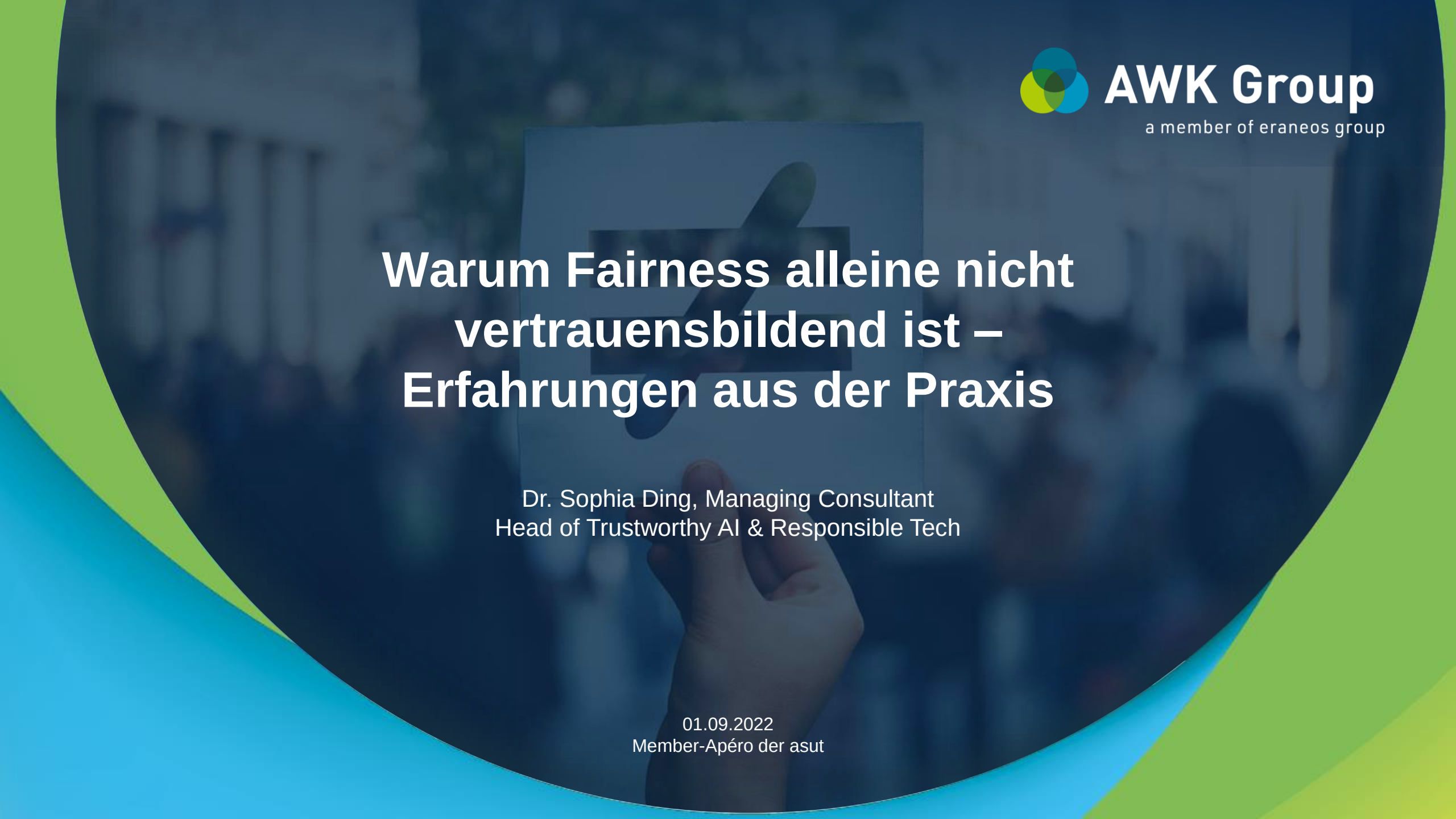


AWK Group
a member of eraneos group

Warum Fairness alleine nicht vertrauensbildend ist – Erfahrungen aus der Praxis

Dr. Sophia Ding, Managing Consultant
Head of Trustworthy AI & Responsible Tech

01.09.2022
Member-Apéro der asut

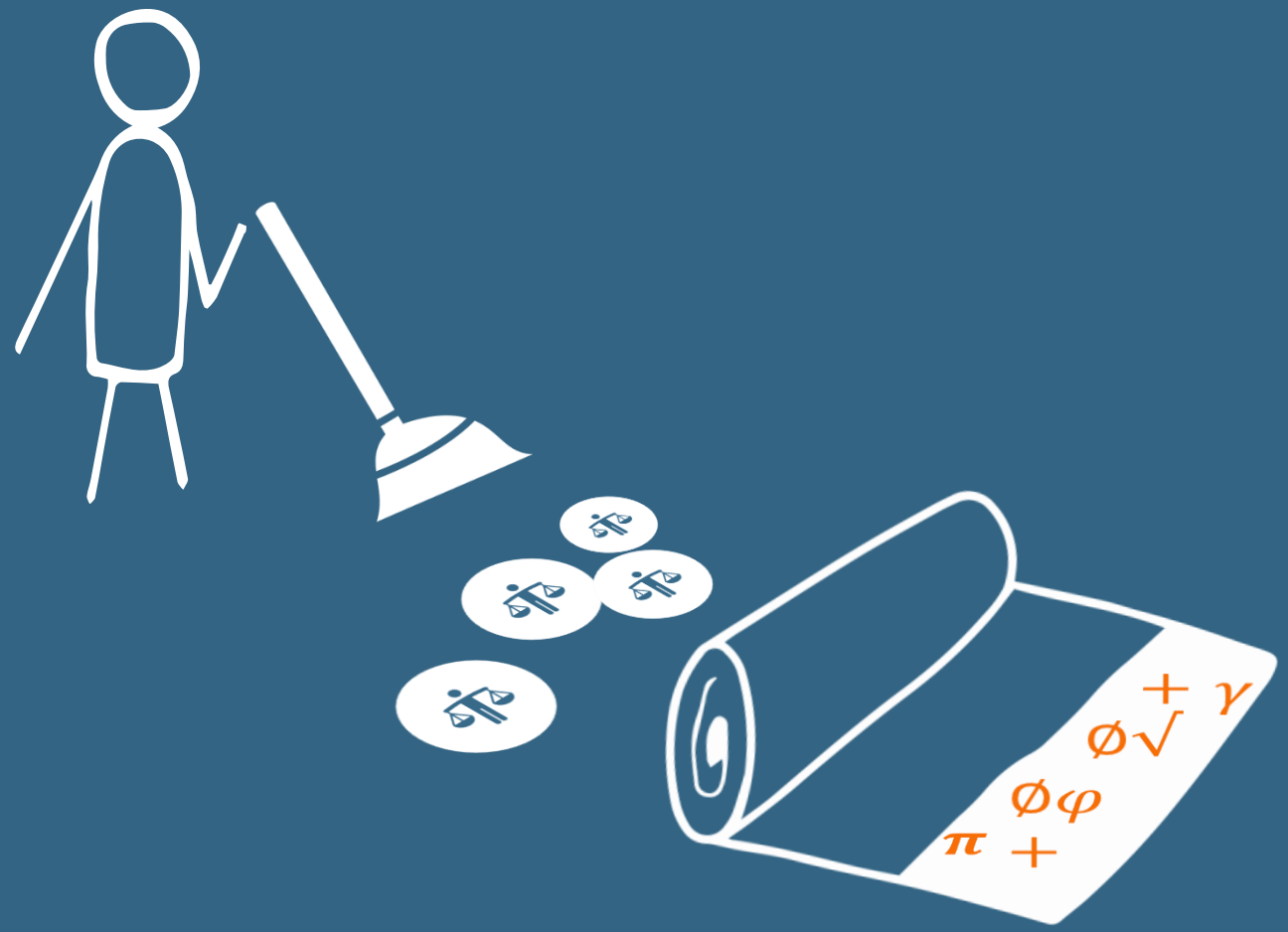
Warum ist AI Fairness in aller Munde?

Viele öffentlichkeitswirksame “AI Fails” haben mit scheinbar diskriminierenden Algorithmen zu tun – deshalb ist geplant AI Fairness regulatorisch zu verankern

Click on picture to learn more

AI Fairness ist quantifizierbar und erscheint deshalb einfach umsetzbar

- **Starke Auswirkungen**
- **Vergleichsweise leicht messbar**
- **Vergleichsweise leicht zu adressieren**



Was benötigen wir neben AI Fairness?

