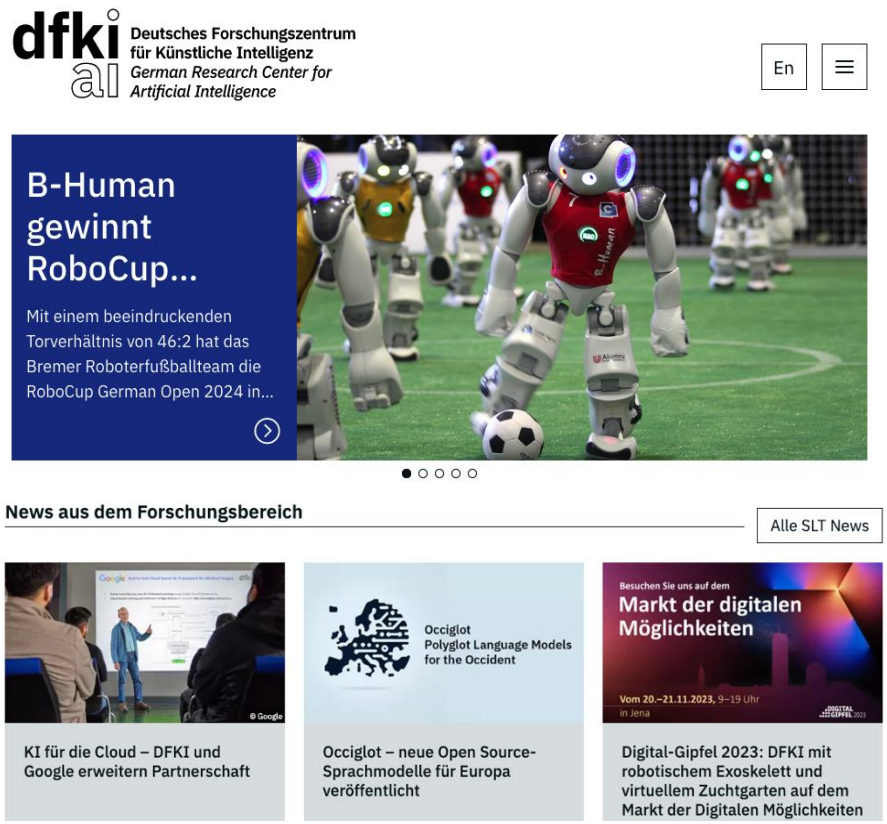

Deepfakes
Reality?

Innovative KI-Lösungen für Medienverifikation und Desinformationsbekämpfung

Dr.-Ing. Tim Polzehl
Senior AI R&D
DFKI
CEO Gretchen AI

DFKI – Deutsches Forschungs-zentrum für Künstliche Intelligenz

- Größtes öffentliches KI-Forschungszentrum Deutschlands, "Center of Excellence"
- 26 Forschungsbereiche mit ca. 1.500 Mitarbeitern aus 76+ Ländern
- SLT “Speech and Language Technology” ca. 60+ Mitarbeiter



www.dfki.de

DFKI is a Joint Venture of:

dfki ai

Source: https://konradweber.medium.com/8-steps-to-verify-deep-fake-videos-1dfc408568c4

Deepfake – was ist das?

*Wiki: "… **realistisch** wirkende **Medieninhalte** (Foto, Audio, Video usw.), die durch Techniken der **künstlichen Intelligenz abgeändert**, **erzeugt** bzw. **verfälscht** worden sind."*

Vorsatz*:*

- Desinformation als wirtschaftliche Manipulation, bspw.: Manipulation von Aktien-Kurse und/oder Rufschädigung
- Desinformation als politische Manipulation
- Straftaten: Betrug, Erpressung, etc.
- Deepfake als Kunst und Unterhaltung

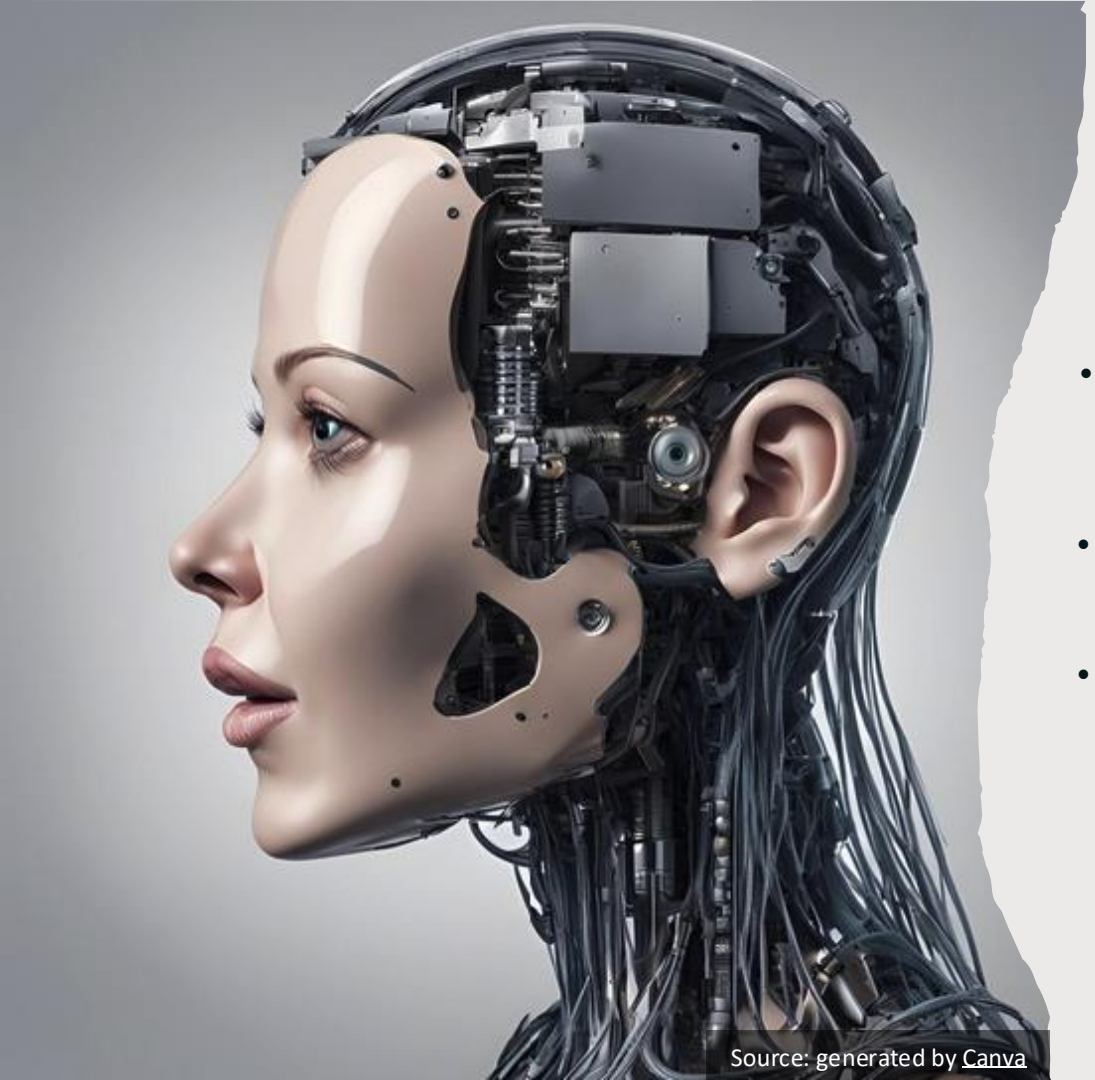
Source: https://www.reddit.com/r/midjourney/comments/1430b26/best_ai_generated_faces_i_got_so_far

KI-generierte Bilder für alle!

