DEFAME FAKES

Detektion von Deepfakes und medialen Manipulationen
von Bildern und Videos

Mina Schütz, Scientist am AIT Austrian Institute of Technology, 11.09.2024

FFG Forum – Digitale Sicherheit in Europa – Herausforderungen und Chancen für die Gesellschaft

Bundesministerium Inneres
Bundesministerium Finanzen
Bundesministerium Landesverteidigung
KIRAS Sicherheitsforschung
FFG Forschung wirkt.

KEY FACTS

DEFAME FAKES

defame Fakes

- Forschungsprojekt finanziert durch das Sicherheitsforschungs-Förderprogramm KIRAS des Bundesministeriums für Finanzen

Bundesministerium Finanzen KIRAS FFG Forschung wirkt.

- Laufzeit: Februar 2024 – Jänner 2026

- Website: www.defamefakes.at
- Koordination: AIT, Martin Boyer <martin.boyer@ait.ac.at>

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY TOMORROW TODAY APA KSÖ pwc oiat Bundesministerium Inneres Bundesministerium Landesverteidigung

ZIELE DES PROJEKTS

DEFAME FAKES

Erforschung und Entwicklung von **Assessment Werkzeugen** zur **Erkennung von Deepfakes**

Konzeptionierung und Initiierung von gesamtgesellschaftlichen **Awareness-Maßnahmen**

Analyse von Bedrohungsszenarien, gesellschaftlichen, ethischen und rechtlichen Implikationen

DEEPFAKES

- KI-basierte Manipulation von Audio, Bild, Video, Text
- Zunehmend einfach zugänglich
- Moderne Generatoren können sehr realistische Ergebnisse liefern

Bedrohungsszenarien

- Kriminelle Gruppen als Early Adopters
- Desinformation und Fake News
- Image- und Rufschädigung
- Identitätsdiebstahl
- Aushöhlung des Vertrauens in digitale Inhalte
- Betroffene: Privatpersonen, Regierungen, Unternehmen

