Risiken von KI-Technologien

Ein systematischer Überblick aus ethischer Perspektive

Impuls im Rahmen des Digi-Dienstag-Formats

Deutscher Paritätischer Wohlfahrtsverband

Dienstag 17. Oktober 2023 (online)

Überblick

I. Wissenschaftlicher Hintergrund

II. Was ist ein „Risiko“?

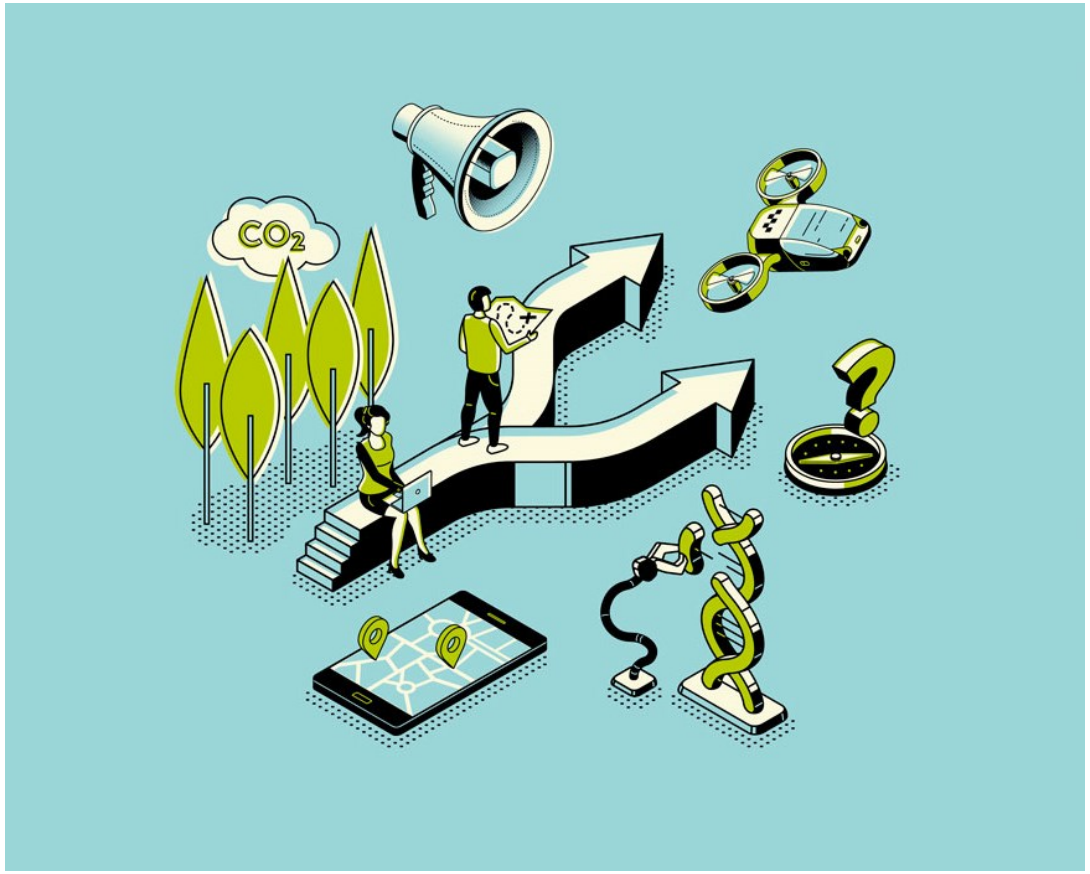
III. Überblick über KI-Risiken

IV. Wie mit KI-Risiken umgehen?

I. Wissenschaftlicher Hintergrund

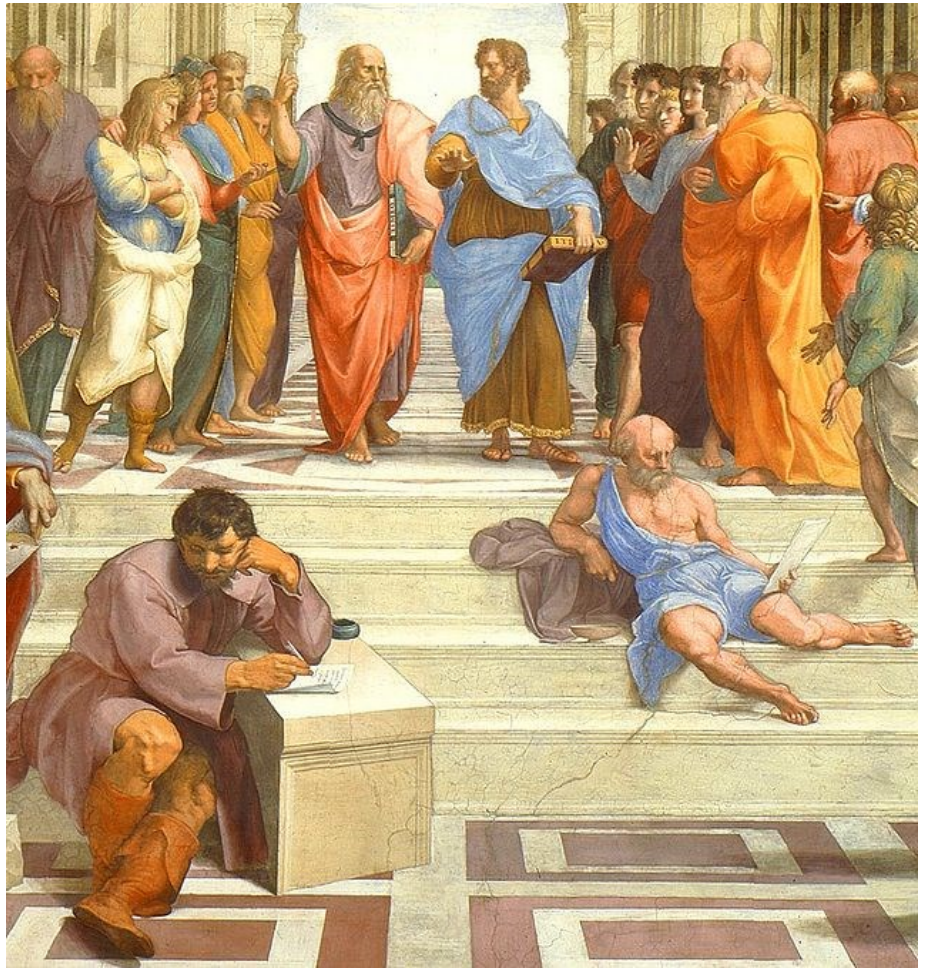
Institutioneller Hintergrund

- Postdoktorand am „Institut für Technikfolgen-abschätzung und Systemanalyse“ (ITAS) des KIT
- Forschungsgruppe „PhilETAS“ (Prof. Dr. Dr. Rafaela Hillerbrand)



Fachlicher Hintergrund

- Philosophie
  - Politische Philosophie
  - Erkenntnistheorie
  - Technikethik



Forschungsschwerpunkte mit Bezug zu KI

- Ethik des Autonomen Fahrens

Dietmar Rabich / Wikimedia Commons / „San Francisco (CA, USA), California Street, autonomes Fahrzeug (Waymo) -- 2022 -- 2925“ / CC BY-SA 4.0

Forschungsschwerpunkte mit Bezug zu KI