DALL·E, Imagen, Stable Diffusion, Midjourney, Sora, Veo, etc.:

- **KI trainiert mit riesigen Datensätzen (>10 Mrd Text-Bild Paare, Videos)** generiert realistische und surrealistische Bilder
- Realistische Bilder zeigen teilweise noch logische Fehler
- Sehr gute Ergebnisse erreichbar
- Video folgt Bild.

Source: https://www.youtube.com/watch?v=G4wJ4WeJrz4



Source: generated by Canva

KI-generierte Sprache für alle!

- *ElevenLabs, Microsoft, OpenAI, etc. basierend auf Sprachsynthese (TTS), Voice Cloning*
- Realistisch klingende Ergebnisse sind der Regelfall
- SOTA: *mehrsprachig, emotions-/ und akzentrealisierende, hintergrund-geräuschadaptierende, zero-shot TTS (bspw: 3 Sekunden Aufnahmen genügen Vall-E, 30 Sekunden openAI)*

KI generiert Text für alle!

Synthetische Texte

Mit Modellen wie **GPT5** (OpenAI), **Gemini (Google), Llama 3** (Meta), **Claude 3** (Anthropic) kann über ein Prompt-based Interface interagiert werden, bspw:

- **Lernen**: Schreibassistent, Übersetzungen, Ideengebung
- **Business**: Customer Service, Content-Generierung
- **Creative Arts**: Inspiration, kollaborative Kunstwerke
- **Research**: simple Datenanalyse, Schreibhilfe, Coding
- **Zugänglichkeit**: kann neue Interaktionswege für Menschen mit Behinderungen eröffnen

news-polygraph – R&D Projekt

12 Mio €, 10 Partner,
Team >35, 5/2023 – 5/2026

Intelligente Entscheidungsunterstützung

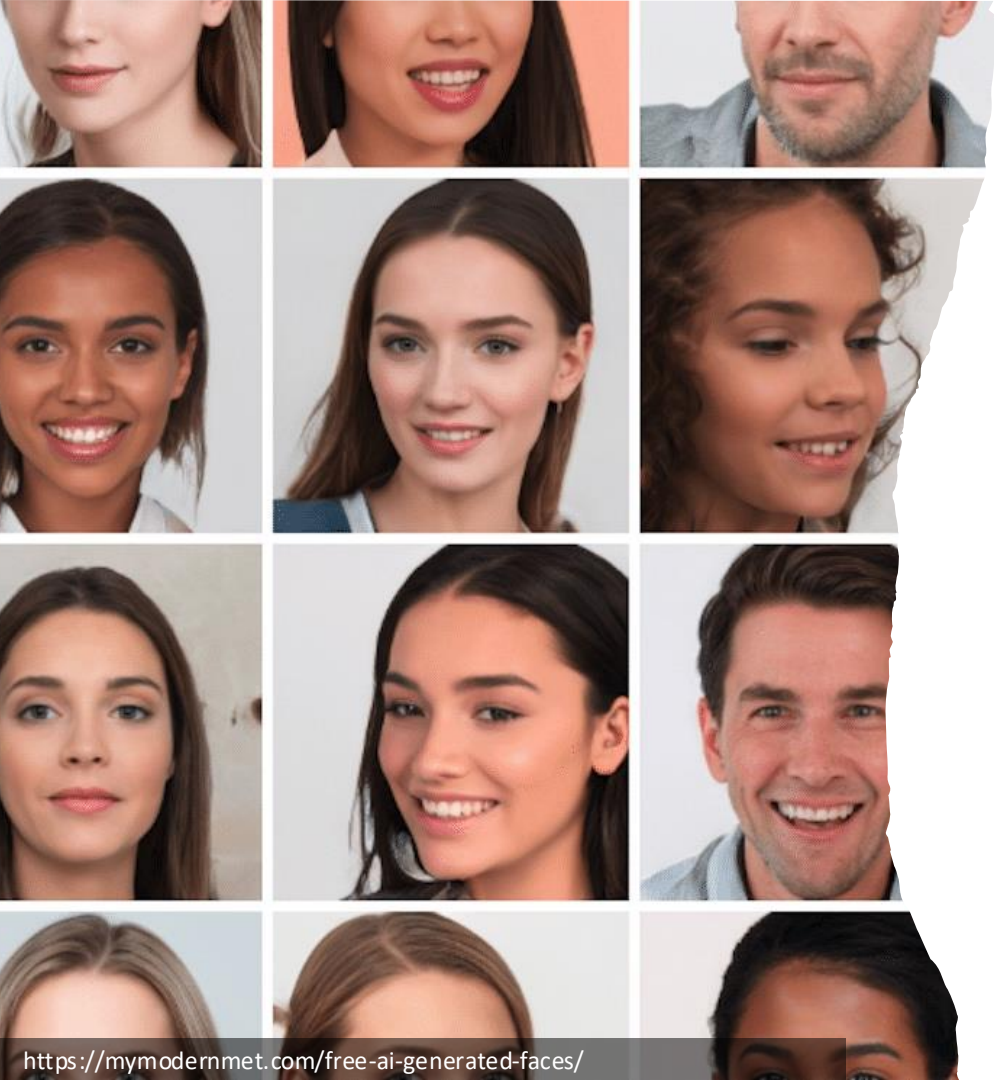
- Deep- and Cheap-Fake Detektion, Fact-Checking
- Schnelle automatische Annotation, Search, Monitoring
- LLM reasoning und RAG (Retrieval Augmented Generation)
- Semantische Analyse

Medienanalyse:

