Decoding Inequality:

Kritische Perspektiven auf Machine Learning und gesellschaftliche Ungleichheit

Forschungskolloquium „Critical Algorithm Studies“

Digital Humanities, Universität Bern

FS25

Dr. Rachel Huber und Dr. Moritz Mähr

Rachel Huber

- Studium Kulturmanagement, Kulturwissenschaften, Digital History in Zürich, Luzern und Hamburg
- Promotion in Digital History, Uni Luzern
- Assoziierte Forscherin, Uni Bern
- Projektleiterin für diverse Projekte zu digitalen Erinnerungskulturen
- Wissenschaftliche Mitarbeiterin und Projektleiterin im Statistischen Amt beim Kanton Zürich

Überblick Lebenszyklus ML-Systeme

- 1.) **Architekturauswahl:** Diskussion verschiedener ML-Architekturen und ihrer Auswirkungen auf Modellkapazitäten und -grenzen. Kritische Betrachtung, wie architektonische Entscheidungen bestimmte Voreingenommenheiten einbetten können.
- 2.) **Datensammlung:** Untersuchung von Datenquellen, Kuratierungs- und Filterprozessen. Kritische Perspektiven auf Repräsentationsprobleme, Copyright-Fragen und Umweltkosten der Datenspeicherung.

Überblick Lebenszyklus ML-Systemen

- 3.) **Training:** Technische Aspekte des Trainingsprozesses und Auswahl von Hyperparametern. Kritische Betrachtung der Umweltauswirkungen, Arbeitsbedingungen in der KI-Industrie und Machtkonzentration bei ressourcenstarken Unternehmen.
- 4.) **Anwendung:** Analyse verschiedener Anwendungsfälle von ML-Systemen, Feinabstimmung für spezifische Aufgaben und Bereitstellungsstrategien. Kritische Diskussion ethischer Überlegungen, potenzieller Missbrauchsszenarien und Fragen der Transparenz und Erklärbarkeit.

Überblick Lebenszyklus ML-Systemen

- 5.) **Evaluation und Überwachung:** Methoden zur Bewertung von Modellleistung und Verzerrungen. Kritische Perspektiven auf die Grenzen aktueller Evaluierungsmetriken.
- 6.) **Governance und Regulierung:** Diskussion aktueller und vorgeschlagener Regulierungsrahmen, ethischer Richtlinien und Herausforderungen bei der Steuerung sich schnell entwickelnder KI-Technologien.

Administratives

- Leistungsnachweis:
  - ❖ Blogbeitrag für Github
  - ❖ Poster
  - ❖ Präsentation
- Deadline der Beiträge: 16. Mai
- Vorstellung der Beiträge: 16. und 23. Mai
- Wahl des Themas bis: 7. März (bitte Dozierenden mündlich oder per E-Mail mitteilen)
- Infos und Updates auf: https://dhbern.github.io/decoding-inequality-2025/contents/home.html

Grundlagen Diskriminierung

- **Lernziele**

1. **Begriffsverständnis**: Diskriminierung, strukturelle Diskriminierung, unbewusste Diskriminierung (unconscious bias).
2. **Sozialer Kontext**: Erkennen, wie Sozialisierung und gesellschaftliche Machtverhältnisse Bias in Daten und Algorithmen einschleusen.
3. **Praxisbezug**: Verstehen, wie KI-Systeme Ungleichheiten reproduzieren können.
4. **Reflexion**: Eigene Vorannahmen hinterfragen und erste Schritte kennenlernen, um Bias in algorithmischen Systemen zu verringern.


MENTIMETER

Einführungsworkshop zu Formen der Diskriminierung

- Begriffe, die ihr mit Diskriminierung in Verbindung bringt

Definition Diskriminierung

- Diskriminierung ist eine qualifizierte Art von Ungleichbehandlung. Sie liegt vor, wenn drei Elemente gegeben sind:
- eine Ungleichbehandlung von Personen in vergleichbaren Situationen, …
- ... die an ein gruppenbezogenes Unterscheidungsmerkmal anknüpft und …
- ... eine Benachteiligung und/oder Herabsetzung beinhaltet

Diskriminierungsformen

- Welche Diskriminierungsformen kennt ihr?
  - Bsp.: Rassismus ist eine Diskriminierungsform
  - >>> andere Formen?

  - Schritt 1: je eine Form auf ein Post-it schreiben (so viele, wie ihr wollt)
  - Schritt 2: vorlesen und auf Whiteboard kleben

Bundesverfassung

2. Titel: Grundrechte, Bürgerrechte und Sozialziele

1. Kapitel: Grundrechte

Art. 7 Menschenwürde

Die Würde des Menschen ist zu achten und zu schützen.

Art. 8 Rechtsgleichheit

[1] Alle Menschen sind vor dem Gesetz gleich.