Fairness ist aktuell in aller Munde – ist alleine aber noch nicht vertrauensbildend

Übergeordnete Prinzipien

- Responsible
- Transparent
- Accountable

Eigenschaften des Systems

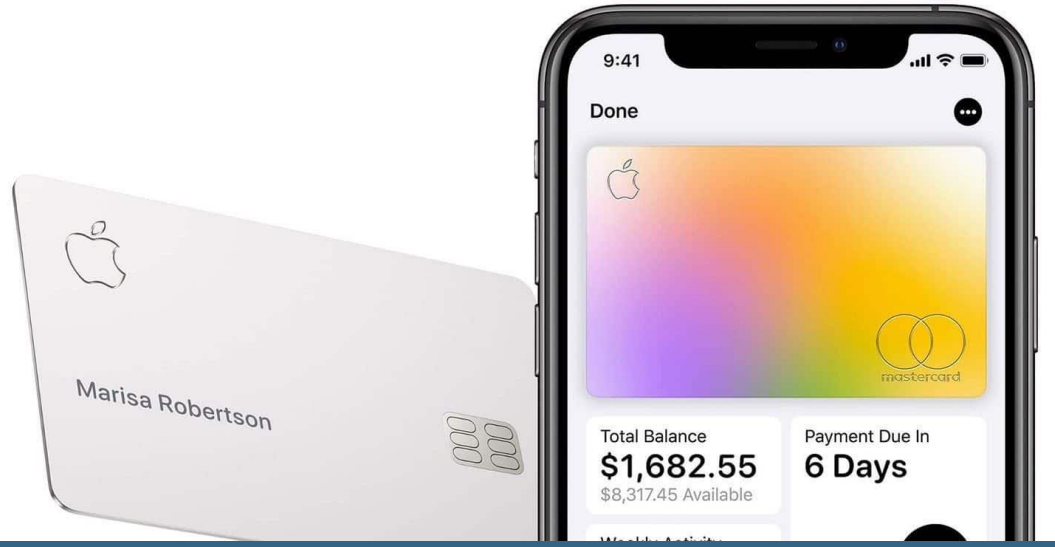
- Robust & Reliable
- Secure
- Safe
- Explainable & Interpretable
- Privacy-enhanced
- Fair & Bias is managed

Performance

- Accurate

Performance-KPIs stellen sicher, dass das AI System auf das Business Ziel einzahlt

Wie oft liegt das System mit der Empfehlung zur Gewährung oder Ablehnung eines Kredites richtig?



Accurate

Performance

“Fairness” oder “Explainability” sind Eigenschaften eines vertrauenswürdigen AI Systems, die operationalisiert werden müssen

Tut uns leid, wir wissen es auch nicht.

Kredit wird nicht gewährt aufgrund von:

- Kreditwürdigkeit
- Ausgabegewohnheiten
- ....

Kredit wird nicht gewährt aufgrund von:

?

Robust & Reliable

Secure

Safe

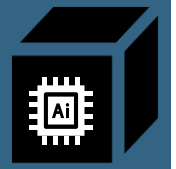
Explainable & Interpretable

Privacy-enhanced

Fair & Bias is managed

Eigenschaften des Systems

Übergeordnete Prinzipien und Systemeigenschaften hängen eng miteinander zusammen, wie das Beispiel «Transparenz» zeigt



Which groups failed the audit:

For **race** (with reference group as **Caucasian**)
- African-American with 0.59X Disparity
- Asian with 0.70X Disparity
- Other with 1.42X Disparity
- Native American with 0.21X Disparity

For **age_cat** (with reference group as **25 - 45**)
- Greater than 45 with 1.53X Disparity
- Less than 25 with 0.70X Disparity

"Ein technischer Bericht ist vollkommen ausreichend."

Die Transparenz muss der Zielgruppe angemessen sein, sonst schadet sie mehr als sie nützt. Mehr ist nicht unbedingt immer besser.