- BMBF-Projekt „KIARA“ – ethische Aspekte bei der Entwicklung und Anwendung eines KI-gestützten Roboters für nukleare Gefahrenlagen

II. Was ist ein „Risiko“?

Was ist eigentlich ein „Risiko“?

- Alltagssprachlich manchmal nur:
  „negativ bewertetes Ereignis“

Was ist eigentlich ein „Risiko“?

- In den Wissenschaften aber auch alltagsprachlich wird „Risiko“ oftmals auch komplexer verstanden

Wahrscheinlichkeit/Möglichkeit eines Ereignisses + negative Bewertung des Ereignisses

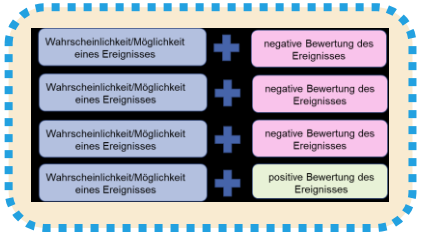
- Entsprechend kann man dann auch „Chancen“ verstehen

Wahrscheinlichkeit/Möglichkeit eines Ereignisses + positive Bewertung des Ereignisses

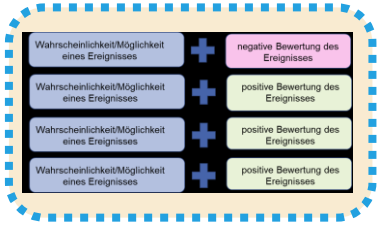
Was ist eigentlich ein „Risiko“?

Wahrscheinlichkeit/Möglichkeit eines Ereignisses	+	negative Bewertung des Ereignisses
Wahrscheinlichkeit/Möglichkeit eines Ereignisses	+	negative Bewertung des Ereignisses
Wahrscheinlichkeit/Möglichkeit eines Ereignisses	+	positive Bewertung des Ereignisses
Wahrscheinlichkeit/Möglichkeit eines Ereignisses	+	positive Bewertung des Ereignisses