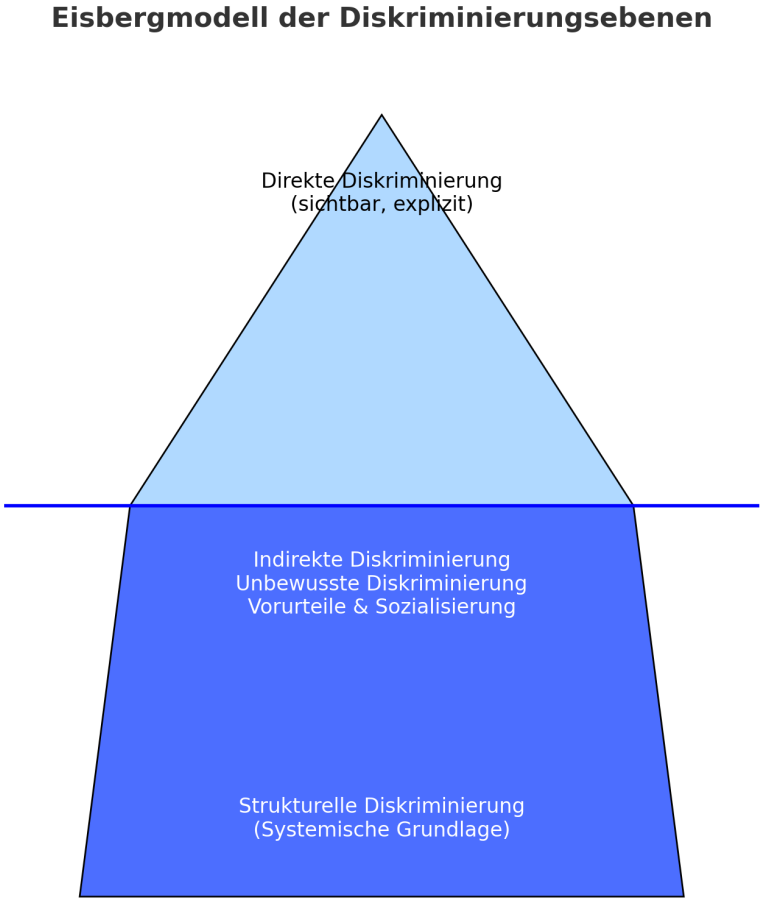
[2] Niemand darf diskriminiert werden, namentlich nicht wegen der Herkunft, der Rasse, des Geschlechts, des Alters, der Sprache, der sozialen Stellung, der Lebensform, der religiösen, weltanschaulichen oder politischen Überzeugung oder wegen einer körperlichen, geistigen oder psychischen Behinderung.

[3] Mann und Frau sind gleichberechtigt. Das Gesetz sorgt für ihre rechtliche und tatsächliche Gleichstellung, vor allem in Familie, Ausbildung und Arbeit. Mann und Frau haben Anspruch auf gleichen Lohn für gleichwertige Arbeit.

[4] Das Gesetz sieht Massnahmen zur Beseitigung von Benachteiligungen der Behinderten vor.

Auf dem Flipchart sind unterschiedliche Personae. Ordnet zusammen die Diskriminierungsformen den Personae zu.

Ebenen der Diskriminierung



Direkte Diskriminierung: Sichtbar und bewusst.

Indirekte Diskriminierung: Versteckt in Regeln und Praktiken.

Unbewusste Diskriminierung & Vorurteile: Implizite, oft unbemerkte Denkmuster.

Sozialisierung: Gesellschaftlich erlernte Werte und Normen.

Strukturelle Diskriminierung: Grundlage im System, schwer erkennbar.

- Content created with ChatGPT

Direkte Diskriminierung (sichtbar, explizit)

- **Was ist das?**
- Direkte Diskriminierung bezeichnet Handlungen, Äußerungen oder Verhaltensweisen, die unmittelbar gegen bestimmte Personen oder Gruppen gerichtet sind.
- Sie ist meist klar erkennbar und bewusst: Jemand wird offen benachteiligt, beleidigt oder ausgeschlossen, basierend auf einem Merkmal wie Hautfarbe, Geschlecht, sexueller Orientierung, Religion, Alter etc.
- **Beispiel im Studium:**
- Eine Studentin wird von einem Dozenten wiederholt ignoriert oder herabgewürdigt, weil er offen Frauen für „weniger kompetent“ hält.
- Ein Kommilitone wird aufgrund seiner Herkunft bei Gruppenarbeiten laufend nicht mit einbezogen.
- **Warum ist das relevant?**
- Direkte Diskriminierung kann das Studienerlebnis stark beeinträchtigen und zu Stress, Leistungsabfall oder sogar Studienabbruch führen.
- Es ist wichtig, direkte Diskriminierung zu erkennen und aktiv dagegen vorzugehen (z. B. melden an zuständige Stellen, sich solidarisieren mit Betroffenen).

Unbewusste Diskriminierung & Vorurteile

- **Was ist das?**
- Unbewusste Diskriminierung (auch „implizite“ Diskriminierung genannt) liegt vor, wenn Menschen ohne böse Absicht bestimmte Personengruppen benachteiligen, z. B. aufgrund von Stereotypen oder verinnerlichten Denkmustern.
- Vorurteile sind vorgefasste Meinungen oder Urteile über Menschen, die meist auf Stereotypen beruhen und selten hinterfragt werden.
- **Beispiel:**
- Man geht automatisch davon aus, dass eine Person mit Kopftuch eher konservativ eingestellt ist und vermeidet deshalb den Kontakt.
- **Warum ist das relevant?**
- Unbewusste Diskriminierung und Vorurteile beeinflussen das Lern- und Arbeitsklima, weil sie subtile Ausschlüsse erzeugen.
- Durch Reflexion (z. B. in Diversity-Workshops) und Feedback kann man diese Mechanismen erkennen und abbauen.

Sozialisierung

- **Was ist das?**
- Sozialisierung meint den Prozess, in dem wir, seit wir auf der Welt sind, gesellschaftliche Normen, Werte und Verhaltensweisen erlernen.
- Viele unserer Einstellungen, Stereotype und Vorurteile entstehen durch die Familie, das Umfeld, die Medien und das Bildungssystem, das wir (lange) nicht hinterfragen.
- **Beispiel im Studium:**
- Aufgewachsen in einer Gesellschaft, in der Geschlechterrollen klar verteilt sind („Frauen sind emotional, Männer rational“), kann man diese Vorstellungen unreflektiert ins Studium mitnehmen.
- Auch Hochschulen tragen zur Sozialisierung bei, indem sie bestimmte Leitbilder vermitteln (z. B. wer als „typische Führungskraft“ gilt).
- **Warum ist das relevant?**
- Was wir als „normal“ empfinden, ist oft ein Ergebnis unserer Sozialisierung. Wer bestimmt, was «normal» ist?
- Wenn wir uns dessen bewusst werden, können wir Stereotype und Vorurteile hinterfragen und uns für ein inklusiveres Miteinander einsetzen.