- Fokus: Text, Audio, Sprache, Bilder, Videos, Metadaten

Trust / Vertrauen

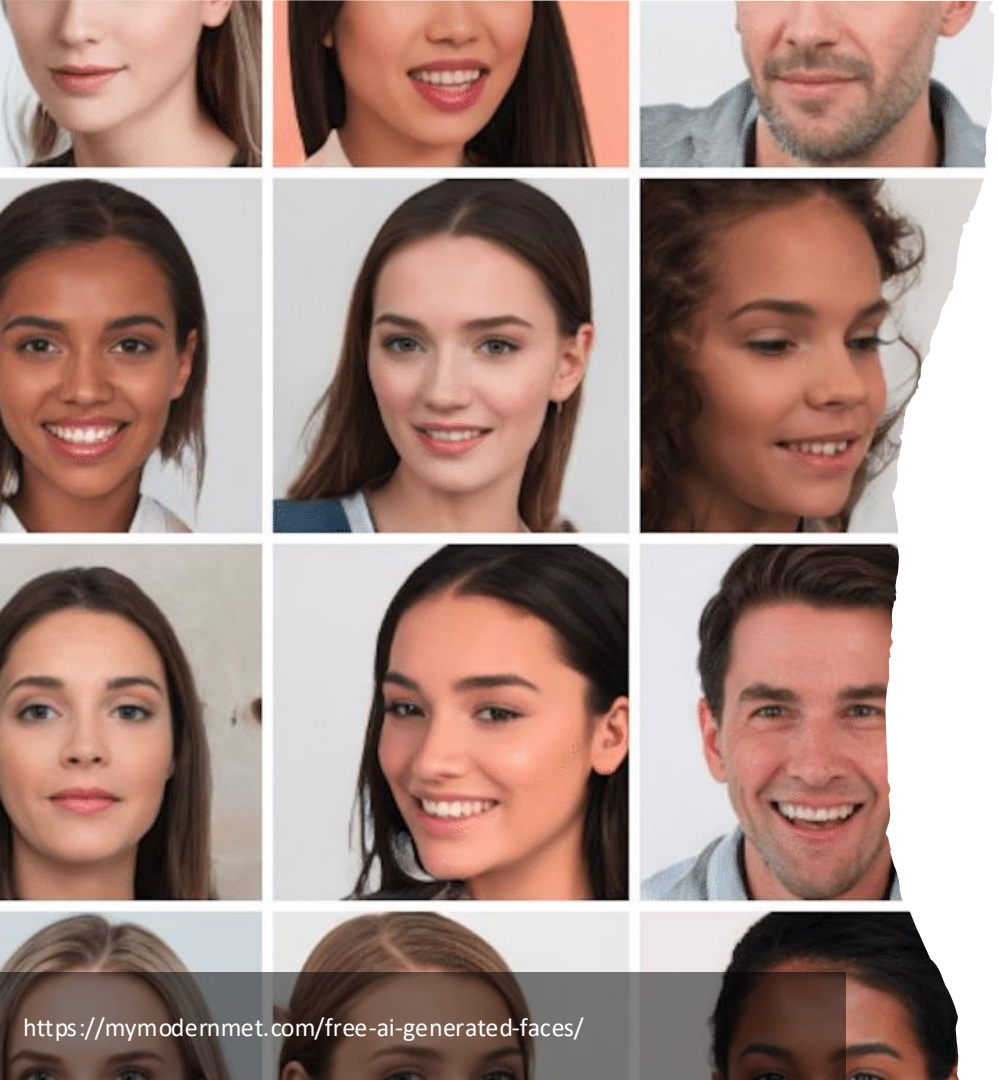
- Erklärbarkeit (XAI) und Transparenz
- Robustheit und Faktentreue (faithfulness)
- Bias, Fairness: Mitigation und Counter-Measures
- Gesetzgebung und Regulatorik: GDPR, AI-Act, DSA



https://mymodernmet.com/free-ai-generated-faces/

KI-generierte Bilder mit KI erkennen?

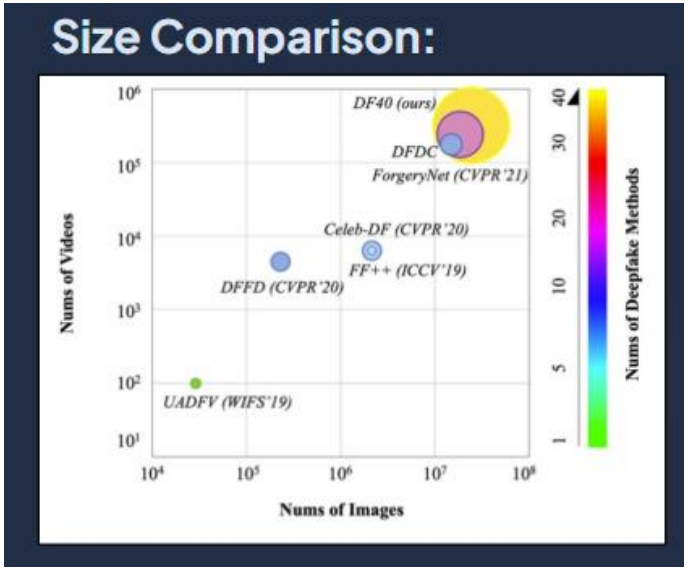
- ***Tools:*** *sentinel, sensity, hive, duckduckgoose, etc. höchst unreliabel*
- **Problem 1:** Fake vs. True problematisch: 51%→ true
- **Problem 2:** “Generalisierung”
- **Problem 3**: viele Fehler (FP, FN)
- Video = Option zu mehr Fehler in Bewegungen + Synchronisation



https://mymodernmet.com/free-ai-generated-faces/

KI-generierte Bilder mit KI erkennen?

- Evaluation in der Spitzenforschung: Wissenschaftliche Benchmarks



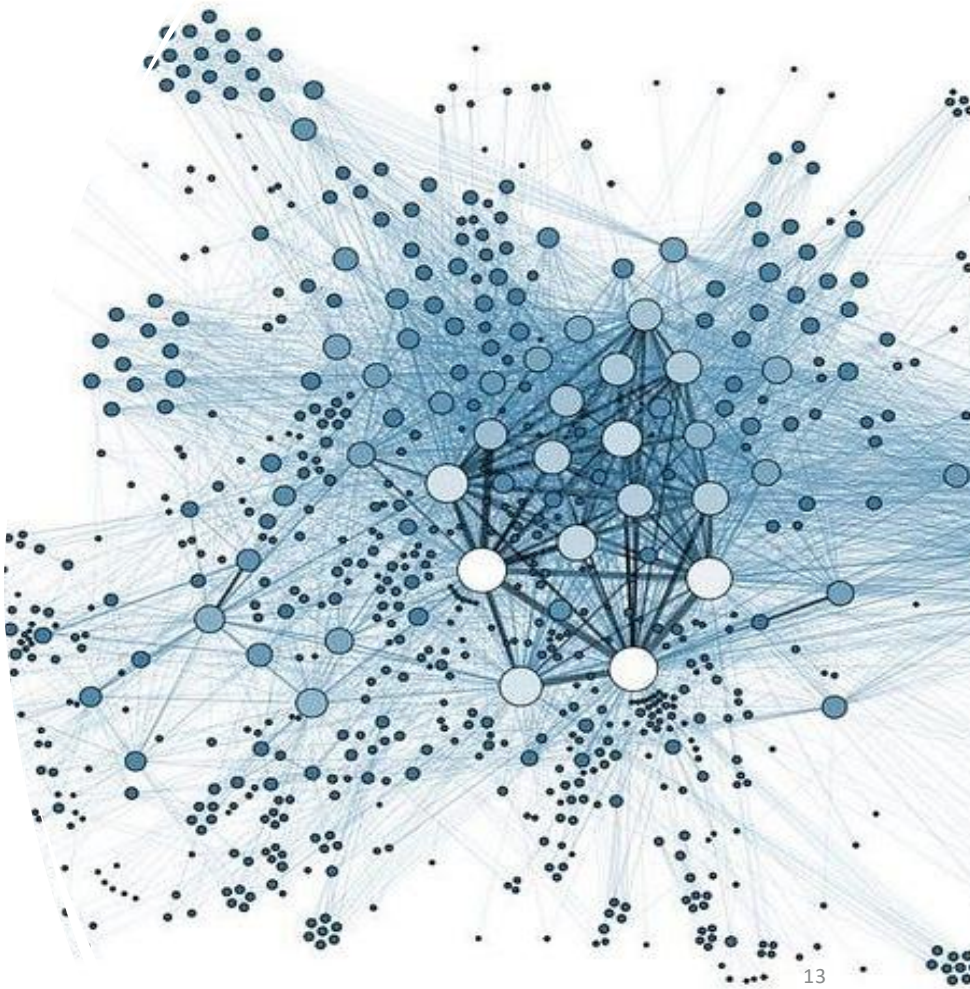
KI-generierte Sprache mit KI erkennen?

