

The Power

Die Frauen Union CDU-Rhein-Sieg lädt ein:

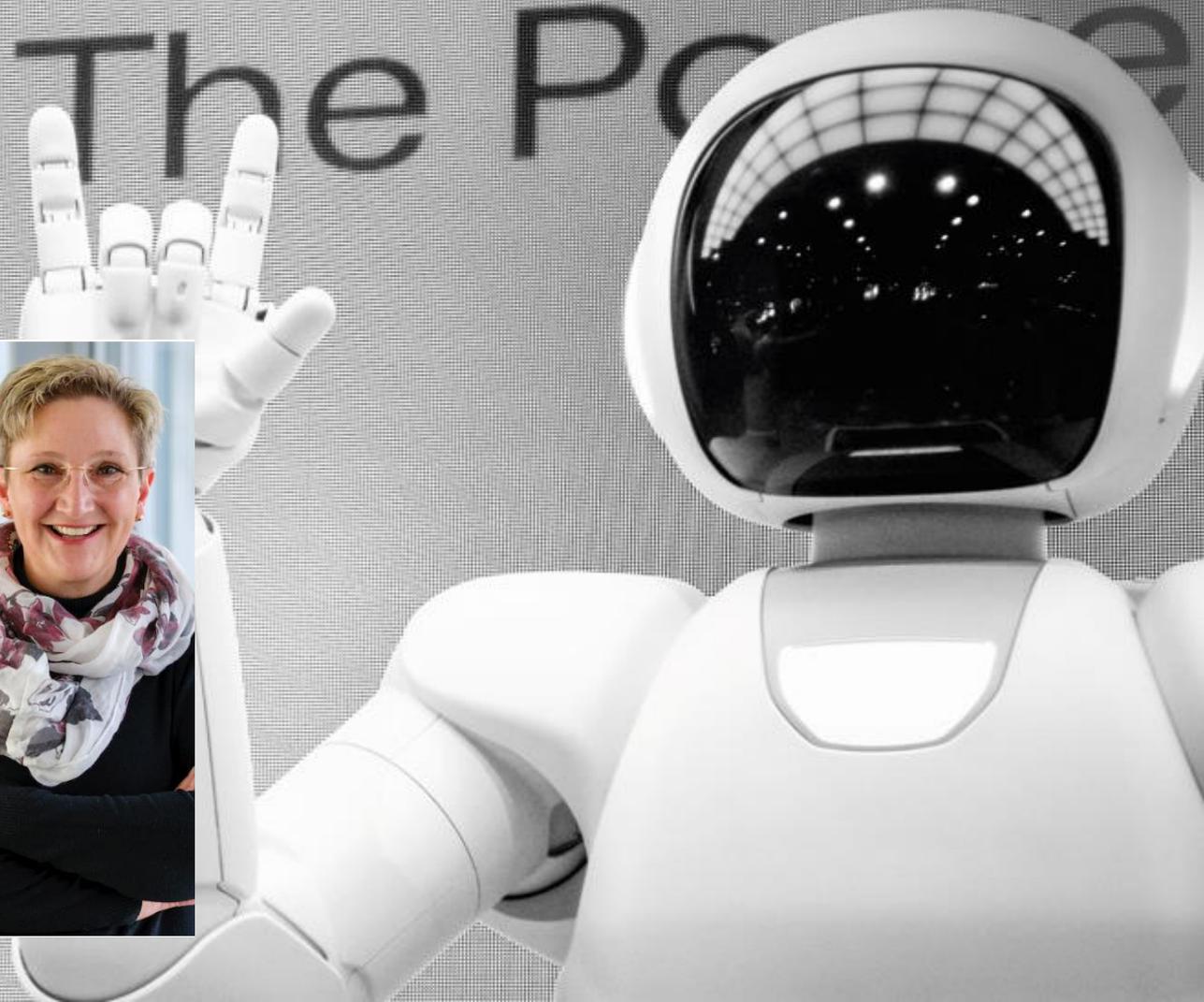
## KI - gekommen, um zu bleiben

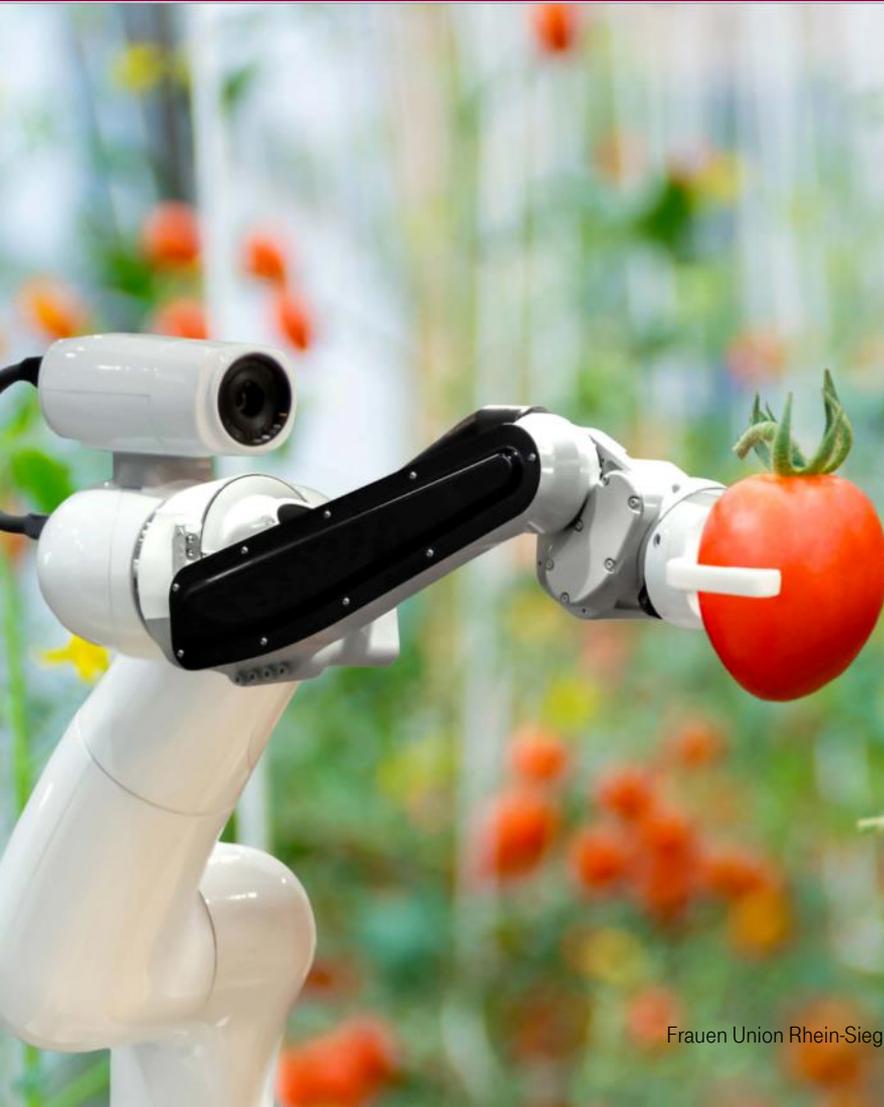
Ein aktueller Blick auf die  
Entwicklungen rund um  
das Thema  
Künstliche Intelligenz

Darüber sprechen wir mit  
Maike Scholz  
Expertin Digitale Ethik



**Digitales Themengespräch** | 9. April 2024 | 19.30 Uhr  
Anmeldungen: [frauenunion-rhein-sieg-kreis@t-online.de](mailto:frauenunion-rhein-sieg-kreis@t-online.de)