Indirekte Diskriminierung (versteckt in Regeln und Praktiken)

- **Was ist das?**
- Indirekte Diskriminierung liegt vor, wenn scheinbar neutrale Regelungen oder Abläufe bestimmte Gruppen systematisch benachteiligen.
- Die Benachteiligung ist nicht immer beabsichtigt, ergibt sich aber aus strukturellen oder organisatorischen Entscheidungen.
- **Beispiel im Studium:**
- Ein Seminar wird ausschließlich zu Zeiten angeboten, in denen Studierende mit Pflegeaufgaben (z. B. Kinderbetreuung) kaum teilnehmen können. Obwohl das Seminar für alle offen ist, werden Menschen mit Care-Verantwortung indirekt ausgeschlossen.
- Eine Fachschaftsveranstaltung findet in einer Kneipe statt, was Studierende mit religiösen Hintergründen oder Alkoholverzicht unwohl fühlen lässt oder ausschließt.
- **Warum ist das relevant?**
- Indirekte Diskriminierung ist oft schwerer zu erkennen, da sie nicht auf offensichtlicher Ablehnung beruht.
- Sie zeigt sich jedoch in Regelungen, die auf den ersten Blick „neutral“ wirken, aber bestimmte Personengruppen de facto benachteiligen.

Strukturelle Diskriminierung (systemische Grundlage)

- **Was ist das?**
- Strukturelle Diskriminierung ist in den Institutionen und gesellschaftlichen Systemen selbst verankert.
- Sie manifestiert sich in Gesetzen, Richtlinien, Organisationsstrukturen und kulturellen Normen, die bestimmte Gruppen systematisch benachteiligen.
- **Beispiel im Studium:**
- Studienordnungen oder Finanzierungssysteme (BAföG, Stipendien) können Anforderungen stellen, die bestimmte Gruppen (z. B. Personen aus einkommensschwachen Familien, Menschen mit Behinderung) nur schwer erfüllen können.
- Hochschulstrukturen können so gestaltet sein, dass die Teilhabe von marginalisierten Gruppen (z. B. Studierende mit Migrationsgeschichte, People of Color, queere Personen) erschwert wird.
- **Warum ist das relevant?**
- Strukturelle Diskriminierung bildet die Grundlage, in der alle anderen Formen eingebettet sind.
- Um Diskriminierung langfristig abzubauen, müssen auch die Strukturen verändert werden (z. B. gerechtere Zulassungsverfahren, barrierefreie Hochschulgebäude, Sensibilisierung des Lehrpersonals).

Warum ist das relevant für ML-Systeme?

- Wie gelangen diese gesellschaftlichen Vorurteile und Machtverhältnisse in Daten?

- In diesem Forschungskolloquium geht es darum einen Überblick zu bekommen, wie KI-Systeme daraus Diskriminierung reproduzieren können. Und was es braucht, damit KI-Systeme keine Nachteile für die Gesellschaft (re)produzieren.

Leseauftrag auf die nächste Session, 28.02.

- Fabian Offert und Ranjodh Singh Dhaliwal, «The Method of Critical AI Studies, A Propaedeutic», 10. Dezember 2024, https://doi.org/10.48550/arXiv.2411.18833.

Leseauftrag

- Zhisheng Chen, Ethics and discrimination in artificial intelligenceenabled recruitment practices, Humanities and Social Sciences Communications, 10, Article 567 (2023), S. 1-12.

Inhaltliche Auseinandersetzung mit KI als Ganzes

- Was bedeutet die Vierte Industrielle Revolution?
- Was bedeutet sie für das Thema KI und Gesellschaft?
- Wie ordnet der Autor KI als Ganzes ein?

Text- und Autorenkritik

- Autoren-Backgroundcheck
- Wie findet ihr den Text?

Textanalyse KI und Recruitment

1. Applications and benefits of AI-based recruitment
2. Factors contributing to algorithmic recruitment discrimination
3. Types of discrimination in algorithmic recruitment
4. Measures to mitigate algorithmic hiring discrimination

Biases in Algorithmic Recruitment

- Welche Diskriminierungsdimensionen erwähnt der Autor?
- Race, Gender, Personality
  - Warum können diese Diskriminierungsdimensionen in automatisierten Bewerbungsprozessen vorkommen?
- Auf welche geht er nicht ein?
- Alter, sozialer Status
  - Welche Variablen / Labels, Charaktereigenschaften im Bewerbungsprozess führen zu Diskriminierung für diese Bevölkerungsrguppen?

- Facebook Microtargeting
- Jobinserate werden von Facebook an gezielte Zielgruppen angezeigt, welche der Kunde sucht. Beispielsweise Alterskohorte 25-36
- ProPublica-Recherche zeigte, dass grosse Stellenanbieter wie Amazon, Verizon UPS damit arbeiteten und gegen das Altersdiskriminierungsverbot in den USA verstiessen
- Auch Google und LinkedIn diskriminierten so potenzielle Bewerber, die älter als 40 waren
- Beispiel in der Schweiz?

Beispiel für Altersdiskriminierung

Beispiel für sozialer Status

- Aufnahmeprüfung für Unis in England:
- "As someone from a poor family in a poor postcode I dread to think what would have happened to my grades if I'd been victim to this flawed system.“
- Daten: frühere Prüfungen, durchschnittliche Noten der Unis pro Stufe
- Algorithmus stufte Noten von Studierenden aus ärmeren Gegenden/öffentlichen Unis tendenziell runter und Noten von Studierenden aus privilegierten Gegenenden/privaten Unis tendenziell hoch

Wieso sehen viele automatisiertes Recruiting als fairer an?

Mythos KI ist unfehlbar und unvoreingenommen

- Der Mythos, dass künstliche Intelligenz unfehlbar und unvoreingenommen ist, ist eine weit verbreitete Annahme, die die Objektivität von KI-Systemen überschätzt und die möglichen Fehler und Voreingenommenheit ignoriert, die bei der Entwicklung und Anwendung von KI auftreten können. KI-Systeme, insbesondere solche, die maschinelles Lernen und neuronale Netze nutzen, sind von der Menge und Qualität der Daten abhängig, mit denen sie trainiert und gefüttert werden. Daher können sich Fehler und Verzerrungen in den Daten auch in den Ergebnissen der KI-Systeme niederschlagen.

