Information security risks of artificial intelligence systems

Marjo Hanhikoski
Centria UAS
ISSUES-project funded by Interreg Aurora

Artificial intelligence has been a hot topic in recent years and its implementation has increased significantly. Artificial intelligence is harnessed to automate various tasks and processes. Many previously impossible things have become possible with the use of artificial intelligence. (Oseni et al. 2020.) Consequently, people are interested in the benefits of artificial intelligence and different organizations are looking for ways to utilize it in their own operations. However, alongside the benefits obtained, it should be remembered that artificial intelligence systems also contain vulnerabilities that expose them to cyber attacks and thus information security risks. Consequently, the information security of these systems should also be considered in the same way as ordinary information systems.

Artificial intelligence is defined as the operation of machines whose operation resembles human intelligence. There are several different ways to implement artificial intelligence, of which machine learning is the most common implementation method. (Finnish Transport and Communications Agency Traficom 2021, 1.) Artificial intelligence is currently divided into two different categories based on their functions: Predictive AI, such as machine vision, and Generative AI, such as ChatGPT (Vassilev et al. 2024, 3). Predictive artificial intelligence creates predictions, recommendations and decisions based on current and historical information. Generative AI creates content such as source code, music, text by learning from existing data models. (Lawton 2023.)

Regarding the information security risks of artificial intelligence systems, it should be noted that most of the information security risks they contain are the same as the information security risks of traditional systems and can be prevented with the information security solutions of traditional systems (Finnish Transport and Communications Agency Traficom 2021, 12, 21). However, artificial intelligence systems are also associated with information security risks specific to them, which are related to their special features and require information security solutions tailored to them. These risks are related to the data and models used by artificial intelligence systems (Finnish Transport and Communications Agency Traficom 2021, 12, 21; Vassilev et al. 2024, 8–9). In January 2024, NIST published a report regarding cyber attacks that can manipulate the operation of artificial intelligence systems. The report discusses four main types of attacks: evasion attacks, poisoning attacks, privacy attacks and abuse attacks. (NIST 2024.)

Evasion attacks do not actually aim to change the operation of the system, but the attacks take advantage of the system's vulnerabilities by using adversarial examples to create errors in the system (Oseni et al. 2020). In practice, this means that adversarial inputs are fed to the artificial intelligence system, the purpose of which is to make the system make mistakes (Finlayson et al. 2019). This kind of attacks target artificial intelligence systems in use and are based on the inputs given to the artificial intelligence system and the outputs of the system. Let's use a stop traffic sign as an example. In an evasion attack, the adversary could modify the stop sign by adding perturbations to it, in which case the robot car's artificial intelligence system would interpret the stop sign as, for example, a speed limit (Finnish Transport and Communications Agency Traficom 2021, 17–18).

Poisoning attacks are divided into data and model poisoning attacks. (Vassilev et al. 2024, 8–9.) In data poisoning attacks, the training data used by the artificial intelligence system is poisoned by either changing, adding, or removing its content. In model poisoning attacks, the adversary tries instead to change the learning algorithm of the artificial intelligence system. The goal of poisoning attacks is to disrupt the learning process of an artificial intelligence system, as a result the machine learning model becomes unreliable or unable to produce the output for which the artificial intelligence system was designed. (Oseni et al. 2020; Hu & Hu 2020, 629; Vassilev et al. 2024, 8–9.) The stop traffic sign can also be used as an illustrative example of a poisoning attack. An adversary could poison the training data of the robot car's artificial intelligence system by adding photos of stop signs, which the adversary has marked as speed limits, for example (Finnish Transport and Communications Agency Traficom 2021, 17). This kind of data poisoning could lead to the robot car interpreting the stop sign it sees as a speed limit.

Privacy attacks are attacks during the use of an artificial intelligence system that aim to gain access to sensitive data about the system itself or the training data used in its learning process for misuse (NIST 2024). The adversary collects information and based on the information, deduces the model of the artificial intelligence system or the content of its training data (Oseni et al. 2020). Privacy attacks include model stealing and training data inference attacks (Traficom 2021, 15–17; Oseni et al. 2020).

Model stealing attacks are either concrete model stealing or artificial intelligence system copying. The adversary tries to copy the model of the artificial intelligence system with input-output mapping. In practice, the adversary tries to copy the model of the artificial intelligence system by teaching its own copy to make the same decisions as the original system. This is done by making inputs into the original system and storing the received answers in the copy system. (Finnish Transport and Communications Agency Traficom 2021, 16.) This type of attack is usually used as a step to carry out other types of and more effective attacks (Vassilev et al. 2024, 32).

