Fragen und Lösungen zur ethischen Digitalisierung

Sitzung des Landrates für digitale Entwicklung und Kultur
14.5.2019

Prof. Dr. K.A. Zweig

TU Kaiserslautern

Algorithm Accountability Lab

Konstituierende Sitzung der Enquete-Kommission „Künstliche Intelligenz“ am 27.9.

Aus der Rede von Bundestagspräsidenten Dr. Schäuble:

- „Die künstliche Intelligenz gilt Vielen als neue Zauberformel des technischen Fortschritts, …
- … sie wird dichten, …
- … sie wird belohnen und bestrafen …“

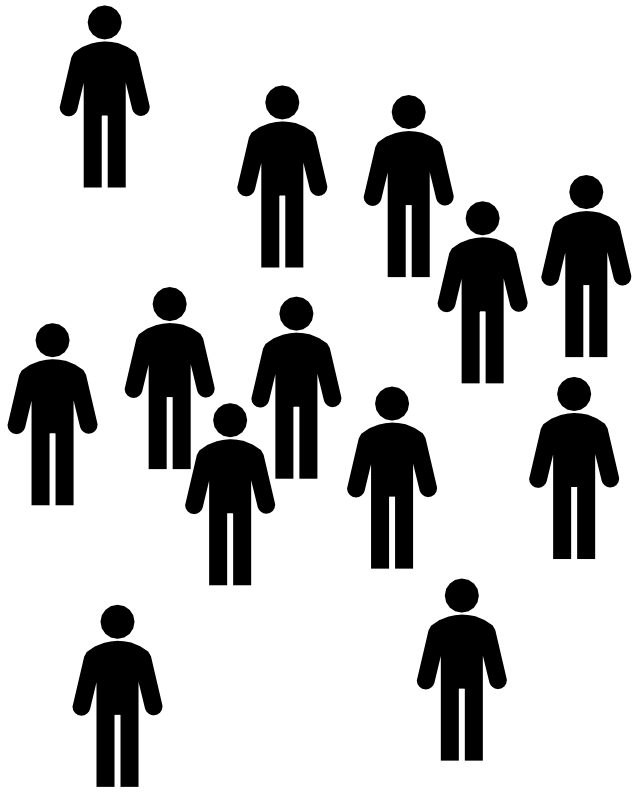
https://www.bundestag.de/gebaerdensprache/mediathek?videoid=7272804#url=bWVkaWF0aGVrb3ZlcmxheT92aWRlb2lkPTcyNzI4MDQ=&mod=mediathek

