Deutsche Stiftung für Recht und Informatik

KI-generierte Fakes – wem müssen wir jetzt vertrauen?

Herstellung und Schutz von Vertrauen in und Vertrauenswürdigkeit von Informationen im Schatten von ChatGPT & Co

Linda Seyda & Tahireh Panahi
Wissenschaftliche Mitarbeiterinnen
Fachgebiet Öffentliches Recht, IT-Recht und Umweltrecht
Universität Kassel

Herbstakademie 2023

Privatheit und Vertrauen für mobile Nutzer

DYNAMO

"Häufig besteht das Ziel weniger darin, eine konkrete falsche Information zu verbreiten.

Vielmehr ist beabsichtigt, das Vertrauen in den Wahrheitsgehalt von Informationen und verlässliche Informationsquellen insgesamt zu unterwandern."

Unzicker, Desinformation: Herausforderung für die Demokratie Einstellungen und Wahrnehmungen in Europa, S.7

Überblick

I. Einführung und Definitionen

II. News-Recap & Law Review

III. Close-Up: Informationsmanipulation durch KI

IV. Close-Up: Vertrauen

V. Rechtliche Lösungsvorschläge und Regulierungsversuche

VI. Fazit

Künstliche Intelligenz (KI)

„Ein maschinengestütztes System, das so konzipiert ist, dass es mit unterschiedlichen Autonomiegraden arbeitet und das zu expliziten oder impliziten Zwecken Ergebnisse wie Vorhersagen, Empfehlungen oder Entscheidungen erzeugen kann, die physische oder virtuelle Umgebungen beeinflussen.“