DEFAME FAKES

DEEPFAKE DETEKTION

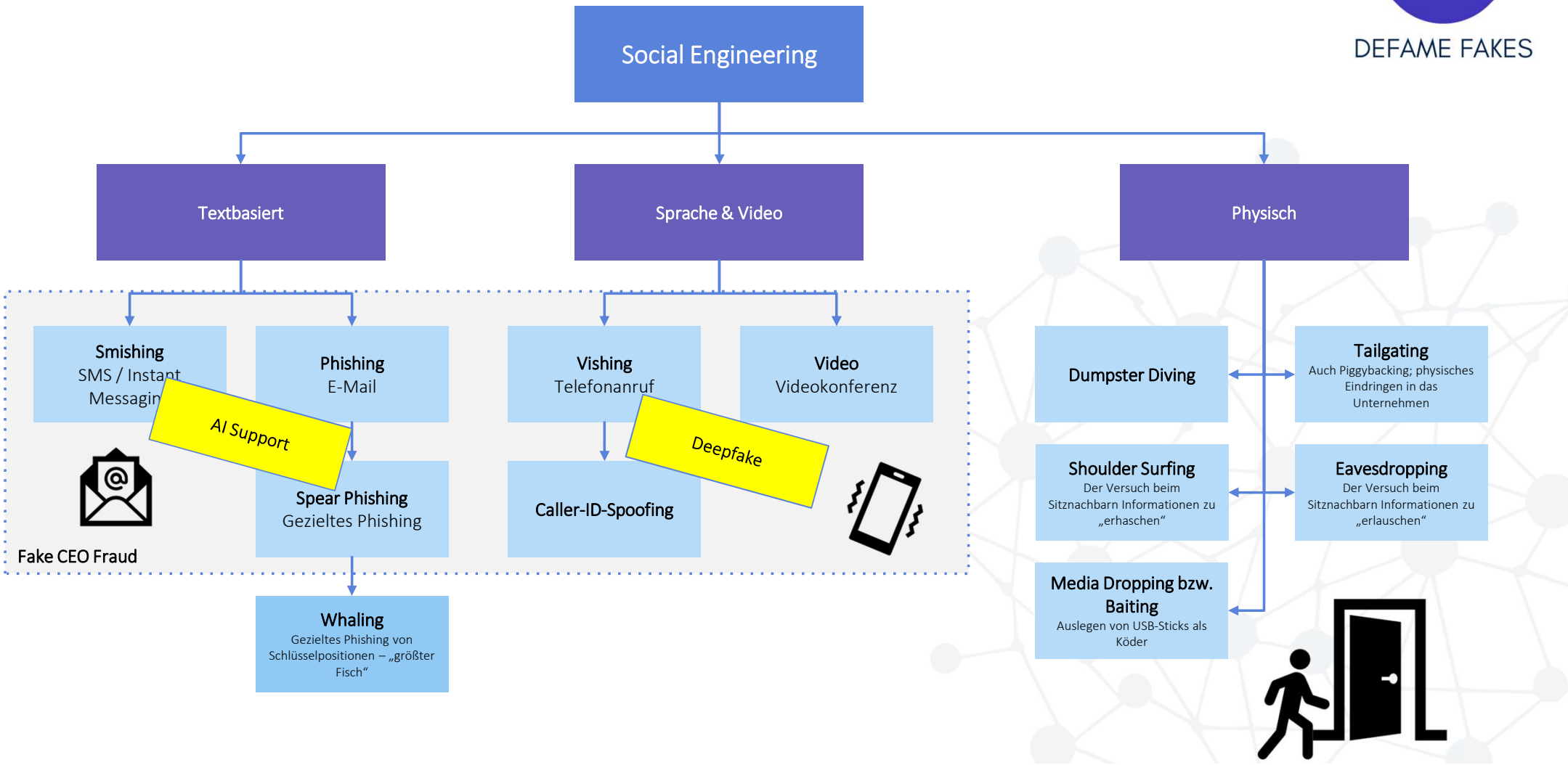
DEFAME FAKES

- Herausforderungen:
  - Kontinuierliche Weiterentwicklung der Deepfake-Technologie
  - Dringlichkeit von Detektionslösungen
  - Bedarf an verallgemeinerbaren Erkennungsmethoden für Deepfakes, unabhängig der Techniken die zur Generierung verwendet wurden

- Kernpunkte:
  - Konzentration auf bestimmte Anwendungsfälle und Szenarien
  - Erklärbare und interpretierbare Ergebnisse der Erkennung
  - Sammlung repräsentativer, Bias-freier Daten für Training und Evaluierung
  - Skalierung und Integration der Lösung in bestehende Arbeitsabläufe

