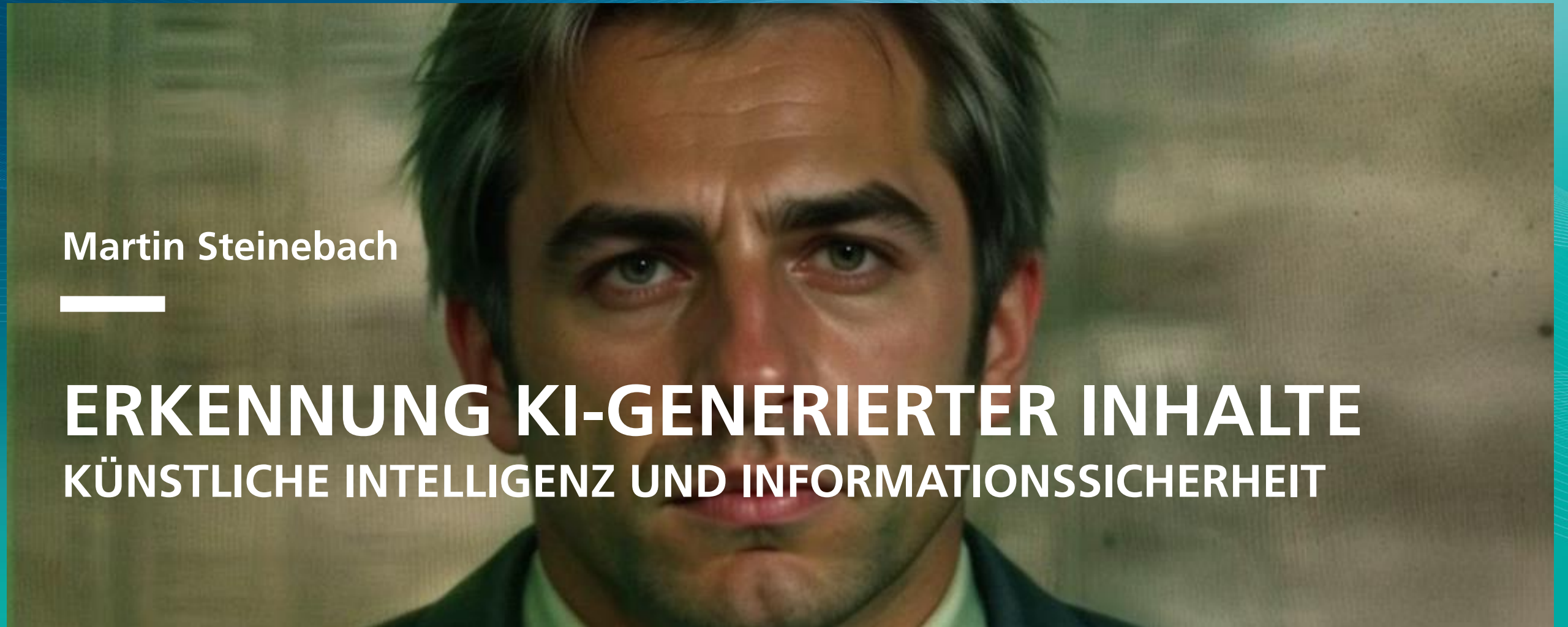




Martin Steinebach

ERKENNUNG KI-GENERierter INHALTE

KÜNSTLICHE INTELLIGENZ UND INFORMATIONSSICHERHEIT

KI-basierte Manipulationen

https://www.reddit.com/r/StableDiffusion/comments/14svyoa/sdxl09_can_it_do_nsfw/



Image

- Neural Editing
- Text-to-Image
- Image-to-Image



Video

- Deepfakes
- Text-to-Video
- Video-to-Video

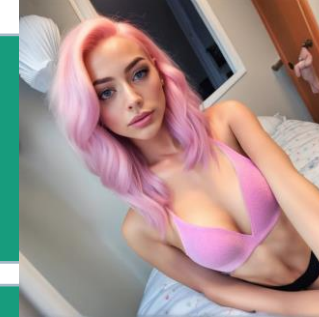


Image-to-Video



Audio

- Text-to-Speech / Text-to-Sound
- Speech-to-Speech
- Voice Cloning

Lipsync

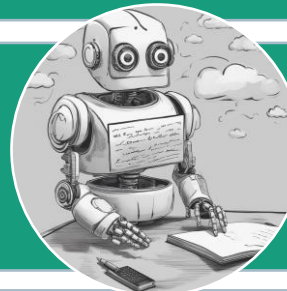


Overdub



Text

- LLMs

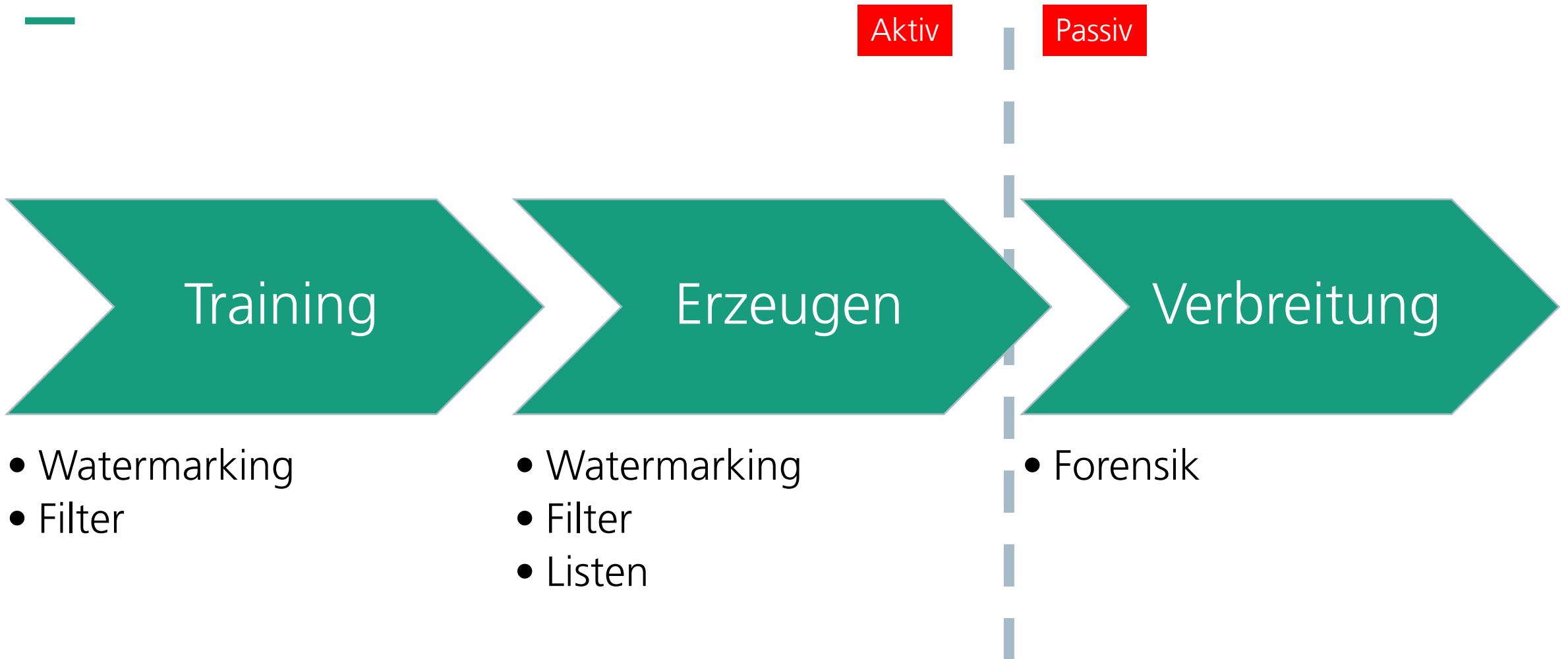


Warum beschäftigen wir uns damit?



- A) Mage.space / stable diffusion
- B) <https://aimojo.pro/top-10-free-undress-ai-tools-exploring-safe-practices-for-image-manipulation>
- C) X/Eliot Higgins
<https://www.20min.ch/story/donald-trump-wurde-mithilfe-kuenstlicher-intelligenz-ki-verhaftet-fake-bilder-auf-twitter-viral-530578576348>