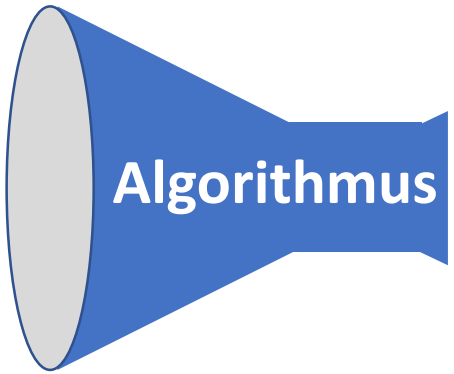
Die zwei Ängste

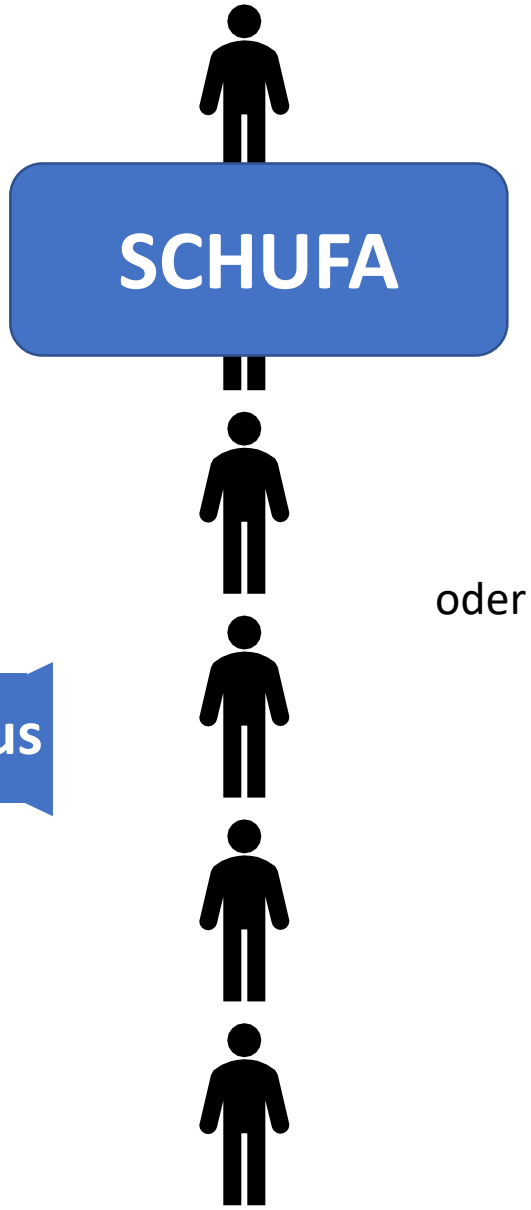
Sie wird dichten

Sie wird richten

Algorithmische Entscheidungssysteme (ADM Systeme)







Scoring-Verfahren

oder

Klassifikation

Maschinelles Lernen

Software, die aus Daten der Vergangenheit Entscheidungsregeln ableitet für zukünftige Daten.

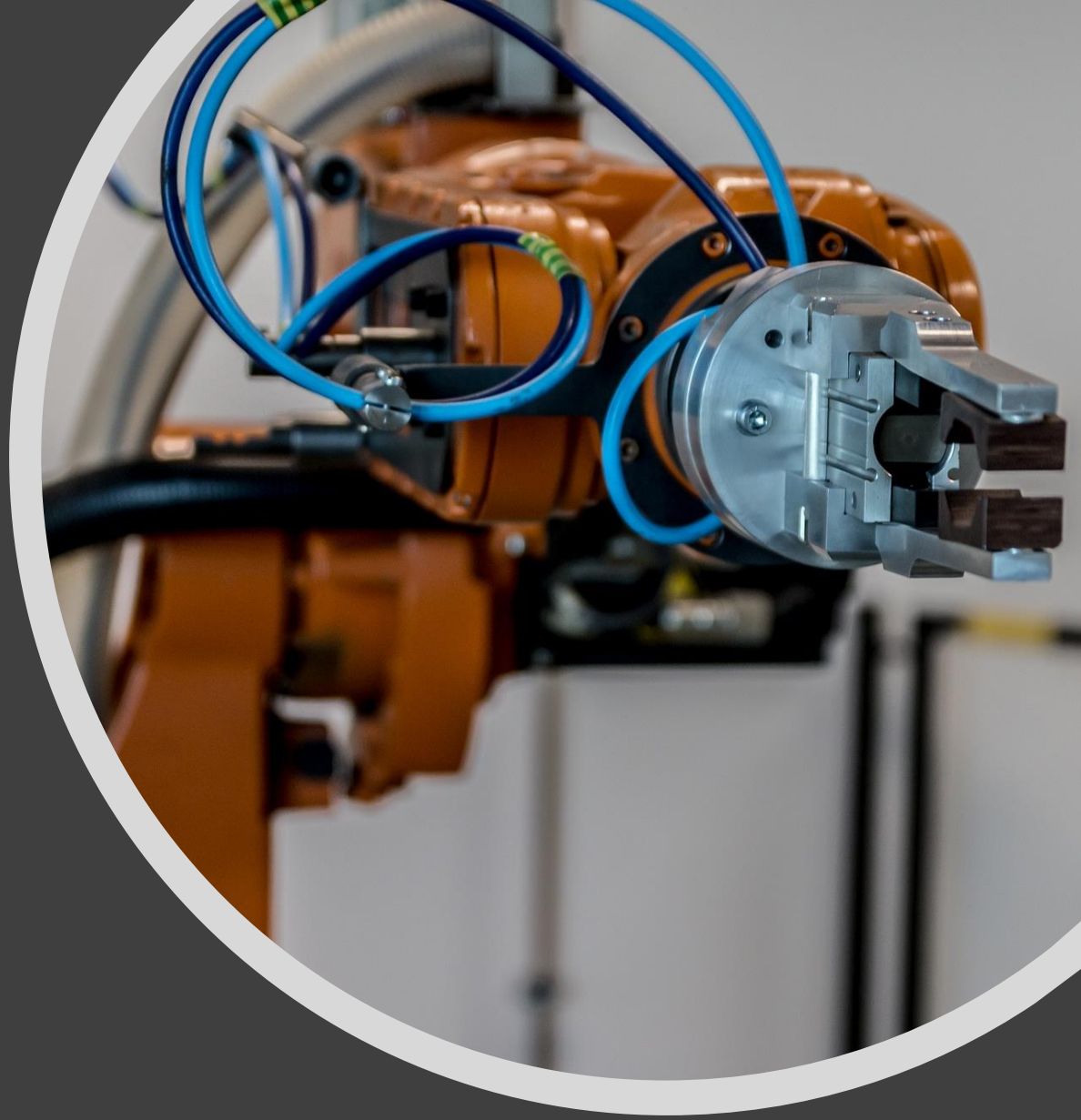
Die Software trifft dann mit den gelernten Regeln Entscheidungen über neue Situationen.