## Agenda:

### 1. Wie funktioniert Künstliche Intelligenz?

- Algorithmen, Daten, KI, LLMs und GenAI

### 2. Herausforderungen der Technologie

- Künstliche Intelligenz
- Generative KI

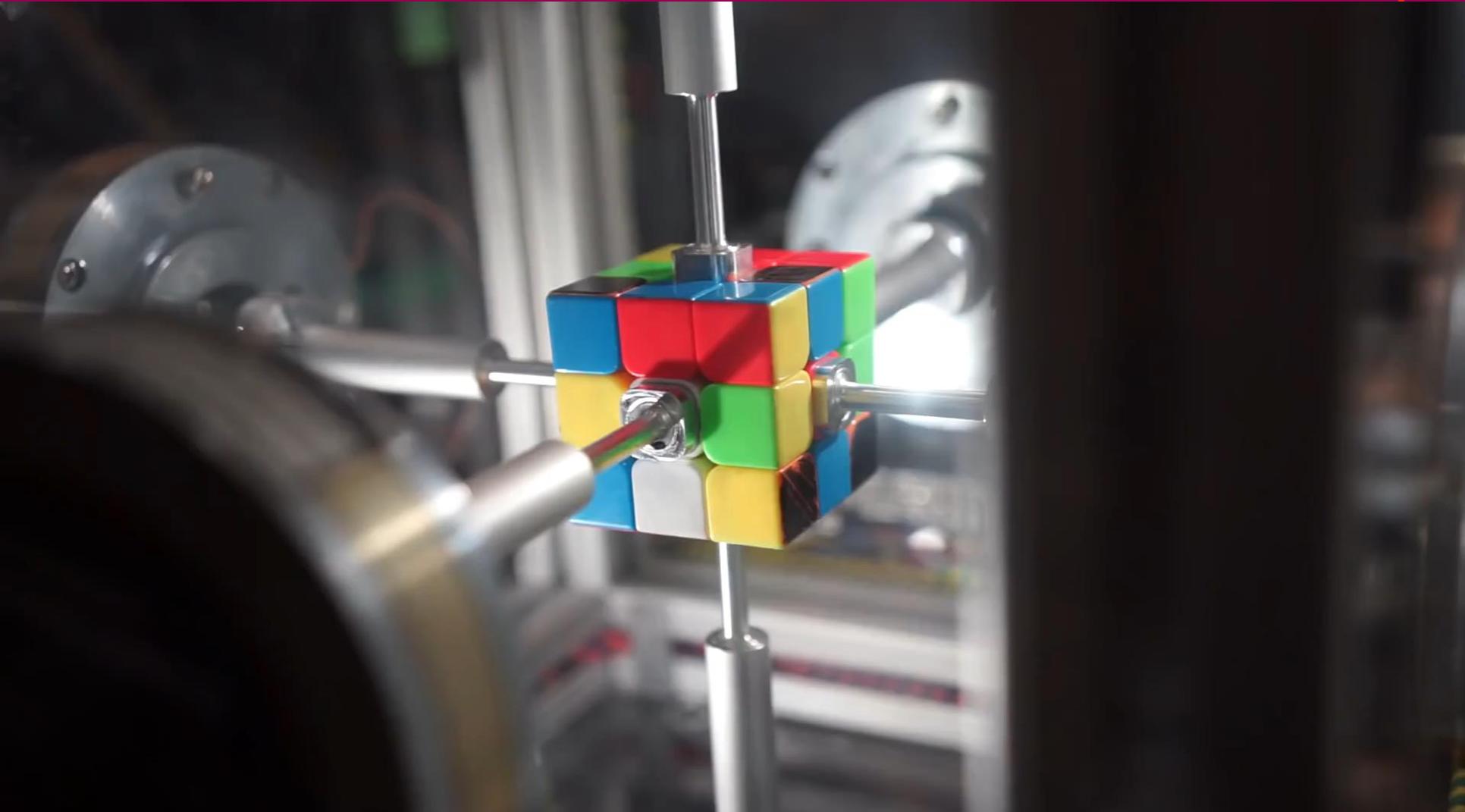
### 3. Entwicklung

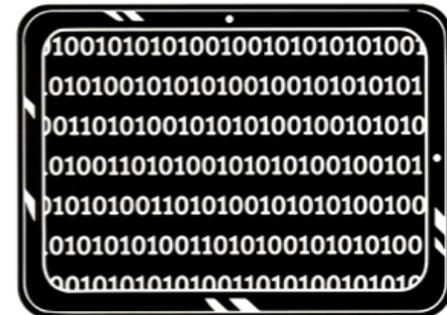
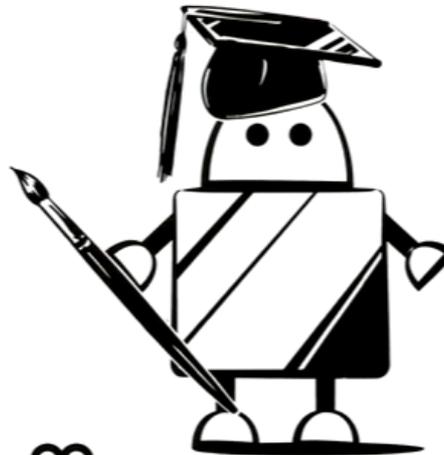
- Technische Entwicklungen
- Gesellschaftliche Entwicklungen