- *Tools: ressemble, AI or Not, hive, etc.* oftmals unreliabel!
- Gleiche Probleme wie bei Bild / Video
- Sprachaufnahmen = Option zu mehr Fehler je länger das Material

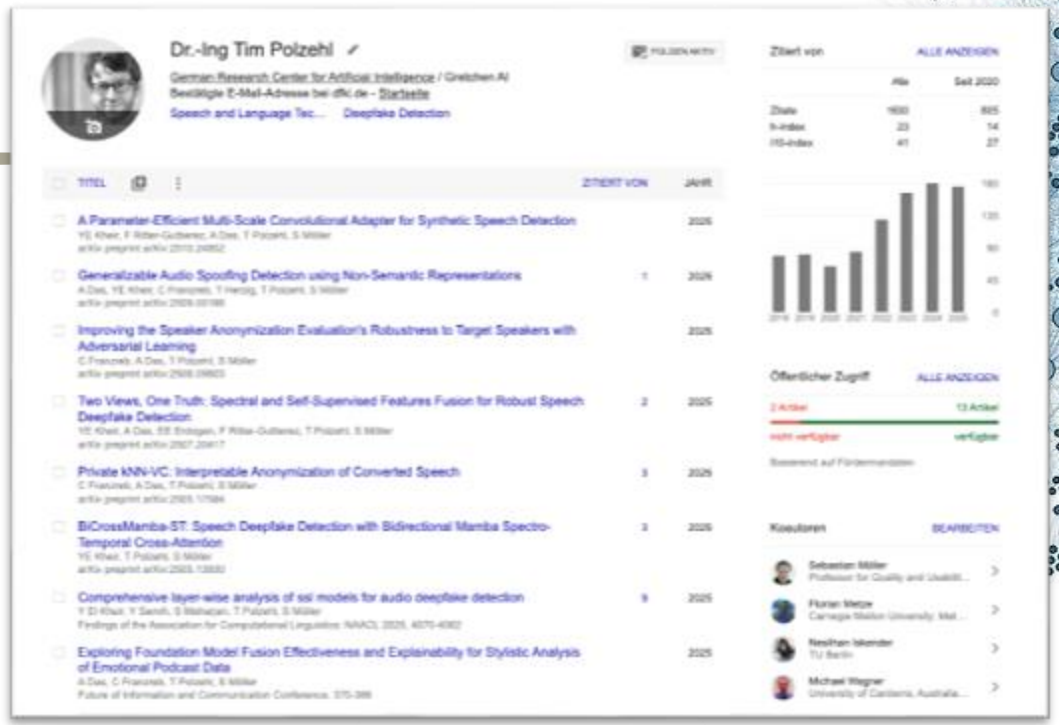
Source: generated by Firefly

DFKI Forschung zu Deepfake Erkennung: Sprache / Bild

- Forschung zu neuen Trainingsmethoden und KI-Architekturen, *Generalization* Problem
- Bestes Modelle mit derzeit ca. 5 Mio. Stichproben trainiert, Akkuratheit bis zu **98%** in-domain
- 20% Fehlerreduzierung out-of-domain im Vergleich zum derzeitigen SOTA KI-Modell
- Eigene (De-) Anonymisierungs- und Voice-Cloning-Forschung, Sprachanalyse, Synthese, Emotionen, Persönlichkeit, etc.



Papers R&D Tim Polzehl

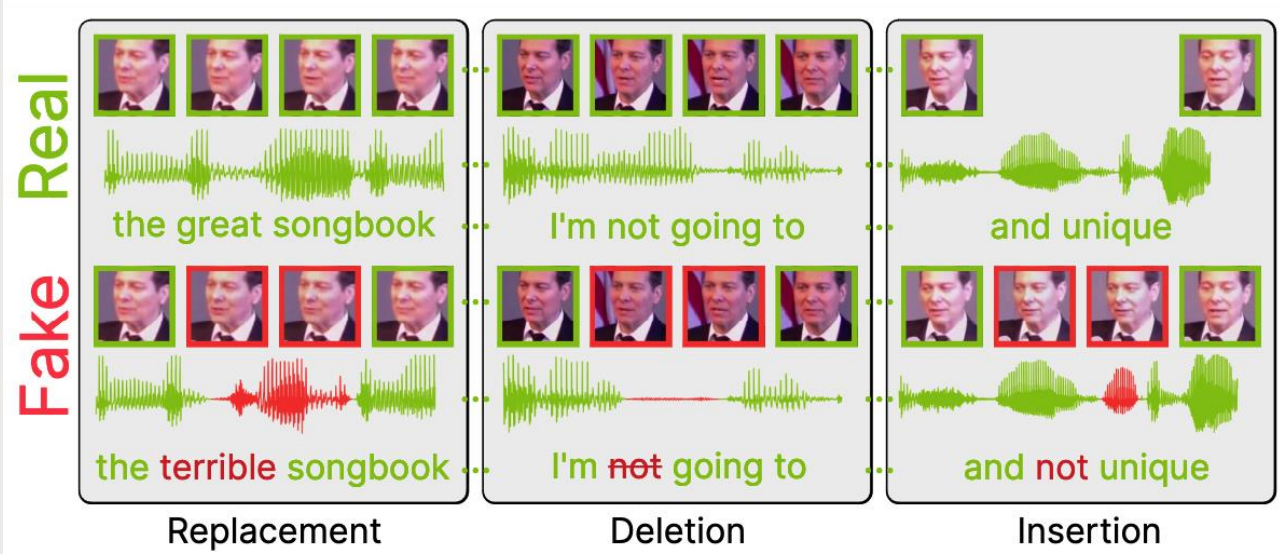


Google Scholar: https://scholar.google.com/citations?hl=de&user=tImnUh0AAAAJ&view_op=list_works&sortby=pubdate