Handlungsoption A



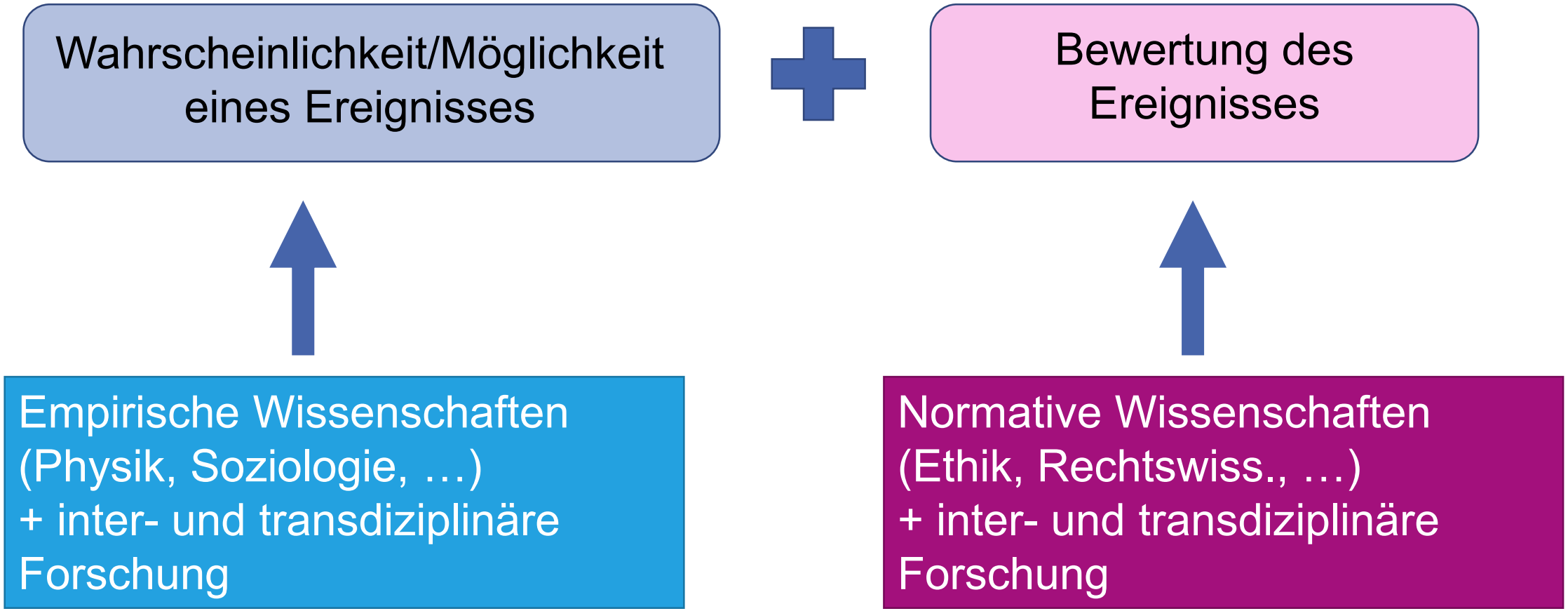
Handlungsoption B



Handlungsoption C

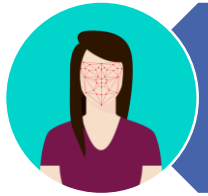
Risiken und Chancen sind Teil von alternativen Handlungsoptionen (mit entsprechenden Pfaden und Abhängigkeiten)

Wie lassen sich Risiken wissenschaftlich einschätzen?



III. Überblick über KI-Risiken

Systematischer Überblick über Risiken durch KI



Gefährdung von Individualrechten

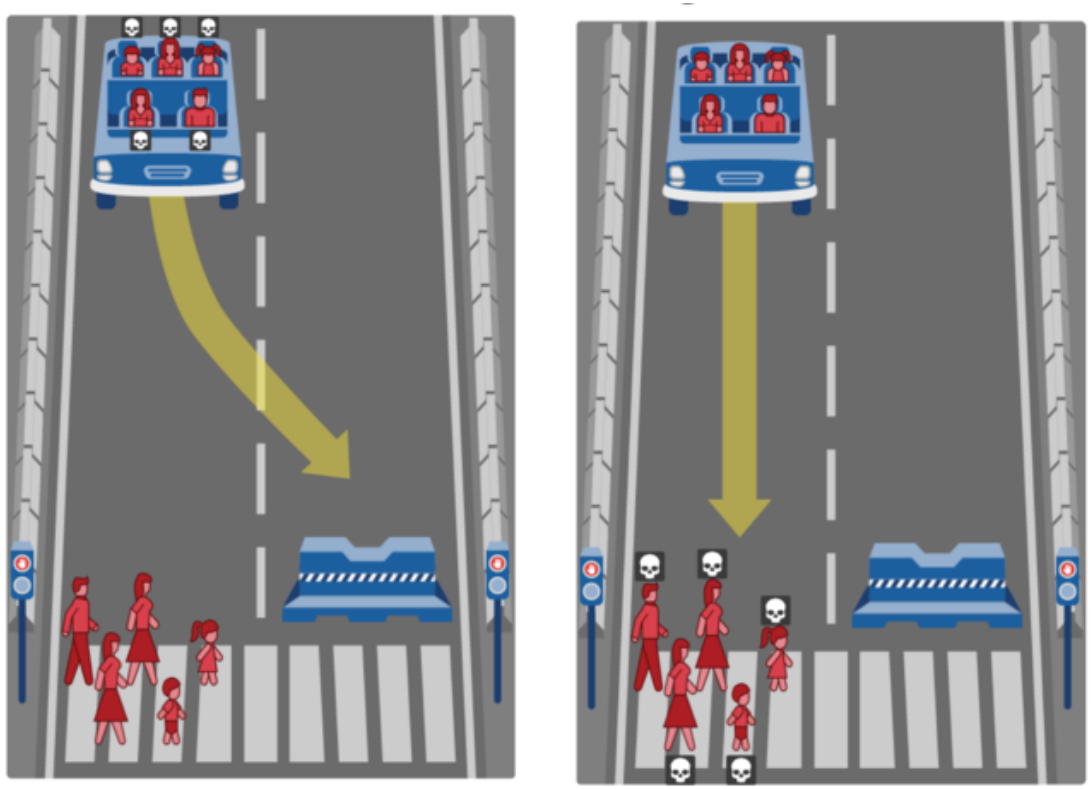
Gefährdung von Wohlergehen

Gefährdung von kollektiven Rechten und Gütern

Existenzielle Gefährdung der Menschheit?