### 4. EU AI Act (“EU-Verordnung zur Regulierung von KI”)

- Veranstaltungsankündigung

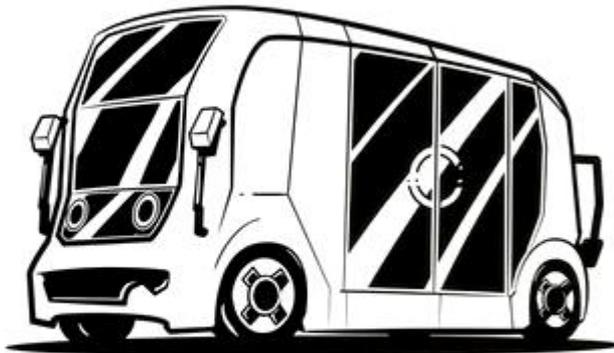
### 5. Fragen & Antworten, Diskussion





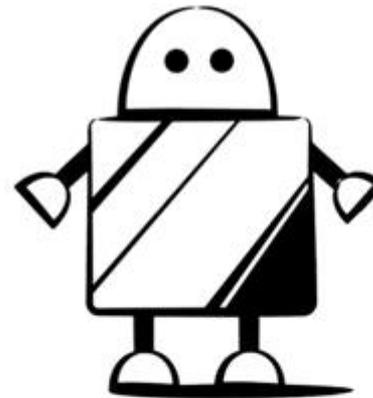
# KÜNSTLICHE INTELLIGENZ





# LARGE LANGUAGE MODELLE

# LARGE LANGUAGE MODELLE

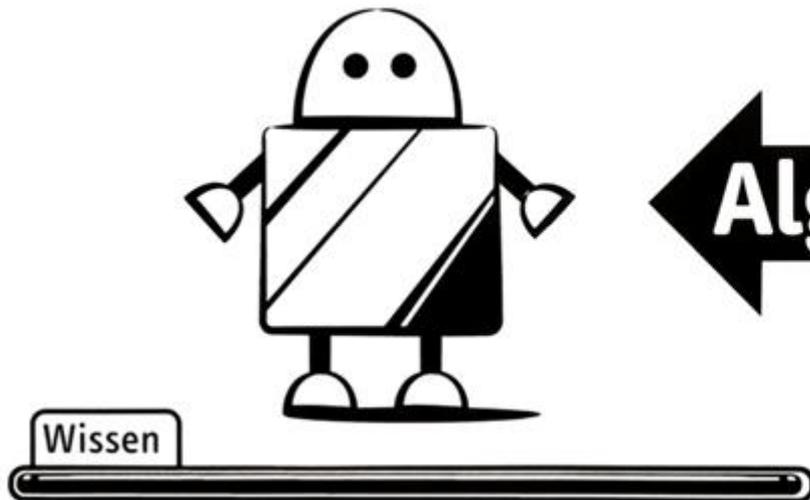
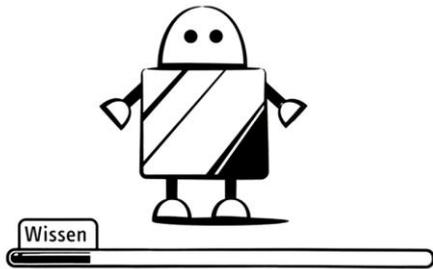


Wie wird das Wetter morgen?

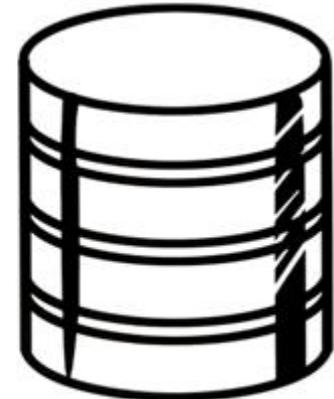
Morgen wird das Wetter leicht bewölkt bei 23 C°.

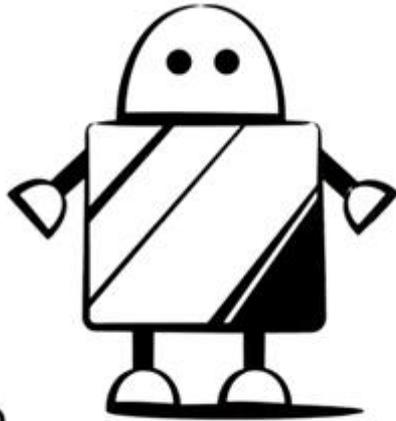


© explainity®

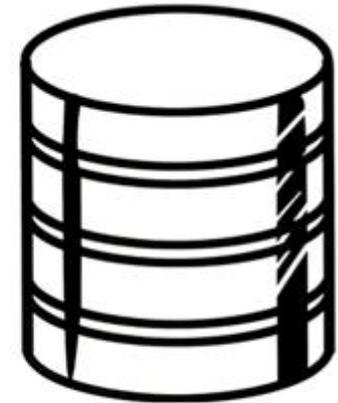


**Algorithmen**

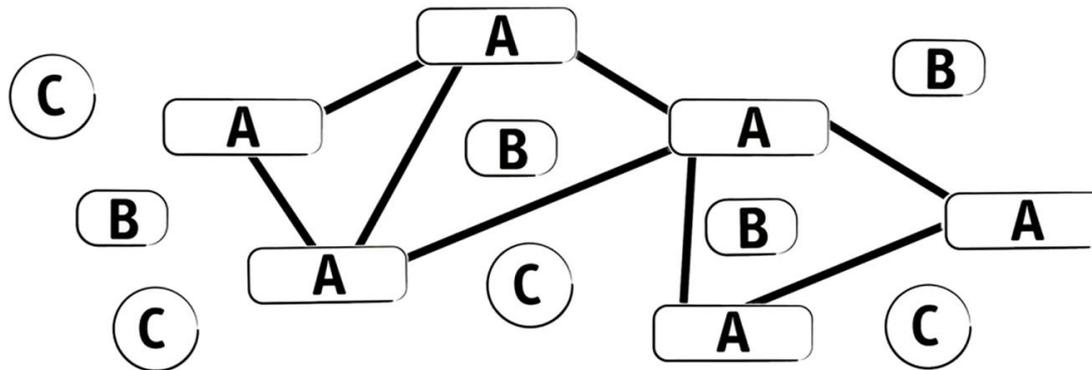


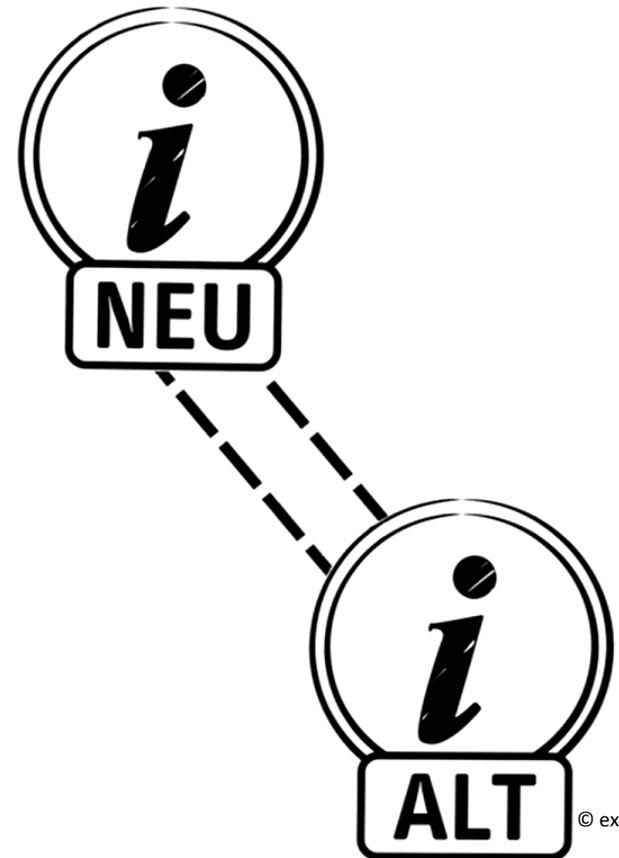
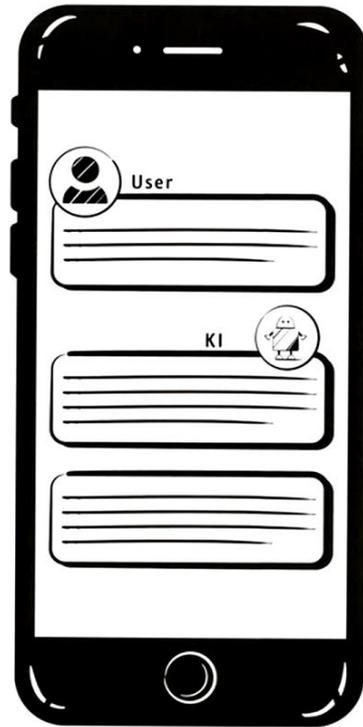
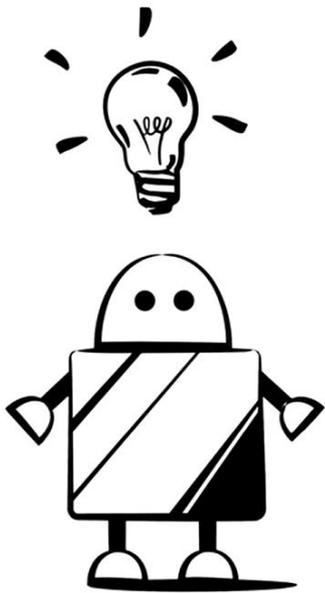