Wann muss das auf technischer Ebene kontrolliert und reguliert werden?

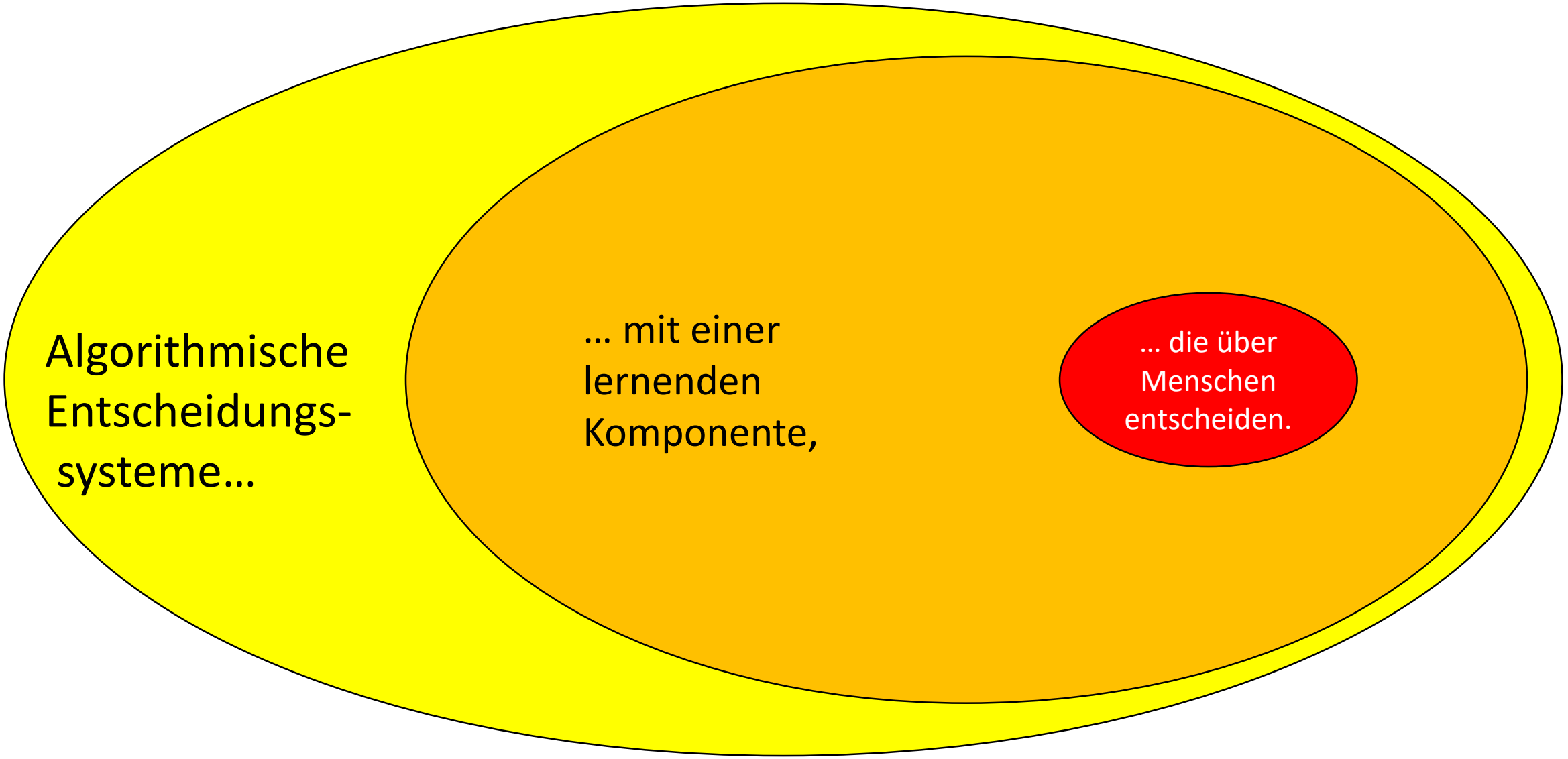
Kontrolle von algorithmischen Entscheidungssystemen

