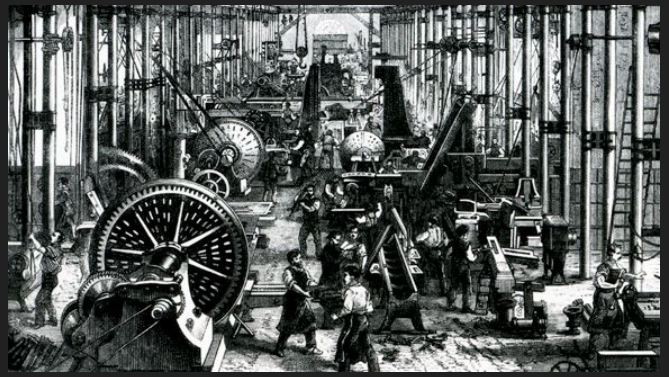
Ethische Künstliche Intelligenz

Marina Moreno
Ludwig-Maximilians-Universität München &
Solon Center for Policy Innovation

Die historische und globale Bedeutung von Technologie



Industrialisierung

- Produktivität
- Innovation

Globalisierung & Digitalisierung

- Kooperation
- Komplexe Abhängigkeiten

Die historische und globale Bedeutung von Technologie

Chancen: Weltarmut ⬇ Population ⬆ Wohlstand ⬆

Risiken: Kriegsinnovation ⬆ Umweltzerstörung ⬆ Resilienz ⬇

Technologie entwickelt sich oft exponentiell; die Fähigkeit der Weitsicht der Menschen aber nicht.

KI kurzfristig: Chancen

Vorteile KI gegenüber Menschen heute:

- Statistische Verarbeitung riesiger Datenmengen
- Schnelligkeit und Komplexität
- Freiheit von Denkfehlern und menschlichem Versagen

Beispiele dieses Potentials:

- Selbstfahrende Autos: Sicherer und zuverlässiger
- Trading Algorithms: Effizienter globaler Finanzmarkt
- Deep Medicine: Hilfestellung bei Diagnosen; individualisierte Medizin

KI kurzfristig: Risiken & Herausforderungen

Selbstfahrende Autos

- Moraltheoretische und rechtliche Probleme:

  Wie entscheidet es im Konfliktfall? Wer ist im Zweifelsfall verantwortlich?
  → Auflösung von stabilen ethischen und rechtlichen Konzepten

- Militärisches Potenzial:

  Lahmlegung von ganzen Versorgungswegen in Städten durch gezielten Computerangriff auf Autos
  → Komplexität und Abhängigkeit reduzieren Resilienz

KI kurzfristig: Risiken & Herausforderungen

Trading Algorithms

Flash Crash 2010: Unvorhergesehene Interaktion von Algorithmen
→ Komplexität reduziert Verständlichkeit
→ Abhängigkeit reduziert Resilienz

Deep Medicine
Biased Algorithms: Rassismus in der Datenmenge pflanzt sich fort
→ Ethisch problematische Folgerungen werden technologisch “legitimiert”

Solidarität: Individualisierte Risikoprämien bei der Gesundheitsversicherung?

KI kurzfristig: 4 Prinzipien

Funktionsweise einer KI muss…

- … nachvollziehbar sein.
- … prinzipiell vorhersagbar sein.
- … manipulationsresistent sein.
- … Bestimmung von Verantwortlichkeit unterstützen.

KI mittelfristig: Chancen

Trend: Steigende Automatisierung und steigender persönlicher Gebrauch

Beispiele dieses Potentials:

- Arbeitsautomatisierung: Steigende Produktivität und Effizienz (Wohlstand)
- Soziale Vernetzung: Schnelle Kommunikation und Information

KI mittelfristig: Risiken & Herausforderungen

Arbeitsautomatisierung

Arbeitslosigkeit ⬆ Ungleichheit ⬆
→ Die gesellschaftlich-kapitalistische Kooperation hängt wesentlich davon ab, dass Arbeit für (fast) alle verfügbar ist.

Einwand: Ökonom*innen argumentieren bisweilen, dass ähnliche Prognosen bisher noch nie eingetroffen sind.

Prinzip der Risikoabsicherung

Harvard und Oxford progonostizieren: Bis zu 40% Arbeitslosigkeit in 20 Jahren.

	Gesellschaftliche Vorbereitung	**Keine gesellschaftliche Vorbereitung**
Hohe Arbeitslosigkeit	Glatter Systemübergang	Chaos + Ungleichheit
Tiefe Arbeitslosigkeit	Systeminnovative Forschung ohne direkten Gegenstand	Status Quo

Wir können Zeilen nur bedingt wählen - Spalten aber schon.

KI mittelfristig: Risiken & Herausforderungen

Soziale Vernetzung

- Politischer Einfluss von Technokraten durch gezielte individualisierte Empfehlungen des Medienkonsums.
- Echochamber verstärken Polarisierung und Radikalisierung.
- DeepFakes verbreiten sich schnell.

KI langfristig: Chancen

Vorteile von KI gegenüber Menschen:

- Schnelligkeit
- Selbständige Optimierung
- Vervielfältigung
- Zugriff auf grosse Datenmengen

→ *Falls* generelle künstliche Intelligenz möglich ist, wird KI den Menschen in der Zukunft wohl überlegen sein.

Chance: Ein intelligenteres Wesen kann unsere Probleme lösen?

KI langfristig: Risiken und Herausforderungen

Alignment Problem: Welche Ziele sollte eine superintelligente KI verfolgen und wie stellen wir sicher, dass sie es tut?

Einige Probleme:

- Wir sind uns unseren praktischen und ethischen Zielen nicht einig.
- Wir wissen nicht, wie sich Ziele robust formalisieren lassen.
  → Büroklammernbeispiel
- KI als selbstoptimierende Blackbox
- KI ist intelligenter und damit mächtiger

KI langfristig: Risiken und Herausforderungen

Die grosse Mehrheit der relevanten Experten prognostizieren künstliche Superintelligenz vor dem Ende des 21. Jahrhunderts.

	Alignmentproblem gelöst	Alignment Problem ungelöst
Superintelligente KI	Utopie?	Dystopie?
Keine Superintelligente KI	Forschung ohne direkten Gegenstand	Status Quo

Wiederum: Zeilen können wir nur bedingt wählen - Spalten aber schon.

KI langfristig: Risiken und Herausforderungen

Zeilen beeinflussen: **Superintelligenz Arms Race**

First Mover Vorteil: Die erste genuine Superintelligenz könnte entscheidende Vorteile haben.

→ Sicherheit wird ohne internationale Regulierung dem Tempo geopfert.

Aber internationale Regulierung ist langsam und Technologieentwicklung häufig schnell.