- D) <https://www.vice.com/en/article/g5xagy/this-open-source-program-deepfakes-you-during-zoom-meetings-in-real-time>

Sicherheitskonzepte



Sicherheitskonzepte – Training

Wasserzeichen

- Nicht wahrnehmbare Wasserzeichen
 - Konzept: Rauschen wird auf erzeugte Inhalte übertragen
- Netzwasserzeichen
 - Konzept: Wasserzeichen beeinflusst erstellte Inhalte nachvollziehbar
- Idee: Nicht entfernbare Kennzeichnung

Filterung

- Entfernen problematischer Trainingsdaten durch Filterung
 - Personen des öffentlichen Interesses, NSFW, Gewalt,...
- Idee: Was ein Model nicht gelernt hat, kann es nicht erzeugen



DEEPPFAKE



Training lokal
Gaming Computer
Kosten: wenige Euro

GENAI



Training global
Server Farm
Kosten: Millionen von Euro

Sicherheitskonzepte – Erstellen

Wasserzeichen

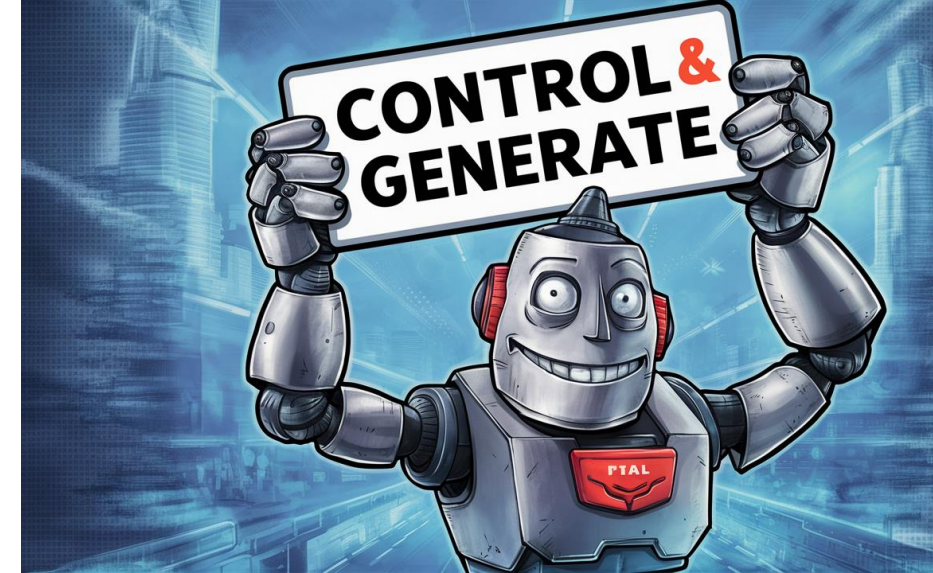
- Einbettung eines Wasserzeichens während oder nach der Erstellung von Inhalten
 - Sichtbare Wasserzeichen (z.B. Logo der Firma)
 - Nicht wahrnehmbare Wasserzeichen

