A Technological Perspective on Misuse of Available AI

Lukas Pöhler[1,2], Valentin Schrader[1,2], Alexander Ladwein[1,2], Florian von Keller[1,3]

CONSCIOUSCODERS

Introduction

Malicious misuse of civilian AI poses serious threats to security on a national and international level. Besides defining **autonomous systems from a technological viewpoint**, we show how already existing and openly available AI technology could be misused. We highlight the **importance to discuss and mitigate threats from misused AI** in order to **ensure open civilian AI development in the future** without unilateralism.

What are Autonomous and Intelligent Systems?

General terms

Algorithms	• Procedure for solving a problem • Calculation rules
Machine Learning	• Data processing • Generalization, prediction
Artificial Intelligence	• Subjective: outside view ➢ no clear definition
Autonomy	• Unstructured environment • Without explicit control

Fig. 1: Overview of different terms in the field of AI

Schematics of an Autonomous System

It seems obvious to assume that autonomous systems based on AI are primarily defined by their algorithms. However, each system is a complex composition of **input data**, a **predefined goal**, the **underlying code**, and the **Interface** (both hardware or software) to interact with either the physical or digital world.

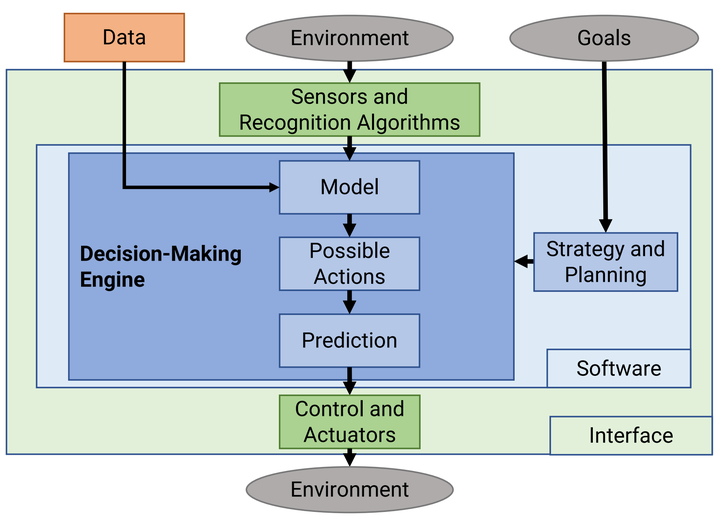

Fig. 2: Autonomous System

Malicous Use

Misuse of AI: Misuse of AI is the use of AI for applications that were not intended originally.

Malicious use of AI: Malicious use of AI technology to an end that threatens security.

Many innovations are based on misuse of open technology from a global community. In our work, we do not consider this benign misuse but explicitly focus on malicious misuse of civilian AI, because it threatens security and lacks attention in the current debate. Figure 3 visualizes the different modes of AI use, which define the scope of this paper.

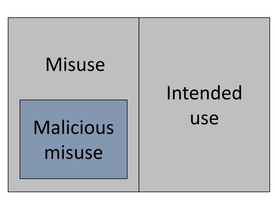

Fig. 3: Schematic of AI modes of use with malicious misuse as scope of this work

Openness

Degrees of Openness

Openness as a continuous scale: the level of openness can vary anywhere from a vague or abstract description to fully functional source code, trained models, detailed tutorials, files for 3D printing or full datasets. The usefulness of resources varies together with the openness level.

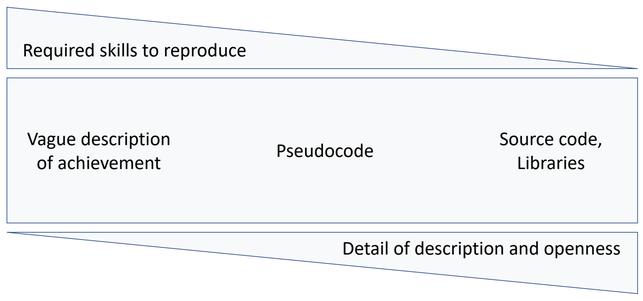

Fig. 4: Degrees of Openness for an Algorithm, taken from [2]

Platforms and Sources for Open Content

Large online platforms have been key for accessibility and usability of open resources. They provide content and also allow project management, collaboration and social interactions. Examples:

- SourceForge (`sourceforge.net`) and GitHub (`github.com`) for software
- Hackaday.io (`hackaday.io`) for hardware
- DBpedia (`wiki.dbpedia.org`) or Wikidata (`www.wikidata.org`) for data
- arXiv (`arxiv.org`) for scientific publications

Openness and AI

AI research and development is characterized by a high degree of openness, which enables widespread **access**, rapid **diffusion** and in the end an increasing **development speed**.