Prominentes Beispiel für Genderdiskriminierung

Prominentes Beispiel

- Was war die Grundlage für Bias im Amazonbeispiel?
- Daten: Bewerbungsläufe von Amazon der letzten 10 Jahre
- Was war das Problem in den Daten?
- Es wurden in den letzten 10 Jahren vornehmlich Männer eingestellt
- Wieso war das ein Problem für den Algorithmus?
- Der Algorithmus hat gelernt, dass Männer das präferierte Outcome sind.

Variablen

- Geschlecht
- Alter
- Herkunft

  - Man muss die rausnehmen, um Algorithmen fairer beurteilen zu lassen
  - aber …

Proxy-Variablen

- In der Regel werden automatisierte Recruiting-Systeme so trainiert, dass Faktoren wie Herkunftsland, Lebensalter oder Geschlecht für die Auswahl nicht beachtet werden sollten. Das Problem: Es gibt auch subtilere Attribute in Bewerbungen, sogenannte «Proxies» (deutsch: Stellvertretervariablen), die auf diese demografischen Eigenschaften hinweisen können, wie etwa die Sprachkompetenz, ausländische Arbeitserfahrung oder ein Studium im Ausland. So hat dieselbe Studie aufgedeckt, dass ein Auslandstudium bei 80% der Bewerber:innen dazu führte, dass sie mit weniger Punkten bewertet wurden. Letztlich kann dies dazu führen, dass Bewerber:innen, die nicht in dem Land aufgewachsen sind und studiert haben, in dem die Stelle ausgeschrieben ist, ungerecht behandelt werden.

Proxy Variablen

- https://www.youtube.com/watch?v=NslS2klFTEY

Monitoring / Lösung beinhaltet

1. Rigorous Dataset Auditing: Ensuring datasets are diverse, representative, and free from harmful biases through continuous review and updates.
2. Algorithmic Transparency: Designing AI models with clear documentation and explainability to uncover and address biases in decision-making.
3. Fairness as a Core Principle: Embedding fairness metrics during the development and deployment stages to actively identify and counteract discrimination.
4. Regular Monitoring: Establishing feedback loops to monitor AI outputs and rectify emerging biases dynamically.

Lösungen

- Problem beschrieben im Text:
  - „Computer scientists are often not trained to consider social issues in context“
- Lösung
  - Trading Zones: Computer Scientists, Data Engineers **and** Humanists
  - Diverse programmer teams

Situation in der EU

- Der EU AI-Act der Europäischen Union (EU) stuft KI-gestützte Einstellungssoftware und Mitarbeiterverwaltungssysteme als „hohes Risiko“ ein (EUR-Lex, 2021). Das bedeutet, dass Systeme mit hohem Risiko eine Vielzahl von Anforderungen erfüllen müssen, um in der EU eingesetzt werden zu können, und wenn sie diese nicht erfüllen, muss der betreffende Mitgliedstaat das System einschränken, verbieten oder zurückrufen.

Situation in der Schweiz

- **Algorithmische Diskriminierung in der Schweiz:** Der bestehende Diskriminierungsschutz in der Schweiz bietet keinen wirksamen Schutz vor Diskriminierung durch algorithmische Systeme und muss verstärkt werden.
- Es gibt kein Pendant zum EU-AI-Act

Situation in der Schweiz

- Das Diskriminierungsverbot betrifft grundsätzlich nur staatliche Akteure wie etwa Behörden. Denn in der Schweiz existiert **kein allgemeines Gesetz, das in genereller Weise Diskriminierung durch Private untersagt**. Algorithmische Systeme verbreiten sich jedoch schnell in der gesamten Gesellschaft und werden in grosser Zahl von Privaten entwickelt und eingesetzt. Daher braucht es hier eine gesetzliche Anpassung.

Situation in der Schweiz

- Der bestehende Schutz gegen Diskriminierung reicht zudem nicht aus, um den **besonderen Merkmalen der algorithmischen Diskriminierung** zu begegnen (wie etwa Skalierungseffekte und Rückkopplungsschleifen)

Vorbereitung für den 28.03.25

- Thema CCTV / Facial Recognition
- Leseauftrag: Joy Buolamwini und Timnit Gebru, «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification», in Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Conference on Fairness, Accountability and Transparency, PMLR, 2018), 77–91.

Decoding Inequality:
Kritische Perspektiven auf Machine Learning und gesellschaftliche Ungleichheit

28.03.25

Application (CCTV/Facial Recognition)

Rachel Huber

Leseauftrag

- Leseauftrag: Joy Buolamwini und Timnit Gebru, «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification», in Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Conference on Fairness, Accountability and Transparency, PMLR, 2018), 77–91.

Kontext für Text

https://www.youtube.com/watch?v=UG_X_7g63rY

Application (CCTV/Facial Recognition)

- Autorinnen-Backgroundcheck
  - Joy Buolamwini
  - Timnit Gebru

Application (CCTV/Facial Recognition)

- Was sind Lösungen, welche die Autorinnen beschreiben?

Application (CCTV/Facial Recognition)

- Was ist im Text mit high stake-sector gemeint?
- Was kennt ihr noch für high stake sectors?

Application (CCTV/Facial Recognition)

