Lernende Systeme

GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

Protecting AI systems, preventing misuse

White Paper

Beyerer, J. & Müller-Quade, J. et al.
Working Group Hostile-to-Life Environments
Working Group IT Security, Privacy, Legal and Ethical Framework

Executive Summary

Artificial intelligence (AI) is already being used in many areas of society, be it in the health and working sectors, in road traffic or in public spaces. Despite the many opportunities that AI technology bears, such as improved healthcare or an attractive, individualized workplace design, the potential for misuse of AI systems should be considered at an early stage. In this way, appropriate measures can be taken strategically to prevent misuse or to minimize potential risks form early on. Ultimately, this can also ensure and strengthen trust in the reliability and security of AI systems.

In this white paper, experts led by the Working Groups Hostile-to-Life Environments and IT Security, Privacy, Legal and Ethical Framework of Plattform Lernende Systeme focus on the misuse of AI systems and present suitable measures for effectively preventing misuse. The authors examine the topic primarily from a technological perspective. They recommend being unbiased about the vulnerabilities of AI technology in specific areas of application and at the same time keeping in mind possible motives and perspectives of perpetrators. Based on this, it is possible to derive the necessary protective measures that can prevent misuse when embedded in an overall strategy. In doing so, they provide a fundamental contribution to openly addressing the issue of misuse related to AI technologies. The theoretical considerations are substantiated and illustrated using realistic application scenarios, in which a "worst case" is contrasted to a "best case".

Misues – Theoretical considerations

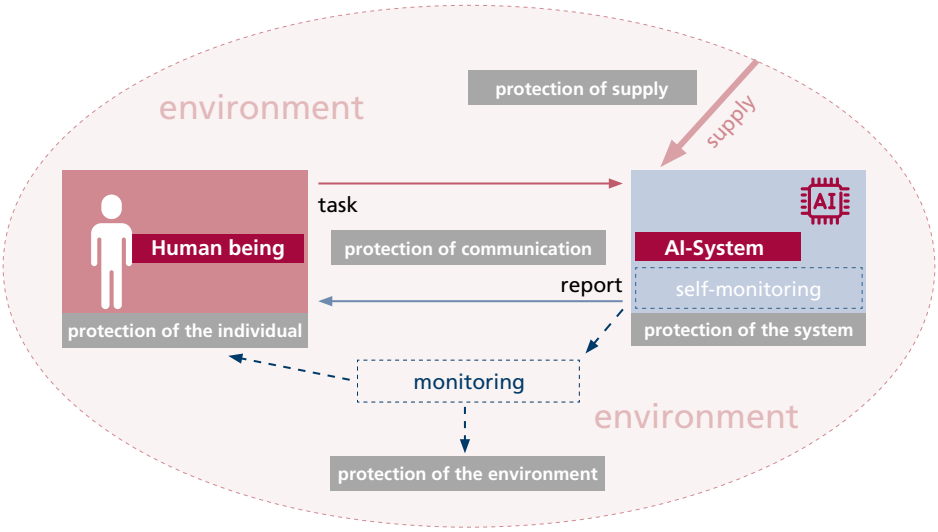
The authors first address the question of what the misuse of technologies is to create a basic understanding of misuse in the context of AI technologies (chapter 2.1). Misuse, and thus its execution, can be specified as "misappropriation with negative consequences" in which fundamental values, such as physical and psychological integrity, (democratic) freedoms and rights, privacy or material and immaterial values and the environment, are violated.

First and foremost, the following questions need to be answered to fundamentally address the topic of protection against misuse. What needs to be protected? Which attack scenarios are conceivable? Which protective measures are possible, adequate, and permissible? Based on the resulting answers and the subsequent analysis, specific technical and organizational protective measures can be derived. It is advisable to take a closer look at the various types of perpetrators, their motives, and the targets of their attacks. After all, misuse is triggered by human action, which can be driven by the most diverse motives of different actors, such as criminals or terrorists (chapter 2.2). It is possible to identify concrete protection goals within attack scenarios by changing the perspective, i.e., putting oneself in the position of different types of perpetrators and anticipating how they might try to misuse AI systems to their advantage or even use it as a weapon (chapter 2.3).

Protective and defensive measures

Since AI systems are always embedded in specific application contexts (see figure 1), the measures to prevent misuse should always start with the main protection targets – the AI system itself, the human being, and the environment.

