AISEC

Secure and Controllable AI

Programm / Ausschreibung	KS 24/26, KS 24/26, BRIDGE 2025/01	**Status**	laufend
Projektstart	01.01.2026	**Projektende**	31.12.2028
Zeitraum	2026 - 2028	**Projektlaufzeit**	36 Monate
Projektförderung	€ 288.946		
Keywords	Security; AI; Threat Hunting; Proactive Measures		

Projektbeschreibung

AI wird zu einer allgegenwärtigen Technologie, die nicht nur durch Spezialist*innen eingesetzt, sondern in klassische Anwendungen und alltägliche Produkte integriert werden wird, implementiert durch Entwickler*innen ohne AI-Kenntnisse, oder sogar extern zugekauft. Gleichzeitig entstehen immer mehr Richtlinien und Gesetze, die den Einsatz von AI regeln, wie bspw. der AI-Act oder der Cyber Resilience Act. Diese verlangen ein hohes Maß an Kontrolle und (Verarbeitungs-) Transparenz von Inverkehrbringer*innen und Betreiber*Innen über ihre Systeme, sowie die Absicherung der Systeme gegen Cyber-Angriffe. Dabei geht der Trend immer weiter Richtung kontinuierlichem Monitoring. Abseits theoretischer Arbeiten existiert derzeit kein praktisch nutzbares Instrumentarium für moderne AI-basierte Systeme.
Im Projekt AISEC wird dieses benötigte Instrumentarium aufgebaut, geleitet von Use-Cases aus Wirtschaft und Anwendung. Das Hauptziel dabei:
„Erforschung von Methoden zur Kontrollierbarmachung und Resilienzerhöhung von AI-Systemen.“
Die Forschung im Projekt fokussiert sich dabei auf die praktische Umsetzung des Controllable-AI-Ansatzes basierend auf den folgenden drei Objectives:
• Aufbau einer dynamischen Wissensbasis zur Modellierung von Angriffen, deren Seiteneffekten und Maßnahmen.
• Erforschung von Indikatoren für Angriffe auf und Kompromittierung von AI-Systemen, nutzbar als Measures in Controllable AI.
• Transfer des theoretischen Ansatzes der Controllable AI in die Praxis mit Fokus auf Schaffung proaktiver Resilienz von der Datenanlieferung bis hin zur Security Operations.
Durch die hohe Entwicklungsdynamik des Forschungsumfelds, sowohl methodisch im Machine Learning, in Bezug auf neue Anwendungen, aber auch neue Angriffsmethoden werden zwei flankierende Maßnahmen benötigt, damit die Ergebnisse nicht veralten und immer in Bezug zur realen Anwendung stehen:
1. Der Stand der Technik, Regularien, aber auch AI-Security-Wissen werden nicht statisch abgelegt, bspw. in Form einer Taxonomie, sondern in einer dynamischen Knowledge-Base einfach erweiterbar vorgehalten.
2. Problemanalyse und Evaluierung von Ergebnissen erfolgt systematisch anhand realer Use-Cases des Wirtschaftspartners und der Forschung. Somit kann jederzeit auf neue Entwicklungen in Technologie, aber auch Regularien reagiert und diese möglichst frühzeitig einbezogen werden.

In Bezug auf die akademische Forschung ist das Projekt hochinnovativ und verfolgt viele neue Ansätze, wir erwarten daher auch eine sehr gute Reflexion innerhalb der AI- und Security-Communities und entsprechend gut zitierte Publikationen aus den Projektergebnissen. Da es sich um ein Projekt der Grundlagenforschung handelt, ist eine direkte Verwertung der Ergebnisse in der Form von Services und Produkten nicht möglich, lediglich der Transfer in die akademische Lehre ist ohne weitere große Investitionen möglich. Grundsätzlich gehen wir aber davon aus, dass eine Verwertung in marktfähiger Produkte und Services in einem Zeithorizont von 3-5 Jahren gut möglich ist.

Abstract

AI is becoming a ubiquitous technology that will not only be used by specialists, but will also be integrated into traditional applications and everyday products, implemented by developers without AI expertise or even bought in externally. At the same time, more and more guidelines and laws are being created that regulate the use of AI, such as the AI Act or the Cyber Resilience Act. These require a high degree of control and (processing) transparency from distributors and operators regarding their systems, as well as the protection of systems against cyber attacks. The trend is increasingly moving towards continuous monitoring. Apart from theoretical work, there are currently no practical tools for modern AI-based systems.
In the AISEC project, this necessary set of tools is being developed, guided by use cases from industry and application. The main objective:
"Research into methods for making AI systems controllable and increasing their resilience."
The research in the project focusses on the practical implementation of the Controllable AI approach based on the following three objectives:
• Development of a dynamic knowledge base for modelling attacks, their side effects and measures.
• Research into indicators for attacks on and compromise of AI systems, usable as measures in Controllable AI.
• Transfer of the theoretical approach of Controllable AI into practice with a focus on creating proactive resilience from data delivery to security operations.
Due to the highly dynamic development of the research environment, both methodologically in machine learning and in relation to new applications, but also new attack methods, two accompanying measures are required to ensure that the results do not become outdated and are always related to the real application:
1. the state of the art, regulations, but also AI security knowledge are not stored statically, e.g. in the form of a taxonomy, but are kept in a dynamic knowledge base that can be easily expanded.
2. problem analysis and evaluation of results is carried out systematically based on real use cases of the business partner and research. This allows us to react to new developments in technology and regulations at any time and incorporate them as early as possible.
In terms of academic research, the project is highly innovative and pursues many new approaches, so we also expect a very good reflection within the AI and security communities and correspondingly well-cited publications from the project results. As this is a basic research project, it is not possible to directly utilise the results in the form of services and products; only the transfer to academic teaching is possible without further major investment. In principle, however, we assume that utilisation in the form of marketable products and services is quite possible within a time horizon of 3-5 years.

Projektkoordinator

- Hochschule für Angewandte Wissenschaften St. Pölten GmbH

Projektpartner

- XSEC infosec GmbH