therefore, both data privacy and user awareness need to be improved.

As already mentioned above, AI will at some point require regulations about its behaviour to prohibit bias or discrimination. Interdisciplinary discussions about the nature and context of this regulated behaviour are necessary to design rule sets that are interpretable by a machine and that can be verified in the case of doubt.

Given the quick development of AI and the growth of its use cases, an interdisciplinary discussion should address these basic issues as soon as possible.

Otherwise, technical regulations will be implemented that are based on engineering concepts but do not consider ethical or legal aspects. This will considerably reduce public acceptance and will thereby hinder or slow down utilising the advantages that AI will offer. For example, it is obvious that the heavy amount of social media traffic cannot be effectively monitored by human observers when it comes to filtering out fake news or hate speech. AI can assist these human observers by identifying modified versions of already known and filtered content, or by pointing to the content that most likely needs moderation. This will raise the accusation of censorship . A political and legal discussion is therefore required on which role AI can take in the control of communication. An accepted trade-off between freedom and regulation is necessary before the actual technical solution evaluates social media traffic.

To summarise, we want to point out that machine learning is not a recent trend in computer security but an already established set of methods in some areas like spam detection. Still, recent advances in AI brings algorithms, which are able to significantly improve security solutions , to domains not addressed by AI so far. From a system security perspective, we will see the need to harden AI applications in the future, as the fast development of software using AI will inevitably bring numerous design and implementation risks with it. Nevertheless, the most important challenge will be the way in which AI is used in the future and which decisions it is allowed to make on its own. This is a question to be addressed by the whole of society; the tasks of IT security will be to provide means of protection, verification and privacy.

Further Reading

Chio, Clarence; Freeman, David (2018). Machine Learning and Security: Protecting Systems with Data and Algorithms. Sebastopol, CA: O'Reilly Media.

Barreno, Marco; Nelson, Blaine; Joseph, Anthony D.; Tygar, J. D. (2010). The security of machine learning. Machine Learning, 81(2), pp. 121-148.

Hitaj, Briland; Ateniese, Giuseppe ; Perez-Cruz, Fernando (2017). Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17). New York: ACM, pp. 603-618. Available at: https://dl.acm.org/citation.cfm?doid=3133956.3134012.

Shokri, Reza; Shmatikov, Vitaly (2015). Privacy-Preserving Deep Learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15). New York: ACM, pp. 1310-1321. Available at: https://dl.acm.org/citation.cfm?doid=2810103.2813687.

Montavon, Grégoire; Samek, Wojciech; Müller, Klaus-Robert (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, Vol. 73, pp 1-15.

Stone, Nathan; Ngoc, Tran Nguyen; Phai, Vu Dinh; Shi, Qi (2018). A Deep Learning Approach for Network Intrusion Detection System. IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 1, pp. 41-50.