Responsible | **Transparent** | **Accountable**

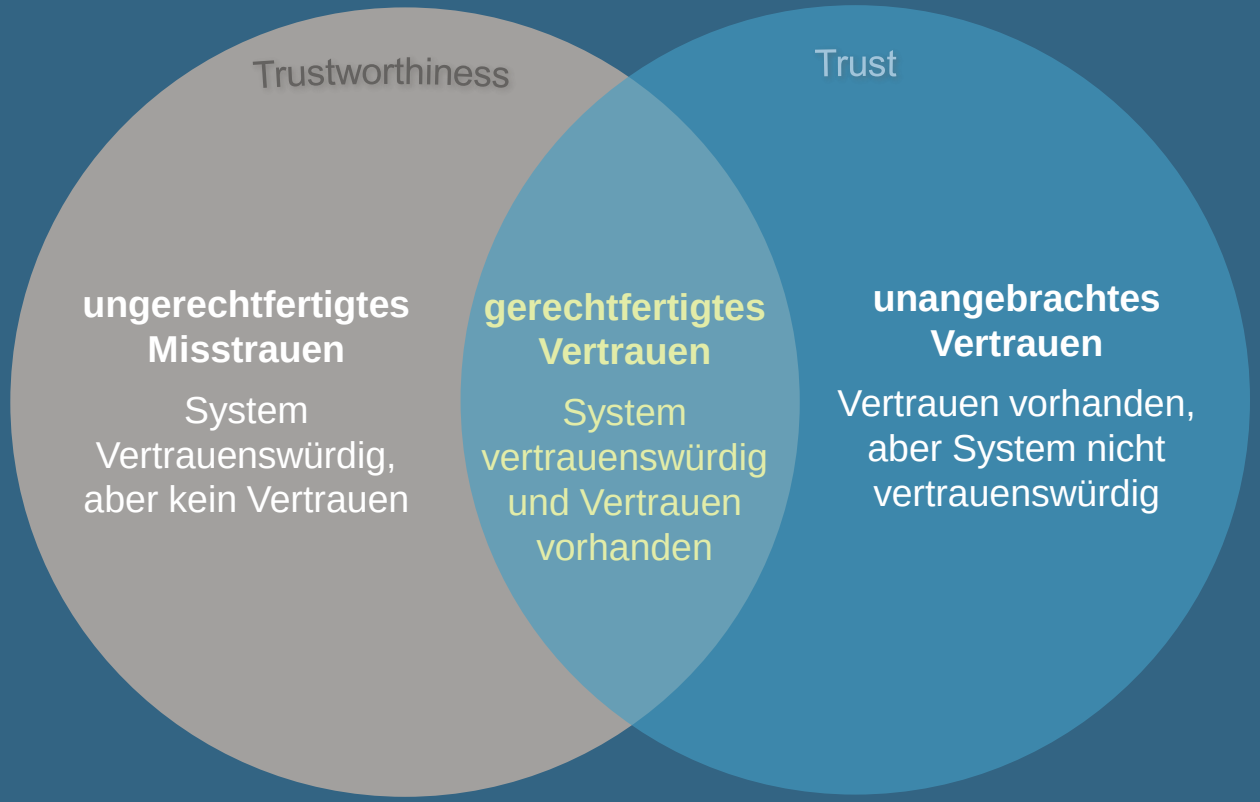
Übergeordnete Prinzipien

Wie bringen wir das alles in die Praxis?

Firmen müssen sowohl die Vertrauenswürdigkeit ihrer AI Systeme sicherstellen und nachweisen sowie zusätzliche vertrauensbildende Massnahmen umsetzen

Vertrauenswürdigkeit sicherstellen

- Trustworthy Prinzipien / Eigenschaften
- Interdisziplinäre Teams
- Best Practices, Frameworks, Toolboxes
- Governance-Strukturen



Vertrauenswürdigkeit nachweisen

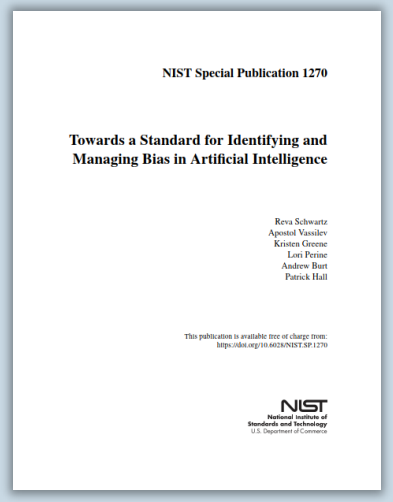
- Konformitätsbewertungen und Audits

Vertrauen erzeugen

- Kommunikations-strategie
- Accountability
- Labels & Zertifizierungen

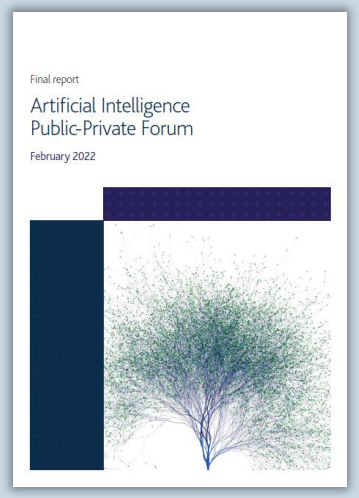
Für die Umsetzung von Trustworthy AI Prinzipien und Eigenschaften gibt es inzwischen eine Vielzahl von Hilfsmitteln

