Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

Secure, robust and transparent application of AI

Problems, measures and need for action

Bundesamt für Sicherheit in der Informationstechnik
Postfach 20 03 63
53133 Bonn
Tel.: +49 22899 9582- 0
E-Mail: ki-kontakt@bsi.bund.de
Internet: https://www.bsi.bund.de

1 Preface

Artificial intelligence (AI) methods perform very well in many applications, for instance in object detection in images, and are increasingly used in areas which influence our daily life. Some of these applications may have *critical consequences*, as for example (semi-)autonomous driving, face recognition or the analysis of medical data. Still, there are many *unsolved problems* with respect to the *secure, robust and transparent application* of AI.

This document presents *selected problems* as well as *measures* for such an application with regard to so-called connectionist AI methods and shows the *need for action* from the BSI's point of view. Ethical and legal questions are outside the scope of this document.

2 Definition and technical introduction

Based on the definition[1] of the *European Commission's High-Level Expert Group on AI*, the BSI understands the term *artificial intelligence* to mean the technology and scientific discipline that includes several approaches and techniques, such as machine learning, machine reasoning and robotics. AI systems are software and hardware systems that use artificial intelligence to act "rationally" in the physical or digital dimension. Based on perception and analysis of their environment, they act with some degree of autonomy to achieve specific goals.

So-called *connectionist AI methods* are based on models consisting of many simple and strongly interconnected processing elements, similar to neurons in human brains. The most well-known example is given by deep *neural networks*, which generally have millions of parameters. These parameters describe the properties of the neurons and their connections. The structure of such a network and the parameter values define the *AI model* and implicitly encode its possible reactions to inputs. The values are not determined manually, but automatically by means of *optimisation methods* using training data. The performance of such a method is quantified via *metrics* using test data. The metrics used and the properties of training and test data (quality and quantity) determine the functionality of the model substantially. In addition, models are often very *sensitive*, i. e., small changes to the input data can significantly affect their behaviour. This has severe consequences: Due to the implicit encoding of possible reactions to inputs, the functionality and outputs of such a model are often extremely difficult to comprehend, i. e., there is a *lack of transparency and explainability*. Because of this lack of transparency and explainability as well as the sensitivity, it is hard to verify the *robustness* of the model with respect to random or targeted perturbations. Therefore, formal guarantees can only be achieved to a limited extent. This gives rise to novel attack vectors, which are described in the next section.

[1] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

3 Novel attacks

Connectionist AI methods are vulnerable to qualitatively new attacks, which we will subsequently call *AI-specific attacks*. Some of these attacks can be carried out relying only on legitimate user-queries to the AI model. Currently, the most relevant AI-specific attacks are:

- **Evasion/Adversarial Attacks**: By manipulating the input data of the AI model during operation, attackers induce it to provide outputs not intended by the developer. The model itself is not changed in the process. Minor changes in input data, which may be hard to detect and are not immediately apparent even to humans, or are interpreted by them as irrelevant, can suffice to achieve significant effects.
- **Data Poisoning Attacks**: By manipulating the training data of the model, attackers induce it to react to (specific) inputs in a way not intended by the developer. Due to the large amount of data and the lack of transparency, these attacks are generally hard to detect.
- **Privacy Attacks**[2]: Attackers extract information concerning training data from the model. *Model inversion attacks* extract training data and *membership inference attacks* determine whether a data item has been used for training.
- **Model Stealing Attacks**: Attackers extract the functionality of the model. In doing so, they extract information on the model structure, e.g. relevant decision parameters, or (approximately) copy the functionality of the model under attack. The goal of these attacks is to steal intellectual property or to prepare other attacks.

[2] The terms are not used consistently within the literature. Depending on the perspective, privacy attacks are considered a subset of model stealing attacks or an independent category.

4 Measures for increasing IT security

Countermeasures against the AI-specific attacks are actively being investigated. However, the measures known so far mostly provide only limited protection. For instance, all known measures against adversarial attacks can be circumvented by so-called *adaptive attacks*[3] according to the current state of research. Existing measures can nevertheless be useful, and can at least hinder the execution of attacks or lessen their impact. The question whether these improvements suffice must be assessed in context of the respective use case. The risk potential and the resulting need for protective measures exhibit a high variance between use cases, since the effects of malfunctions and successful attacks as well as the ambient conditions in operation may strongly differ between them.