AI Video Manipulation Pipeline

AV-Deepfake1M:
1 Mio Fakes

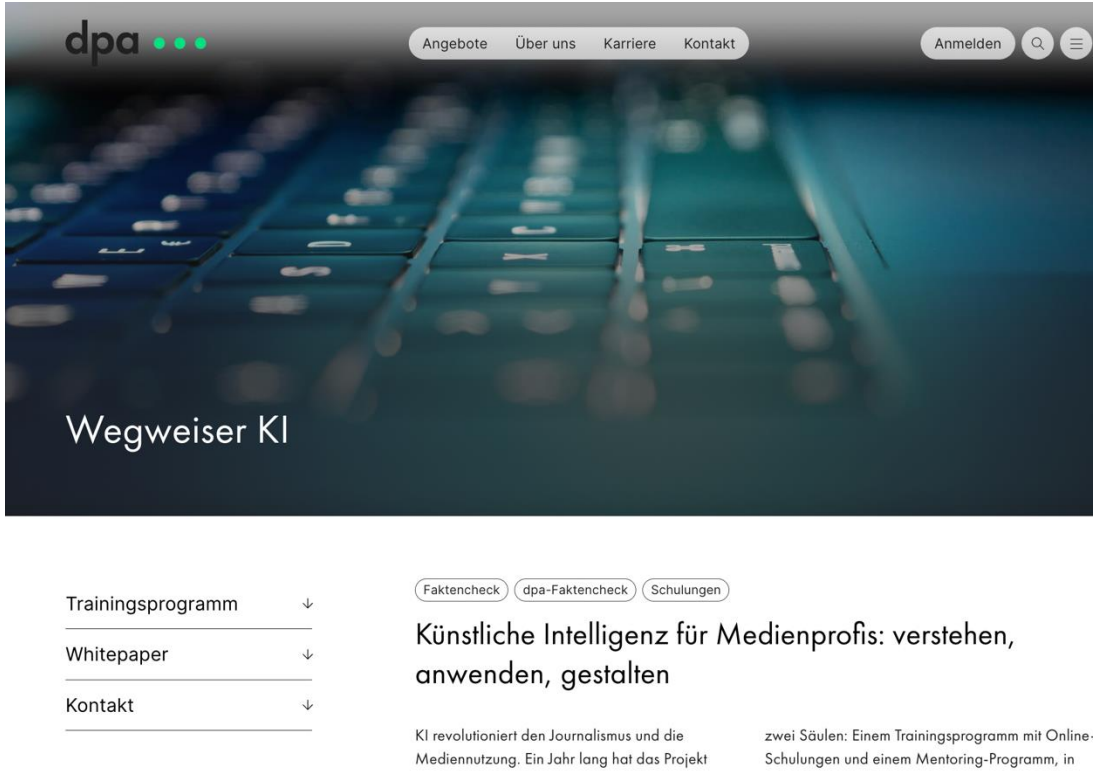
- Eigene High-Performance Modelle **>98%** Akkuratheit



https://arxiv.org/pdf/2311.15308

Wegweiser KI - dpa

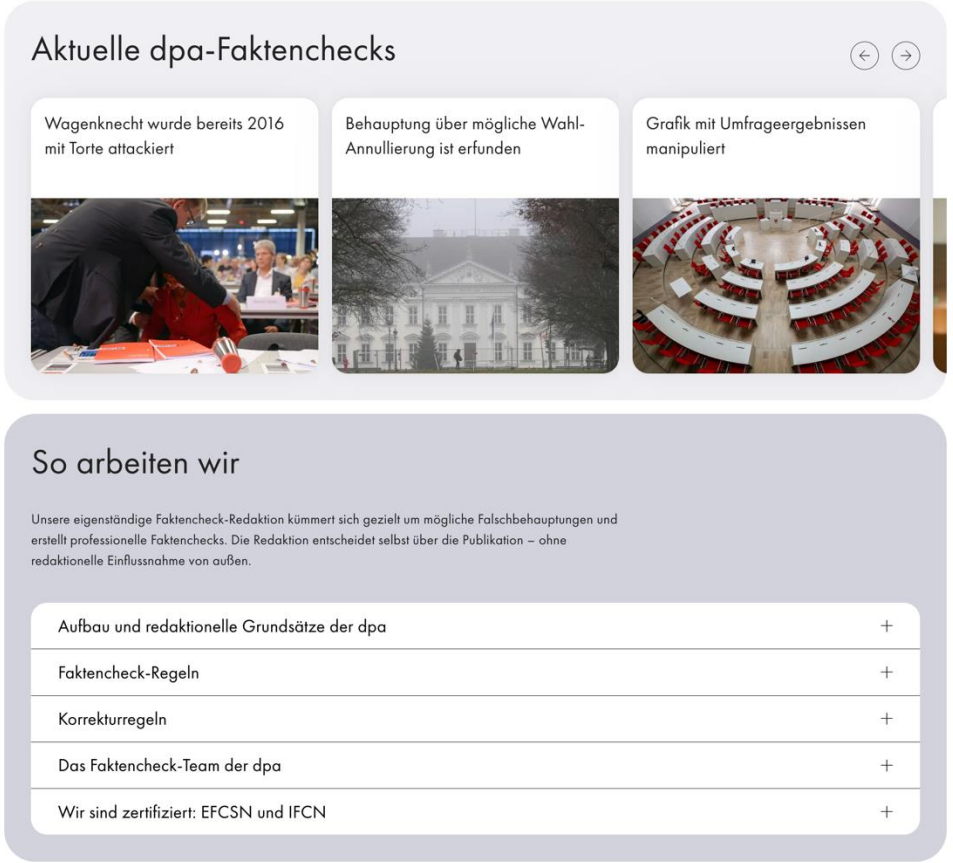
- Schulungsprogram mit über 100 Medienhäusern Deutschlands
- Laufzeit 6 Monate
- Inhalte:
  - KI in Redaktionen
  - KI für Faktencheck



dpa Faktencheck

- Deutschlands größtes Faktencheck-Team

- Zertifiziert / Kodex
- Menschliche Verifikation
- Abläufe, Best Practice
- Vertrauen / 4-Augen-Prinzip
- Effizienzen und Volumen



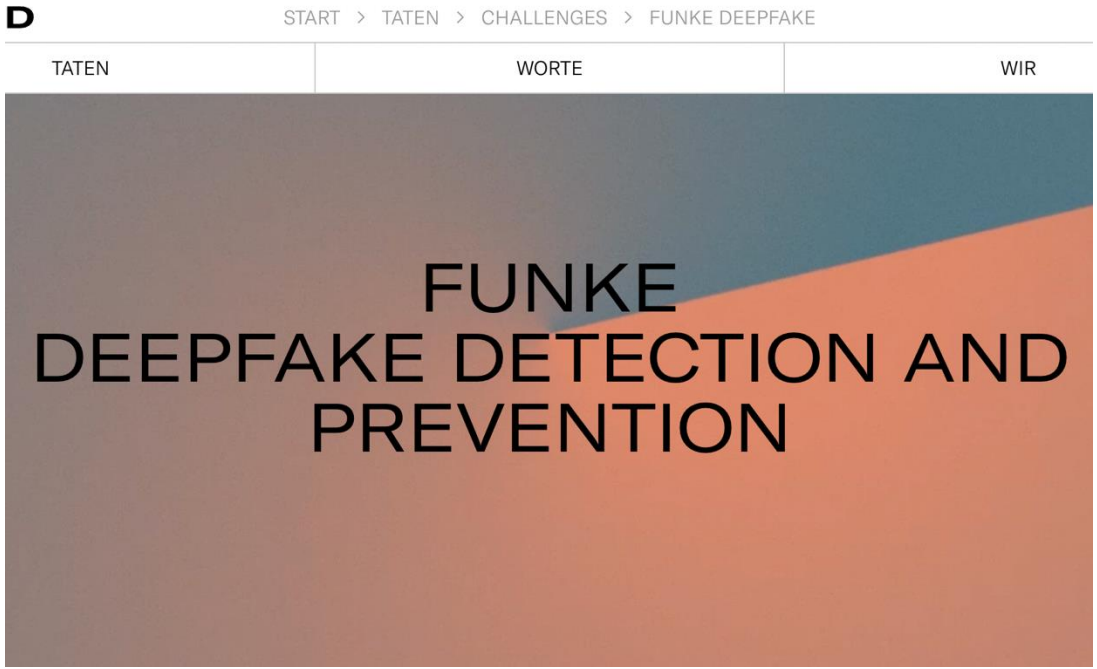
SPRIND

Deepfake Detection Challenge in 2 Tracks

- Track 1: Prevention
- Track 2: Detection

Resultate

- Erfolg in Phase 1 + 2 (Nov. 2025)
- Gretchen AI GmbH Gründung aus DFKI im Nov. 2024
- Performance: Leaderboard Führung seit Aug. 2025