1) Gefährdung von Individualrechten

- Recht auf Leben und körperliche Unversehrtheit
  - z.B. autonomes Fahren
  - z.B. autonome Waffensysteme



https://commons.wikimedia.org/wiki/File:Moral_Machine_Screenshot.png, CC BY-SA 4.0 Deed

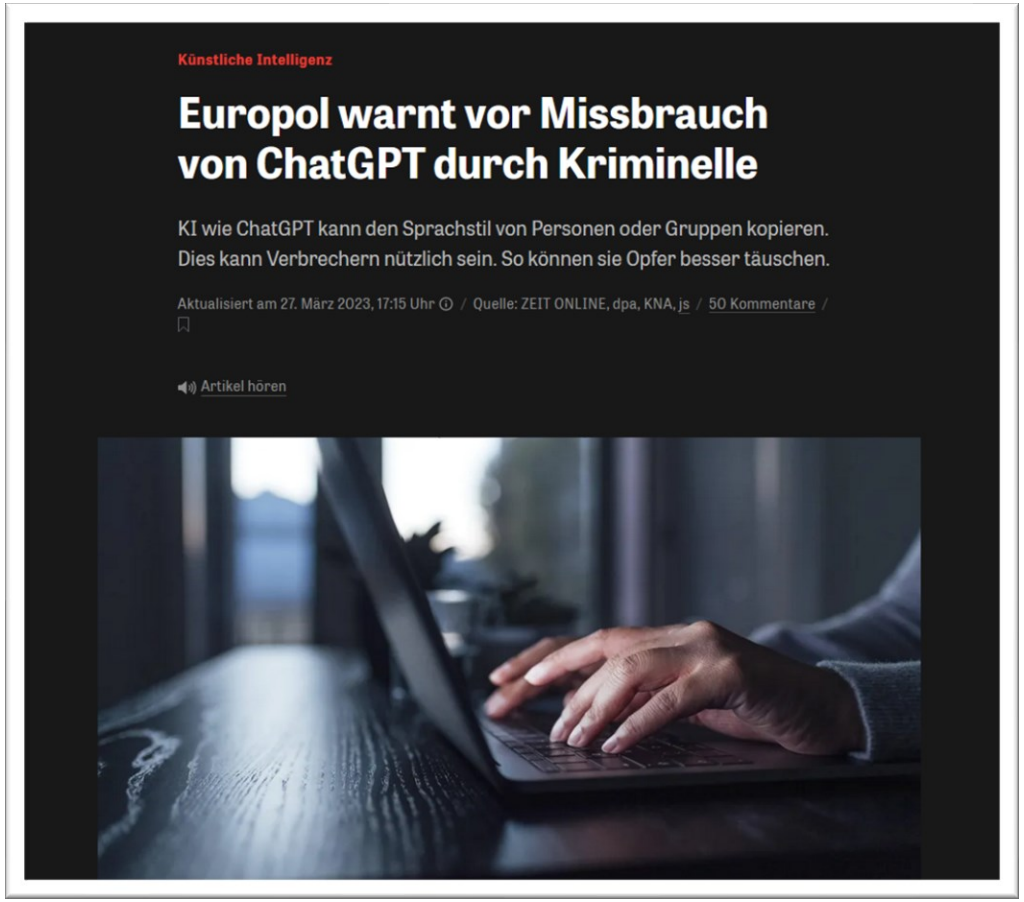
1) Gefährdung von Individualrechten

- Nichtdiskriminierung / Faires Verfahren
  - z.B. COMPAS-Algorithmus zur Einschätzung der Rückfallsgefährdung
  - z.B. Amazons Algorithmus bei Anstellungsverfahren
  - z.B. automatisierte Entscheidungen im Sozialbereich (NLD, AUS)

Screenshot from: Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

1) Gefährdung von Individualrechten

- Eigentumsrechte / Privatheit
  - Cybercrime



Screenshot from: Stephanowitz, J., dpa, & KNA, K. K. N.-A. (2023, March 27). Künstliche Intelligenz: Europol warnt vor Missbrauch von ChatGPT durch Kriminelle. *Die Zeit*. https://www.zeit.de/digital/internet/2023-03/chatgpt-europol-missbrauch-kriminelle-kuenstliche-intelligenz

1) Gefährdung von Individualrechten

- Persönlichkeitsrechte
  - Deepfakes

https://commons.wikimedia.org/wiki/File:Pope_Francis_in_puffy_winter_jacket.jpg

2) Gefährdung des Wohlergehens

- Selbstachtung
  - Arbeitslosigkeit durch Automatisierung

https://commons.wikimedia.org/wiki/File:Kodiak_Robotics_self-driving_truck_02.jpg, CC BY-SA 4.0 Deed