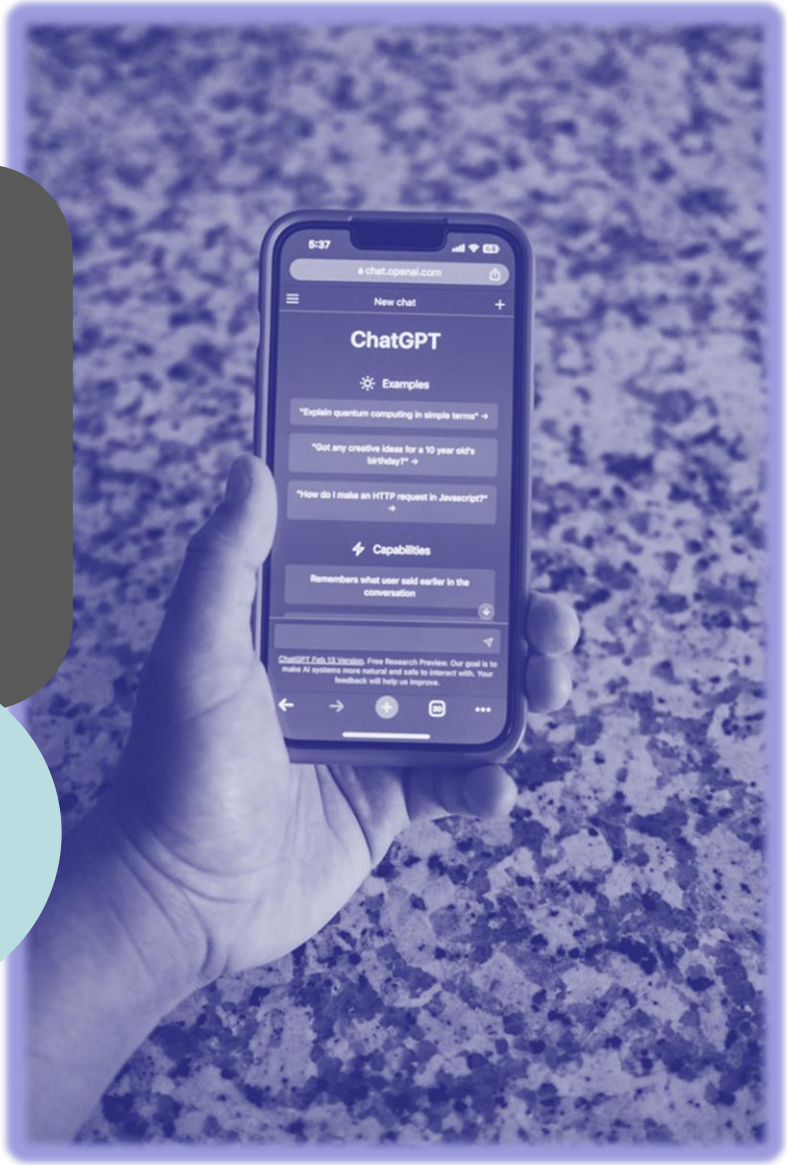
Art. 3 Nr. 1
KI-VO-E

ChatGPT

- generativ vortrainierter Transformator (Sprach-KI)
- trainiert mit Textdaten (Foren, Artikeln, Büchern, gesprochener Sprache)
- Trainingsziel: selbst Text zu generieren
- anthropoid

Veröffentlicht am 30.11.2022

Entwickelt von OpenAI



Generative KI
Technologien, mit denen synthetische Texte, Bilder, Videos, Audios, Codes oder sonstige Daten produziert werden können.

Song “Heart on my sleeve” von Drake feat. The Weeknd (generiert vom TikTok-User @ghostwriter977)

Deepfakes
- Subkategorie von generativer KI
- umfasst täuschend echt wirkende Bild-, Video- und Audiomanipulationen

Ein “falscher Klitschko” kontaktiert Franziska Giffey per Video-Anruf?

News-Recap:

KI macht Schlagzeilen

!!Contentwarnung!!

In diesem Vortrag werden menschenverachtende Inhalte und Verschwörungsnarrative reproduziert.

Die Vortragenden distanzieren sich persönlich ausdrücklich davon.

11.04.2023, 19:03

So will eine KI jetzt die gesamte Menschheit auslöschen

Löscht KI die Menschheit aus? (Symbolfoto) © Dmitry Rukhlenko/imago

An Open

6 months the

ChatGPT:

Law-reviewed

ChatG
Nun r
28.04.2023
Im März ha
ChatGPT v
wieder zu

Council of the European Union

Brussels, 10 July 2023
(OR. en)

11320/1/23
REV 1

LIMITE

TELECOM 221
JAI 955
COPEN 237
CYBER 174
DATAPROTECT 191
EJUSTICE 28
COSI 129
IXIM 150
ENFOPOL 320
RELEX 828
MI 584
COMPET 717
CODEC 1255

PUBLIC

Interinstitutional File:
2021/0106(COD)

NOTE

From:	Presidency
To:	Delegations
No. Cion doc.:	8115/21
Subject:	Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - Preparation for the trilogue

DOCUMENT PARTIALLY ACCESSIBLE TO THE PUBLIC (02.08.2023)

INTRODUCTION

1. The Commission adopted the proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) on 21 April 2021.

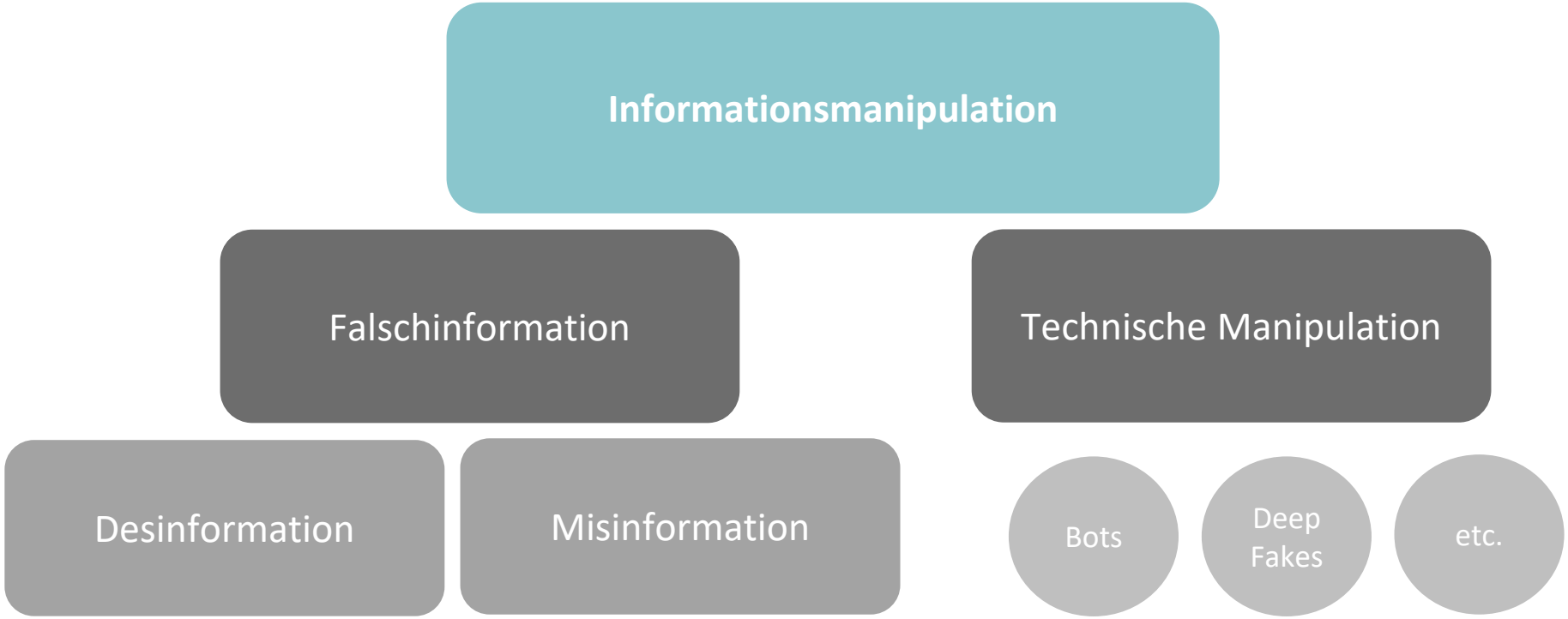
2. The Council unanimously adopted its General Approach on the proposal on 6 December 2022.