Gretchen AI

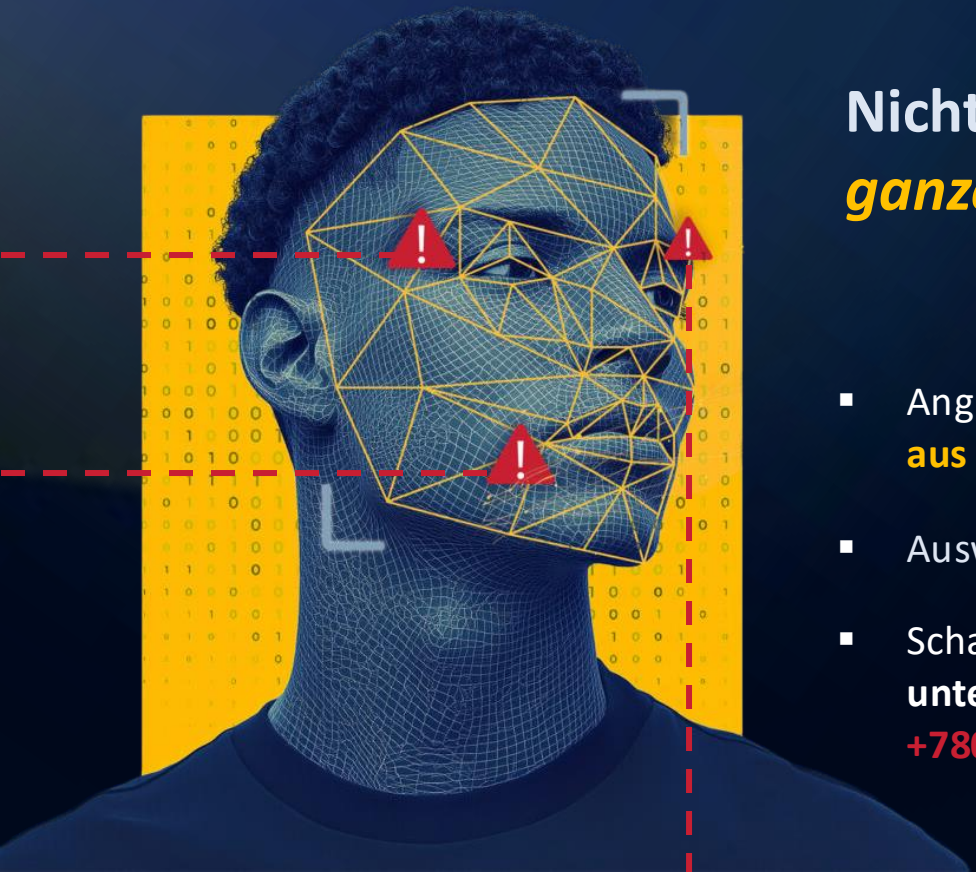
Gretchen AI –

von **manipulierten Pixeln**

zu **verlässlichen Fakten.**

Falschinformationen auf dem Vormarsch!

Gretchen AI



Nicht „nur“ *Deepfake* – sondern die ***ganze Narrative dahinter!***

- Angreifer kombinieren **Deepfake**-Bilder mit **aus dem Kontext gerissenen** Medien
- Auswirkung: **lange, manuelle** Verifikation- sprozesse
- Schaden: Öffentlichkeit **getäuscht,** Entscheidungsträger **untergraben**,
  +780% Betrugsfälle **in EU**

Gretchen AI: Fakes entlarven – Wahrheit beweisen!

Mit KI-Agenten *Narrative* finden und verifizieren!

1. **Bildanalyse**
   Eigene KI Deepfake-Erkennung besser als SOTA
2. **Analyse von Bildkontexten und Narrativen**
   Eigene KI-Agenten extrahieren web-weit Kontext
3. **Redaktions-Dashboard**
   Interagieren & Analysieren: Vertiefen, Belegen, Vertrauen

1. Analyze Image

Gretchen AI

Deepfake Detection Engine

Deepfake Detection Models

Gretchen AI

Eigene kommerzielle Detektionsmodelle für Deepfake und Synthese

01. **Novel architectures (CLIPBi-Mamba, Multi-view)**

02. **Ensemble model (Raw, Spectral, Spatial) & mixture-of-experts**

03. **Novel data augmentation & fine-tuning strategy**

Detection Performance

~97.8% AUC

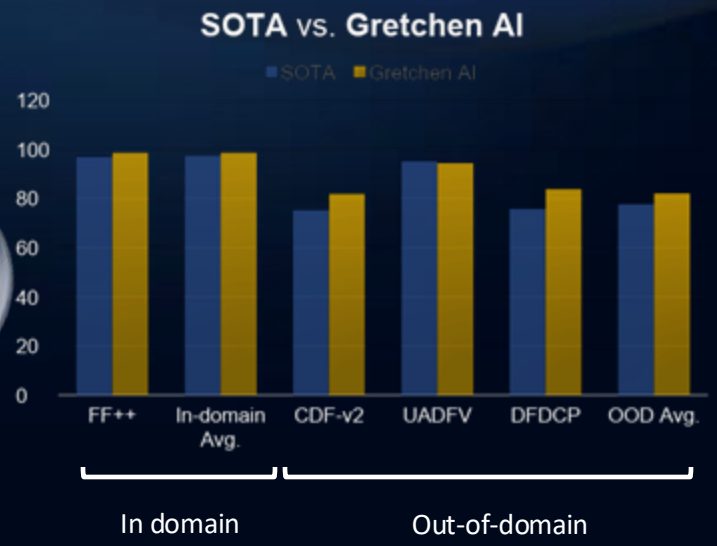
Improving by

>12% over SOTA

On avg. OOD detection

Ultrafast Models:

~1000 img / min



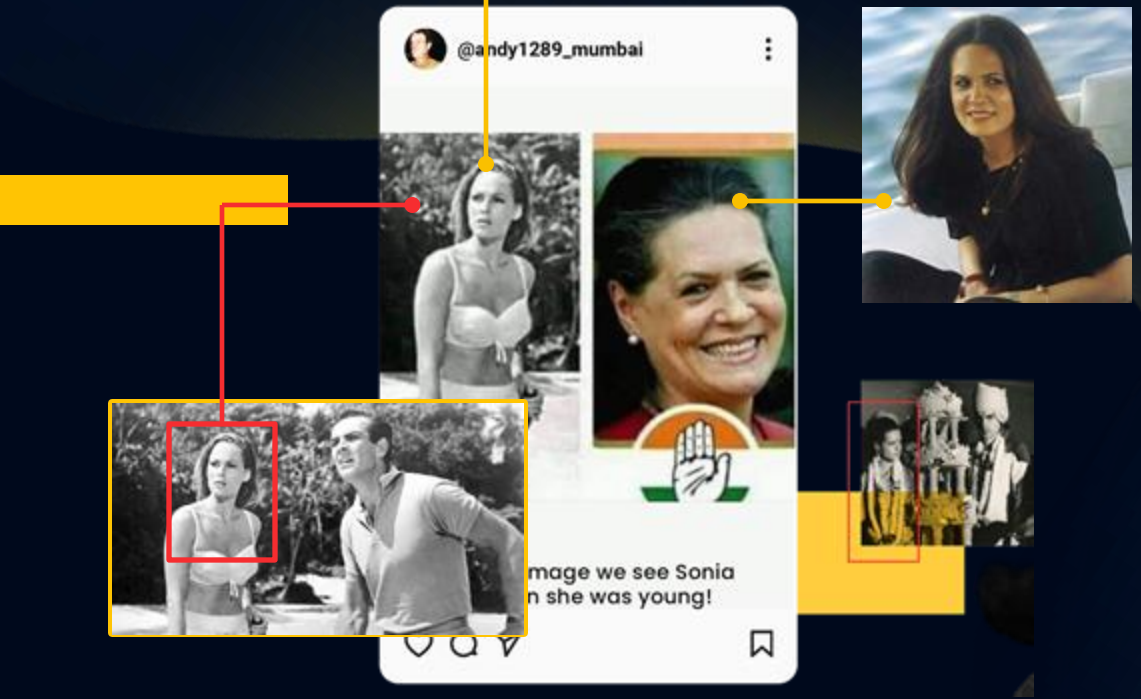
2. Analyze the Context

Gretchen AI

Gretchen Context Engine

Dekontextualisierung

"Is this **Sonia Ghandhi** when she was young?"



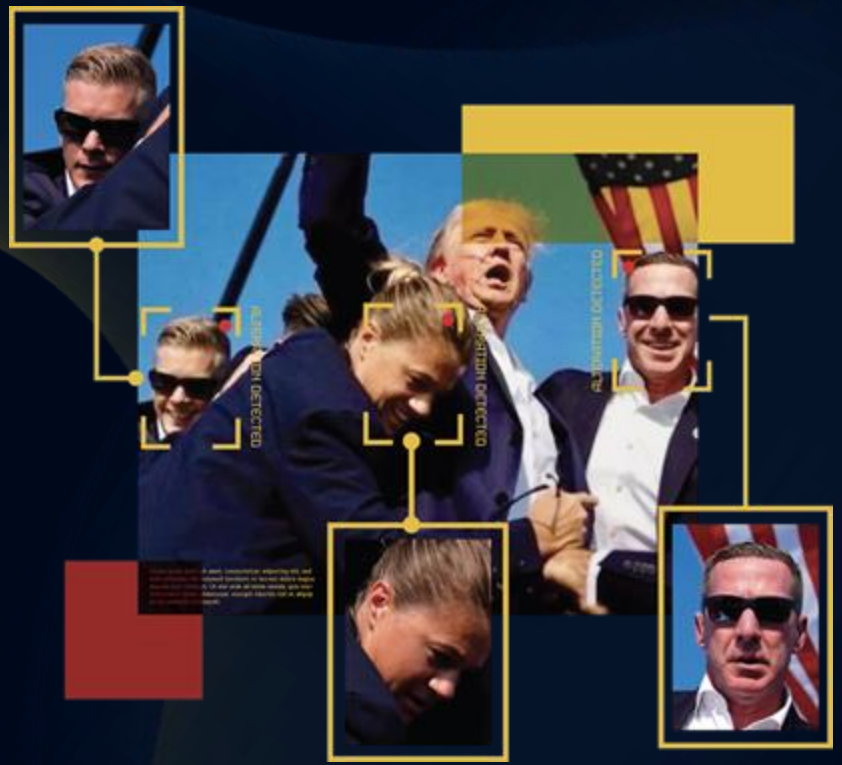
- Aus dem Zusammenhang gerissene Bilder, falscher Kontext
- Oftmals Wiederverwendung des Originalbildes + Deepfakes
- Entlarvung (*Debunk*) durch Herkunftsanalyse (*Provenance*) und semantisch-logische Suche

Medienverifikation im Journalismus: Belegbarkeit?

01. **Deepfake Erkennung**

02. **Frage nach Ursprung und belegbarem Kontext**

- **DALPS**
  - Date
  - Action
  - Location
  - People
  - Source



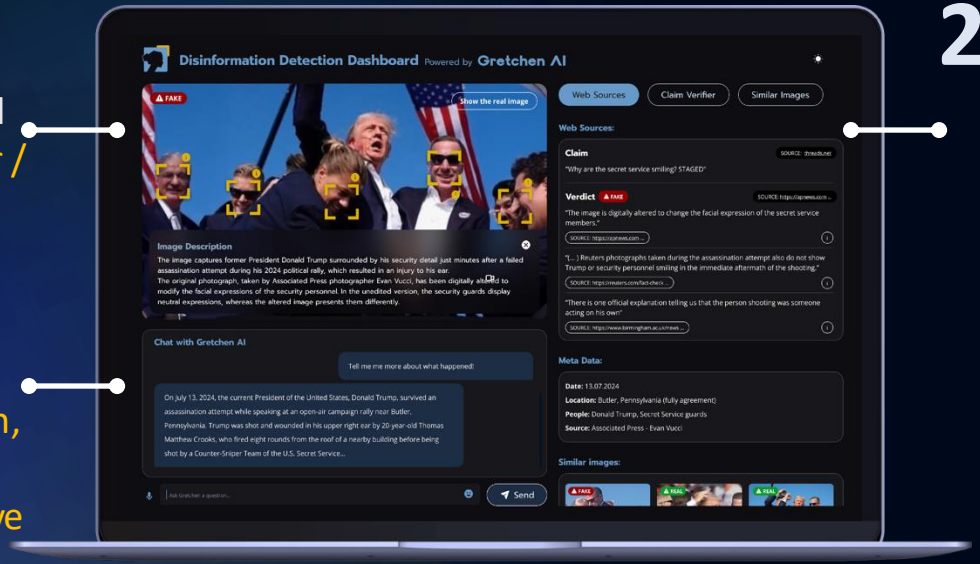
Dashboard Medienverifikation

Gretchen AI

PoC mit Partner *dpa*: ”*500% Zeiteinsparung*” ggü. manueller Verifikation

1. Erkenne
Manipulierung und synthetische Bilder / Video

3. Erstelle
belegbare Bildbeschreibungen, Berichte, und verifizierte Narrative in Sekunden!



2. Enthülle
öffentliche Narrative und Plots, z.B. Datum, Handlung, Ort, Personen, Quelle, etc. aus verlässlichen Quellen

Traction

Gretchen AI

VIVA TECHNOLOGY

“Most Promising Cyber-Security Startup” Auszeichnung auf FGTL Vivatech

Erfolgreiche Partnerschaft mit ***dpa*** *Deutschlands größtem Faktencheck-Team*

1. European Market
2. Branch-off Solutions MVPs

Gründung und **SPRIND** *Deepfake Detection Challenge* **Award**

SPRIN-D

Erfolgreiche Partnerschaft mit **DFKI** *Deutschlands größtem Forschungszentrum für KI*

dfki ai

MvP roll-out ***dpa***, ***BR24 & WDR in PoC***, *RBB DW, etc.*

Team

Gretchen AI

- **Senior AI and Dev. Expertise**
  >30y R&D, >40y AI (top-tier), >80y Dev.
- **Senior Entrepreneurial Expertise**
  3 Startups, >20y Wirtschaft, Pre-Seed Invest.
- **Großes Netzwerk, Exzellente Partner**

mth MEDIATECH HUB ACCELERATOR | BABELSERG
de:hub digital ecosystems
BR24
dpa
transfer media
TECHNISCHE UNIVERSITÄT BERLIN
Ubermetrics A UNICEPTA COMPANY
SPRIN-D
delphai
DW
rbb
dfki
Fraunhofer IDMT


Tim


Arnab


Daniel

C-Level


Jakob


Ayswarya


Lisa

Business


Lisa


Heitke


Sonia

Design, Comm.