Filtern

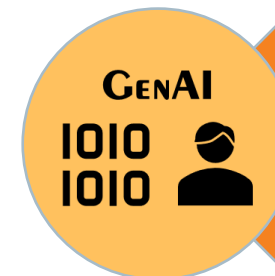
- Erkennung von problematischen Inhalten in Prompts (Text) oder nach der Erstellung (Bild)

Auflistung

- Erstellte Inhalte werden einer öffentlich zugänglichen Liste hinzugefügt
 - Hash (robust, unscharf oder kryptografisch)



Erstellen lokal
Gaming Computer
Kosten: wenige Euro



Erstellen lokal oder in der
Cloud
Gaming Computer
Kosten: wenige Euro

Sicherheitskonzepte – Verbreitung

Forensik

- Erkennung von Erstellungsspuren
 - Upscaling-Muster
 - Hochfrequente Audiomuster
- Erkennung von typischen Stilen
 - NLP ChatGPT-Erkennung (plain vanilla)
- Erkennung von Erstellungsfehlern
 - Unstimmigkeiten bei der Hautfarbe (Deepfakes)
 - Unstimmigkeiten bei der Schärfe (Text zu Bild)

Herausforderungen: Fehlerquoten (Fehlalarme)



Sicherheitskonzepte – Positivlisten

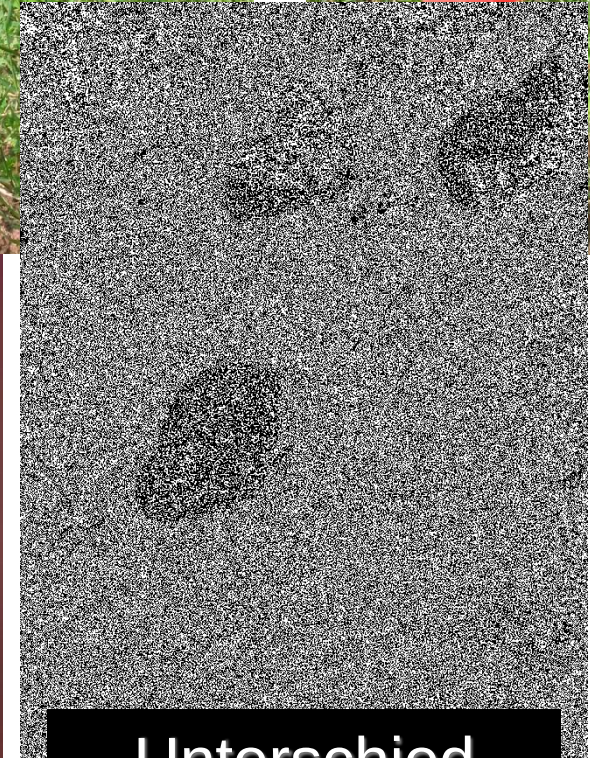
Können als Alternative vertrauenswürdige Inhalte gekennzeichnet werden?

- Zwei Ansätze
 - Vertrauenswürdige Liste
 - Signatur des vertrauenswürdigen Urhebers
 - plus durchsuchbare Abstraktion (~Hash)
 - Reden
 - Nachrichteninhalte
 - Signierter Inhalt
 - „Vertrauenswürdige Kamera“
- Wer entscheidet über vertrauenswürdige Ersteller?
- Wie geht man mit nutzergenerierten Inhalten um?



Digitale Wasserzeichen

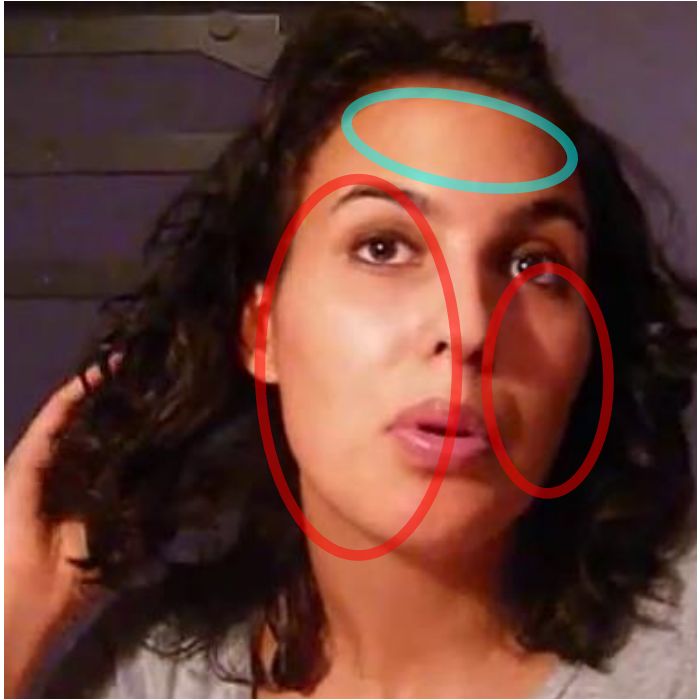
- Wahrnehmbar oder nicht wahrnehmbar
- Nicht.-wahrnehmbare WZ erfordern Detektor-Strategie
- Sicherheitsprobleme im Szenario nicht gelöst