3. The European Parliament (hereinafter: EP) confirmed its position in a plenary vote on 14 June 2023.

oar –
gen die KI-Anwendung
t hat, ist er in Italien nun

Close-up:

Informations-manipulation durch KI


Informationsmanipulation
Falschinformation
Technische Manipulation
Desinformation
Misinformation
Bots
Deep Fakes
etc.

Informationsmanipulation durch ChatGPT

Misinformation

- Falsche Korrelationen und Kausalitäten
- Ergebnisse, die erfunden wirken

Desinformation

- falsche Informationen zu einem bestimmten Zweck erzeugen und z.B. in einem anderen Kontext veröffentlichen
- in wenigen Augenblicken ganze Kampagnen
- Sprachstil des Erstellers verschleiern; Sprachstile imitieren

- Einige Prompts, die auf Verschwörungsnarrative und Desinformation hindeuten, werden nicht ausgeführt
- Hasserfüllte Inhalte sollen nicht generiert werden
- Prompt „fiktiv“ führt dennoch zu gewünschten Inhalten
- Kein Schutz vor gänzlich neuen Theoriebausteinen

Ausreichende Schutzmaßnahmen in ChatGPT?

Vertrauenserosion?

Neue Einsatzmöglichkeiten von KI führen zu:

- unabsichtlicher Fehlinformation
- bewussten Täuschungsversuchen
- gesamtgesellschaftlichem Vertrauensverlust

Desinformation und Deepfakes bergen erhebliche Risiken für das Vertrauen in Mensch und Maschine.

Close-up:

„Vertrauen“

Was ist Vertrauen?

Der Zustand, in dem jemand auf jemand anderen oder etwas bauen kann; die Sicherheit, sich auf jemanden oder etwas verlassen zu können (Duden).

Wie ist Vertrauen rechtlich einzuordnen?

Zivilrecht

- § 242 BGB
- § 122 Abs. 1 BGB (Vertrauensschaden)

Strafrecht

- Abgrenzung bewusste Fahrlässigkeit vs. Eventual-Vorsatz

Öffentliches Recht

- § 48 Abs. 2 Satz 3 VwVfG

Vertrauen und Recht

- Vertrauen in Personen, Sachen oder in einen Erfolgseintritt oder Zustand in nahezu allen Rechtsgebieten
- Aber keine Legaldefinition

Vertrauen in der juristischen Literatur

- Rekurriert vor allem auf soziologische Erkenntnisse, insbesondere von *Luhmann*

Luhmann: „Ohne jegliches Vertrauen aber könnte er morgens sein Bett nicht verlassen. Unbestimmte Angst, lähmendes Entsetzen befielen ihn“. Daraus resultiere eine „unvermittelte Konfrontierung mit der äußersten Komplexität der Welt“, welche kein Mensch aushalten könne“