Benedict


Rajeshwari


Yassine


Mohamed

AI-Engineers

Gretchen AI

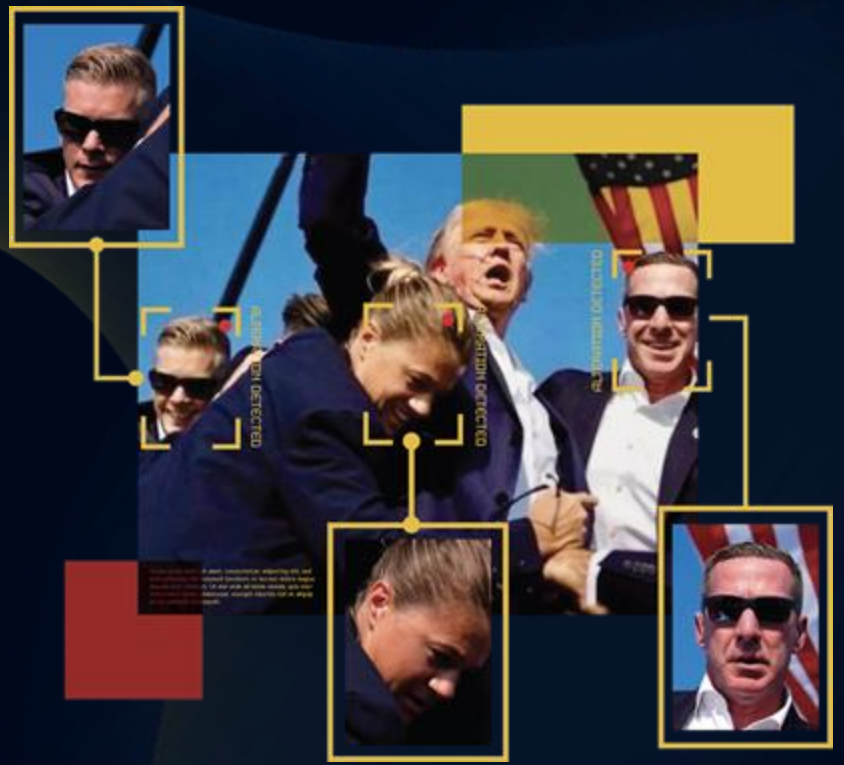
Lunch

Medienverifikation im Journalismus: Belegbarkeit?

01. **Deepfake Erkennung**

02. **Frage nach Ursprung und belegbarem Kontext**

- **DALPS**
  - Date
  - Action
  - Location
  - People
  - Source



Verifikation News / Deepfake in der DZ-Bank Gruppe

01. **Deepfake Erkennung**

02. **Frage nach Ursprung und originalem Kontext**

DALPS

- Date
- Action
- Location
- People
- Source
- → Was noch?!

Use Case #1

"Die Presseabteilung bekommt Fake News zugespielt und reagiert fälschlicherweise darauf, bspw. gibt eine Pressemitteilung zu falschen Fakten bekannt."

Folge: Reputationsschäden?!

Verifikation News / Deepfake in der DZ-Bank Gruppe

01. **Deepfake Erkennung**

02. **Frage nach Ursprung und originalem Kontext**

DALPS

- Date
- Action
- Location
- People
- Source
- → Was noch?!

Use Case #2

"Händler im Haus reagieren auf das Trading Geschehen und verfolgen permanent Nachrichten. Wenn Fake News zu falschen Entscheidungen führt, könnten damit auch falsche Geschäfte getätigt werden."

Folge: wirtschaftliche Schäden?!

Verifikation News / Deepfake in der DZ-Bank Gruppe

01. **Deepfake Erkennung**

02. **Frage nach Ursprung und originalem Kontext**

DALPS

- Date
- Action
- Location
- People
- Source
- → Was noch?!

Use Case #3

“Es könnte sich jemand als Vorstand mittels Deepfake ausgeben und die Presseabteilung zu falschen Pressemitteilungen auffordern, oder interne Prozesse auslösen.”

Folge: wirtschaftliche Schäden + Reputationsschäden?!

Deepfake-Erkennung für Marketing- und Marktbeobachtung

Gretchen AI

Medien → Modalitäten

- **Text**
  - Aussagen, Behauptungen, Konnotationen, etc.
- **Bild / Video**
  - Logo / Marke
  - Produkte / Anzeigen
  - Events
  - Personen (C-Level)

Ergebnis:
neuartiges multimodales Lagebild

Deepfake-Erkennung für Marketing- und Marktbeobachtung

Gretchen AI

Erweiterungen zu DALPS^{++}

01: Neuigkeitswert und Überraschungseffekt

- Märkte reagieren weniger auf Bekanntes, eher auf Unerwartetes?

Deepfake-Erkennung für Marketing- und Marktbeobachtung

Gretchen AI

Erweiterungen zu DALPS^{++}

02: Relevanz für Marktteilnehmer

- Je direkter eine Nachricht die Gewinnerwartung, Zinsen, oder Risiken betrifft, desto größer die Wirkung?

Deepfake-Erkennung für Marketing- und Marktbeobachtung

Gretchen AI

Erweiterungen zu DALPS^{++}

03: Tonlage (Sentiment)

- Nachrichten mit stark positivem oder negativem Sentiment beeinflussen Anlegerstimmung stärker?

Deepfake-Erkennung für Marketing- und Marktbeobachtung

Gretchen AI

Erweiterungen zu DALPS^{++}

04: Emotionalität und visuelle Wirkung

- Nachrichten mit emotionalen Bildern oder Videos (Explosion, CEO-Aussage, Katastrophen- bilder) erzeugen stärkere Marktreaktionen als rein textbasierte Meldungen?

Deepfake-Erkennung für Marketing- und Marktbeobachtung

Gretchen AI

Erweiterungen zu DALPS^{++}

05: Glaubwürdigkeit und Quelle

- Wenn eine Nachricht vermeidlich von anerkannten Medien, offiziellen Stellen oder CEOs stammt, entfaltet sie mehr Wirkung?

Deepfake-Erkennung für Marketing- und Marktbeobachtung

Gretchen AI

Erweiterungen zu DALPS^{++}

0X: Blickwinkel:

- Geo. Märkte (Europa, Asien, etc…)
- Indizes: DAX40, TecDax, SSE, NYSE
- Markt: Deutsche Bank, KfW, UniCredit, Commerzbank, etc.

Gretchen AI

Von manipulierten Pixeln zu verlässlichen Fakten!

Danke!

tim@gretchen-ai.com

CEO Gretchen AI