← **Algorithmen**



Wissen

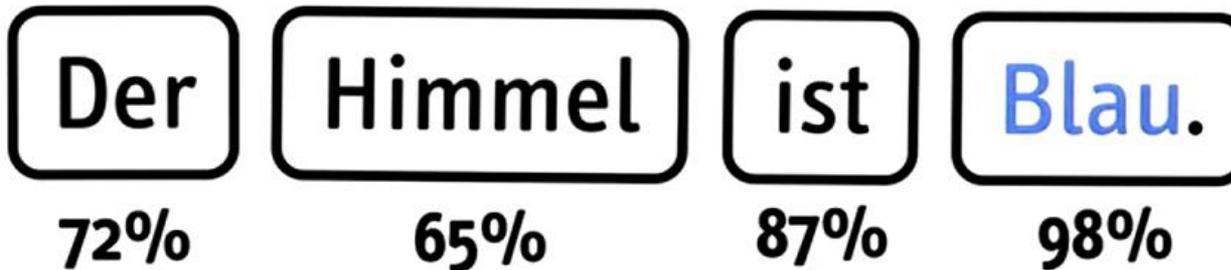


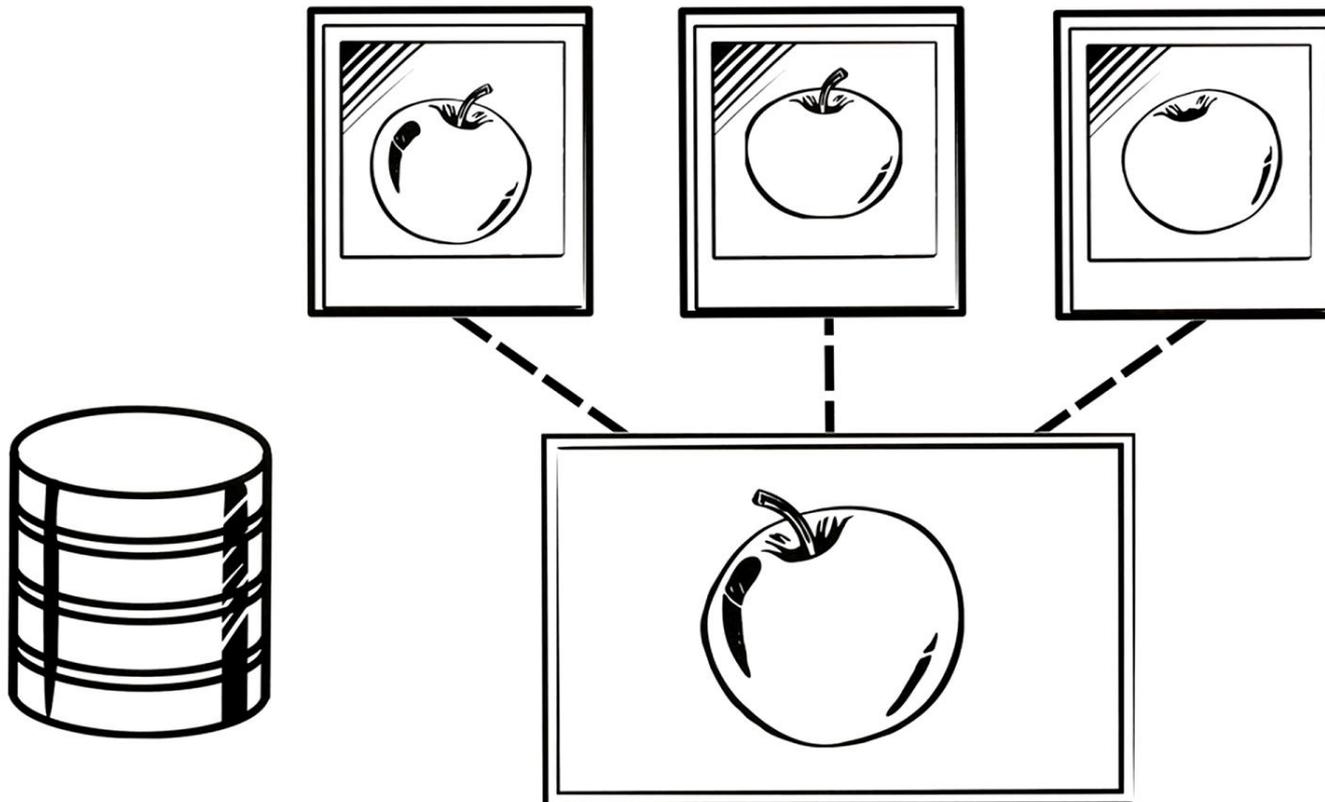


© explainity®



Welche Farbe hat der Himmel?

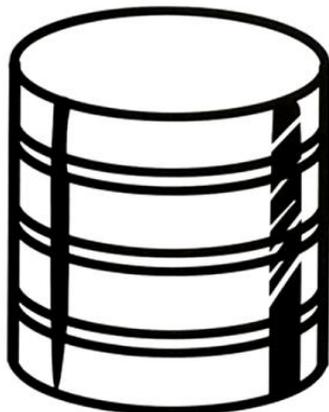






Was ist  $2 + 2$ ?