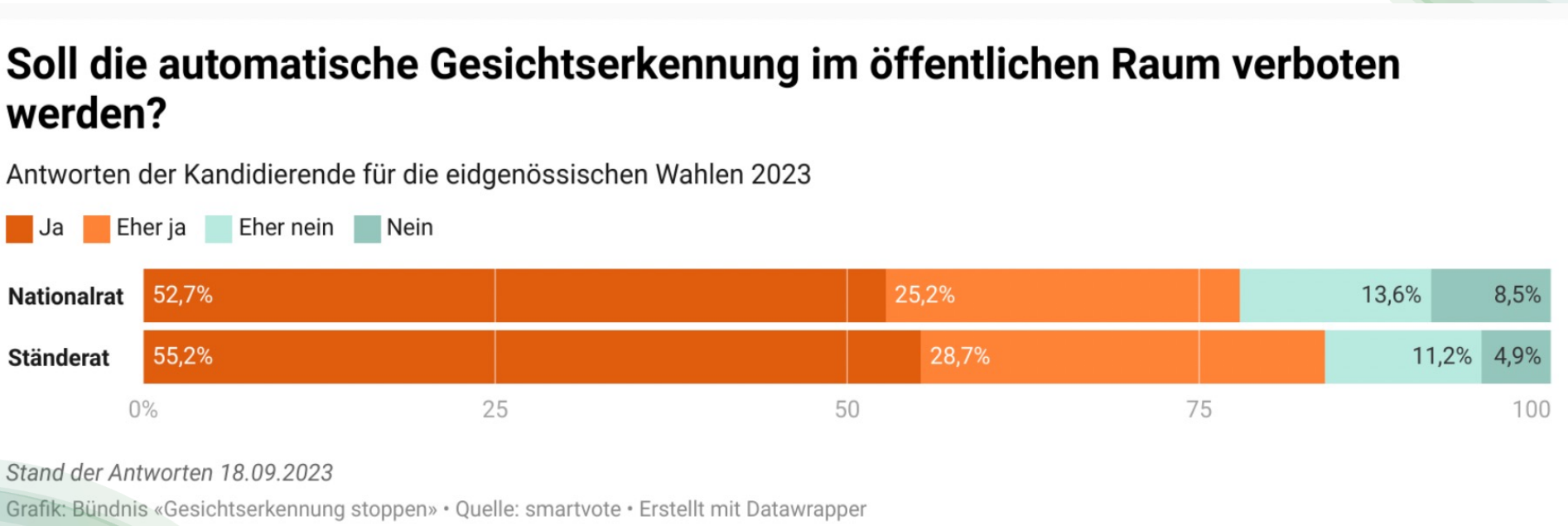
- Wo wird Face detection, classification und Recognition eingesetzt in den USA?
- Wie sieht die Situation in der Schweiz aus?

Application (CCTV/Facial Recognition)

- Die Autorinnen reden von intersektionaler Vorgehensweise. Welche Intersektion beschreiben und untersuchen sie?

Application (CCTV/Facial Recognition)

Film «The Coded Bias» von Shalini Kanayya
auf Netflix (20 Min.)



Situation in der Schweiz?

Situation in der Schweiz?

- In den Städten Zürich, St. Gallen und Lausanne sowie im Kanton Basel-Stadt haben die Parlamente Vorstösse für ein Verbot der Gesichtserkennung bereits angenommen, Digitale Gesellschaft. Ähnliche Vorstösse sind noch in Bearbeitung in den Städten Luzern und Genf sowie in den Kantonen Zürich und Basel-Landschaft. Dies zeigt eine aktuelle Übersicht des Bündnisses «Grundrechte schützen – Gesichtserkennung stoppen» der Organisationen Digitale Gesellschaft, AlgorithmWatch CH und Amnesty International.

Quelle: Webseite Digitale Gesellschaft, Kampagne «Gesichtserkennung stoppen», 16.09.2023

Situation in der Schweiz?

- Was ist problematisch an biometrischen Erkennungssystemen?
- Kann Menschen davon Abhalten, im öffentlichen Raum an Demonstrationen oder Kundgebungen ihre Meinung zu äussern, oder an Demonstrationen oder Kundgebungen teilzunehmen.
- Damit sind Grundrechte wie Meinungsäusserungsfreiheit und Versammlungsfreiheit beschnitten

Situation in der Schweiz?

- Case: SBB wollte ab Herbst 2023 in 50 Bahnhöfen der Schweiz die Kundenströme messen. Dabei sollte biometrische Erkennung eingesetzt werden, welche die Personen in Alter, Geschlecht und Grösse einordnen soll. Damit sollten die Umsätze in den Bahnhofsops gesteigert werden.
- Digitale Gesellschaft Schweiz, Algorithm Watch lancierten eine Kampagne und sammelten Unterschriften von über 17‘000 Personen, NGOs und Parteien, die sich dagegen aussprachen.
- Auf den zivilgesellschaftlichen Druck hin, nahm die SBB die Erkennung von biometrischen Daten aus dem Kundenfrequenzmesssystem raus.

Quelle: Gesichtserkennung-stoppen.ch

Leseauftrag auf den 4. April 2024

- Lauren Kirchner, Mattu Jeff Larson, «Machine Bias» (ProPublica)
- «Jobs für Flüchtlinge - Algorithmus verteilt neu Asylbewerber auf Kantone» (Schweizer Radio und Fernsehen (SRF), 10. Mai 2018)
- «Algorithmus verbessert Erwerbschancen von Flüchtlingen» (ETH Zürich, 18. Januar 2018)

Decoding Inequality: Kritische Perspektiven auf Machine Learning und gesellschaftliche Ungleichheit

24.04.25

Application (Predictive Policing USA und Schweiz/ Migrationsalgorithmus)

Rachel Huber

Textanalyse Pro Publica

- Background-Check ProPublica

RISK Assessment System USA

- COMPAS
- Correctional Offender Management Profiling for Alternative Sanctions
- Der Fall ist von 2016, wird COMPAS noch benützt in den USA?
- Wo wird es noch eingesetzt?
- New York, Wisconsin, California, Florida
- 2021 wurde COMPAS in 46 US-Bundesstaaten eingesetzt

RISK Assessment System USA

- COMPAS tends to make the reverse error with whites: they are more prone to being labeled as lower-risk yet end up committing more crimes.
- A mere 20 percent of those who were projected to engage in violent criminal activities really followed through with such actions

COMPAS USA

- Wieso ist der Algorithmus rassistisch?
- Wie kamen die dem Algorithmus zugrunde liegenden Daten zustande?

Hauptprobleme beim Einsatz des Risk Assessment Systems?

- Datenprobleme
- Intransparenz
- Fehlende Kontextualisierung

Stand heute?

- Wird noch eingesetzt
- Immer mehr Bezirke verzichten aufgrund Protests darauf

Lösungen?

- **Algorithmic transparency** — making it clearer how risk scores are calculated.
- **Human-in-the-loop systems**, where risk assessments inform but don’t dictate outcomes.
- **Open-source or explainable AI** alternatives to proprietary systems like COMPAS.