Maschinelles Lernen muss um so stärker kontrolliert und reguliert werden, je höher das durch die Software mögliche individuelle und gesamtgesellschaftliche Schadenspotenzial ist.

I.A. sind Entscheidungen über Objekte, z.B. im Produktionsprozess, nicht kritisch und bedürfen keiner technischen Kontrolle und Regulierung.



Welche ADM-Systeme sind problematisch?



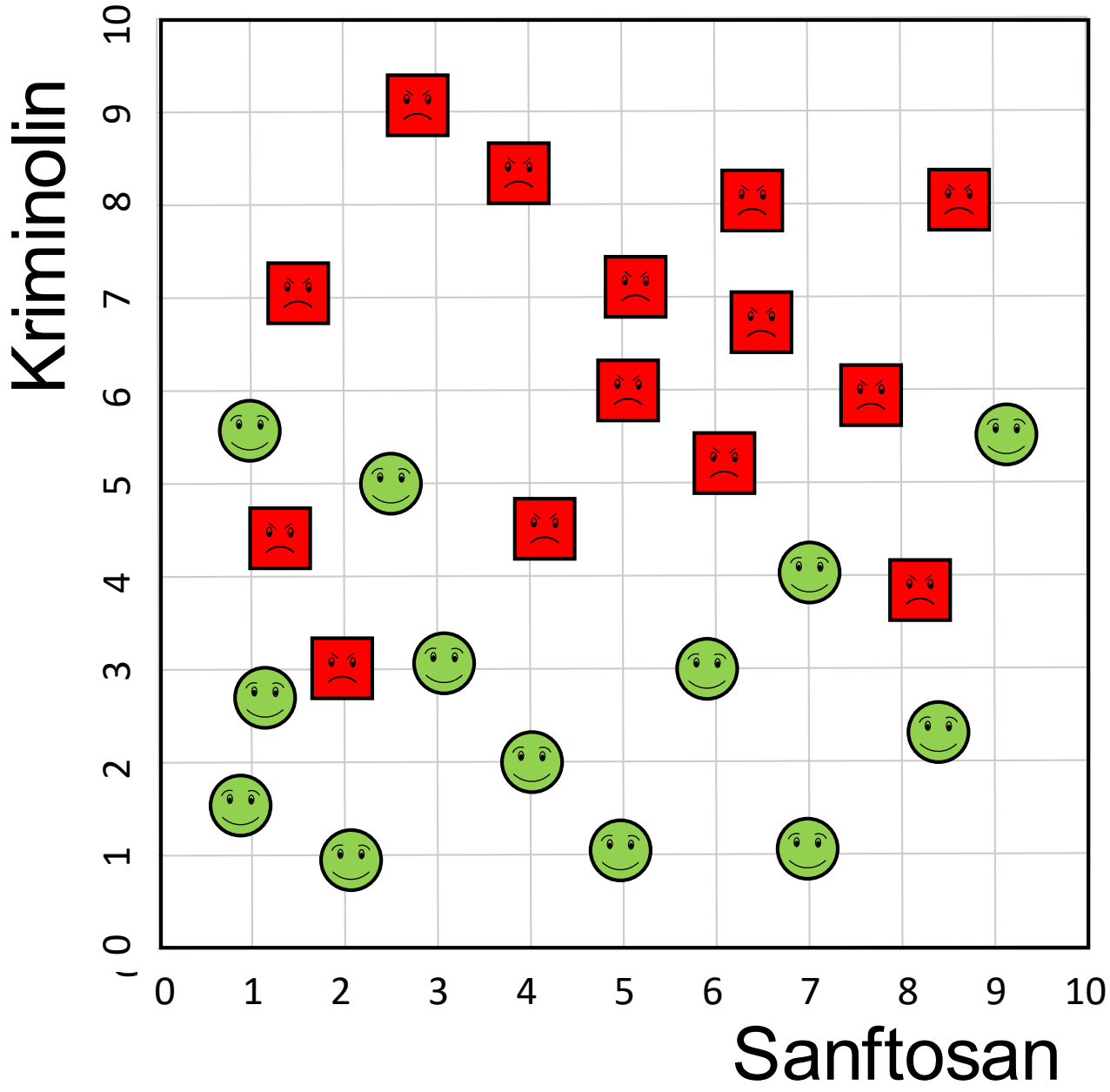
Wie „lernt“ das System von Daten?