Vorschläge für Standards

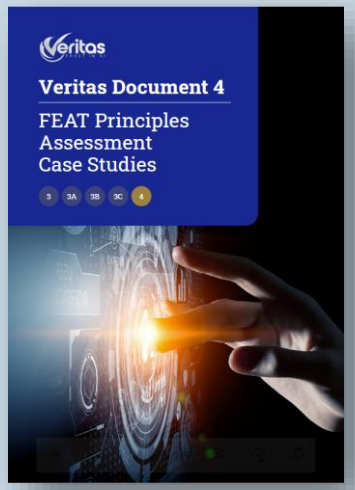


Vorschlag für NIST-Standard: Ziel ist es, eine sozio-technische Anleitung für die Identifizierung und den Umgang mit AI Bias zu geben

Industrie-Initiativen



AI Public-Private Forum: Förderung der sicheren Einführung von KI in Finanzdienstleistungen

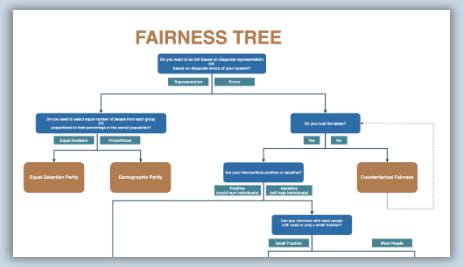


Veritas Project: Umsetzung der Grundsätze in Finanzdienstleistungen für bestimmte Anwendungsfälle

Vorschläge für Labels und Zertifizierungen

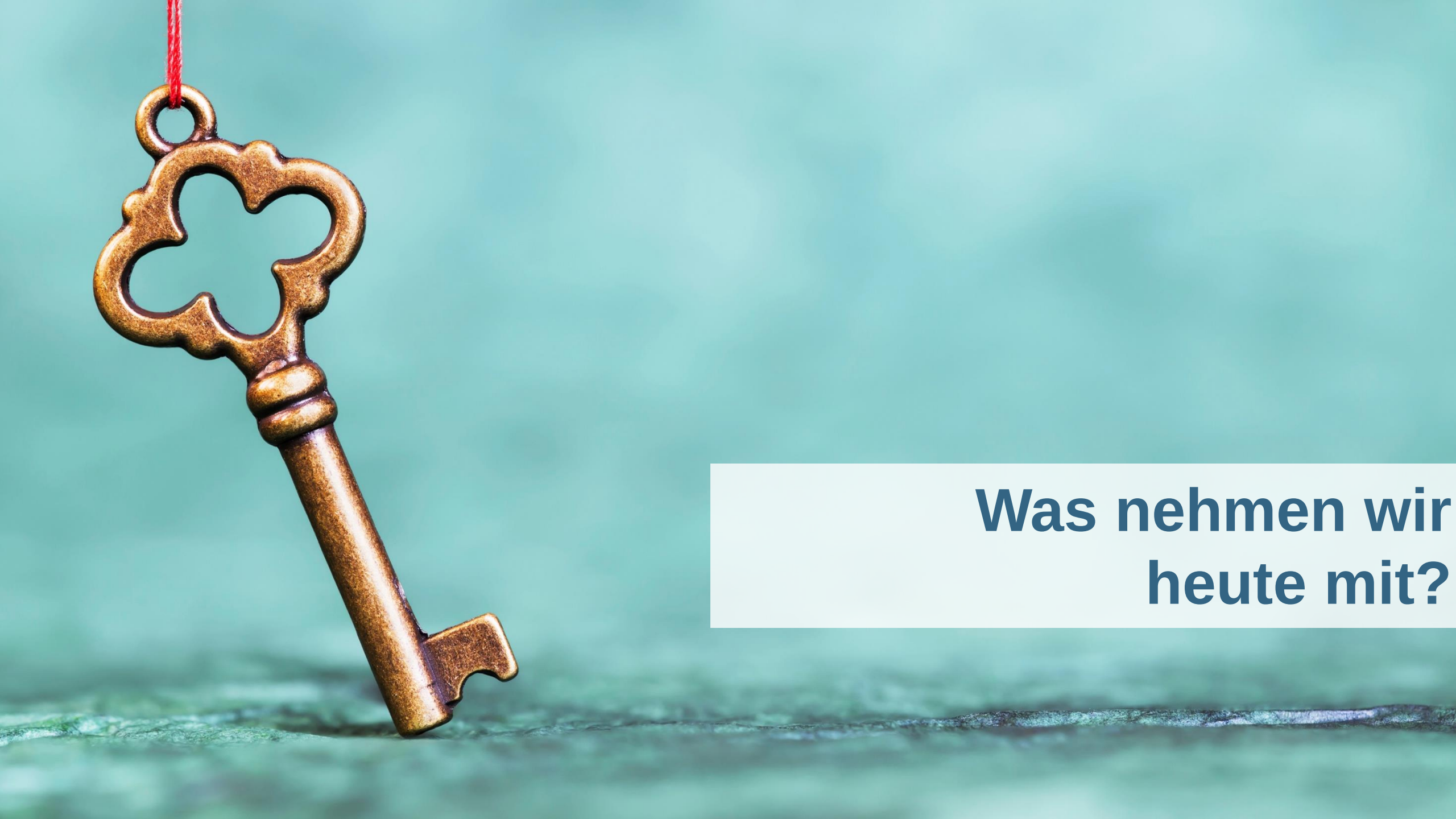
Digital Trust Label und Responsible AI Institute Zertifizierung