Predictive Policing in der Schweiz?

- FOTRES Kanton Zürich (ähnliches System wie COMPAS)
  - Aus wissenschaftlicher Perspektive kritisch, weil es zu wenig untersucht wurde.
  - Es gab bisher keine unabhängige Studie zu FOTRES

Predictive Policing in der Schweiz

- FOTRES ist nicht ein fester, unveränderter Algorithmus, sondern wird ständig weiterentwickelt. Ab einem bestimmten Grad der Weiterentwicklung haben wir es gewissermassen mit einem neuen Algorithmus zu tun, und dieser müsste auf seine korrekte Funktionsweise und potenziell diskriminierenden Auswirkungen hin erneut überprüft werden. Allerdings ist dies im Fall von FOTRES seit 2011 nicht geschehen. Gegenwärtig wird FOTRES in der Version 4.0 angewendet, in der Studie von 2011 wurde aber FOTRES 2.0 untersucht.

Positive Aspekte: Data Mining

- Hinsichtlich der Ermittlungsarbeit steht das Potenzial für die Beweismittelauswertung durch Bild- und Texterkennung im Mittelpunkt, etwa für die strafrechtliche Verfolgung von Kinderpornografie. Wird bei Ermittlungen in diesem Kriminalitätsbereich potenzielles Beweismaterial beschlagnahmt, stehen die Ermittlungsbehörden vor der Aufgabe, immer grössere Datenmengen auswerten zu müssen. Hier könnten KI-Systeme eine Hilfe sein, um das Beweismaterial vorzusortieren. So würden nicht nur die Ermittlungspersonen psychisch entlastet, sondern es könnten auch mehr Verfahren mit den zur Verfügung stehenden personellen Ressourcen betrieben werden.

Positive Aspekte: Data Mining

- Texterkennungssysteme bei umfangreichen Ermittlungen im Bereich der Wirtschafts- und Steuerkriminalität.

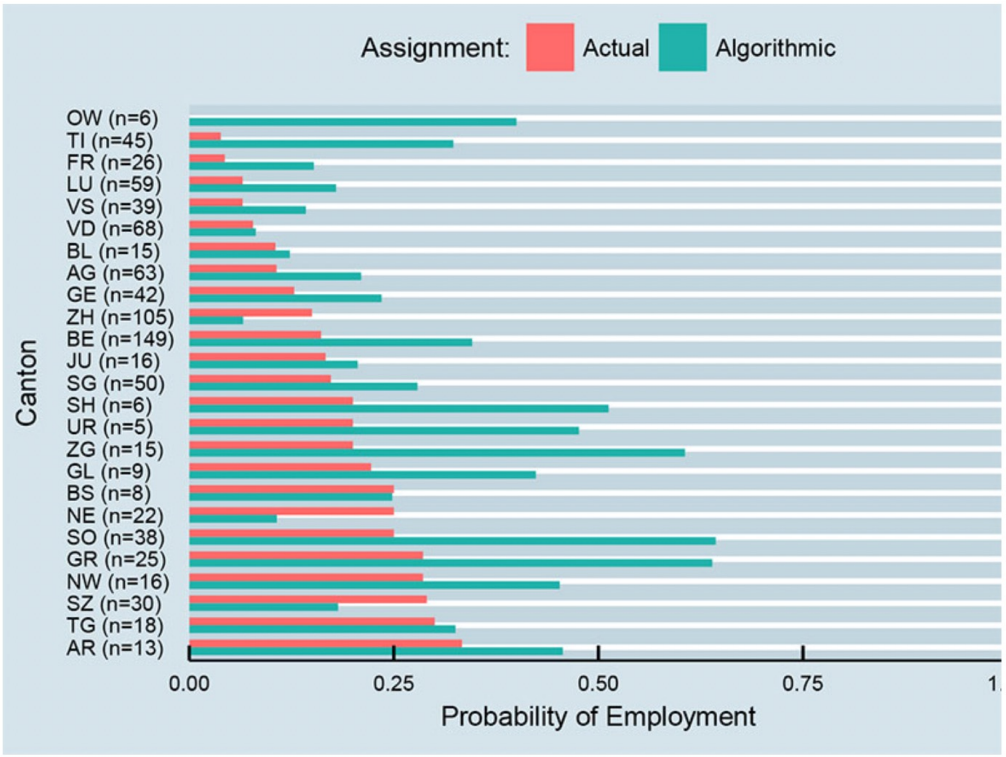
Migrationsalgorithmus

Schweiz

Migrationsalgorithmus

1. Wie wird der Algorithmus entwickelt?
2. Datengrundlage: Staatsekretariat für Migration, Datensatz mit Angaben zu ca. 22’000 Asylsuchenden
3. Was sind die Vorteile?
4. Was könnten Nachteile sein?
5. Wie ist eure Haltung dazu?

Vorteile Migrationsalgorithmus



Bildquelle: https://ethz.ch/de/news-und-veranstaltungen/eth-news/news/2018/01/algorithmus-verbessert-erwerbschancen-von-fluechtlingen.html

Migrationsalgorithmus

- Berücksichtigt wird Bevölkerungsgrösse eines Kantons, und dass sich die Nationalitäten gleichmässig über alle Kantone verteilen.
- Die Erwerbstätigkeit von Asylsuchenden in der Schweiz könnte mit einem datengestützten Ansatz von 15 auf 26 Prozent erhöht werden

11. April 2025 mit Dr. Moritz Mähr

- Application: ChatGPT
- Emily M. Bender u. a., «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜», in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–23

Decoding Inequality: Kritische Perspektiven auf Machine Learning und gesellschaftliche Ungleichheit

09.05.25

Governance and Regulation

Rachel Huber

Lernziele KI-Regulierung und Governance

1. Die Notwendigkeit und Herausforderungen einer umfassenden KI-Regulierung erklären können
2. Die zentralen Risiken unregulierter KI-Systeme reflektieren.
3. Die vier Risikokategorien des EU AI Acts verstehen und anwenden können.
4. Den EU-Ansatz mit den Regulierungsmodellen der Niederlande und des UK vergleichen.
5. Die Schweizer Position zur KI-Regulierung einordnen

Die Notwendigkeit einer umfassenden KI-Regulierung

1. Schutz von Grundrechten und Demokratie

- KI kann Diskriminierung verstärken (z. B. bei Bewerbungen, Strafjustiz, Migration).
- Persönlichkeitsrechte, Privatsphäre und Menschenwürde müssen geschützt werden.