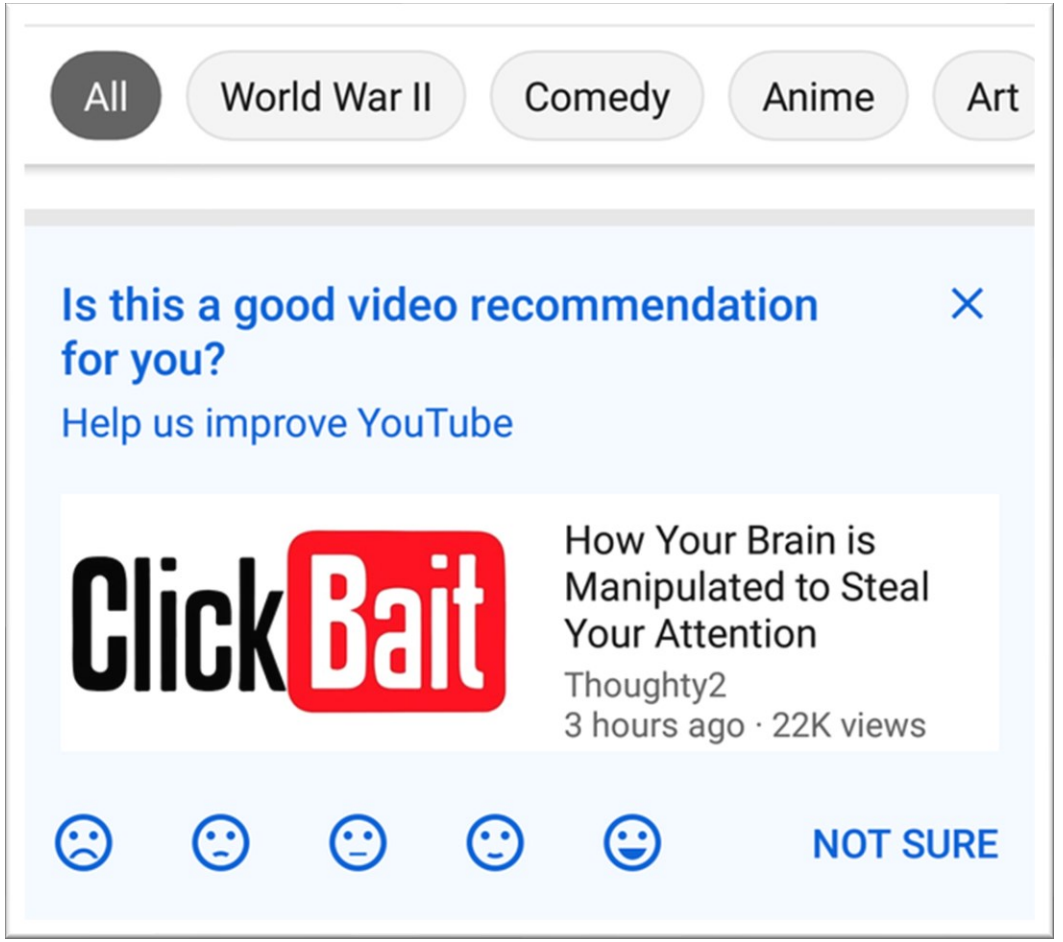
2) Gefährdung des Wohlergehens

- Bedürfnis nach menschlichen Beziehungen
  - Isolierung durch Automatisierung

Screenshot from: NDR. (n.d.). *Mit Roboter Greta gegen den Fachkräftemangel in der Pflege*. Retrieved October 16, 2023, from https://www.ndr.de/nachrichten/schleswig-holstein/Mit-Roboter-Greta-gegen-Fachkraeftemangel-in-der-Pflege,roboter784.html

2) Gefährdung des Wohlergehens

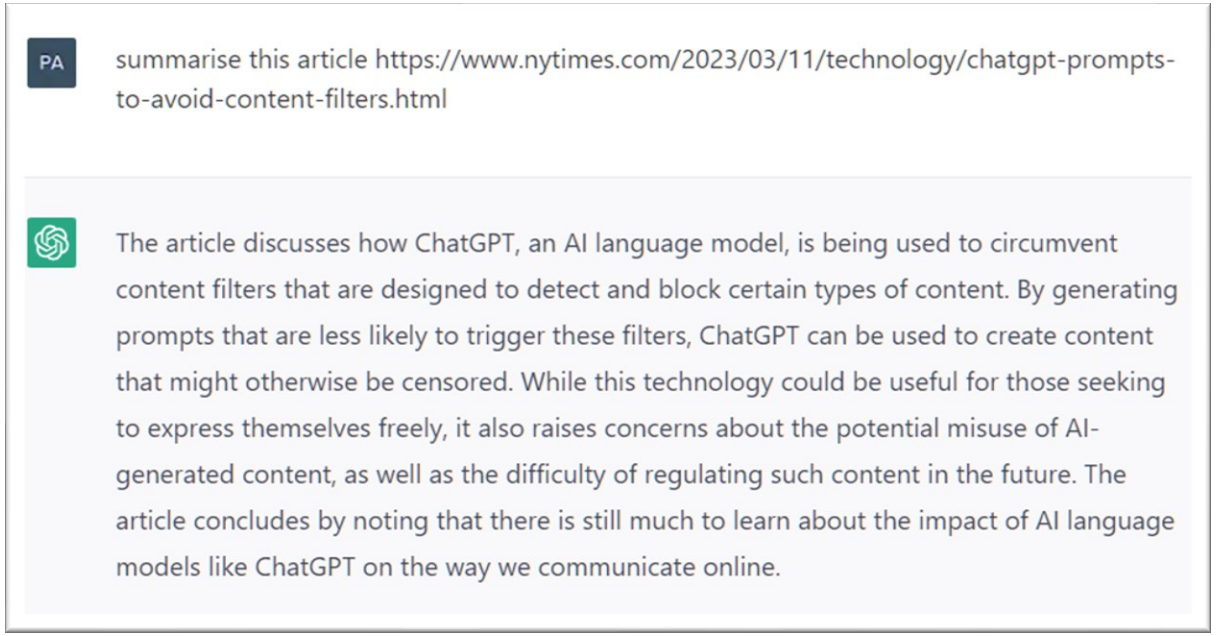
- Autonomie und mentale Gesundheit
  - Manipulation z.B. durch personalisierte Empfehlungsalgorithmen
  - Verbreitung von Vorurteilen
  - Arbeitsbedingungen von „Click workern“



https://commons.wikimedia.org/wiki/File:YouTube_video_recommendation_feedback_feature_(2020).png