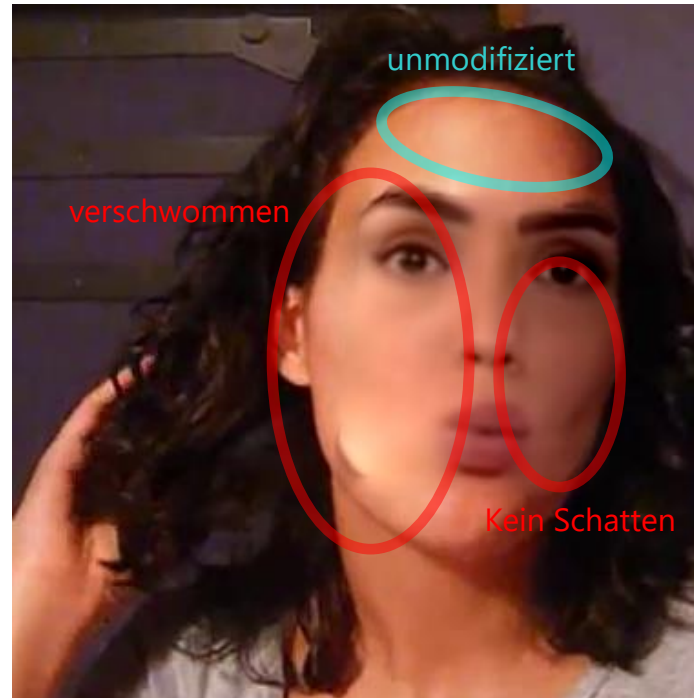
Unterschied
(verstärkt)

Deepfake Erkennung in Gesichtsregion

Original Video Frame



Deepfaked Video Frame



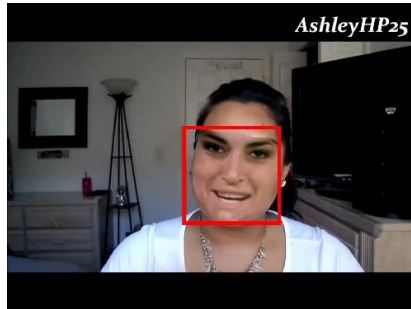
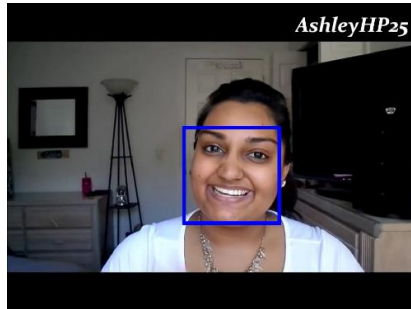
Difference Mask



- Vergleich von Eigenschaften in ähnlichen (benachbarten) Regionen
- Analyse von Textureigenschaften
- 3 bis 5 Frames pro Sekunden können auf Spiele-PC so analysiert werden

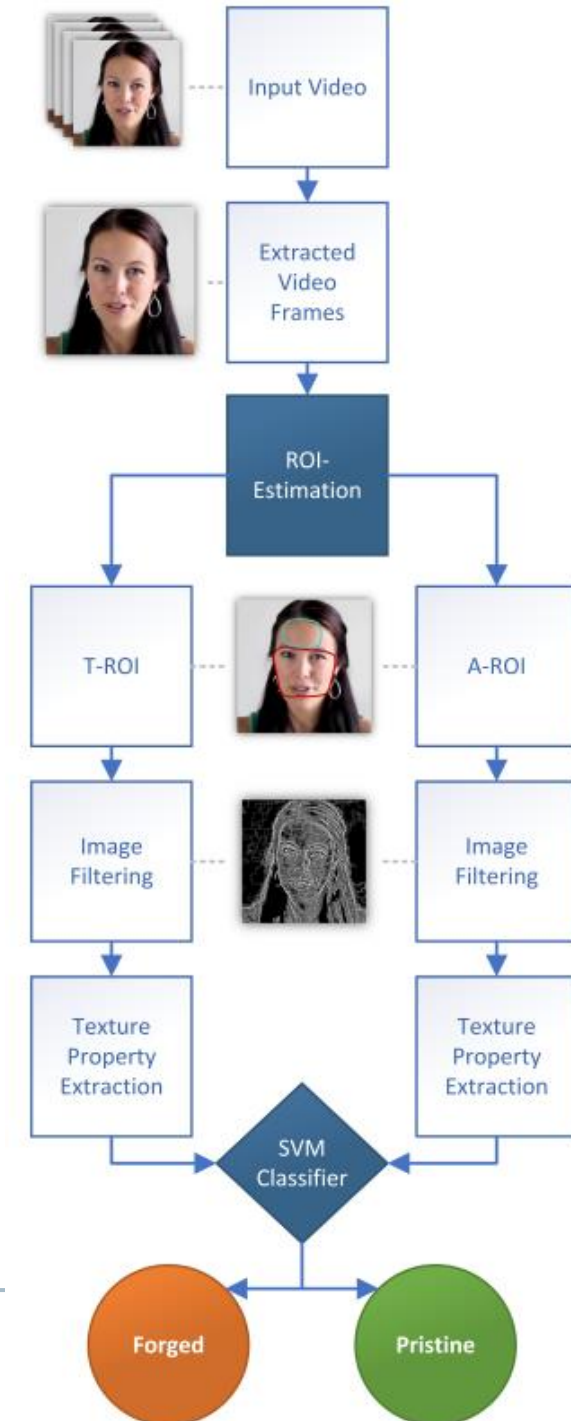
Source: FaceForensics++ Dataset

Deepfake-Erkennung



Auswertung der Kompressionseigenschaften
(blau = authentisch, rot = Deepfake)

Quellen: FaceForensics++ Dataset; Detecting "DeepFakes" in H.264 Video Data Using
Compression Ghost Artifacts, Frick et al.