2. Transparenz und Nachvollziehbarkeit

- Viele KI-Systeme sind Black Boxes – Entscheidungen sind schwer nachvollziehbar.
- Regulierung soll sicherstellen, dass Menschen verstehen, wie und warum Entscheidungen getroffen werden.

3. Vermeidung von Schäden

- KI kann physische, psychologische oder wirtschaftliche Schäden verursachen (z. B. autonome Fahrzeuge, Fake News, Deepfakes).
- Regulierung ist nötig, um klare Haftungsfragen zu klären und Sicherheitsstandards festzulegen.

Die Notwendigkeit einer umfassenden KI-Regulierung

4. Vertrauensbildung in KI-Technologien

- Klare Regeln fördern Akzeptanz und verantwortungsvolle Innovation.
- Unternehmen und Nutzer:innen brauchen Sicherheit im Umgang mit KI.

5. Faire wirtschaftliche Rahmenbedingungen

- Große Tech-Konzerne haben strukturelle Vorteile.
- Regulierung soll Wettbewerbsverzerrungen verhindern (z. B. Datenzugang, Open-Source-Förderung).

Herausforderungen einer umfassenden KI-Regulierung

1. Schneller technologischer Wandel

- KI entwickelt sich rasant weiter (z. B. Generative KI).
- Regulierungsprozesse sind oft langsamer als die Technik selbst.

2. Abgrenzung und Definition von KI

- Was genau fällt unter „KI“?
- Viele Anwendungen sind **hybride oder basieren auf traditionellen** Algorithmen.

3. Internationale Fragmentierung

- Unterschiedliche Standards in EU, USA, China, Schweiz etc.
- Risiko von „Regulierungsflucht“ und Inkompatibilitäten.

Herausforderungen einer umfassenden KI-Regulierung

4. Balance zwischen Innovation und Kontrolle

- Zu strikte Regeln könnten Innovationen hemmen.
- Zu lockere Regeln könnten Risiken vergrößern.

5. Komplexität der Systeme

- Technische Tiefe erschwert effektive Kontrolle und Überwachung.
- Behörden benötigen Expertise, Ressourcen und geeignete Prüfmechanismen.

Was sind die zentralen Risiken unregulierter KI-Systeme?

1. Diskriminierung und soziale Ungleichheit

- KI kann bestehende Vorurteile aus Trainingsdaten übernehmen und verstärken
- Risiken in sensiblen Bereichen wie Justiz, Polizei, Migration, Arbeitswelt
- Beispiel: Algorithmische Systeme benachteiligen Bewerber:innen aufgrund von Geschlecht, Herkunft oder Alter

2. Intransparente Entscheidungen (Black Boxes)

- Viele KI-Systeme sind nicht nachvollziehbar
- Betroffene erfahren nicht, wie Entscheidungen zustande kommen – etwa bei Krediten, Sozialleistungen oder Bewerbungen
- Das untergräbt Rechtsstaatlichkeit und individuelle Autonomie

Was sind die zentralen Risiken unregulierter KI-Systeme?

Desinformation und Manipulation

- Generative KI (z. B. Deepfakes, Chatbots) kann gezielt eingesetzt werden, um Wahlen zu beeinflussen, Hass zu schüren oder Vertrauen in Institutionen zu untergraben
- Ohne Regulierung fehlt es an Transparenzpflichten und Herkunftskennzeichnung

4. Sicherheitsrisiken

- Unkontrollierte KI in autonomen Systemen (Fahrzeuge, Drohnen, Waffen) kann physische Schäden verursachen
- Auch in der Cybersicherheit entstehen neue Angriffsvektoren (KI-generierte Malware, Phishing)

Was sind die zentralen Risiken unregulierter KI-Systeme?

5. Verlust menschlicher Kontrolle

- In komplexen automatisierten Systemen (z. B. im Finanzwesen oder in der Überwachung) werden Entscheidungen immer häufiger automatisiert ohne menschliches Eingreifen getroffen.
- Gefahr: Entscheidungen ohne Verantwortung und Kontrolle.

6. Machtkonzentration und wirtschaftliche Abhängigkeit

- Große KI-Modelle sind meist in der Hand weniger Konzerne (z. B. OpenAI, Google, Meta).
- Ohne Regulierung entsteht ein ungleicher Zugang zu Technologie und Märkten – insbesondere für KMU und den globalen Süden.

7. Ökologische Auswirkungen

- Große KI-Modelle (z. B. GPT, Stable Diffusion) verbrauchen massive Mengen an Energie und Wasser – oft ohne ökologischen Ausgleich.
- Fehlende Umweltstandards fördern eine nicht nachhaltige Skalierung.

Vier Risikokategorien des EU AI Acts 1

1. **Unzulässiges Risiko (Unacceptable Risk)**

KI-Systeme in dieser Kategorie sind verboten, da sie als unvereinbar mit den Werten und Grundrechten der EU gelten. Beispiele umfassen:

- Soziales Scoring durch staatliche Stellen, das Personen basierend auf ihrem Verhalten oder persönlichen Merkmalen bewertet.
- Echtzeit-Biometrie zur Fernidentifikation in öffentlich zugänglichen Räumen (z. B. Gesichtserkennung).
- Manipulative KI, die menschliches Verhalten durch subliminale Techniken beeinflusst.
- Ausnutzung von Schwachstellen bestimmter Gruppen, wie Kinder oder Menschen mit Behinderungen.

Solche Systeme sind grundsätzlich verboten, mit wenigen Ausnahmen, beispielsweise für Strafverfolgungsbehörden unter strengen Bedingungen.