2) Gefährdung des Wohlergehens

- Wissen und Verstehen
  - Überschätzen der Verlässlichkeit von generativer KI



https://commons.wikimedia.org/wiki/File:ChatGPT_hallucination.png

3) Gefährdung kollektiver Rechte und Güter

- Gefährdung unserer freiheitlich demokratischer Grundordnung

3) Gefährdung kollektiver Rechte und Güter

- Gefährdung unserer freiheitlich demokratischer Grundordnung
  - Erodierung von Konsense und Normen durch Filterblasen und Echokammern über Social Media

https://commons.wikimedia.org/wiki/File:Eli_Pariser,_author_of_The_Filter_Bubble_-_Flickr_-_Knight_Foundation_(1).jpg, CC BY-SA 2.0 Deed

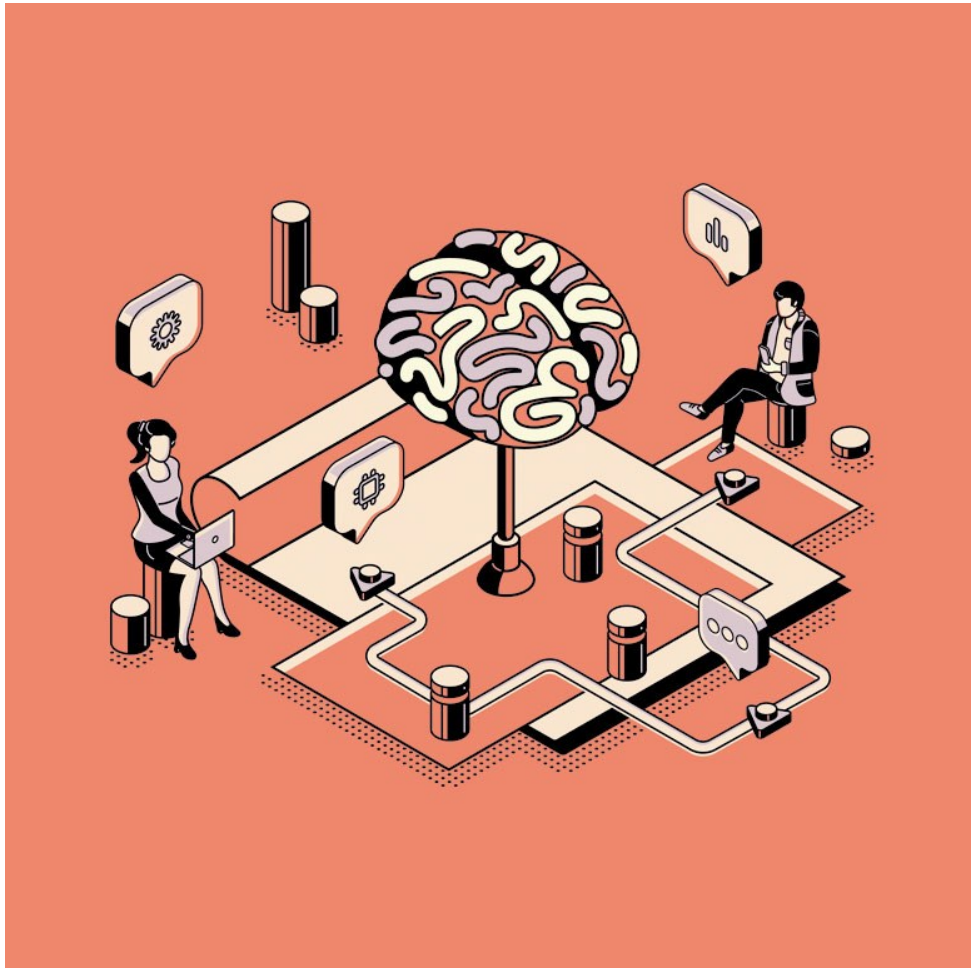
3) Gefährdung kollektiver Rechte und Güter

- Gefährdung unserer freiheitlich demokratischer Grundordnung
  - Erodierung von Konsense und Normen durch Filterblasen und Echokammern über Social Media
  - Manipulation der Meinungsbildung durch politische Deepfakes

https://commons.wikimedia.org/wiki/File:Trump%E2%80%99s_arrest_(2).jpg

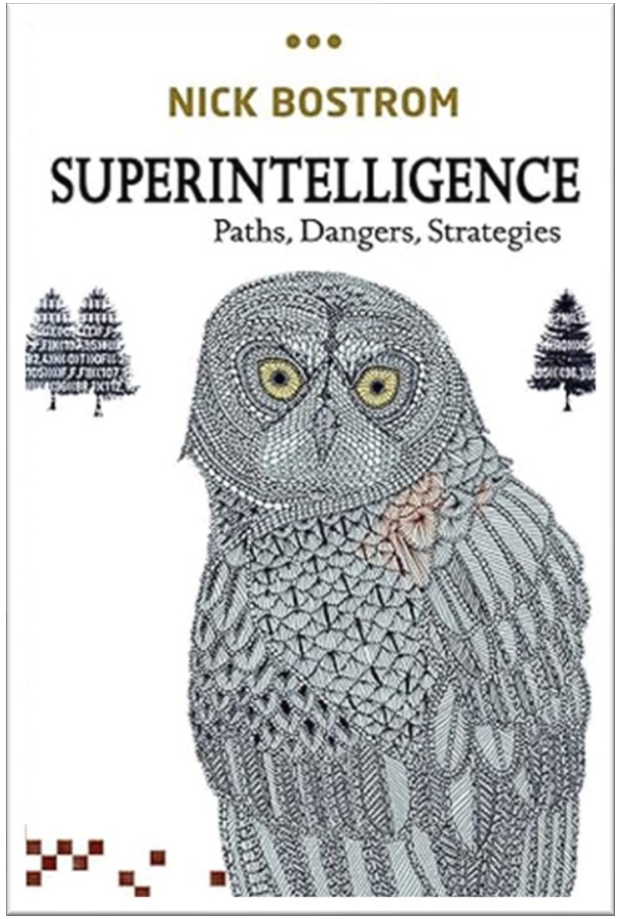
3) Gefährdung kollektiver Rechte und Güter

- Gefährdung unserer freiheitlich demokratischer Grundordnung
  - Erodierung von Konsense und Normen durch Filterblasen und Echokammern über Social Media
  - Manipulation der Meinungsbildung durch politische Deepfakes
  - Verantwortlichkeitslücken



4) Existenzielle Gefährdung der Menschheit?

