# Cybersecurity for AI Systems: A Survey

Raghvinder S. Sangwan, Youakim Badr, Satish M. Srinivasan (2023-05-04)

November 10, 2025

- **Purpose** **(p. 166):** This research examines the landscape of these cyber attacks and organizes them into a taxonomy. It further explores potential defense mechanisms to counter such attacks and the use of these mechanisms early during the development life cycle to enhance the safety and security of artificial intelligence systems.
- 🟨 **(p. 166):** Therefore, it is important that we start thinking about designing security into AI systems, rather than retrofitting it as an afterthought. This research addresses the following research questions

## Literature Review (p. 167)

- 🟥 **(p. 167):** This survey was founded on searching, by keywords, to find related articles to cybersecurity of AI systems. The top most used keywords are as follow: cybersecurity, cyberattack, and vulnerabilities. We searched Scopus, an Elsevier abstracts and citation database, for articles having titles that matched the search query (“cyber security” OR “cybersecurity” OR “security” OR “cyberattack” OR “vulnerability” OR “vulnerabilities” OR “threat” OR “attack” OR “AI attack”) AND (“AI” OR “ML” OR “Artificial Intelligence” OR “Machine Learning”) AND (“system”)).
- 🟦 **(p. 172):**

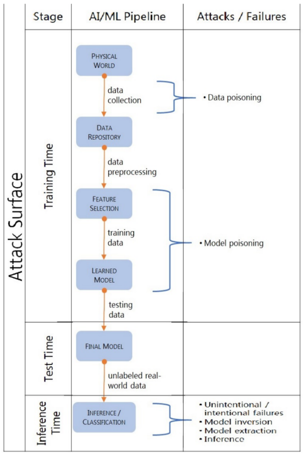

Figure 1: Anatomy of Cyberattacks

## Munoz-Gonzalez-illustrated poisoning attacks (p. 174)

- **Definition** **(p. 174):** Munoz-Gonzalez et al. [38] illustrated poisoning attacks on multi-class classification problems. The authors identified two attack scenarios for the multi-class problems: (1) errorgeneric poisoning attacks and (2) error-specific poisoning attacks. In the first scenario, the adversary attacks the bi-level optimization problem [40,82], where the surrogate data is segregated into training and validation sets. The model is learned on the generated surrogate training dataset with the tampered instances. The validation set measures the influence of the tampered instances on the original test set, by maximizing the binary class loss function.