DIY:

Sie sind heute meine „Support Vector Machine“

Bösartige Kriminelle

Unschuldige Bürger

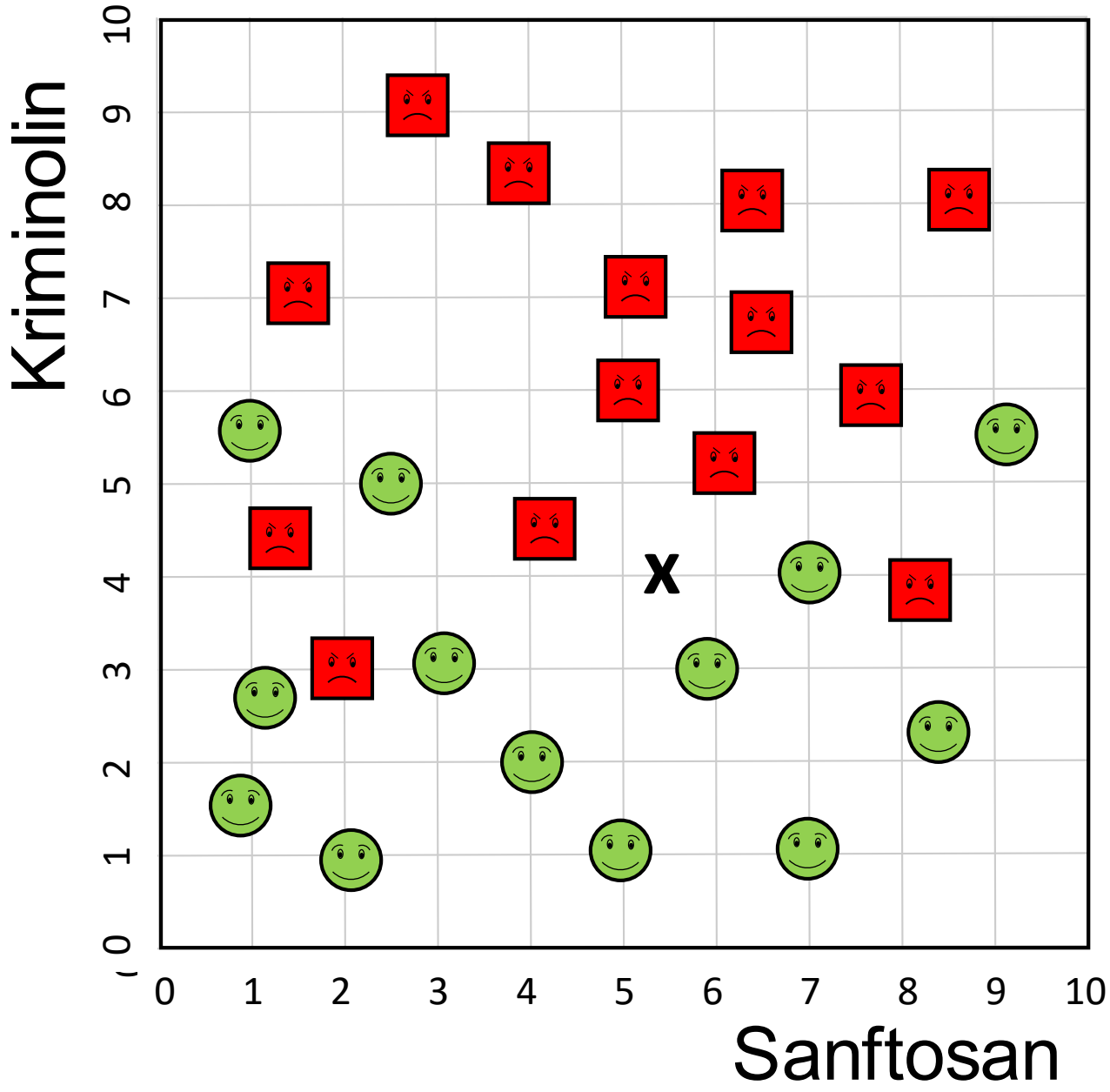


Bösartige Kriminelle

Unschuldige Bürger

Bewerten Sie Frau Müller:

5.5 Sanftosan
4.0 Kriminolin

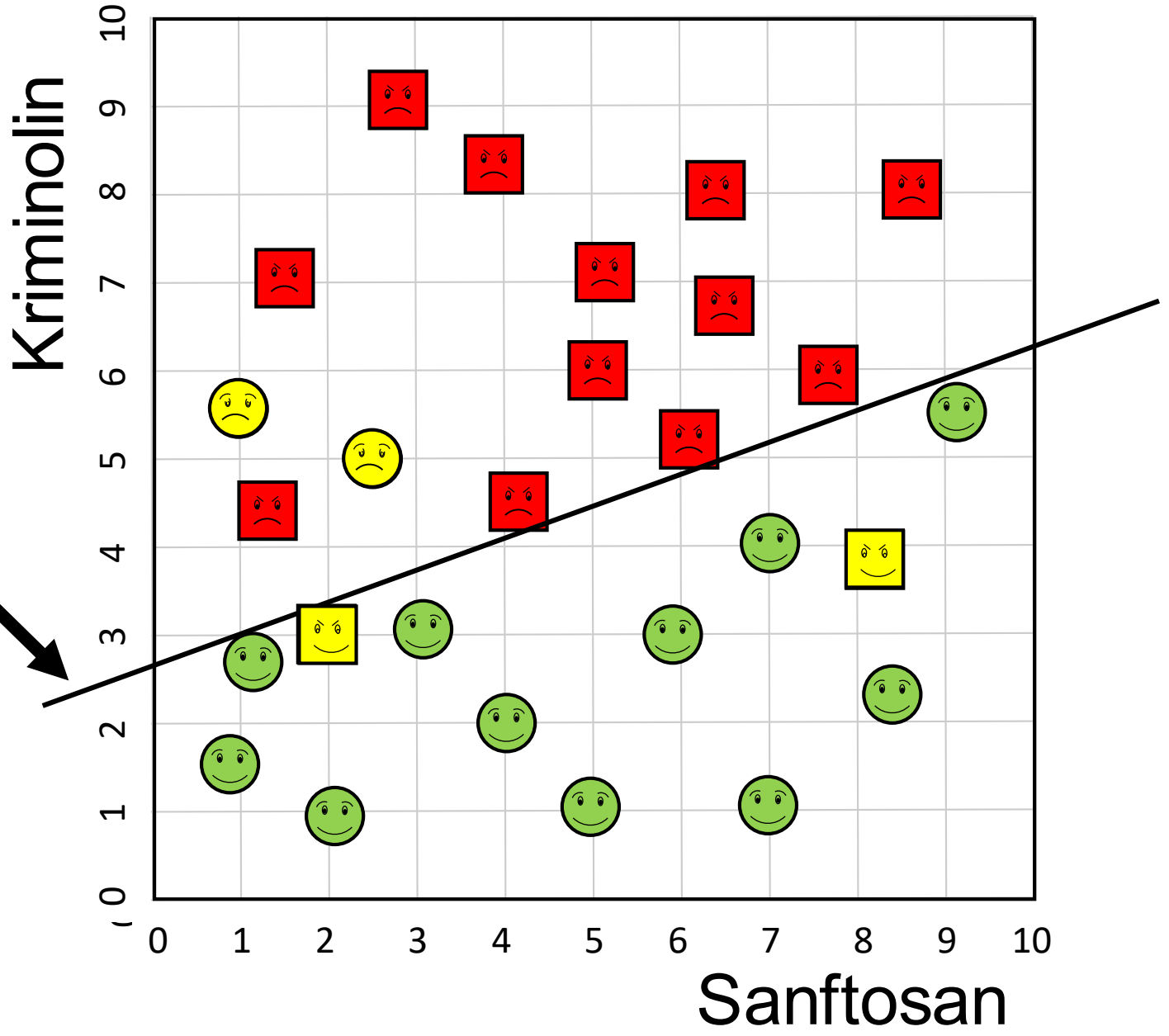


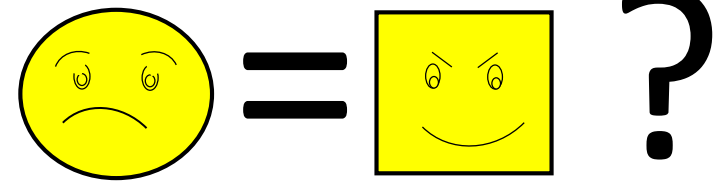
Eine der möglichen Trennlinien