Beispiel

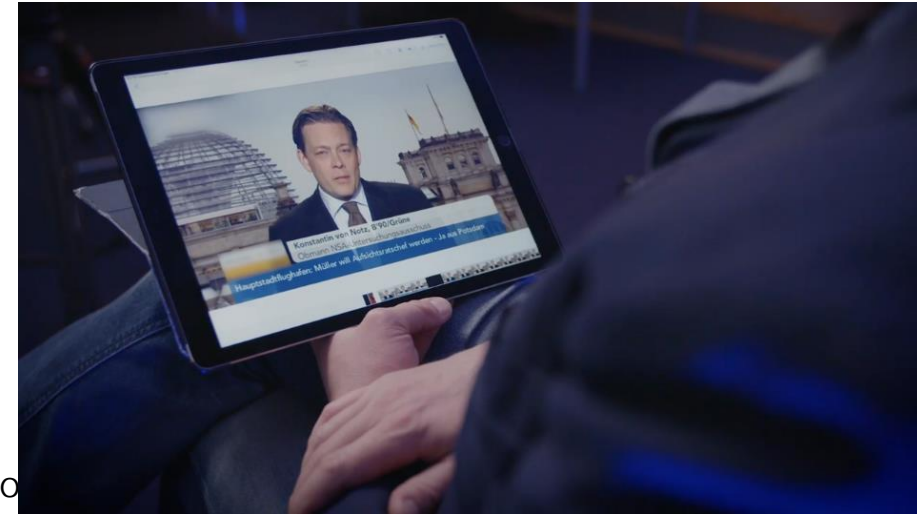


L:80 R:75

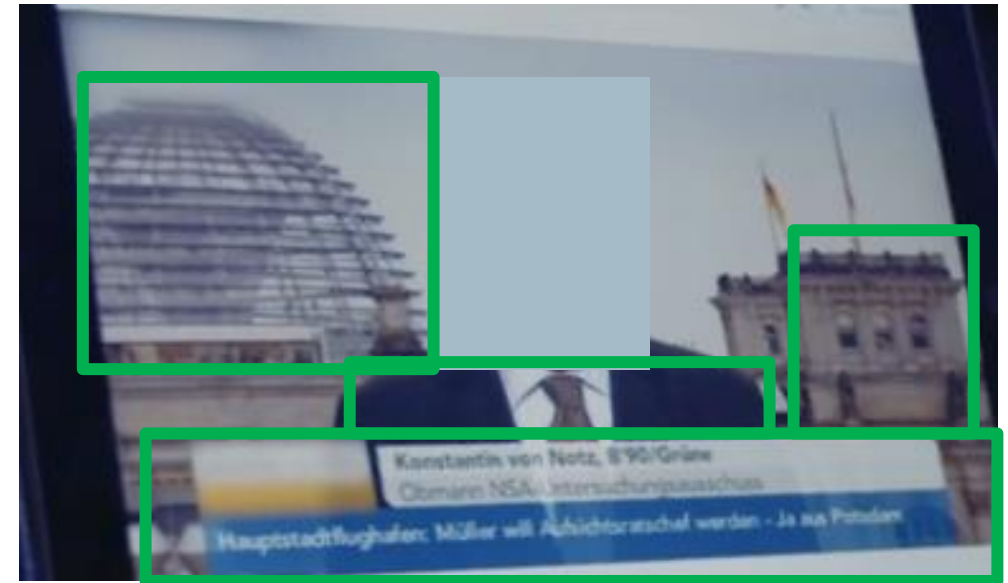
Alternativen zur Deepfake-Erkennung

Erkennen von Quellvideos

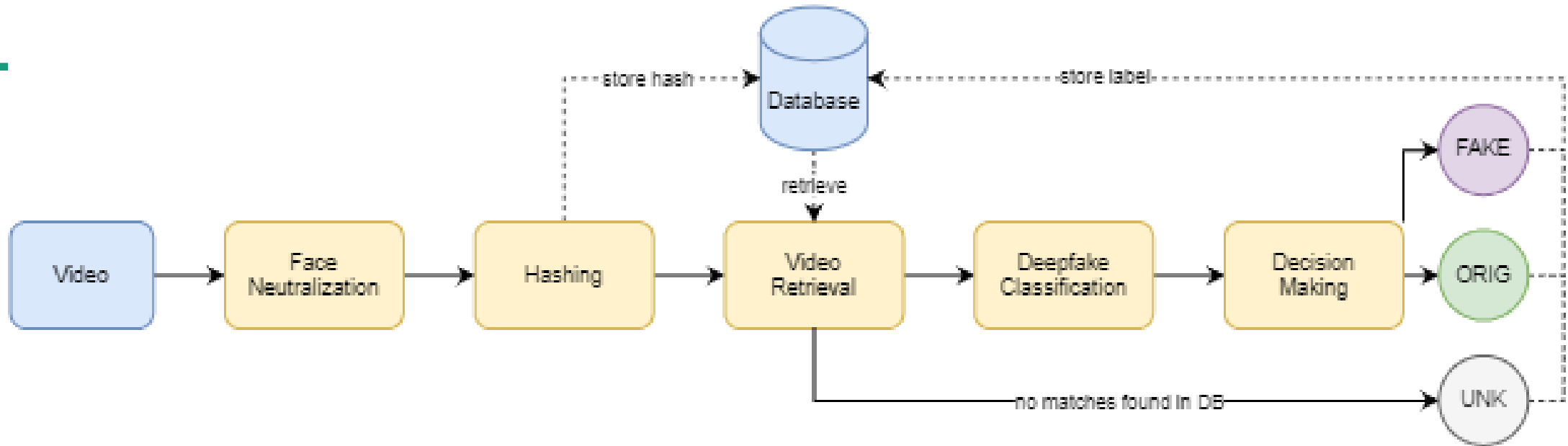
- Geeignet bei
 - Steuerung von Gesichtsausdrücken und Gesprochenem bei vorhandenem Video einer Person
 - Einfügen von Gesichtern in Erotikvideos
- Herangehensweise
 - Bibliotheken von Segmenten von Quellvideos
 - Robuste Hashs
 - Merkmalsextraktion
 - Abgleich von Fake-Video mit Bibliothek



<https://www.thedeepfake.report/stories/the-deep-experiment>



Alternativen zur Deepfake Erkennung



fake	985	15	1000
real	15	985	1000
Σ predicted	1000	1000	2000

Results of the proposed approach, FaceForensics++

	fake	real	Σ actual
fake	943	99	1000
real	57	985	1000
Σ predicted	1000	1000	2000

Results of the built-in deepfake detector, FaceForensics++

Blümer, S., Steinebach, M., Frick, R. A., & Bunzel, N. (2023). Detection of deepfakes using background-matching. *Electronic Imaging*, 35, 381-1.

