DesinFact

DESINformations Früh erkennung von gefährdenden online nAChrichten Trends

Programm / Ausschreibung	KIRAS, Kooperative F&E-Projekte, KIras Kooperative CS F&E Projekte (KFE CS_2022)	**Status**	laufend
Projektstart	01.01.2024	**Projektende**	31.12.2025
Zeitraum	2024 - 2025	**Projektlaufzeit**	24 Monate
Keywords	Künstliche Intelligenz Desinformation Trustworthiness		

Projektbeschreibung

Desinformation stellt eine große Herausforderung für unsere Gesellschaft dar. Konzertierte Desinformations-Kampagnen sind ein Aspekt hybrider Bedrohungen, welche einerseits darauf abzielen können, konkrete kritische Infrastrukturen zu stören oder zu beschädigen – wie z.B. die Verteilungssicherheit von Energieträgern, Rohstoffen oder Medikamenten – oder andererseits, erweiterte kritische Infrastrukturen, wie demokratische Einrichtungen zu unterminieren und das Vertrauen in sie oder deren Vertreter zu zerstören.

Ein rechtzeitiges Erkennen von Desinformations-Kampagnen stellt somit einen essenziellen Beitrag zur Resilienz gegenüber solcher Bedrohungslagen dar. Aktuell stehen jedoch kaum Hilfsmittel zur Verfügung, um aktiv und frühzeitig Desinformations-Kampagnen zu erkennen. Betroffene erfahren oft viel zu spät über deren Involvierung, was deren Möglichkeiten einschränkt, wirksam darauf reagieren zu können. Oft bleibt nur noch Schadensbegrenzung übrig. Eine frühzeitige Erkennung solcher Trends würde einen Handlungsspielraum verschaffen, um z.B. geeignete Gegendarstellungen auszuarbeiten. DesinFact hat zum Ziel, den Forschungsstand zur automatischen Erkennung von Desinformationstrends zu verbessern, Lücken in technischen, rechtlichen und ethischen Bereichen zu identifizieren und geeignete Ansätze zu entwickeln, um ein solches System zu ermöglichen.

Um Desinformation-Kampagnen zu erkennen, müssen unterschiedliche Datenquellen überwacht werden, um Trends identifizieren zu können. Um diese dann automatisiert als Desinformations-Kampagne zu bewerten, müssen Ansätze zur Anwendung kommen, welche höchsten Qualitätsstandards entsprechen, da eine Fehlentscheidung – nämlich dass hier tatsächlich Fake News Inhalte verbreitet werden - auf die Autor:innen oder deren Institutionen zurückfällt, und somit deren Ruf beträchtlich schädigen können. Gleichfalls schädigt dies das Ansehen der Anbieter solcher Technologien – also der beteiligten Konsortialpartner - sowie das öffentliche Vertrauen in eine solche Technologie.

Deshalb liegt der Fokus in DesinFact auf der Steigerung der Vertrauenswürdigkeit (Trustworthiness) in technischen, rechtlichen und ethischen Belangen ein Hauptaugenmerk der Forschungstätigkeit. Es sollen Methoden zur messbaren Qualitätssteigerung bzw. Erklärbarkeit der Entscheidungen erforscht werden. Diese Methoden sollen sowohl für Expert:innen

als auch für operative Betreiber:innen verständlich sein.

Ein Aspekt zur Steigerung der Genauigkeit ist die Verknüpfung der Analyse von Netzwerkstrukturen und Kommunikationsmuster mit Inhaltsbasierter Analyse. Hierfür sollen in DesinFact Methoden zur Erkennung von Verbreitungswegen und Schlüssel-Aktoren in Desinformations-Netzwerken erforscht und mit Inhaltsbewertenden Verfahren verknüpft werden. Ein weiterer Fokus von DesinFact besteht in der Erforschung einer möglichen öffentlichen Bereitstellung eines Systems zur Erkennung von Desinformation. Ein solches System soll es Bürger:innen ermöglichen, online Inhalte auf Desinformation hin untersuchen zu lassen. DesinFact wird dabei sozio-technische Aspekte erforschen, welche für eine adäquate Einführung einer solchen Technologie relevant sind.

Da Desinformation jedoch eine hochkomplexe Aufgabenstellung ist, deren Einschätzung von zahlreichen Faktoren, wie z.B. Alter, Allgemeinbildung, oder kulturellem, politischem sowie religiösem Hintergrund, abhängt, sind kontroversielle Entscheidungen kaum vermeidbar. Dementsprechend müssen sowohl die Bewertungssysteme als auch die Ergebnispräsentation klar und verständlich sein. Entsprechende interdisziplinäre Studien sind zentrale Inhalte von DesinFact.

Abstract

Disinformation poses a major challenge to our society. Concerted disinformation campaigns are one aspect of hybrid threats, which can aim to disrupt or damage specific critical infrastructures - such as the distribution security of energy sources, raw materials or medicines - or to undermine broader critical infrastructures such as democratic institutions and destroy trust in them or their representatives.

Timely detection of disinformation campaigns is therefore an essential contribution to resilience against such threats. Currently, however, there are hardly any tools available to actively detect disinformation campaigns at an early stage. Those affected often learn about their involvement far too late, which limits their ability to respond effectively. Often, only damage limitation remains. Early detection of such trends would provide room for maneuver, e.g. to prepare appropriate counter-narratives. DesinFact aims to improve the state of research on automatic detection of disinformation trends, to identify gaps in technical, legal and ethical areas, and to develop suitable approaches to enable such a system.

To detect disinformation campaigns, different data sources need to be monitored to identify trends. In order to then automatically assess these as disinformation campaigns, approaches must be applied that meet the highest quality standards, since an erroneous decision - namely that fake news content is actually being disseminated here - can rebound on the authors or their institutions, and thus considerably damage their reputation. Likewise, this damages the reputation of the providers of such technologies - i.e., the consortium partners involved - as well as the public trust in such a technology.

Therefore, the focus in DesinFact is on increasing trustworthiness in technical, legal and ethical aspects as a main focus of the research activity. Methods for measurable quality improvement or explainability of decisions are to be researched. These methods should be understandable for experts as well as for operational operators.

One aspect of increasing accuracy is to link the analysis of network structures and communication patterns with content-based analysis. To this end, DesinFact will explore methods for detecting dissemination channels and key actors in disinformation networks and combine them with content assessment methods. Another focus of DesinFact is the research of

a possible public provision of a system for the detection of disinformation. Such a system should enable citizens to have content checked for disinformation online. DesinFact will explore socio-technical aspects that are relevant for an adequate implementation of such a technology.

However, since disinformation is a highly complex task whose assessment depends on numerous factors, such as age, general education, or cultural, political, and religious background, controversial decisions can hardly be avoided. Accordingly, both the evaluation systems and the presentation of results must be clear and understandable. Corresponding interdisciplinary studies are central contents of DesinFact.

Projektkoordinator

- AIT Austrian Institute of Technology GmbH

Projektpartner

- Bundeskanzleramt
- leiwand AI gmbh
- Universität für Weiterbildung Krems
- Complexity Science Hub Vienna CSH - Verein zur Förderung wissenschaftlicher Forschung im Bereich komplexer Systeme
- Bundesministerium für Landesverteidigung
- X-Net Services GmbH