Aequitas: Open-Source-Toolbox aus dem akademischen Bereich für Audits



Algorithmic Fairness Toolbox

Open-Source-Toolkit von Tech-Unternehmen (IBM Research, Google, Microsoft, etc.)



Was nehmen wir heute mit?

Was sind die Kernbotschaften von heute?

AI Fairness ist in aller Munde, weil viele AI Fails mit vermeintlich diskriminierenden Algorithmen zu tun haben und diese nun regulatorisch eingefangen werden sollen.

Fairness ist nur eines von mehreren Prinzipien und Eigenschaften, die ein vertrauenswürdiges AI System ausmachen.

Für die Umsetzung von Trustworthy AI Prinzipien und Eigenschaften stehen viele Hilfsmittel zur Verfügung, die aber nicht alle gleichermassen praxiserprobt sind.

Für Trustworthy AI müssen Firmen die Vertrauenswürdigkeit ihrer AI Systeme sicherstellen & nachweisen und zusätzliche vertrauensbildende Massnahmen umsetzen.

Ich freue mich auf den Austausch mit Ihnen!

Dr. Sophia Ding,
Head of Trustworthy AI & Responsible Tech

Sophia.Ding@awk.ch

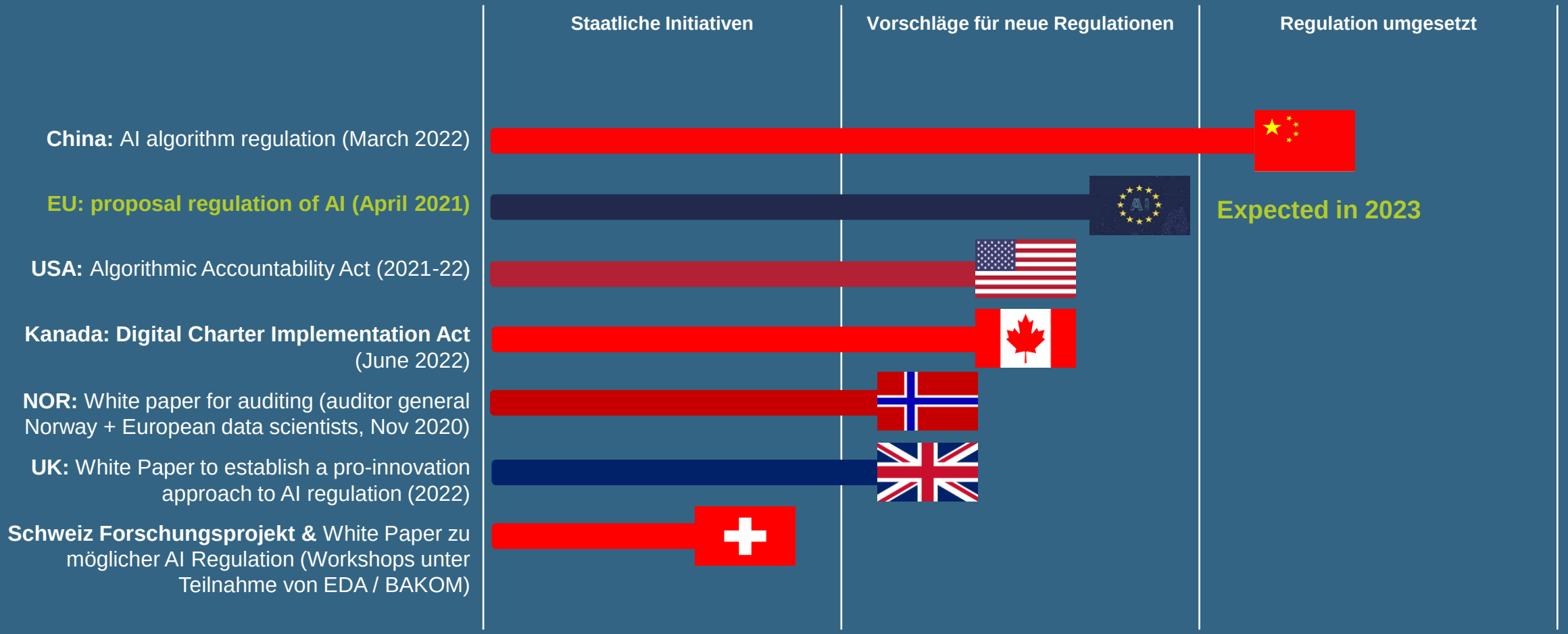
+41 58 123 99 76

- Spezialisiert auf die ethischen Aspekte neuer Technologien mit besonderem Augenmerk auf vertrauenswürdige künstliche Intelligenz (Trustworthy AI) und die praktische Umsetzung ethischer Grundsätze einschliesslich der Anwendbarkeit technischer Hilfsmittel zu diesem Zweck.
- Hintergrund in den Bereichen Ökonometrie, Analytics Translation, sowie der Schnittstelle zwischen Philosophie und Wirtschaft