Training data inference attacks include model inversion, attribute inference, and member/membership inference attacks. Model inversion is very similar to model stealing attacks. However, the difference between them is that in model inversion, the adversary tries to re-model the training data of the artificial intelligence system using input-output mapping, with which the adversary tries to build a copy of the system's machine learning model. This type of attack is very damaging in terms of data protection because the training data of the artificial intelligence system can contain very sensitive data. (Traficom 2021, 16; Oseni et al. 2020; Vassilev et al. 2024, 29–30.)

Attribute inference attacks are a lighter form of attack than model inversion. In this attack, the adversary tries to deduce some features of the training data from the machine learning model of the artificial intelligence system. (Finnish Transport and Communications Agency Traficom

2021, 16.) In the member/membership inference attacks, the adversary tries instead to find out whether a certain data point belonged to the training data of the artificial intelligence system. A member/membership inference attack is a very damaging attack method from the point of view of data protection because it can be used to reveal information related to, for example, a single person. (Finnish Transport and Communications Agency Traficom 2021, 16; Oseni et al. 2020; Vassilev et al. 2024, 29–30.)

Abuse attacks only apply to generative AI. The goal of these attacks is to indirectly include incorrect information in data sources used by generative AI, such as websites or online publications. Abuse attacks differ from poisoning attacks in that the adversary includes their corrupted information in legitimate sources to manipulate the generative artificial intelligence to produce output suitable for the adversary's goals, such as disinformation, links to malware-infected web pages, redirecting the user to search results suitable for the adversary's isms, spreading malicious source code, or redirecting to fraudulent websites. (NIST 2024; Vassilev et al. 2024, 38, 47–48.)

Most of the attack methods presented above are relatively easy to implement and require little knowledge of artificial intelligence systems and the ability to execute the attack (NIST 2024). Attacks on artificial intelligence systems have been studied diligently in recent years from different perspectives, and based on the studies, several ways to secure these systems have been presented and the effectiveness of these security measures has been evaluated (Oseni et al. 2020). However, the developed security measures are still incomplete (NIST 2024). Consequently, organizations should be aware of the information security risks of artificial intelligence systems when considering the implementation of the system in their own operations.

REFERENCES

Finlayson, S., Bowers, J., Ito, J., Zittrain, J., Beam, A. & Kohane, I. 2019. Adversarial Attacks on medical machine learning: Emerging vulnerabilities demand new conversations. *Science,* 363(6433), 1287–1289. Available at: https://doi.org/10.1126/science.aaw4399. Retrieved March 6, 2024.

Finnish Transport and Communications Agency Traficom. 2021. *Tekoälyn soveltamisen kyberturvallisuus ja riskienhallinta.* Available at: https://www.traficom.fi/sites/default/files/media/publication/Teko%C3%A4lyn%20soveltamisen%20kyberturvallisuus%20ja%20riskienhallinta.pdf. Retrieved March 6, 2024.

Hu, C. & Hu, Y-H. 2020. Data Poisoning on Deep Learning Models. *2020 International Conference on Computational Science and Computational Intelligence (CSCI),* 628–632. Available at: https://doi.org/10.1109/CSCI51800.2020.00111. Retrieved March 6, 2024.

Lawton, G. 2023. *Generative AI vs. predictive AI: Understanding the differences.* TechTarget. Available at: https://www.techtarget.com/searchenterpriseai/tip/Generative-AI-vs-predictive-AI-Understanding-the-differences. Retrieved March 6, 2024.

NIST. 2024. *NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems.* Available at: https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems. Retrieved March 6, 2024.

Oseni, A., Moustafa, N., Janicke, H., Liu, P., Tari, Z. & Vasilakos, A. 2020. Security and Privacy for Artificial Intelligence: Opportunities and Challenges. *J. ACM,* 37(4). Available at: https://doi.org/10.48550/arXiv.2102.04661. Retrieved March 6, 2024.

Vassilev, A., Oprea, A., Fordyce, A. & Anderson, H. 2024. *Adversarial Machine Learning – A Taxonomy and Terminology of Attacks and Mitigations.* Gaithersburg: National Institute of Standards and Technology. Available at: https://doi.org/10.6028/NIST.AI.100-2e2023. Retrieved March 6, 2024.