Alle möglichen Trennlinien erzeugen Fehler:

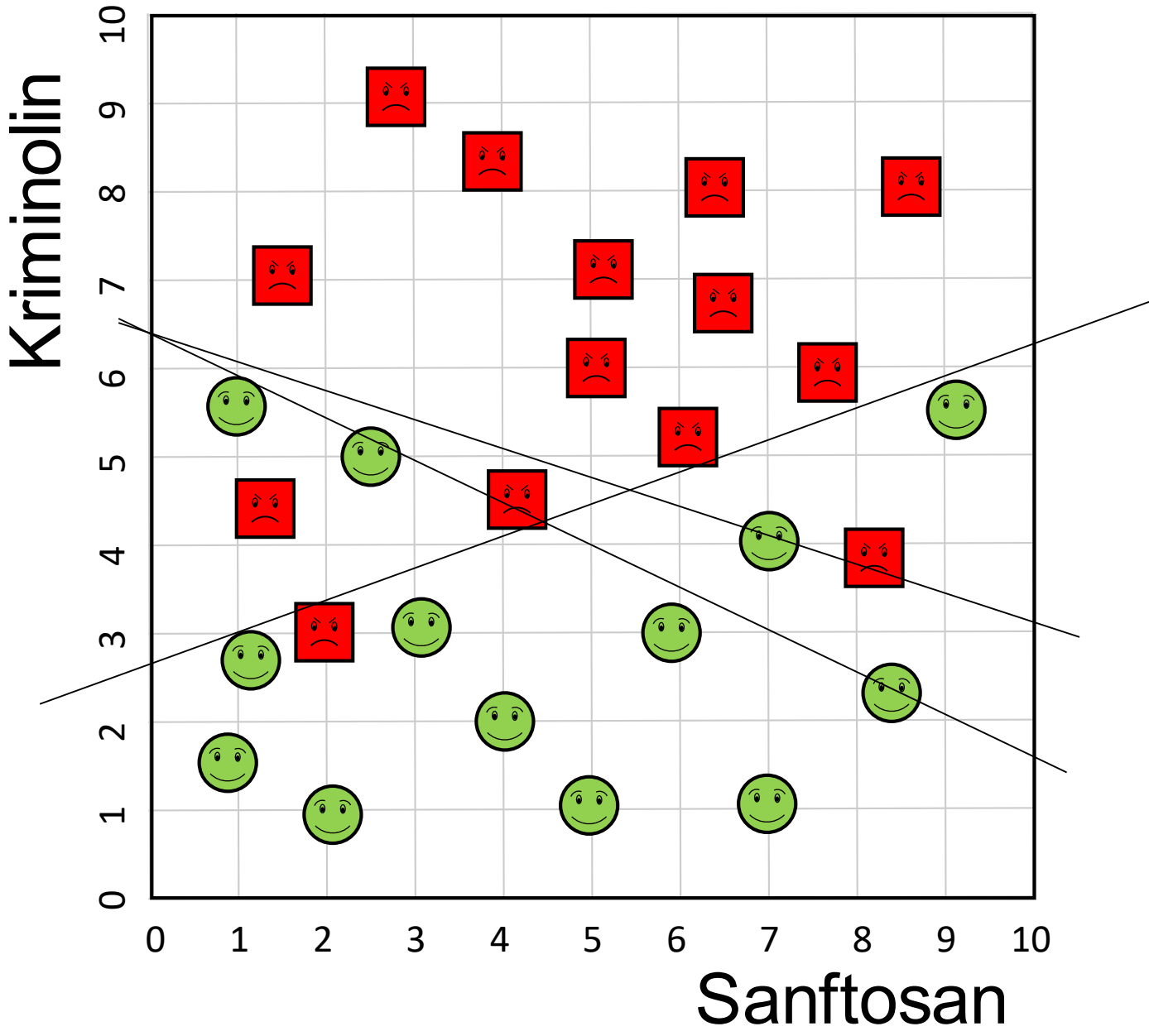
Bösartige Kriminelle, die unentdeckt bleiben