Beispiel

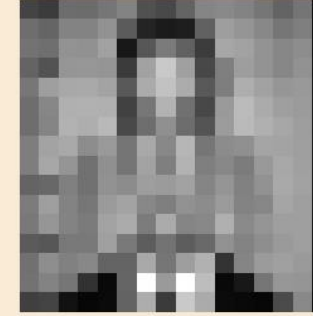
Robuster Hash

Face Swap

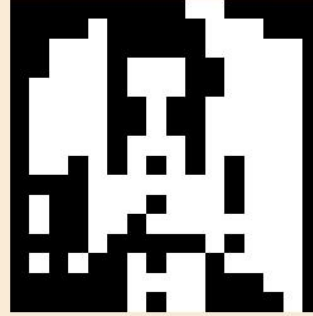
C:/Users/steineba/Downloads/\$8b257956ff0b1ac07843bcbc7f2b8427.jpg



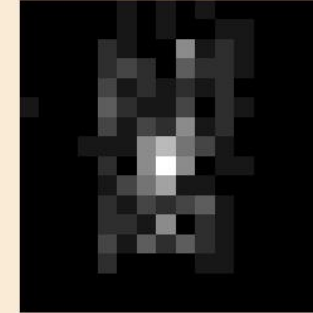
16x16 grayscale image 1



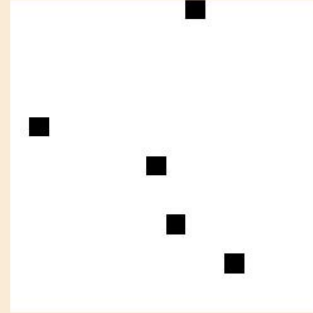
Blockhash image 1



Abs difference image 1 image 2



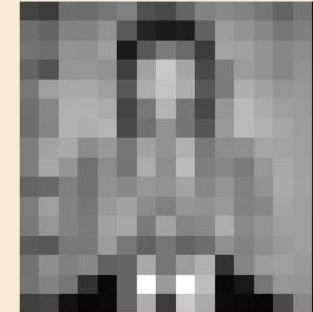
Blockhash difference



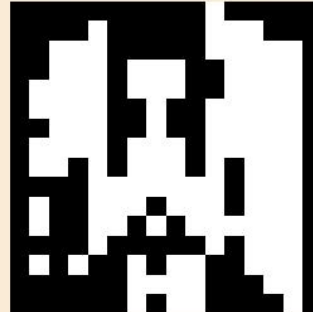
C:/Users/steineba/Downloads/ex_df.jpg



16x16 grayscale image 2



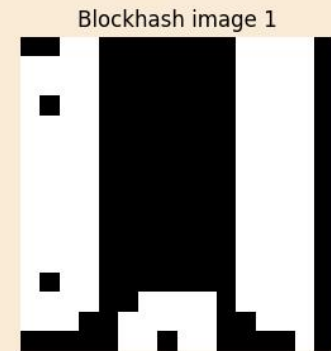
Blockhash image 2



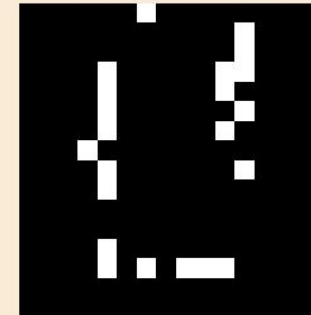
Beispiel

Robuster Hash

Face Swap und Maskierung



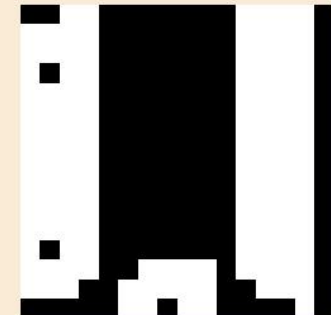
Abs difference image 1 image 2



Blockhash difference



Blockhash image 2



Erkennen von Text-to-Image

- Automatisiert durch forensische Methoden
 - Basierend auf dem Erzeugungsprozess
- Automatisiert durch maschinelles Lernen
- Digitale Wasserzeichen
 - Beim Training
 - Nach dem Erzeugen
- Manuelles Suchen nach Fehlern

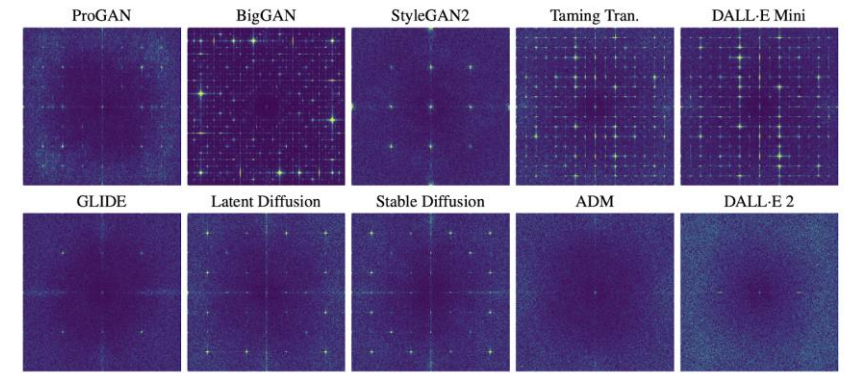


Fig. 2: Fourier transform (amplitude) of the artificial fingerprint estimated from 1000 image residuals. Top row: from left to right ProGAN [20], BigGan [21], StyleGAN2 [22], Taming Transformers [23], DALL-E Mini [24]. Bottom row: GLIDE [5], Latent Diffusion [25], Stable Diffusion [4], ADM [26], DALL-E 2 [3]