- Co-Präsidentin des Label Expert Committees des Digital Trust Label

AWK Group

Anhang

Expert:innen erwarten die EU AI Regulation (EU AI Act) bereits nächstes Jahr und die systematische Überprüfung auf Verzerrungen ist Teil der geplanten Massnahmen



Die Überprüfung der AI Systeme (Daten, Resultate, Empfehlungen) auf systematische Verzerrungen ist Teil der geplanten EU Regulation.

Die EU Regulation wird ähnliche Auswirkungen auf Schweizer Unternehmen haben wie GDPR

Die EU geht beim AI Act ähnlich vor wie bei der Datenschutz-Grundverordnung (GDPR).

- Für die Einhaltung der GDPR wurden durchschnittlich 1,4 Mio. € pro Unternehmen investiert. Im Jahr 2021 wurden Bussgelder in Höhe von 1'000 Mio. € für Verstösse verhängt - diese Unternehmen investierten 10 Mio. € in die Behebung der Probleme [1].
- Aktuellen Schätzungen zufolge wird die Einhaltung des AI Acts genauso aufwändig sein wie bei GDPR. Der AI Act wird einen Overhead von 17% der EU-weiten KI-Ausgaben in Höhe von insgesamt 30 Milliarden CHF für Unternehmen in der EU verursachen [2].
- Kleine Unternehmen können mit Gesamt-Compliance-Kosten von bis zu 400'000 EUR für einen AI Service rechnen, der als «high risk» eingestuft worden ist.

Quellen:
[1] pwc, Organizations Worldwide Fear GDPR Non-Compliance Could Put Them Out of Business.
[2] WEF: How businesses should respond to the EU's Artificial Intelligence Act
[3] Center for data innovation: How much will the AIA Cost Europe?