- Kontrollproblem entweder durch Superintelligenz oder aber durch Intransparente KI
  - Nicht auszuschließendes Zukunftsszenario, das zu reflexivem Handeln verpflichtet
  - Allerdings Befürchtung das mit einem öffentlichen Fokus auf dieses Problem (u.a. von Vertreter:innen der KI Industrie) von den anderen schon jetzt konkreten Gefährdungen abgelenkt werden soll



III. Wie mit KI-Risiken umgehen?

Wie soll man mit den Risiken umgehen?

- Ganzheitliche Betrachtung der Risiken
  - im Zusammenhang mit Chancen
  - als Teil von Handlungsoptionen und -pfaden mit entsprechenden Abhängigkeiten
  - vor dem Hintergrund unterschiedlicher Entscheidungskontexte

Wie soll man mit den Risiken umgehen?

- KI-Ethik Richtlinien mit Fokus auf Prinzipien für Design und Anwendung
  1) „Gerechtigkeit“
  2) „Nichtschädigung“
  3) „Wohltätigkeit“
  4) „Respekt für persönliche Autonomie“
  5) „Erklärbarkeit“

Minds and Machines (2018) 28:689–707
https://doi.org/10.1007/s11023-018-9482-5

CrossMark

AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations

Luciano Floridi[1,2] · Josh Cowls[1,2] · Monica Beltrametti[3] · Raja Chatila[4,5] · Patrice Chazerand[6] · Virginia Dignum[7,8] · Christoph Luetge[9] · Robert Madelin[10] · Ugo Pagallo[11] · Francesca Rossi[12,13] · Burkhard Schafer[14] · Peggy Valcke[15,16] · Effy Vayena[17]

Received: 28 October 2018 / Accepted: 2 November 2018 / Published online: 26 November 2018

Abstract
This article reports the findings of AI4People, an Atomium—EISMD initiative designed to lay the foundations for a “Good AI Society”. We introduce the core opportunities and risks of AI for society; present a synthesis of five ethical principles that should undergird its development and adoption; and offer 20 concrete recommendations—to assess, to develop, to incentivise, and to support good AI—which in some cases may be undertaken directly by national or supranational policy makers, while in others may be led by other stakeholders. If adopted, these recommendations would serve as a firm foundation for the establishment of a Good AI Society.

Keywords Artificial intelligence · AI4People · Data governance · Digital ethics · Governance · Ethics of AI

Wie soll man mit den Risiken umgehen?

- KI-Ethik Richtlinien mit Fokus auf Prinzipien für Design und Anwendung
  1) „Gerechtigkeit“
  2) „Nichtschädigung“
  3) „Wohltätigkeit“
  4) „Respekt für persönliche Autonomie“
  5) **„Erklärbarkeit“**

Minds and Machines (2018) 28:689–707
https://doi.org/10.1007/s11023-018-9482-5

CrossMark

AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations

Luciano Floridi[1,2] · Josh Cowls[1,2] · Monica Beltrametti[3] · Raja Chatila[4,5] · Patrice Chazerand[6] · Virginia Dignum[7,8] · Christoph Luetge[9] · Robert Madelin[10] · Ugo Pagallo[11] · Francesca Rossi[12,13] · Burkhard Schafer[14] · Peggy Valcke[15,16] · Effy Vayena[17]