ON THE DETECTION OF SYNTHETIC IMAGES GENERATED BY DIFFUSION
MODELS

<https://arxiv.org/pdf/2211.00680.pdf>



Erkennen von Text-to-Image

Beispiele für manuelle Erkennung

Auffällige Augenform bzw. Fehler in der Darstellung von Augen



Erkennen von Text-to-Image

Beispiele für manuelle Erkennung

Anzahl von Fingern nicht korrekt



<https://www.reddit.com/r/StableDiffusion/comments/x9vjka/hand/>



Erkennen von Text-to-Image

Beispiele für manuelle Erkennung

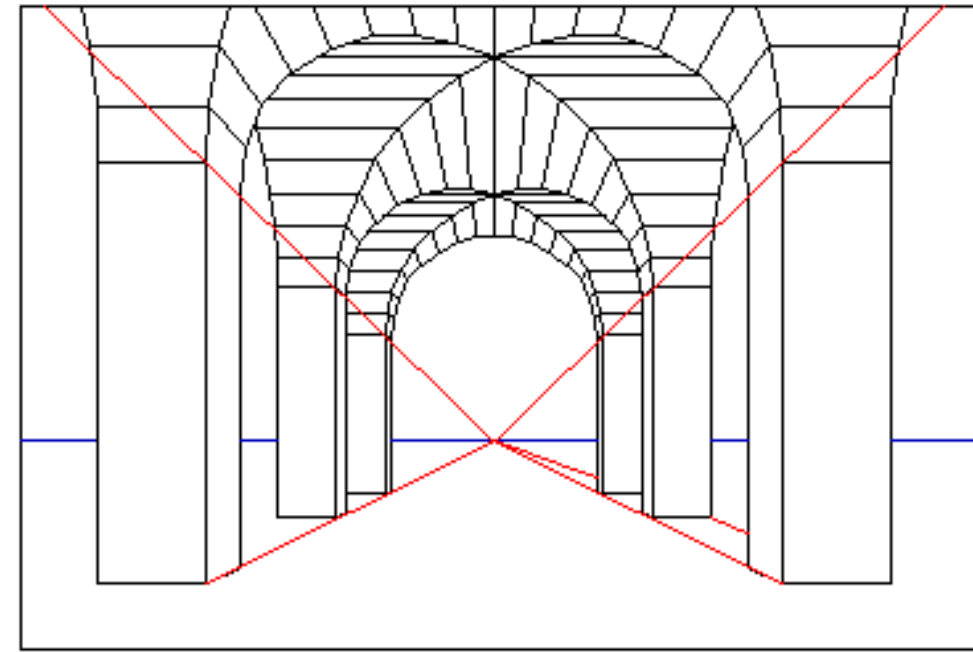
Fehler in
Objektübergängen



Erkennen von Text-to-Image

Beispiele für manuelle Erkennung

Fehler in Fluchtpunkten



<https://commons.wikimedia.org/wiki/File:Zentralperspektive.png>

Erkennen von Text-to-Image

Erkennen von Text-to-Image

Spiegelung nicht korrekt



Erkennen von Text-to-Image

Beispiele für manuelle Erkennung

Fehler in Tiefenschärfe



Hintergrund



Kantenerkennung

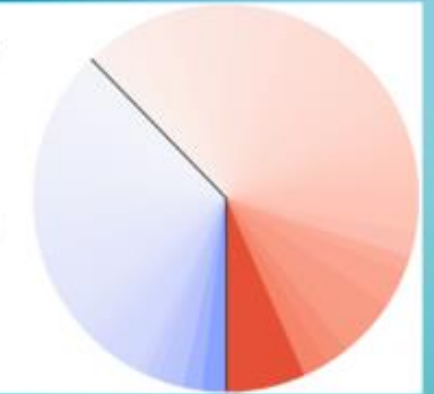
Erkennen von ChatGPT

- Natural Language Processing (NLP) und Forensik kann helfen, künstlich erstellte Texte zu erkennen
- Autorschaftsattributions
 - Lernen, welchen Stil ChatGPT nutzt
 - Textstil vergleichen
 - Aber: Stil kann durch Prompts verändert werden
- Fehlererkennung
 - Fehler, die Menschen machen
 - Fehler, die LLMs machen
- In Summe: Unzuverlässig, erfordert Expertengutachten

