TruthGuard

Bekämpfung von Desinformation durch die Analyse und Detektion von Fake-News-Netzwerken

Programm / Ausschreibung	KIRAS, Kooperative F&E-Projekte, KIRAS-Kybernet-Pass CS Kooperative F&E Projekte (CS KFE_2023)	**Status**	laufend
Projektstart	01.01.2025	**Projektende**	31.12.2026
Zeitraum	2025 - 2026	**Projektlaufzeit**	24 Monate
Keywords	Fake News, Generative Künstliche Intelligenz, Bewusstseinsbildung		

Projektbeschreibung

Fake-News sind falsche oder irreführende Informationen, die bewusst verbreitet werden, um Meinungen zu manipulieren oder bestimmte Interessen zu fördern. Sie gefährden Demokratien, indem sie das Vertrauen in Medien und Institutionen untergraben und die öffentliche Meinung verzerrt beeinflussen. Im Jänner 2024 stellte das Bundesheer das Risikobild 2024 für Österreich vor und hielt darin fest, dass Desinformationskampagnen eine der größten Bedrohungen für die Sicherheit Österreichs als Teil hybrider Bedrohungen sind. Das Bedrohungspotenzial durch Desinformationskampagnen wird durch das Aufkommen von Bots und Fake-News-Netzwerken im Internet verschärft, die mit KI-Unterstützung sehr effektiv und mit wenig Aufwand erstellt werden können.

Obwohl Desinformationskampagnen ein massives Problem für die Zukunft darstellen, werden deren negativen Auswirkungen in der breiten Öffentlichkeit unterschätzt. Auch in der Forschung ist der Wissenstand um Fake-News-Netzwerke und deren technologische Umsetzung gering, da gerade im letzten Jahr massive technologische Fortschritte stattfanden, die solche Fake-News-Kampagnen durch automatisch erstellte Text-, Bild- und Videoinhalte, erst möglich machen.

Um Fake-News-Netzwerken entgegenzuwirken, wird im TruthGuard Projekt erstmals beide Seiten von Fake-News untersucht, nämlich sowohl die Generierung, als auch die Identifikation von Fake-News. Durch die beidseitige Untersuchung kann ein ganzeinheitliches System erforscht werden, dass sowohl State-of-the-Art Methoden verbessert als auch für bewusstseinsbildende Maßnahmen eingesetzt werden kann. Die Maßnahmen des Projekts werden durch einen Oversight Board mit externen Partnern (LoI_DZ, LoI_Bait) begleitet, das sich mit ethischen und rechtlichen Fragestellungen des Projekts befasst und dessen Überlegungen in die wissenschaftlichen und technischen Entwicklungen einfließen. Das Partner-KMU wird einen umfassenden Verwertungsplan erstellen und die wissenschaftlichen Ergebnisse werden international disseminiert (LoI_Athena). Die wissenschaftlichen und technischen Ziele von TruthGuard sind folgend aufgelistet:

Erforschung von Methoden zur wirksamen Regulierung generativer Modelle zur Bekämpfung von Fake-News, einschließlich der Identifizierung von Schwachstellen in der Umsetzung von Ethik-Richtlinien und der Sicherung der Integrität dieser Modelle.

Erforschung von neuen vollautomatischen Methoden zur Identifizierung von Fake-News bzw. Bots basierend auf Daten von verschiedenen sozialen Medien Netzwerken.

Integration von einem generierendem und einem identifizierenden Tool von Fake-News in einer Demonstrator-Anwendung (eigenes Mastodon Netzwerk) für spezifische Einsatzszenarien in einer TRL 4-Testumgebung, die mit dem Bedarfsträger als auch mit anderen Stakeholdern (LoI_BKA, Vertreter der Zivilbevölkerung) getestet werden. Die Szenarien wurden strukturiert mit wissenschaftlichen Frameworks am Anfang des Projektes ermittelt.

Untersuchung zielgruppenspezifischer Herausforderungen und plattformspezifischer Problemlagen, die zur Entwicklung von gezielten Maßnahmen zur Förderung des kritischen Umgangs mit Informationen führen, um damit die Widerstandsfähigkeit der Zivilbevölkerung gegen Desinformation zu stärken.

Abstract

Fake news is false or misleading information that is deliberately disseminated in order to manipulate opinions or promote certain interests. They jeopardise democracies by undermining trust in the media and institutions and distorting public opinion. In January 2024, the Austrian Armed Forces presented the 2024 risk picture for Austria, stating that disinformation campaigns are one of the greatest threats to Austria's security as part of hybrid threats. The potential threat posed by disinformation campaigns is exacerbated by the emergence of bots and fake news networks on the internet, which can be created very effectively and with little effort with AI support.

Although disinformation campaigns represent a massive problem for the future, their negative effects are underestimated by the general public. Even in research, the level of knowledge about fake news networks and their technological implementation is low, as massive technological advances have taken place in the last year that make such fake news campaigns possible in the first place through automatically generated text, image and video content.

In order to counteract fake news networks, the TruthGuard project is the first to analyse both sides of fake news, namely the generation as well as the identification of fake news. By analysing both sides, a holistic system can be researched that both improves state-of-the-art methods and can be used for awareness-raising measures. The project's activities will be supported by an Oversight Board with external partners (LoI_DZ, LoI_Bait), which will deal with ethical and legal issues relating to the project and whose considerations will be incorporated into the scientific and technical developments. The partner SME will draw up a comprehensive commercialisation plan and the scientific results will be disseminated internationally (LoI_Athena). The scientific and technical objectives of TruthGuard are listed below:

Research into methods for the effective regulation of generative models to combat fake news, including the identification of weaknesses in the implementation of ethics guidelines and the safeguarding of the integrity of these models.

Research into new, fully automated methods for identifying fake news and bots based on data from various social media networks.

Integration of a generating and an identifying tool of fake news in a demonstrator application (own Mastodon network) for specific deployment scenarios in a TRL 4 test environment, which are tested with the user as well as with other stakeholders

(LoI_BKA, representatives of the civilian population). The scenarios were determined in a structured manner using scientific frameworks at the beginning of the project.

Investigation of target group-specific challenges and platform-specific problems that lead to the development of targeted measures to promote the critical handling of information in order to strengthen the resilience of the civilian population against disinformation.

Projektkoordinator

- Fachhochschule Salzburg GmbH

Projektpartner

- Bundesministerium für Landesverteidigung
- Österreichisches Institut für angewandte Telekommunikation
- neke-neke GmbH
- AIT Austrian Institute of Technology GmbH
- Austria Institut für Europa- und Sicherheitspolitik (AIES)