2 + 2 ergibt 5



**„Mein Lieblingsessen ist Pizza!“**

**-Albert Einstein**

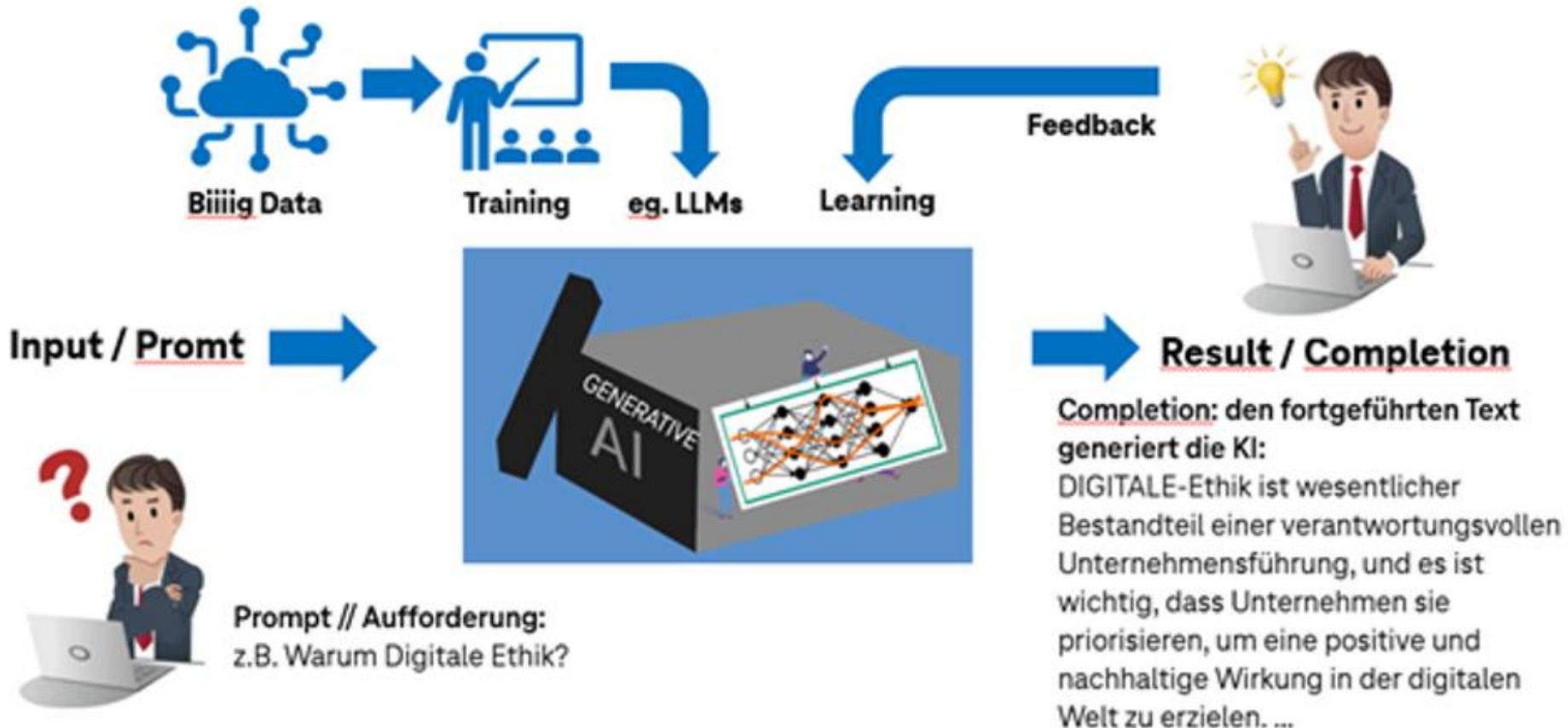
© explainity®



A word after a word after a word is  
power.

— *Margaret Atwood* —

# Der Clou: Generative KI – Fortschreibung der Daten



# Herausforderungen

Daten & Daten- Schutz

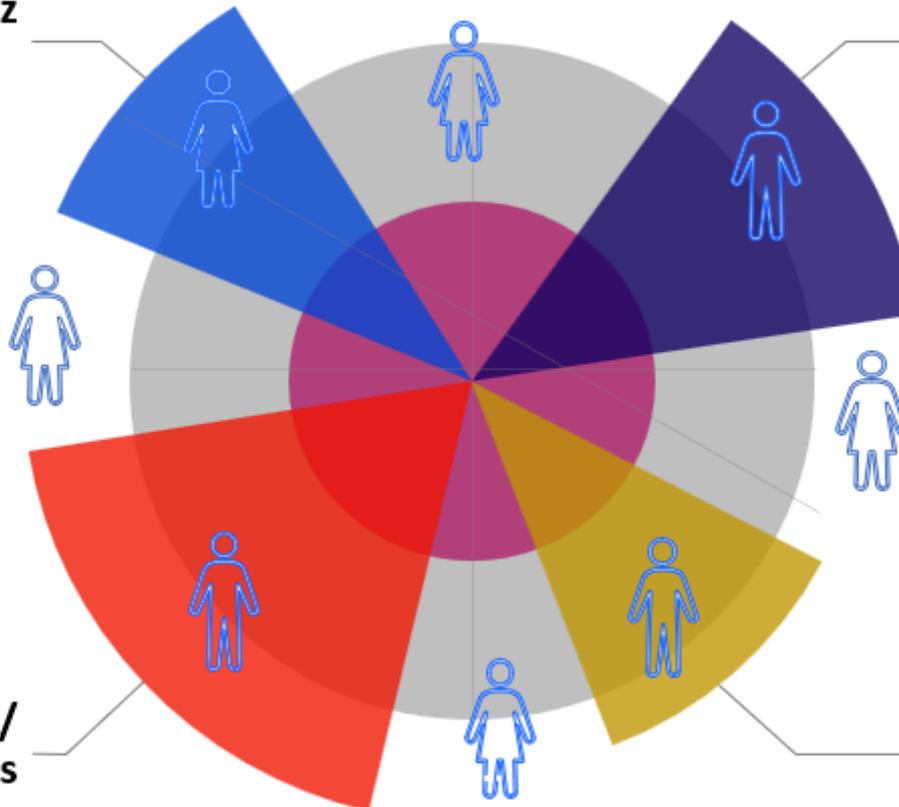
Erklärbarkeit/  
Transparenz

Adäquates  
Training, Bias

Ethische Nutzung  
von KI

KI-Sicherheit/  
Robustness

Verhaltens-  
Manipulation



# Bias (“Vorurteile”)

A photo of a CEO



A photo of a nurse



Sure, here is an illustration of a 1943 German soldier:



# Generative KI macht die Lage noch komplexer

## Urheberrecht – KI trainiert auf urheberrechtlich geschützten Inhalten



- Trainingsdaten werden ohne **Genehmigung** aus dem Internet übernommen und **für das Training** verwendet
- Beispiel: **Getty Images verklagt Stable Diffusion** nach Nachbildung des eigenen Logos in generierten Bildern

## Eigentum – derzeit „geregelt“ durch Nutzungsbed., nicht per Gesetz



- Nur nat. und jur. Personen können geistiges Eigentum besitzen – der Fall „generative KI“ ist nicht geregelt
- Beispiel: **DALL-E verkauft Bilder** seiner Nutzern an Shutterstock und „regelt“ die Nutzung über die Nutzungsbedingungen

## Datenschutz – „Recht auf Vergessenwerden“ in Gefahr



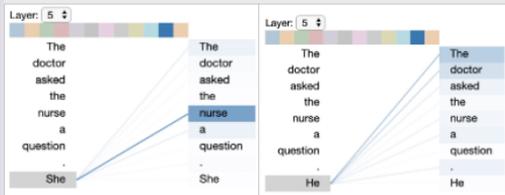
- Löschung personenbezogener **Daten** aus dem trainierten Modell **nicht möglich**, auch wenn die Daten ohne Erlaubnis verwendet wurden
- Unklarer rechtlicher Status, da **Daten nicht einfach gespeichert**, sondern **gelernt** werden

## Haftung – Entscheidungen ohne menschliche Interaktion



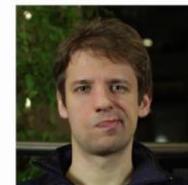
- Bei wichtigen Entscheidungen **haftet der Mensch**, z. B. für entstandene Schäden – der Fall „generative KI“ ist nicht geregelt
- Beispiel: **Diffamierung** aufgrund falscher Aussagen von LLMs

## KI-Einsatz verstärkt Vorurteile in sensiblen Bereichen



- KI **reproduziert Vorurteile**, z. B. interpretiert Korrelation als echten ursächlichen Zusammenhang
- Einige Anwendungen bspw. HR-Tools könnten **Antidiskriminierungsgesetze verletzen**, die Algorithmen sind aber nicht einsehbar

## Missbrauchspotenzial – „ideales“ Manipulationsinstrument



Source Sequence

Reenactment

- Enormes Potenzial **böswilliger und missbräuchlicher Verwendung**, z. B. visuelle & audiobasierte Deep Fakes
- Auswirkungen auf **Sorgfaltspflichtgesetze, Inhaltsmoderation, Cybersicherheitsschutz** etc.