Threats

Aside from "killer robots", autonomous weapons in virtual environments are similarly threatening security and probably more available than their hardware equivalents. The threats from malicious AI can be assigned to three categories[2]:

Digital Security

- Elimination of trade-off between **scale** and **efficacy** of digital attacks
- Less **labor** intensive
- Autonomous systems as targets (vulnerabilities, data poisoning[1])

Fig. 5: Spear-Phishing

Political Security

- Automation of **surveillance**
- Persuasion through **propaganda** and **deception**
- Media **manipulation**
- Classification capabilities of machine learning and **analyze human behaviors**, mood and beliefs

Fig. 6: Deepfake

Physical Security

- **(Lethal) Autonomous Weapon Systems**
- Direct attacks by **AI-managed agents**
- **Subversion** of cyber-physical systems (critical infrastructure)

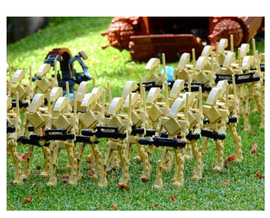

Fig. 7: Swarming

Implications of threats

Expansion of existing threats

Diffusion, efficiency, scalability of AI technology
⇒ Existing attacks become possible
- *for more actors*
- *on a wider scale*
- *on more targets*

⇒ **Power shift: Governments to non-state actors**

New attacks

Attacks that were impossible before, e.g. deepfakes
⇒ **Threat of political security**

Changed character of attacks

- Fine targeting
- Difficult attribution
- E.g. strategic swarm

⇒ **Lower barrier for engaging in violence**

Prevention of malicious misuse

Access Prevention through Points of Control

Malicious misuse of AI can be prevented by **restricting access** to and diffusion of AI functions that can be misused in a malicious way. A threefold approach could be followed here: classify AI system components as critical, require registration and sub-sampling of components.

Attack prevention

However, it is also important to **prevent attacks** of potentially malicious systems. This includes IT security and AI for attack prevention. Further, prohibition of functionalities and externally acessible emergency shutdowns can be imagined. The best attack prevention for digital security is clearly an informed society.

Non-technical measures

A number of **non-technical measures** such as international and interdisciplinary discussion, the integration of a variety of actors, permanent committees and most important, collaboration between academia, private sectors, states and the civil society are of high importance.

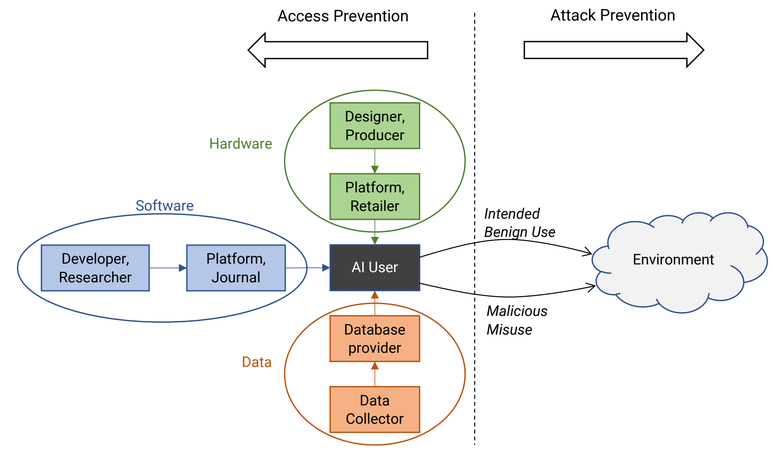

Fig. 8: Proliferation Chain of AI

References

[1] Blaine Nelson Battista Biggio and Pavel Laskov. *Poisoning Attacks against Support Vector Machines*. 2013. url: `https://arxiv.org/pdf/1206.6389.pdf`.

[2] Miles Brundage et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Tech. rep. 2018. arXiv: 1802.07228. url: `http://arxiv.org/abs/1802.07228`.

[1]ConsciousCoders (`www.conscioucoders.io`, ai.misuse@conscioucoders.io)
[2]Technical University of Munich, Germany
[3]Ludwig-Maximilians-Universität München, Germany

https://www.conscioucoders.io/download/paper-gge