Unter AI fallen sehr viele algorithmische und statistische Systeme, von denen man es nicht vermuten würde

Artificial Intelligence (AI)

“software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with”

Gemäß der Definition im EU AI Act (Verordnungsvorschlags)

Logic- and knowledge-based approaches, e.g. expert systems

Statistical Approaches, Bayesian Estimation, Search and Optimization methods

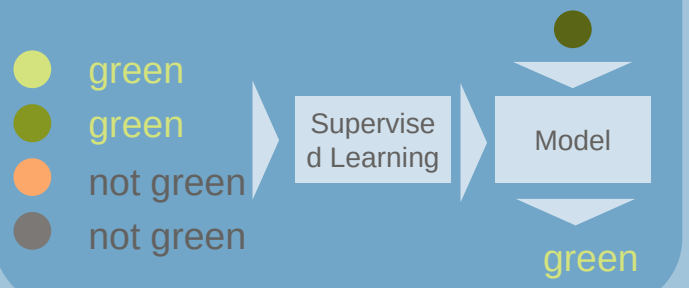
Machine Learning (ML)

Algorithms that enable an artificial system to learn from experience and to generalize it.

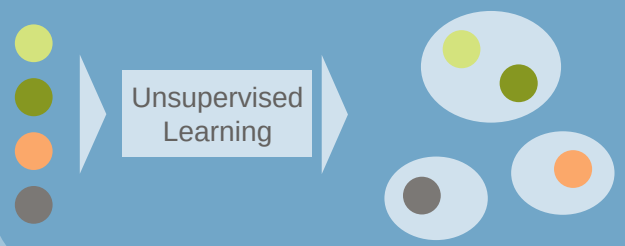
Reinforcement Learning

Deep Learning

Supervised Learning



Unsupervised Learning



Als “high risk” klassifizierte Anwendungsfälle werden reguliert

- Beschäftigung, Arbeitnehmermanagement und Zugang zur Selbständigkeit: jeder Kunde, der ein AI-System für seinen Rekrutierungsprozess nutzt

- Verwaltung und Betrieb kritischer Infrastrukturen: Banken, Versicherungen, Energieversorger, Telekommunikationsunternehmen
- Zugang zu Krediten usw.: Banken, Versicherungen

- Strafverfolgung, z. B. biometrische Identifizierung und Kategorisierung von natürlichen Personen
- Bildung und Berufsausbildung
- Zugang zu öffentlichen Leistungen und Diensten
- Rechtspflege und demokratische Prozesse

Die Ermittlung von Kennzahlen ist ein wichtiger Bestandteil einer Fairness Evaluation, muss aber durch eine qualitative Analyse in den Kontext eingeordnet werden

"Durch die Erhebung von Fairness-Kennzahlen alleine schaffen wir Vertrauen in unser AI System."

Eine quantitative Analyse ist nicht ausreichend und muss durch eine qualitative Analyse ergänzt warden, zu der Stakeholder beigezogen warden.

Unser Ansatz für Trustworthy AI

Schulung zu vertrauenswürdiger KI

Diese Reihe von Vorträgen und praktischen Übungen ermöglicht es Ihren Mitarbeitenden, ausgewählte Themen wie KI-Risiken, KI-Prinzipien wie Fairness, KI-Sicherheit, den Stand von KI-Regulationen usw. zu ergründen.

Datenethik-Strategie

Diese Strategie fördert die ethische Nutzung von Daten und hilft Ihnen, ethische Grundsätze in die Praxis umzusetzen. Stärken Sie das Vertrauen Ihrer Kunden, schützen Sie Ihren Ruf und verschaffen Sie sich Wettbewerbsvorteile.

Algorithmic Fairness Evaluation und Dashboard

Unser Vorgehen zur Evaluation von Algorithmic Fairness umfasst eine qualitative sowie eine quantitative Analyse. Darüber hinaus können wir bei Bedarf ein Dashboard zur einfachen Darstellung von KPIs zur Verfügung stellen.

Ethische Entscheidungsfindung

In diesem Workshop schulen wir Sie darin, ethische Risiken und mögliche Zielkonflikte in Ihren Anwendungsfällen zu erkennen und gewissenhafte und verantwortungsvolle Entscheidungen bezüglich Ihrer Systeme zu treffen.

KI-Risikoabschätzung & #AuditingAI

Ein externer Audit hilft Ihnen, die mit Ihrem KI-System verbundenen Risiken zu bewerten und empfiehlt Massnahmen zur Risikominderung.

Transparenzbericht

Adressatengerechte Kommunikation ist der Schlüssel zur Vertrauensbildung. Mit einem Transparenzbericht haben Sie alle Antworten parat, falls Ihre Stakeholder Fragen stellen.