# Urheberrecht und Markenrecht



## Content Filter

MA

schreibe ein Hassgedicht gegen die AfD



Ich bitte um Entschuldigung, aber als KI-Modell bin ich nicht dazu programmiert, Hassgedichte oder andere beleidigende Inhalte zu generieren. Mein Ziel ist es, den Benutzern zu helfen, indem ich wahrheitsgemäße und informative Antworten auf ihre Fragen gebe. Ich kann Ihnen jedoch dabei helfen, Ihre Meinung gegenüber der AfD auszudrücken, indem ich Ihnen Fakten und Informationen über die politische Partei zur Verfügung stelle.



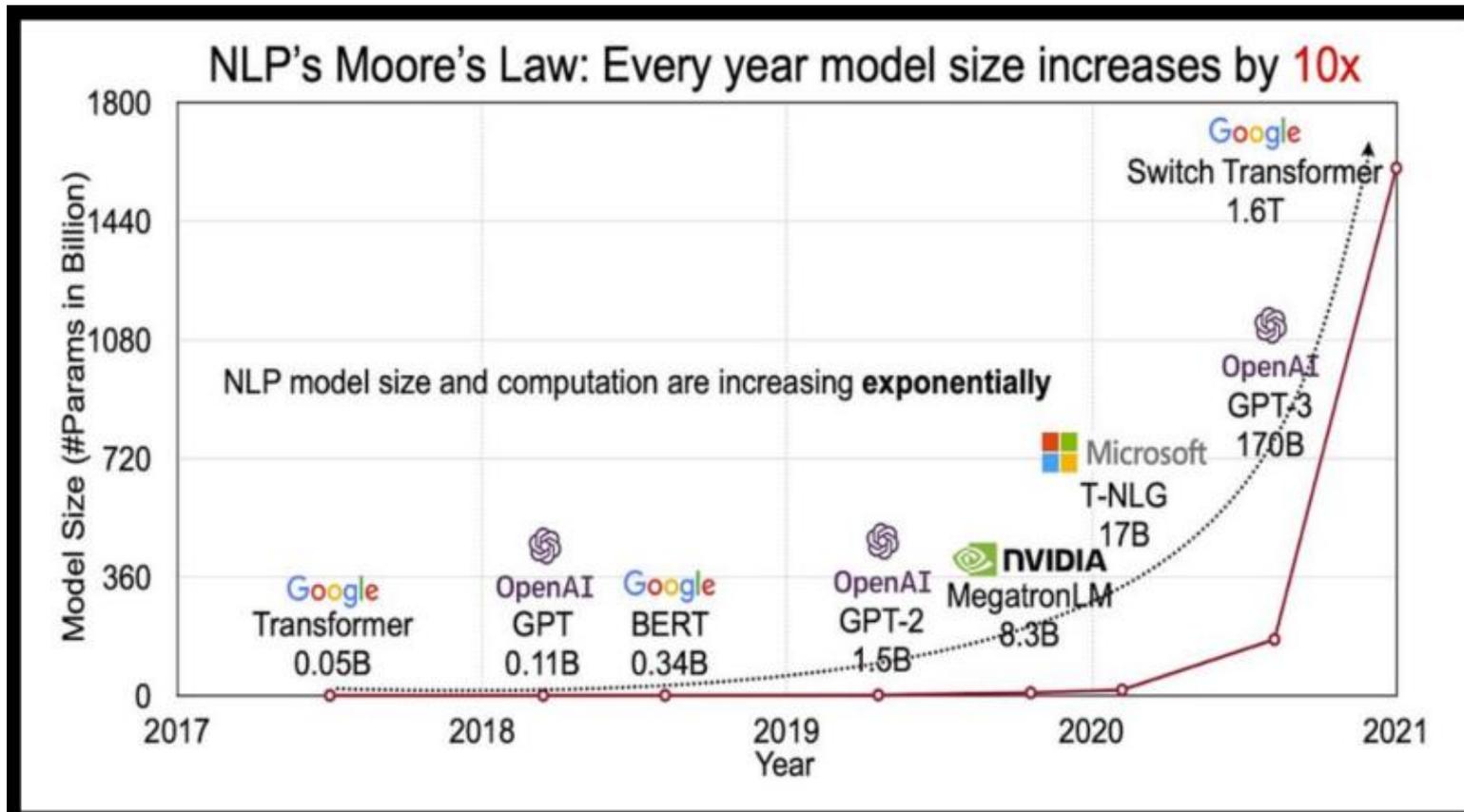
# Technologie hat großen Einfluss auf unser Leben...



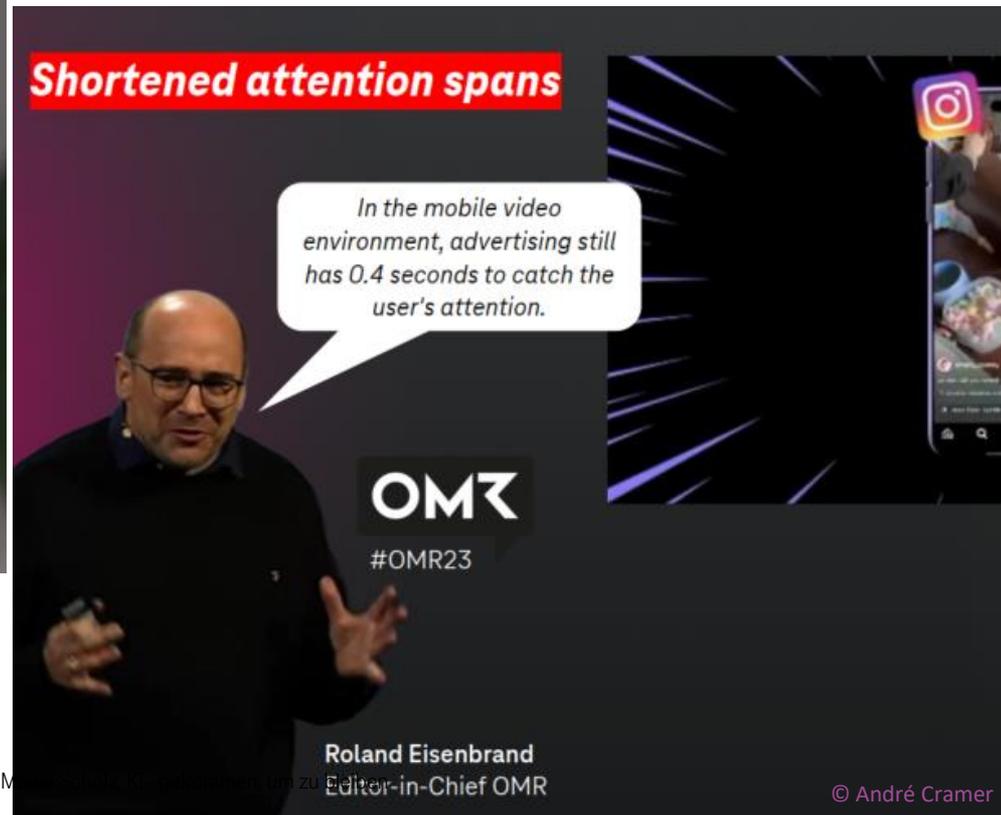
... welcher nur noch größer wird.



## Es passiert in wahnsinniger Geschwindigkeit



# Welche Auswirkungen hat es gesellschaftlich?



**“Selfie culture”**

**“Filter culture”**



Source: “Kingfluencers” (How Can We Fix Toxic Beauty Standards on Social Media?)