Professional producers, providers and developers of connectionist AI systems should currently take into account the following points in order to guarantee a *minimum level* of security for the systems:

- The *classical measures concerning software and systems security* remain unchanged for AI systems and should be implemented[4]. These measures are, however, *not* sufficient on its own.
- Within the scope of an *AI-specific risk management*, the *whole lifecycle of the AI system* should be analysed systematically with respect to relevant risks, including the AI-specific attacks mentioned above. Based on this analysis, mitigation measures on the level of the AI system and further technical as well as organisational measures for changing the ambient conditions can be derived in a risk-based approach. For instance, the robustness to adversarial attacks can be improved using so-called *adversarial training*. It should be judged whether the efficacy of the measures is adequate for the given application scenario. In order to assess risks properly, it may be helpful to *carry out adaptive attacks* on one's own AI systems (red teaming) or to commission external parties to do so. The risk analysis should be *repeated* regularly to take into account the current state of research.
- The *metrics* used for evaluating the quality of AI models should *take into account* the *risk potential* of the respective application. Besides the accuracy on expected input data, other aspects, as e.g. *over-/underfitting*[5], *bias effects*[6] as well as the *robustness to random or targeted perturbations,* should be considered likewise. In an ideal case, different algorithms and model approaches are compared using different metrics in order to judge whether they are suitable for the respective application.
- *Sufficient quality and quantity of training, test* and operational data, if applicable, should be ensured by systematic tests and measures. It is recommended to introduce a *professional data management*. It should include data and model protection against manipulations, logging of changes and the ability to attribute each data item to its source. When using data and models from external sources, special care is needed. A decision on this issue must take into account the application-specific risk.
- In order to *detect* AI-specific *attacks* and to analyse security incidents, queries and access to the AI system should be *logged* in a suitable way. The logs should regularly be checked for anomalies. Processes for quickly reacting to security incidents during operation should be established.
- In order to address *changes in ambient conditions*, the correct functionality of AI systems should be tested *in regular intervals* using the corresponding metrics.
- The criticality with respect to the *lack of transparency and explainability* of connectionist models should be assessed in the context of the respective use case. In order to explain the results to a limited extent,

[3] These are attacks which are specifically adapted to the respective model and the defensive measures used.
[4] Concrete recommendations may for instance be taken from the BSI's IT-Grundschutz-Kompendium.
[5] Overfitting denotes excessive and underfitting insufficient adaptation of an AI model to the training data. Both negatively affect the quality of the model.
[6] These are a consequence of systematic errors within the training data used, which can e.g. contain certain correlations more often than this is really the case or than it is socially desirable.

one can consider using so-called *XAI methods*[7], while bearing in mind their current limitations. From the IT security point of view, it is preferable to use simple and transparent models rather than large and complex ones. It is recommended to check whether it is feasible to reduce the number of parameters or to use intrinsically interpretable AI models (e.g. decision trees) instead of or in combination with connectionist AI methods.

In addition, providers of AI systems should describe in *precise and comprehensible terms* under which *boundary conditions* the AI system exhibits which *functionality,* and what the *limitations* of the system are. Potential users should get access to this description in order to assess the use of the AI system for their use case.

[7] These are methods for interpreting more complex AI models.

5 Need for action and BSI activities

From the BSI's point of view, there is an urgent need to continue investigating the security of AI systems, in order to be able to make reliable statements about their security.

1. **Development of standards, technical guidelines, test criteria and methods**: There are currently no sufficiently applicable standards for reliably evaluating and technically checking the security of AI systems in critical application contexts (as they may arise, for instance, in the automobile and armaments industry, in biometrics, in healthcare as well as in the financial, IT, and telecommunications area). For less critical applications, standards for security are likewise missing (with few exceptions).
2. **Research on effective countermeasures against AI specific attacks**: The existing measures against these attacks are often not sufficient. In order to facilitate the secure and robust operation of AI systems, further countermeasures must be developed with a special focus on their usability.
3. **Research on methods for transparency and explainability**: The prevalent lack of explainability of AI systems significantly influences their IT security and hampers acceptance of the systems by users. Therefore, it is important to continue investigating methods for explainability with a focus on their usability.

The BSI participates in national and international committees and groups with the aim of strengthening the security of AI systems. In addition, the BSI is actively involved in the development of evaluation criteria and methods for AI applications in various domains, in particular in the areas of automotive and cloud services. In July 2020, the BSI authored a scientific paper[8] that treats selected aspects of this document in greater detail. In February 2021, the BSI published the AI Cloud Service Compliance Criteria Catalogue (AIC4)[9]. The AIC4 criteria define a baseline security level for cloud services utilizing AI and can be audited by an independent party. A corresponding audit report can, if issued and applied appropriately, support professional cloud customers to assess the security of the respective cloud service for their use case.

Details on the activities and publications of the BSI are available on the web page[10].

[8] https://doi.org/10.3389/fdata.2020.00023
[9] https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf
[10] https://www.bsi.bund.de/ki