KI UND DEEPFAKES IM SOCIAL-ENGINEERING-UMFELD

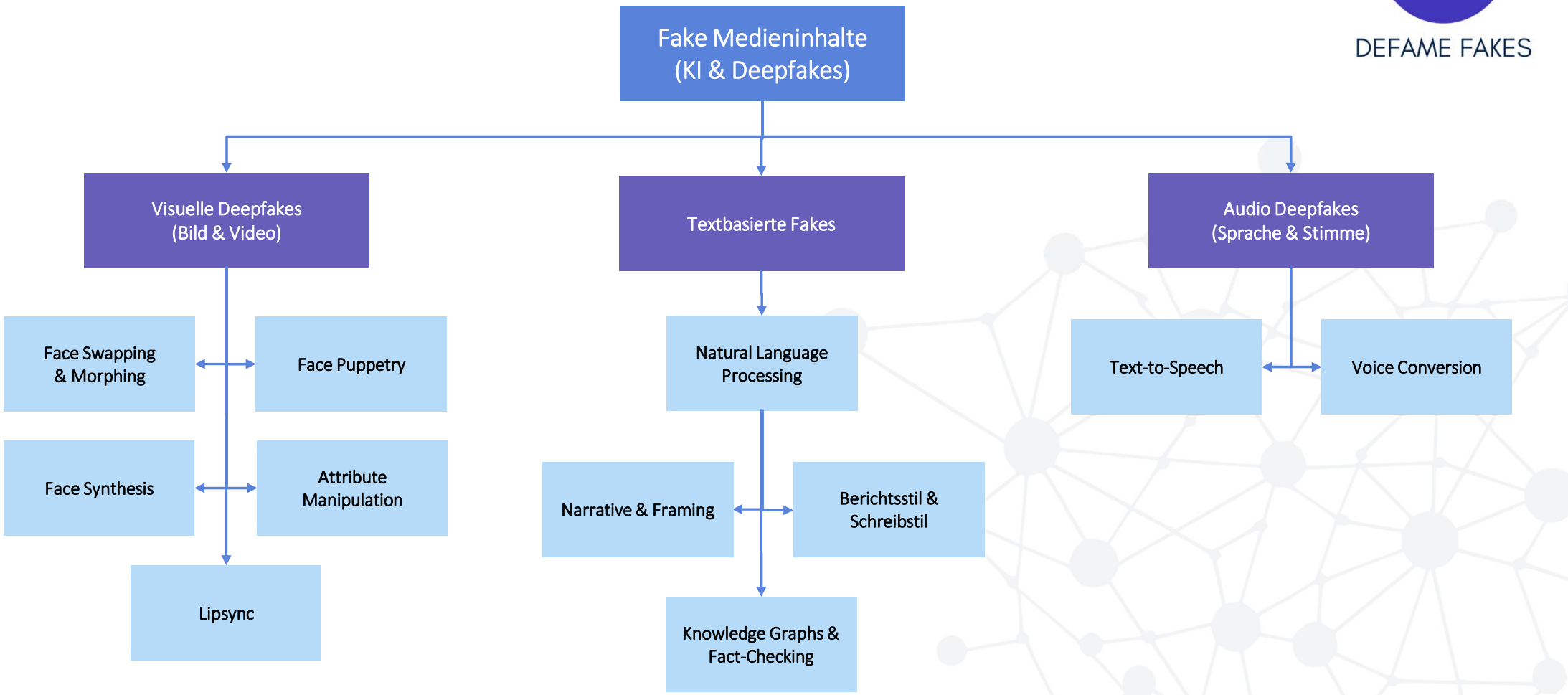
DEFAME FAKES



DEEPFAKE TECHNOLOGIEN

pwc

DEFAME FAKES



PROJEKT-PORTFOLIO (AIT)

EVIDENZ

FFG

STARLIGHT

RAIDAR
RAPID AI BASED DETECTION OF AGGRESSIVE OR RADICAL CONTENT ON THE WEB

FFG

DESINFACT

FFG

defalsif-ai

FFG

HYBRIS

FFG

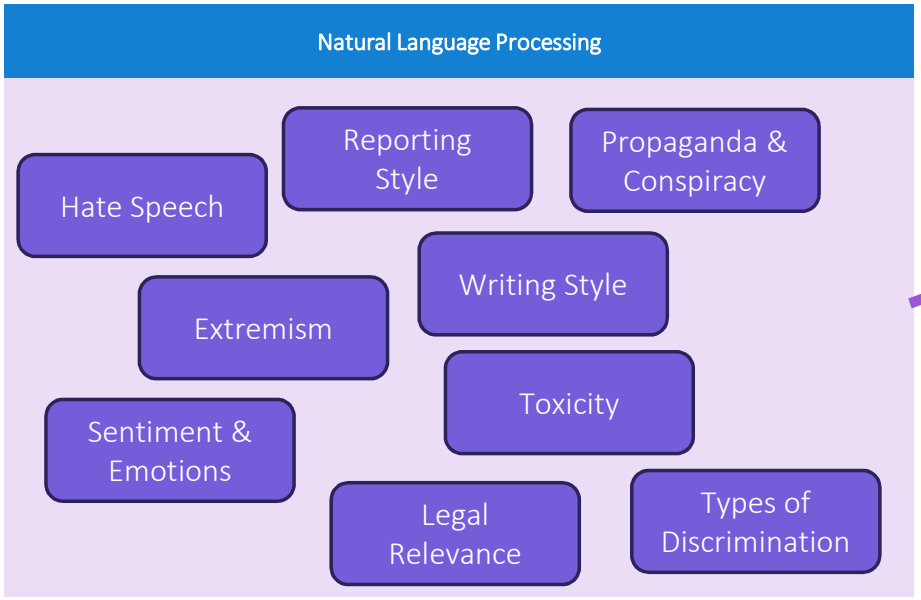
German-Austrian Digital Media Observatory