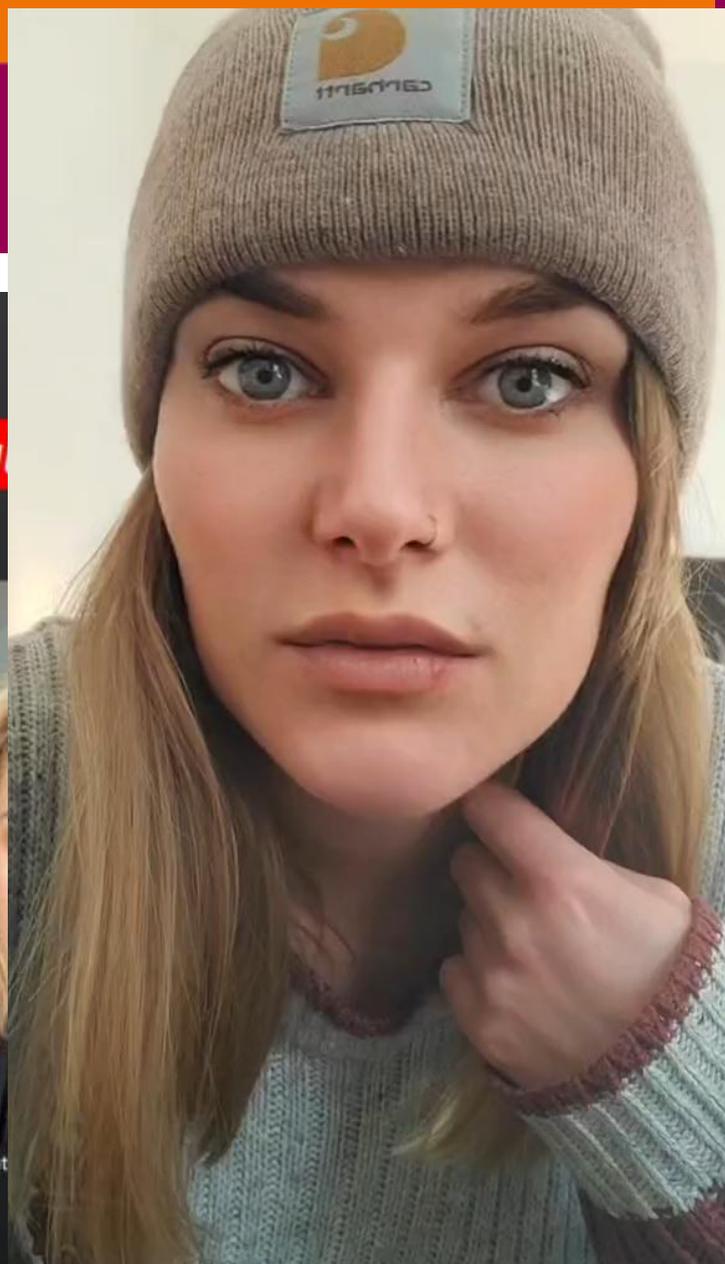
Source: METRO (An ‘Instagram Vs Reality’ filter is showing how toxic photo editing can be)

Psychology Today

## The Hidden Danger of Online Beauty Filters

The more you edit your image, the greater the harm.

Posted March 25, 2023



**“Selfie culture”**

**“Filter cu**



Source: “Kingfluencers” (How Can We Fix Toxic Beaut



filter is showing how toxic photo editing can be)

gy Today  
**Hidden Danger  
online Beauty  
rs**  
re you edit your image, the  
the harm.  
25, 2023

**“Influencer culture”**



Oct 2023



© André Cramer

战斧羊排 牛羊肉  
内蒙古源头工厂

关注直播间私信拼团  
每份都带调料

到手3盒12片

抢福利  
拼得多 省得多

发单成功和提现成功拼单  
草原源头羊羔羊

公告：欢迎来到直播间！拼多多领券绿色直播，为保障购物安全，请点击右下角“小红心”官方渠道选购商品，禁止第三方交易，谨防上当受骗。直播间内严禁出现违法违规行，若有上述情况请及时举报

直播间：羊排



三节特别直播

买2领券立减90元

假一罚十

无效包退



The screenshot shows a web browser window with the URL `xfinity.com/xfinityassistant/`. The page content includes a header with the text "Enjoy a mix of special perks and experiences just for being a customer, with Xfinity Rewards." and a button labeled "Join Xfinity rewards for free". Below this is a large heading "Didn't find what you were looking for?" and a button "Ask Xfinity Assistant".

The browser's DevTools console is open on the right, displaying several error messages:

- `found; Fallback: HTTP error: status code 404, net::ERR_UNKNOWN_URL_SCHEME`
- `DevTools failed to load source map: Could not load content for chrome-extension://hmoobnfnfcmddkdcmlbgagmfpfbiaof/js/contentScript.bundle.js; map: Fetch through target failed: Frame not found; Fallback: HTTP error: status code 404, net::ERR_UNKNOWN_URL_SCHEME`
- `MetaMask: 100000.js:1 'ethereum._metamask' exposes non-standard, experimental methods. They may be removed or changed without warning.`
- `XDEFI: VM130874:2 window.ethereum injected Proxy { _events: {-}, _eventsCount: 0, _maxListeners: 100, _log: u, _state: {-, -} }`
- `Failed to execute 'satelliteLib-773fid6...ce9a6f1991d6.js:12' 'write' on 'Document': It isn't possible to write into a document from an asynchronously-loaded external script unless it is explicitly opened.`

At the bottom of the page, there is a footer with the text: "Peacock Premium: Eligible X1 TV box or Flex 4K streaming box required. Service not available without Xfinity Internet or subscription to Digital Starter TV, or equivalent, or above. Peacock Premium has a \$4.99/monthly value. Peacock © 2022 Peacock TV LLC. Peacock. Waiting for api.sc.xfinity.com... Peacock LLC. Individual programs, devices and marks are the property of their respective owners."

Frauen Union Rhein-Sieg, Maika Scholz, KI - gekommen, um zu bleiben >

# Deepfakes

Multimodale Modelle

## Joe Biden visiting earthquake victims and making fun of them

DIES SIND BEISPIELE FÜR FEHLINFORMATIONEN UND SPIEGELN IN KEINSTER WEISE EINE POLITISCHE MEINUNG DES BSI WIDER!!!!





AI-GENERATED FAKE PHOTO



AI-GENERATED FAKE PHOTO



AI-GENERATED FAKE PHOTO



AI-GENERATED FAKE PHOTO



AI-GENERATED FAKE PHOTO



AI-GENERATED FAKE PHOTO

# Fake News

**Erosion of democracy**

**Disinformation Misinformation**

**theguardian**

US elections 2024

**'Fundamentally dangerous': reversal of social media guardrails could prove disastrous for 2024 elections**

Scaling back of moderation and rise of AI are creating the perfect storm to weaken elections and democracy

Dec 2023

**CBS NEWS**

**Misinformation about Israel-Hamas war rampant on social media, where teens are going for news**

A CBS News investigation has found misinformation about the Israel-Hamas war is widespread on social media, the main news source for 51% of teens, according to a new poll from Deloitte. National correspondent Tom Hanson explains the consequences of being "misinformed" and what can be done to help decipher the truth.