Persönliches Vertrauen und Systemvertrauen

Fuhrmann: Komplexitätsreduktion in Zeiten stetigen Wandels

Jandt: Mechanismus zur Stabilisierung unsicherer Erwartungen

Mischung aus Wissen und Nichtwissen

Abhängigkeit von der Vertrauenswürdigkeit eines Bezugsobjekts

Vertrauenswürdigkeit in der Informationskrise
- philosophischer Exkurs

O'Neill, 2004: Entwicklung und Ablehnung von Vertrauen ist davon abhängig, ob Informationen auf bestimmte Quellen zurückzuführen sind. Die größte Bedrohung für Vertrauen ist Täuschung.

(P) Bewusstsein über massenhafte Verbreitung von Desinformation, KI, Bots

(P) Soziale Medien mit Akteur:innen, die keinen journalistischen Qualitätskriterien unterliegen

Aber: Faktenchecker:innen

Herstellung und Schutz von Vertrauenswürdigkeit durch das Recht

- Nationales Prozessrecht: **Glaubwürdigkeit**
  - Dient lediglich mittelbar dem Schutz der Vertrauenswürdigkeit des Justizsystems

- Neue **EU-Digitalrechtsakte** wollen rechtliche Bedingungen für vertrauenswürdige Technologien schaffen.

KI-VO-E
Vertrauens-
würdige KI

DSA
Vertrauens-
würdiges
Online-
Umfeld

Rechtliche Lösungsvorschläge und Regulierungsversuche

(P) Ausnahme für die Ausübung der Grundrechte auf freie Meinungsäußerung und auf Freiheit der Kunst und Wissenschaft?

Digital Services Act (DSA)

- Art. 34 Abs. 1 lit. b DSA: Risikobewertungspflicht für sehr große Online-Plattformen und Online-Suchmaschinen
  - jedenfalls vor der Einführung neuer Funktionen (wozu auch KI gehört)
  - „Risiken“ umfasst auch Desinformation
  - Beeinflussung durch vorsätzliche Manipulation, unauthentische Verwendung oder automatisierte Ausnutzung des Dienstes (S.2)

- (P) ChatGPT ist keine Online-Suchmaschine i.S.d. Art. 3 lit. j, Art. 33 Abs. 1 DSA
  - Kein Zugriff auf „prinzipiell alle Websites“; ChatGPT berücksichtigt nur Trainingsdaten bis zu einem bestimmten Zeitpunkt

Digital Services Act (DSA)

- **Vertrauenswürdige Hinweisgeber: innen, Art. 22 Abs. 1 DSA**
  - Neue rechtliche Figur mit eigenem Rechte- und Pflichtenkreis
  - Vorrangstellung bei der Meldung rechtswidriger Inhalte

- Kodifikation von Vertrauenskriterien, Art. 22 Abs. 2 DSA
  - besondere Sachkenntnis und Kompetenz in Bezug auf die Erkennung, Feststellung und Meldung rechtswidriger Inhalte (lit. a)
  - Unabhängigkeit von Anbieter: innen von Online-Plattformen (lit. b)
  - sorgfältige, genaue und objektive Übermittlung von Meldungen (lit. c)

Media Freedom Act –E

Ziel

- Zugang zu hochwertigen Mediendiensten und vertrauenswürdigen Informationen, EG 11 S. 2
- Freier Fluss vertrauenswürdiger Informationen für Binnenmarkt der Mediendienste von wesentlicher Bedeutung, EG 13

Inhalt

- Sicherstellung der Unabhängigkeit und Pluralität von Mediendiensten
- Keine Sorgfalts- und Wahrheitsprüfungspflichten, wie z.B. § 19 MStV

VO über die Transparenz und das Targeting politischer Werbung –E

Ziel

- Schutz des Vertrauens in politische Anzeigen und der Integrität des demokratischen Prozesses

Inhalt

- Targeting- und Amplifizierungstechniken i.S.d. Art. 2 Nr. 8 des Entwurfs können durch KI verstärkt werden
- Transparenzpflichten für politische Werbung (Art. 6), Einschränkung von Mico-Targeting auf Grundlage besonderer Kategorien personenbezogener Daten i.S.d. Art. 9 DSGVO (Art. 12 Abs. 1)