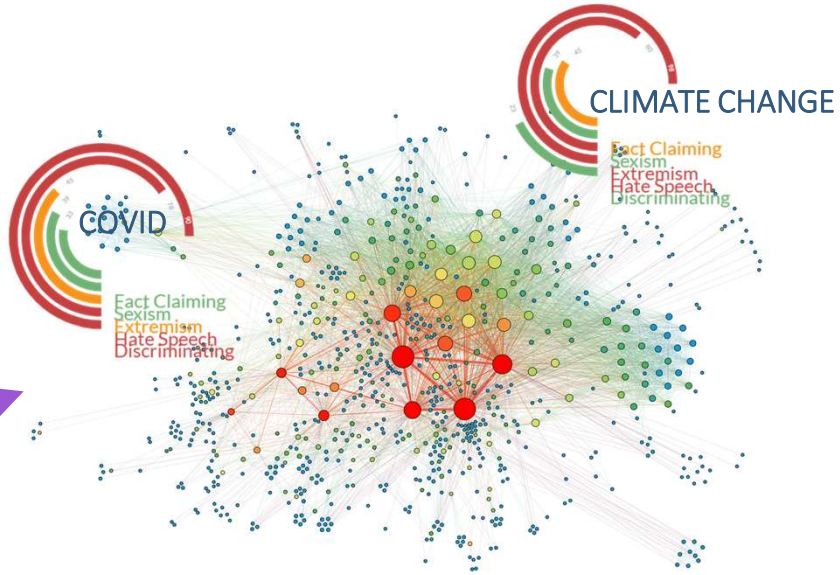
DEFAME FAKES

FFG

INFORMATION NUTRITION LABELS

- Automatische Erkennung von Desinformation mittels KI Modellen
- Einschätzung von über 50 Merkmalen
- Erkennung von Zusammenhängen und Themen





Bundesministerium Finanzen
KIRAS
FFG Forschung wirkt.

HYBRIS

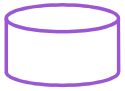
AIT AUSTRIAN INSTITUTE OF TECHNOLOGY · Artificial Researcher · TU WIEN TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology · VSC VIENNA SCIENTIFIC CLUSTER · research institute · thinkers.ai · BOKU · Bundesministerium Landesverteidigung

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY

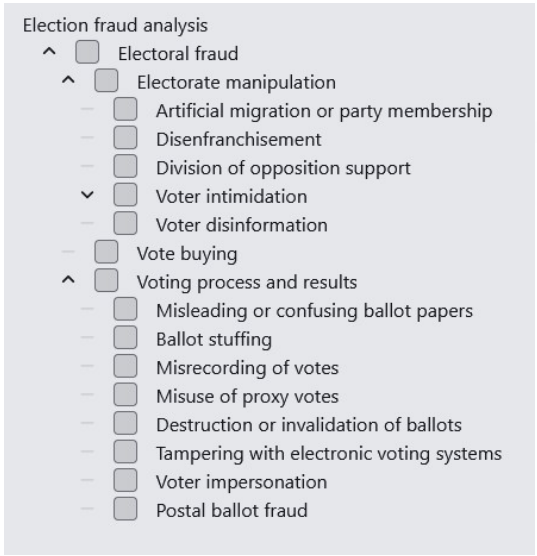
SITUATIONAL AWARENESS

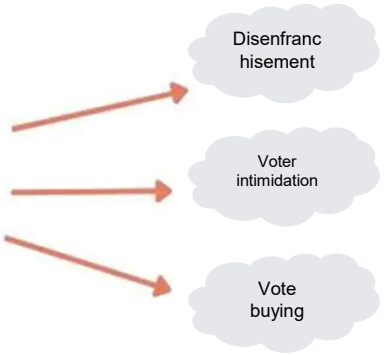
THREAT ESTIMATION & NARRATIVES

Data Source



Taxonomy





Story-Card

Title: <LLM generated accumulated Title>

Num. Articles: 35

Key-Claims:

1. <LLM extracted key-caim>
2. <LLM extracted key-caim>
3. <LLM extracted key-caim>
4. <LLM extracted key-caim>
5. <LLM extracted key-caim>

Indicators:

- Incitement of Volence
- Economic Implications
- Medical Implications

Narratives:

- Central Narrative:
- Further Narrative

Predjudice:

- Antisemitism
- Anti-Feminism

Bundesministerium Finanzen · KIRAS · FFG Forschung wirkt.

Vielen Dank für Ihre Aufmerksamkeit!

Mina Schütz

Mina.schuetz@ait.ac.at

Bundesministerium Finanzen
KIRAS
FFG Forschung wirkt.