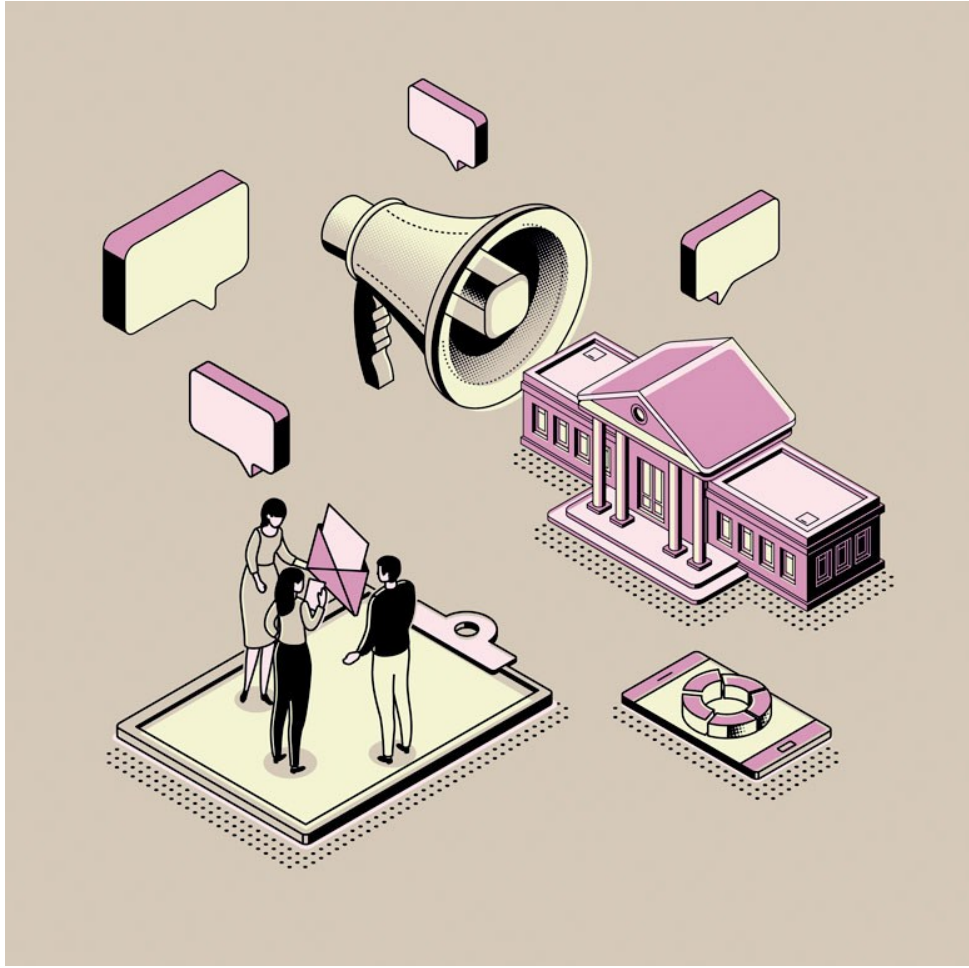
Received: 28 October 2018 / Accepted: 2 November 2018 / Published online: 26 November 2018

Abstract
This article reports the findings of AI4People, an Atomium—EISMD initiative designed to lay the foundations for a “Good AI Society”. We introduce the core opportunities and risks of AI for society; present a synthesis of five ethical principles that should undergird its development and adoption; and offer 20 concrete recommendations—to assess, to develop, to incentivise, and to support good AI—which in some cases may be undertaken directly by national or supranational policy makers, while in others may be led by other stakeholders. If adopted, these recommendations would serve as a firm foundation for the establishment of a Good AI Society.

Keywords Artificial intelligence · AI4People · Data governance · Digital ethics · Governance · Ethics of AI

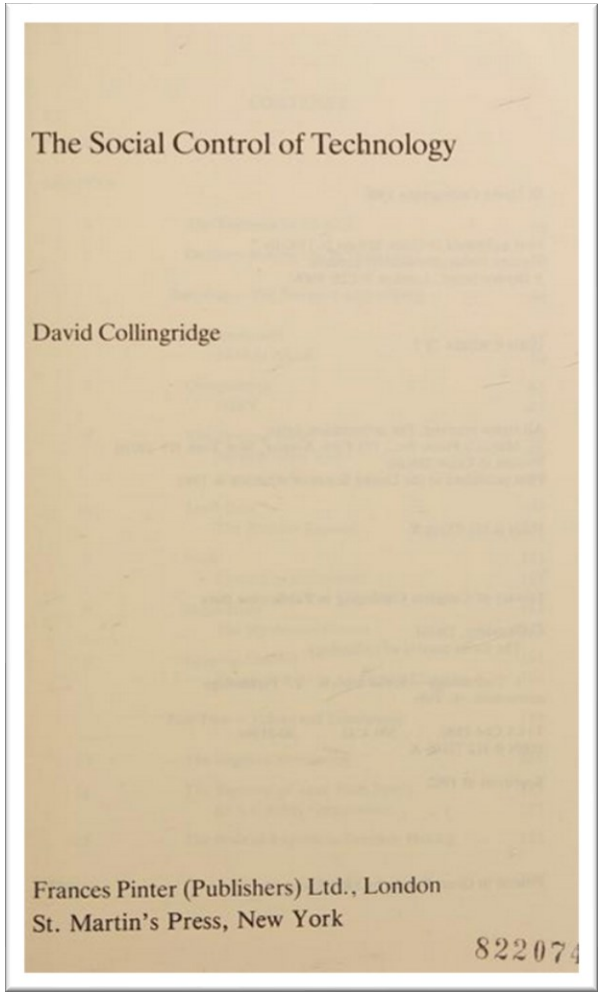
Wie soll man mit den Risiken umgehen?

- Ein Problem mit KI-Ethik Richtlinien: Sie sind freiwillig
- Es sind aber konkrete Rechte gefährdet → Freiwilligkeit ist nicht ausreichend zur Sicherung der Rechte, hier sind Staat und Gesellschaft in der Pflicht
- KI Regulierung und öffentliche Debatte notwendig!
  - EU AI Act
  - Bundesgesetze



Wie soll man mit den Risiken umgehen?

- Bei neuen Technologien gibt es das sogenannte „Kontroll-Paradox“
  - Am Anfang der technologischen Entwicklung sind nicht alle Risiken und unerwünschten Nebeneffekte bekannt
  - Bei fortgeschrittener technologischen Entwicklung weiß man mehr über die Risiken, aber es ist schwieriger die Technologie zu ändern oder zu regulieren



Wie soll man mit den Risiken umgehen?

- Lösungsansatz der Technikfolgenabschätzung:
  - Kritische Öffentlichkeit und fortwährende Debatte
  - Einbezug der Zivilgesellschaft in die Wissenschaft: Input spezieller Problematiken und Perspektiven durch Akteure aus der Praxis ist notwendig

Vielen Dank!

Kontakt: michael.schmidt@kit.edu