Vier Risikokategorien des EU AI Acts 2

2. Hohes Risiko (High Risk)

Diese KI-Systeme dürfen eingesetzt werden, unterliegen jedoch strengen Anforderungen, da sie erhebliche Auswirkungen auf Gesundheit, Sicherheit oder Grundrechte haben können. Beispiele sind:

- KI in kritischen Infrastrukturen (z. B. Energieversorgung, Verkehr).
- Biometrische Identifikationssysteme.
- Systeme zur Bewertung von Schülern oder zur Bewerberauswahl.
- KI zur Kreditwürdigkeitsprüfung oder Versicherungsbewertung.

Anforderungen umfassen unter anderem Risikomanagement, Datenqualitätssicherung, Transparenz, menschliche Aufsicht und Konformitätsbewertung.

Vier Risikokategorien des EU AI Acts 3

3. Begrenztes Risiko (Limited Risk)

KI-Systeme mit begrenztem Risiko unterliegen Transparenzpflichten. Dies bedeutet, dass Nutzer darüber informiert werden müssen, dass sie mit einem KI-System interagieren. Beispiele:

- Chatbots.
- Deepfakes oder KI-generierte Inhalte.

Solche Systeme müssen klar kennzeichnen, dass Inhalte von KI generiert wurden, um Nutzer nicht zu täuschen.

Vier Risikokategorien des EU AI Acts 4

4. Minimales oder kein Risiko (Minimal or No Risk)

Diese Kategorie umfasst die Mehrheit der KI-Anwendungen, die keine spezifischen regulatorischen Anforderungen erfüllen müssen. Beispiele:

- Spam-Filter.
- KI in Videospielen.

Obwohl keine gesetzlichen Verpflichtungen bestehen, wird empfohlen, freiwillige Verhaltenskodizes zu befolgen, um ethische Standards zu gewährleisten.

Den EU-Ansatz und Regulierungsmodelle der Niederlande und des UK

EU: strikte, risikobasierte Regulierung (EU AI Act)

ASPEKT	DETAILS
Gesetzgebung	EU AI Act (2024), direkt anwendbare Verordnung
Regulierungsansatz	Risikobasiert: 4 Kategorien (unzulässig, hoch, begrenzt, minimal)
Ziele	Schutz von Grundrechten, Sicherheit, Marktregulierung, Innovation
Instrumente	Verpflichtende Konformitätsprüfungen, Transparenzpflichten, Sanktionen
Fokus	Einheitlicher Binnenmarkt, Kontrolle von Hochrisiko-Systemen
Besonderheit	Extraterritoriale Wirkung (betrifft auch Drittstaaten wie CH, USA)

Den EU-Ansatz und Regulierungsmodelle der Niederlande und des UK

UK: Pro-Innovation-Ansatz („Agile Regulation“)

Aspekt	Details
Gesetzgebung	Noch kein zentrales KI-Gesetz, sondern sektorspezifisch über Aufsichtsbehörden
Regulierungsansatz	Leichtgewichtig, prinzipienbasiert (z. B. Fairness, Sicherheit, Erklärbarkeit)
Ziele	Förderung von Innovation und Wettbewerbsfähigkeit
Instrumente	Freiwillige Leitlinien, „Sandboxen“, branchenspezifische Empfehlungen
Fokus	Aufsicht durch bestehende Regulierungsbehörden (z. B. ICO, CMA, Ofcom)
Besonderheit	Kein zentraler Regulierer, keine harte Klassifikation nach Risiko

Den EU-Ansatz und Regulierungsmodelle der Niederlande und des UK

Niederlande: Wertorientierter und vorsorgender Ansatz

Aspekt	Details
Gesetzgebung	Keine eigene KI-Verordnung, aber aktive Umsetzung des EU AI Acts
Regulierungsansatz	Kombiniert EU-Vorgaben mit starkem Fokus auf Ethik, Teilhabe, Transparenz
Ziele	Menschzentrierte KI, inklusiv, fair, rechenschaftspflichtig
Instrumente	KI-Ethikrichtlinien, Experimentierräume (z. B. „AI Impact Assessment“), algorithmische Audits
Fokus	Frühzeitige Einbindung von Bürger:innen und zivilgesellschaftlichen Akteuren
Besonderheit	Pionier bei praktischen Prüfwerkzeugen wie dem „Algorithm Register“

Position der Schweiz

- Am 12. Februar 2025 hat der Bundesrat entschieden, die KI-Konvention des Europarats zu ratifizieren. Diese Konvention legt grundlegende Prinzipien für den Umgang mit KI fest, insbesondere im Hinblick auf Menschenrechte, Demokratie und Rechtsstaatlichkeit. Die Umsetzung in der Schweiz soll jedoch nicht durch ein umfassendes, horizontales Gesetz erfolgen, sondern durch gezielte Anpassungen bestehender Gesetze in spezifischen Sektoren wie Datenschutz, Produktsicherheit oder Arbeitsrecht. Dabei sollen auch nicht-bindende Massnahmen wie Branchenstandards oder Selbstregulierung zum Einsatz kommen. Ziel ist es, Innovation zu fördern, Grundrechte zu schützen und das Vertrauen der Bevölkerung in KI zu stärken.

Vergleich: Schweiz vs. EU

Aspekt	Schweiz	EU AI Act
Regulierungsansatz	Prinzipienbasiert, sektor- und themenspezifisch	Detailliert, risikobasiert, umfassend
Zielsetzung	Förderung von Innovation, Schutz der Grundrechte, Vertrauensbildung	Schutz von Grundrechten, Harmonisierung des Binnenmarkts, Innovationsförderung
Anwendungsbereich	Primär staatliche Akteure; private Akteure bei Grundrechtsrelevanz	Staatliche und private Akteure gleichermaßen
Umsetzung	Anpassung bestehender Gesetze, keine neue umfassende KI-Gesetzgebung	Neue, umfassende Verordnung mit direkten Verpflichtungen
Internationaler Fokus	Ausrichtung an Europarats-Konvention, Beobachtung internationaler Entwicklungen	Setzt internationale Standards, extraterritoriale Wirkung

16.05.25 Mai mit Moritz Mähr

- Beiträge präsentieren