General measures to protect AI systems against misuse



Source: Own presentation

Analyzing and identifying possible motives and targets for attacks on AI systems in dependence to the respective area of application provides cues to derive a wide variety of protective and defensive measures of a technical and organizational nature. On the one hand, these measures can be based on systems with artificial intelligence, such as anomaly detection, video and speech recognition, or identity recognition and on the other hand on systems without artificial intelligence (chapter 2.4). The technical measures can be aimed both at the AI system in use and at the technical infrastructure surrounding the AI system, which itself is not AI-based.

For the technical measures to be effective and reliable, they should be supplemented by organizational measures, such as defined rules and processes to be followed or testing and certification measures.

The preceding theoretical considerations on the misuse of AI systems and possible protective measures against misuse are illustrated realistically using seven application scenarios from areas such as health, leisure, mobility and the working environment (chapter 3). Among other things, this white paper describes how criminals use ransomware to extort ransom money with a phishing attack on a company. In the worst case scenario, they manage to infiltrate an email into the communication traffic to all employees, which they use to paralyze the entire accounting system. In the mobility scenario, a terrorist organization carries out an attack on a businesswoman's autonomous vehicle via a CAR-2X interface. In this way, they can intervene directly on the control electronics and steer the car in a targeted manner. A risk of misuse may also arise from the use of drones. Criminals manipulate drones at a World Cup soccer match via malware so that they can be misused as a "weapon" against visitors.

Scenarios in realistic application areas

The scenarios presented illustrate how and where misuse could already become reality today or in the years to come. They show which concrete measures are to be derived in each case in order to turn the "worst case" into a "best case" through suitable and effective protection and defence mechanisms or to prevent misuse in the first place.

The website of Plattform Lernende Systeme presents application scenarios that make the misuse of AI systems vivid and invite to interactively explore this topic. The scenarios each compare a "worst case" with a "best case" and show how appropriate measures can be taken to prevent misuse.

Recommendations

Against this backdrop, a number of concrete recommendations (chapter 4) can be formulated in addition to the aspects already mentioned, which aim to effectively prevent misuse of AI systems. Ultimately, an open societal discourse that focuses on transparency and trust without stoking fears about misuse is needed.

- **Policymakers** should clearly define responsibilities and claims settlements in the event of misuse via governmental measures. They should influence the development of and compliance with standards and certifications for AI systems with a high potential of misuse and promote research projects to identify the opportunities and potential risks of AI systems at an early stage in order to initiate preventative measures and minimize possible risks in good time.

- **Research institutions** should increasingly address the issue of misuse of AI systems, considering the requirements for security and reliability. They should emphasize this aspect in the qualification of specialist personnel, investigate published cases, and communicate appropriate measures transparently to the public.

- **Educational Institutions** should raise awareness about the misuse of AI systems and promote knowledge on AI and how to use it. It is important, to communicate this topic society-wide and to cover it in school and professional education and training from early on.

- **Companies**, also in cooperation with research institutions, are called upon to jointly develop and implement new solutions at an early stage, to assess possible targets of misuse, to ensure the safe and reliable use of AI systems and to take this into account in the continuous education of the staff.

- **Civil society** is called upon to get actively involved to clearly position the social and beneficial perspective about misuse.

By focusing on the topic of misuse of AI systems, this paper makes an important contribution not only to considering early and proactive protection, but also to ‚actually' implementing suitable measures to protect against attempts to misuse AI systems. The basic considerations presented as well as the concrete measures open up perspectives for action and also provide a framework for effectively preventing and counteracting the misuse of AI systems.

Imprint
Editor: Lernende Systeme – Germany's Platform for Artificial Intelligence | Managing Office | c/o acatech | Karolinenplatz 4 | D-80333 München | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de | Follow us on Twitter: @LernendeSysteme | Status: March 2022 | Photo credit: Peera_Sathawirawong/Adobe Stock/Title

This executive summary is based on the white paper *KI-Systeme schützen, Missbrauch verhindern – Maßnahmen und Szenarien in fünf Anwendungsgebieten*, Munich, 2022. The authors are members of the working group Hostile-to-Life Environments and of IT Security, Privacy, Legal and Ethical Framework of Plattform Lernende Systeme.
https://doi.org/10.48669/pls_2022-2

SPONSORED BY THE

Federal Ministry of Education and Research

acatech
NATIONAL ACADEMY OF SCIENCE AND ENGINEERING