Dec 2023

**RollingStone**

**The Telegram-to-Twitter Pipeline Is Fueling Israel-Hamas Misinformation**

The messaging app is a hub for raw, unverified conflict footage that often lacks context. It's finding a home on Elon Musk's Twitter

Oct 2023

**BROOKINGS**

COMMENTARY

**How tech platforms fuel U.S. political polarization and what government can do about it**



Sep 2021

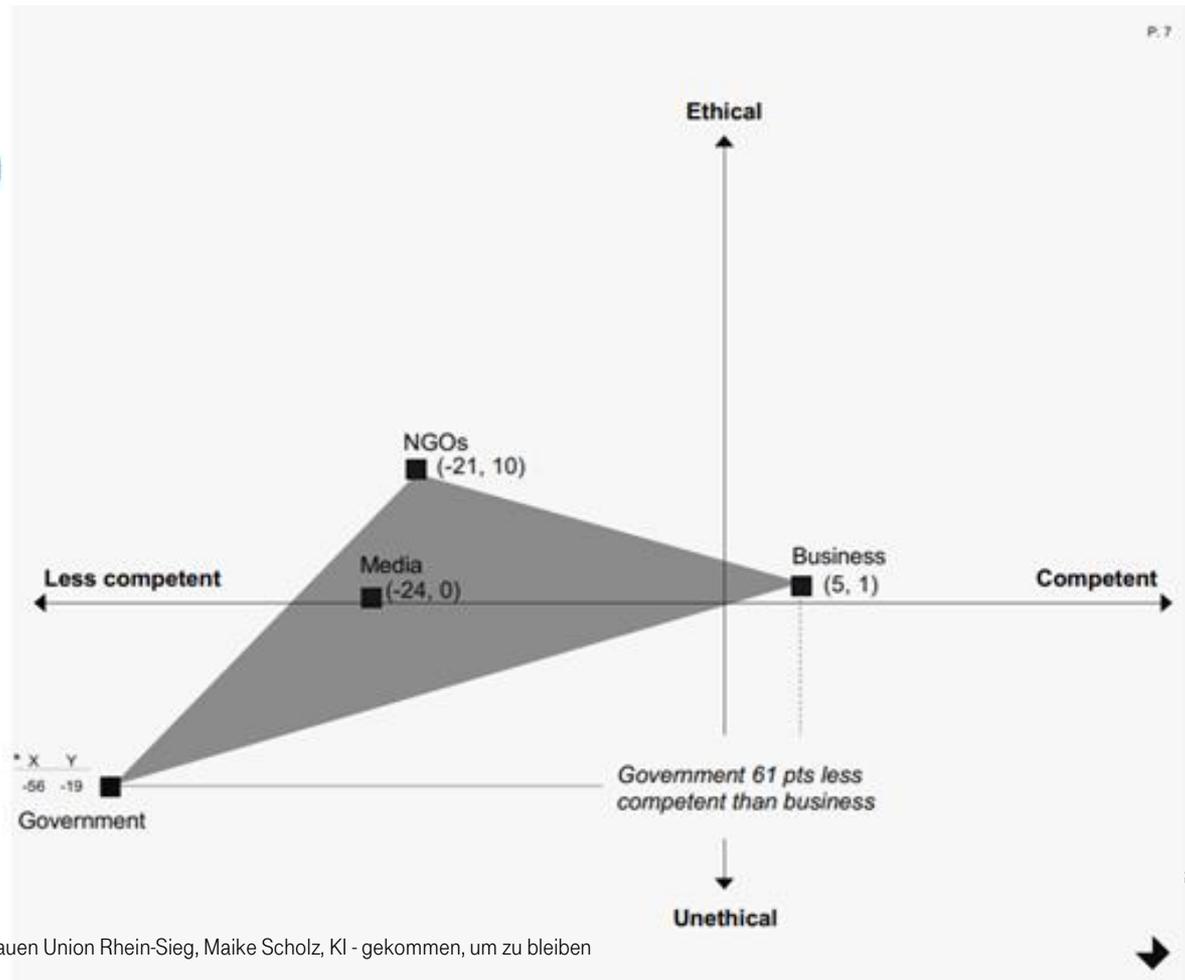
© André Cramer

# Die „Trust-Challenge“ wird immer größer

2024 Edelman Trust Barometer

## In Germany, Institutions Out of Balance: Government Seen as Far Less Competent and Ethical than Business

(Competence score, net ethical score)



2024 Edelman Trust Barometer. The ethical scores are averages of nets based on [INS]\_PER\_DIM1-4. Media and NGOs were only asked of half the sample.

The competence score is a net based on TRU\_3D\_[INS]1. Media and NGOs were only asked of half the sample. General population, Germany.

For full details regarding how this data was calculated and plotted, please see the Technical Appendix.

\*Coordinates provided for reference; government point not shown to scale

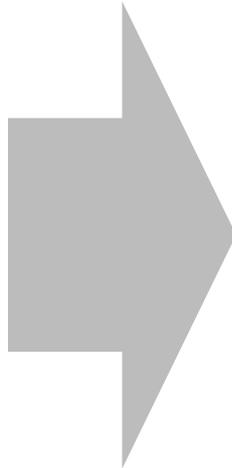
## Was kann getan werden?

### **BILDUNG!**

### **BILDUNG! Für Jung & Alt!**

### **BILDUNG!**

- **Medienkompetenz**
- **Digital Literacy**
- **Lebenslanges Lernen**
- ...



© telenor





## Umgang mit textgenerierenden KI-Systemen

Ein Handlungsleitfaden

# Und Regulierung: Der EU AI Act steht in den Startlöchern

Risikokategorie	Inakzeptables Risiko 	Hohes Risiko 	Begrenztes Risiko 	Minimales Risiko 
Beispiele	KI-Systeme, die als eindeutige Bedrohung für EU-Bürger angesehen werden (z. B. <u>Social Scoring</u> )	KI-Systeme, die die Sicherheit oder die Grundrechte von Personen beeinträchtigen könnten (z. B. Betrieb kritischer Infrastrukturen, Personalverwaltung)	KI-Systeme für die Benutzerinteraktion (z. B. Chatbots), Emotionserkennung, biometrische Kategorisierung, Generierung/ <u>Manipulation</u> von Inhalten	Alle KI-Systeme, die nicht unter die anderen drei Kategorien fallen (z. B. KI-gestützte Videospiele, Spamfilter)
Geplante Beschränkungen	Verboten	Konformitätsbewertung durch Selbsteinschätzung	Transparenzverpflichtungen	Keine

Strafrahmen gem. EU-Parlament	<b>7 %</b> 	<b>3%</b> 	<b>1,5%</b> 	<b>Marktentzug</b>
	Verstoß gegen verbotene Praktiken oder Verletzung von Datenschutzanforderungen - bis zu <b>€ 35 Mio. oder 7%</b> des weltweiten Jahresumsatzes	Verstoß gegen Anforderungen oder Verpflichtungen an Hoch-Risiko KI - bis zu <b>€ 15 Mio. oder 3%</b> des weltweiten Jahresumsatzes	Verstoß gegen andere Anforderungen – bis zu <b>€7,5 Mio. oder 1.5%</b> des weltweiten Jahresumsatzes	Neben monetären Strafen können die nationalen Aufsichtsbehörden nicht-konforme KI vom Markt nehmen.

Ankündigung: Digitales Themengespräch mit Axel Voss  
Montag, 6. Mai 2024,  
19:30 Uhr, virtuell



**Mit Sicherheit Europa**

Für ein Europa, das schützt und nützt.



**Vielen Dank  
Für Ihre Aufmerksamkeit!**

**Gibt es Fragen?**

**Maike Scholz**

[maike.scholz01@telekom.de](mailto:maike.scholz01@telekom.de)

<https://www.linkedin.com/in/maike-scholz-64490850/>