Laschet plädierte für einen behutsamen Ausgleich zwischen der Begrenzung von Zuwanderung und den Möglichkeiten der Integration einerseits sowie den Einzelfällen und humanitären Härtefällen andererseits zu finden. Daneben müsse auch der Familiennachzug für diejenigen Flüchtlinge möglich sein, die eine Wohnung und Arbeit in Deutschland hätten. Im März endet die Aussetzung des Familiennachzugs. Für eine Verlängerung ist das Votum des Bundestages nötig. Das Thema dürfte einer der Knackpunkte in den Sondierungsgesprächen zwischen Union und SPD über eine mögliche Regierungsbildung werden, die am 7. Januar beginnen sollen. Flüchtlinge und Zuwanderer, die das Mittelmeer über die Libyenroute erreichten, sollten aus Sicht der AfD ungeachtet lebensgefährlicher Risiken nach Libyen zurückgebracht werden. Menschen, die freiwillig durch Libyen laufen, um nach Europa zu kommen, setzen sich diesen Menschenrechtsverletzungen aus. sagte AfD - Spitzenkandidat Alexander Gauland am Montag in Berlin. Da kann ich nicht plötzlich sagen, 'ja wenn ihr jetzt im Mittelmeer angekommen seid, dann dürft ihr aber nicht über Libyen zurück, es gibt Menschenrechtsverletzungen'. Die Co - Spitzenkandidatin Alice Weidel forderte die sofortige Schließung der Mittelmeerroute. Die Bundesmarine müsse aufgegriffene Flüchtlinge an nordafrikanische Küsten zurückbringen. Beide forderten, das deutsche Recht auf Asyl müsse abgeschafft werden. Die Spitzenkandidaten bekräftigten die von ihrer Partei angestrebte radikale Reform der Asyl- und Flüchtlingspolitik, um den Zuzug von Ausländern nach Deutschland zu blockieren. Dazu sollten Asylzentren in der Nähe der Krisengebiete eingerichtet werden. Diese Zentren müssten von einem Sonderbeauftragten der Europäischen Union mit den betreffenden Ländern ausgehandelt werden. Sollten die Außengrenzen der Europäischen Union nicht gegen Migranten abgesichert werden können, müsse Deutschland seine Grenzen dichtmachen, sagte Weidel. Die AfD - Spitzenkandidaten wollen mit diplomatischem Druck und einer Abschaffung der Entwicklungshilfe in ihrer bisherigen Form die afrikanischen Staaten dazu bringen, zurückgebrachte Flüchtlinge aufzunehmen. Es muss nach deutschen Interessen gehen, sagte Gauland. Hilfen würden demnach nur Staaten gewährt, die kooperationswillig seien. Um diese Ziele zu erreichen, solle unter anderem das Entwicklungsministerium dem Auswärtigen Amt angegliedert werden. Es muss ja nicht immer 'America first' heißen. Es kann ja auch mal 'Deutschland zuerst' heißen, sagte Gauland in Anspielung auf den amerikanischen Präsidenten Donald Trump.



Laschet forderte eine vorsichtige Balance zwischen Einschränkungen bei der Zuwanderung und den Möglichkeiten der Integration auf der einen Seite und den besonderen Umständen und humanitären Ausnahmen auf der anderen Seite zu finden. Zudem müsste es für Flüchtlinge, die in Deutschland Wohnung und Arbeit gefunden haben, die Möglichkeit des Familiennachzugs geben. Im März läuft die Aussetzung des Familiennachzugs aus und eine Verlängerung erfordert die Zustimmung des Bundestages. Dieses Thema dürfte einer der schwierigsten Punkte bei den Verhandlungen zwischen Union und SPD über eine mögliche Regierungsbildung sein, die am 7. Januar beginnen sollen. Die AfD fordert, dass Flüchtlinge und Zuwanderer, die das Mittelmeer über die Route aus Libyen erreicht haben, unabhängig von lebensbedrohlichen Risiken zurück nach Libyen gebracht werden sollten. AfD - Spitzenkandidat Alexander Gauland sagte am Montag in Berlin, dass Menschen, die freiwillig durch Libyen laufen, um nach Europa zu kommen, sich Menschenrechtsverletzungen aussetzen. Die Co - Spitzenkandidatin Alice Weidel forderte die sofortige Schließung der Mittelmeerroute und dass die Bundesmarine aufgegriffene Flüchtlinge an nordafrikanische Küsten zurückbringen sollte. Beide forderten, dass das deutsche Recht auf Asyl abgeschafft werden sollte. Die Spitzenkandidaten bekräftigten ihre Parteiliche Forderung nach einer radikalen Reform der Asyl- und Flüchtlingspolitik, um den Zuzug von Ausländern nach Deutschland zu blockieren. Dies soll durch die Einrichtung von Asylzentren in der Nähe von Krisengebieten erreicht werden, die von einem Sonderbeauftragten der Europäischen Union mit den betreffenden Ländern ausgehandelt werden müssen. Sollte die Absicherung der Außengrenzen der Europäischen Union gegen Migranten nicht möglich sein, müsse Deutschland seine Grenzen schließen, sagte Weidel. Die Spitzenkandidaten der AfD möchten afrikanische Staaten dazu bringen, zurückgebrachte Flüchtlinge aufzunehmen, indem sie diplomatischen Druck ausüben und die Entwicklungshilfe in ihrer bisherigen Form abschaffen. Gauland sagte, dass die Hilfen nur an kooperationswillige Staaten gewährt werden und dass es nach deutschen Interessen gehen muss. Um diese Ziele zu erreichen, soll das Entwicklungsministerium dem Auswärtigen Amt angegliedert werden. Gauland bezog sich auf den amerikanischen Präsidenten Donald Trump und sagte, dass es nicht immer America first heißen muss, sondern auch mal Deutschland zuerst.



Liar's Dividend

Abstreitbarkeit ist ein wichtiges Risiko von Deepfakes



Herausforderungen

Große Datenmengen
beim Training

Strategien

Klassifizieren und Filtern
von Inhalten

Open Source
Implementierungen

Schnelle technische
Entwicklung

Wasserzeichen

Infrastrukturen zur Suche
nach Inhalten

Vermischen von echten
und künstlichen Daten

Forschung zu
forensischen Lösungen

Vielen Dank



Kontakt

Prof. Dr. Martin Steinebach
Head of Media Security and IT Forensics
Tel. +49 6151 869-349
martin.steinebach@sit.fraunhofer.de

<https://www.sit.fraunhofer.de>