EU-Verhaltenskodex gegen Desinformation

- **Vertrauensförderung in Commitments Nr. 1, 3, 15, 22**
- **eher vage ausgestaltet**
- **Ko-Regulierung: Bindungswirkung nur in Ausnahmefällen (i.V.m. Art. 45 Abs. 2, Art. 74 Abs. 1 lit. a DSA); ansonsten freiwillig**
- **Beitritt zum Verhaltenskodex als Vertrauenswürdigkeit fördernde Eigenschaft**

Fazit

- **KI führt zu einer immer künstlicheren Umgebung, in der Wahrheits- und Vertrauensfragen immer häufiger gestellt werden**
  - **Informationskrise: Vertrauenswürdigkeit von KI-Systemen, Nutzer: innen der KI-Systeme und durch diese hergestellten Informationen steht infrage**
    - **Die EU will das durch Falschinformationen erschütterte Vertrauen in mehreren Rechtsakten und Verordnungsentwürfen schützen**
    - **Vertrauenswürdigkeitskriterien für Hinweisgeber: innen**
      - **Aber: rechtliche Regulierung hinkt der technischen Entwicklung hinterher**
- **Trilog zur KI-VO-E nicht abgeschlossen:**
  - ➢ **Vorschlag: Ausnahme von der Transparenzpflicht für Deepfakes wegen Meinungsfreiheit einschränken**

Referenzen?

siehe Tagungsband oder nächste Folie

Vielen Dank für Ihre Aufmerksamkeit!

Referenzen News-Recap und Law-Review:

- https://www.spiegel.de/netzwelt/apps/chatgpt-scheitert-an-bayerischem-abitur-a-a28cdaa6-cdb9-49a1-9d38-b85e6c24773d Spiegel v. 10.2.2023, ChatGPT scheitert an bayerischem Abitur
- https://www.stern.de/digital/-beaengstigend---chatgpt-besteht-im-zweiten-anlauf-bayerisches-abitur-33499852.html Stern v. 25.5.2023, Haddadian, J., ChatGPT besteht im zweiten Anlauf bayerisches Abitur: Darum ist die KI jetzt "schlauer"
- https://www.ndr.de/nachrichten/info/Hamburg-Mit-Chat-GPT-im-Abi-gemogelt,ndrinfo46278.html , NDR-Info, Sendung v. 26.5.2023, Hamburg: Mit Chat-GPT im Abi gemogelt
- https://www.handelszeitung.ch/tech/dieses-bild-von-papst-franziskus-geht-viral-es-ist-ein-ki-fake-587434 , Handelszeitung v. 27.3.2023, Ruffiner, O., Dieses Fake-Papstbild geht zurzeit viral
- Eliot Higgins, Twitter v. 20.3.2023
- https://www.faz.net/aktuell/politik/inland/afd-mit-ki-fotos-abgeordnete-der-partei-rechtfertigen-taeuschende-bilder-18788651.html , FAZ v. 31.3.2023, Haupt, Friederike, Die AfD macht Stimmung mit Fotos, die keine sind
- https://www.derstandard.de/story/3000000182075/ki-kann-passwoerter-allein-am-geraeusch-der-tastatur-erkennen Der Standard v. 7.8.2023, KI kann Passwörter allein am Geräusch der Tastatur erkennen
- https://futureoflife.org/open-letter/pause-giant-ai-experiments/ future of life v. 22.3.2023, Pause Giant AI Experiments: An Open Letter
- https://www.fr.de/wissen/ki-kuenstliche-intelligenz-ende-menschheit-oxford-studie-liefert-duestere-antwort-91821912.html , Frankfurter Rundschau v. 6.10.2022, Wendt, J.-F., Bringt KI das Ende der Menschheit? Oxford-Studie liefert eine düstere Antwort
- https://www.heute.at/s/so-will-eine-ki-jetzt-die-gesamte-menschheit-ausloeschen-100265464, Heute v. 11.4.2023, So will eine KI jetzt die gesamte Menschheit auslöschen
- https://www.dw.com/de/chatgpt-wieder-in-italien-verf%C3%BCgbar-nun-mit-alterspr%C3%BCfung/a-65468909 DW v. 28.4. 2023, ChatGPT wieder in Italien verfügbar – Nun mit Altersprüfung