defame Fakes

Detektion von Deepfakes und medialen Manipulationen in Bildern und Videos

Programm / Ausschreibung	KIRAS, Kooperative F&E-Projekte, KIras Kooperative CS F&E Projekte (KFE CS_2022)	**Status**	abgeschlossen
Projektstart	01.02.2024	**Projektende**	31.01.2026
Zeitraum	2024 - 2026	**Projektlaufzeit**	24 Monate
Keywords	Medienforensik; Fusion; Fake News; Desinformation; Künstliche Intelligenz;		

Projektbeschreibung

Beispiele wie DeepFaceLab, DALL E 2, Stable Diffusion oder TikTok Gesichtsfilter markieren für viele Menschen einen – positiv wie negativ – aufschlussreichen Moment im Verständnis dessen, was KI derzeit im Bereich der Bild- und Videomanipulation leisten kann. Für ermittelnde Behörden, die öffentliche Verwaltung, Medienorganisationen, die Privatwirtschaft, aber auch für Bürgerinnen und Bürger ergeben sich daraus vielfältige Kontroversen, Bedrohungen, aber auch Herausforderungen, mit denen sich die Gesellschaft sowohl im privaten, wirtschaftlichen, staatlichen und demokratischen Umfeld auseinandersetzen muss.

Im angestrebten kooperativen F&E-Vorhaben defame Fakes befassen wir uns daher mit der Erkennung von Deepfakes und medialen Manipulationen in digitalen Bild- und Videoinhalten, mit dem Ziel der kontinuierlichen Aushöhlung des Vertrauens in digitale Inhalte entgegenzuwirken und die technologischen Fähigkeiten zur Verifikation zu stärken. Dazu konzentrieren wir uns im Kontext des Ausschreibungsschwerpunktes „Cybersicherheit“ 3.3.1 insbesondere auf 2 große thematische Kernbereiche:

Diese sind…
a) die Konzeption und Erforschung geeigneter, verständlicher Assessment-Werkzeuge zur Erkennung von Deepfakes und Manipulationen in großen Sammlungen digitaler Bild- und Videoinhalte – um hiermit die technologischen Fähigkeiten zur Ergreifung geeigneter, reaktiver Maßnahmen zu stärken – sowie…
b) die Initiierung und Gestaltung eines Prozesses zur Awareness-Bildung, um einen breiten, gesamtgesellschaftlichen Zugang zu fördern, die nationale Umsetzung des „Aktionsplan Deepfake“ unter enger Einbindung von Behörden, Medien, GSK-Partnern und relevanten nationalen & europäischen Stakeholdern zu unterstützen und damit letztlich präventive Maßnahmen und Akzente zur Wissens- und Bewusstseinsbildung zu setzen.

Dafür wird ein ausgewogenes Konsortium mit den wissenschaftlichen Partnern AIT und ÖIAT sowie den Industrie- und Medienpartnerinnen und -partnern PwC und APA von den Endanwendern und -anwenderinnen – BMI, BMLV und KSÖ – unter Koordination der größten österreichischen außeruniversitären Forschungseinrichtung AIT geleitet.

Die Forschungsinhalte in defame Fakes erstrecken sich, beginnend mit der gemeinsamen Darstellung von Bedrohungsszenarien, den daraus resultierenden Anforderungen und der gemeinsamen iterativen Konzeption von präventiven und reaktiven Maßnahmen, über die Zusammenstellung und Generierung einer Datensammlung, die die Angriffsvektoren und Bedrohungen in einer für Endanwender:innen und Projektpartner:innen geeigneten Weise abbildet. Darauf aufbauend werden die oben erwähnten Assessment-Werkzeuge erforscht und Awareness-Bildungsprozesse skizziert und evaluiert. Dabei spielt die Analyse der sozialen, ethischen und rechtlichen Implikationen eine permanente und begleitende Rolle in allen Forschungsaktivitäten. Beispielsweise wenn es um die gesellschaftlichen und ethischen Implikationen von Deepfakes, rechtliche Rahmenbedingungen und Regulierungsbedarfe, aber auch um die Bewertung von Bedrohungsszenarien - z.B. Deepfakes im Cybercrime - geht.
Schließlich zeigt ein im Projekt erarbeiteter Verwertungsplan, wie die resultierenden Forschungsergebnisse weiterentwickelt und als produktives System sowohl in öffentlichen als auch in privaten Einrichtungen eingesetzt werden können.

Abstract

Examples such as DeepFaceLab, DALL E 2, Stable Diffusion or TikTok face filters mark for many people a revealing moment – both positive and negative – in understanding what AI can currently do in the field of image and video manipulation. For investigating authorities, public administrations, media organisations, the private sector, but also for citizens, this leads to a multitude of controversies, threats, but also challenges that society has to deal with in private, economic, governmental and democratic environments.

In the envisaged cooperative R&D project defame Fakes, we therefore address the detection of deepfakes and media manipulation in digital image and video content, with the aim of countering the continuous erosion of trust in digital content and strengthening technological verification capabilities. To this end, we focus in particular on 2 major thematic core areas in the context of the call focus "Cybersecurity" 3.3.1:

These are...
a) the design and research of appropriate, understandable assessment-tools for the detection of deepfakes and manipulations in large collections of digital image and video content – in order to hereby strengthen the technological capabilities enabling appropriate, reactive measures - as well as...
b) the initiation and design of an awareness-raising process to promote a broad, society-wide approach, to support the Austrian, national implementation of the “Aktionsplan Deepfake“ with close involvement of authorities, media, SSH partners and relevant national and European stakeholders, and thus ultimately to set preventive measures and emphasize knowledge and awareness building.

For this purpose, a balanced consortium with the scientific partners AIT and ÖIAT as well as the industry and media partners PwC and APA is led by the end users – BMI, BMLV and KSÖ – under the coordination of the largest Austrian research and technology organization AIT.
The research content of defame Fakes ranges from the joint description of threat scenarios, the resulting requirements, and the joint iterative design of preventive and reactive measures to the compilation and generation of a data collection that maps the attack vectors and threats in a way that is relevant to end users and project partners.
On this basis, the above-mentioned assessment-tools will be explored, and awareness-raising processes outlined and evaluated. The analysis of social, ethical, and legal implications plays a constant and accompanying role in all research

activities. For example, the social and ethical implications of deepfakes, legal frameworks and regulatory needs, but also the evaluation of threat scenarios – e.g. deepfakes in cybercrime.
Finally, an exploitation plan developed in the project shows how the research results can be further developed and used as a productive system in both public and private institutions.

Endberichtkurzfassung

WICHTIGE INFORMATION: Sie finden die beiden Kurzfassungen (Deutsch/Englisch) im Anhang des Endberichts als barrierefreie PDF-Dateien.

Projektkoordinator

- AIT Austrian Institute of Technology GmbH

Projektpartner

- PwC Wirtschaftsprüfungs- und Steuerberatungsgesellschaft mbH
- Österreichisches Institut für angewandte Telekommunikation
- Bundesministerium für Landesverteidigung
- APA - Austria Presse Agentur eG
- Bundesministerium für Inneres
- Kompetenzzentrum Sicheres Österreich (KSÖ)