Unschuldige Bürger, die für kriminell gehalten werden





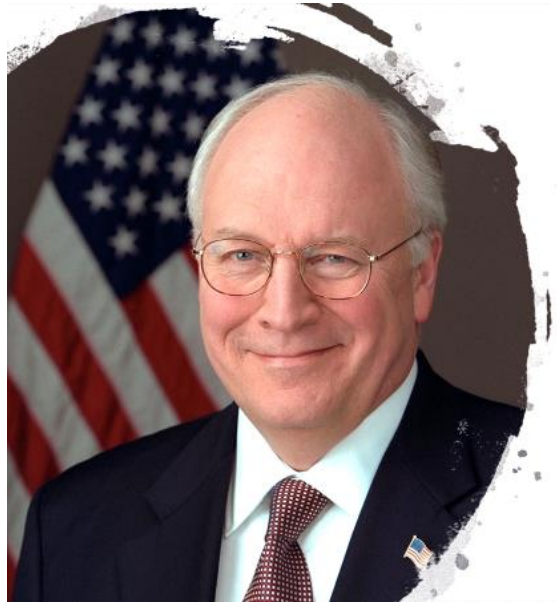
Wenn beide Fehler als gleich schlimm gelten, gibt es mehrere optimale Trennlinien mit möglichst wenigen Fehlern.




Sind beide Arten
von Fehlern
gleich zu bewerten?

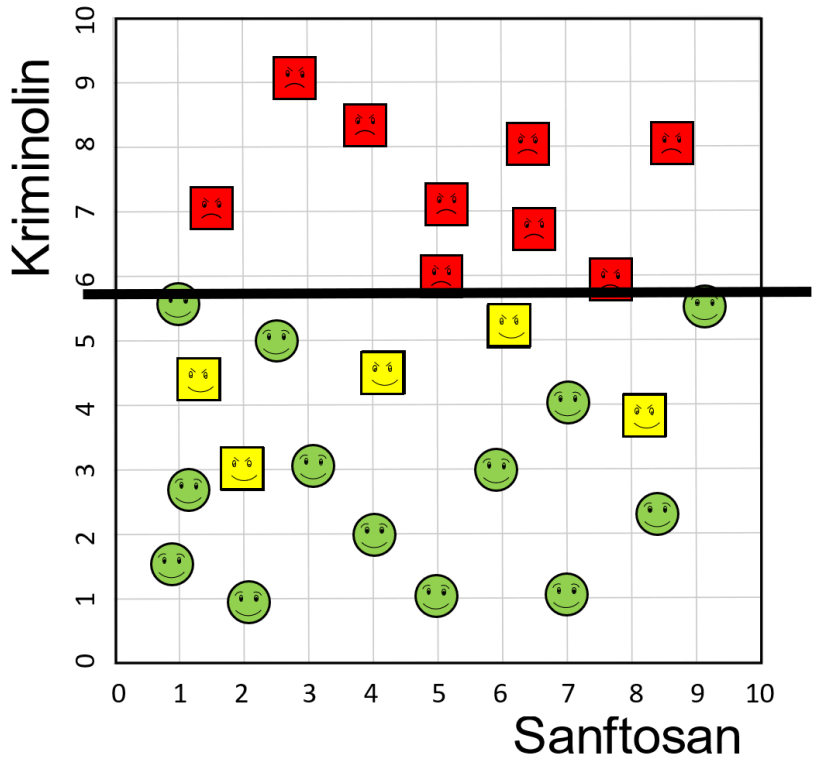
„It is better that ten guilty persons escape than that **one** innocent suffer.“

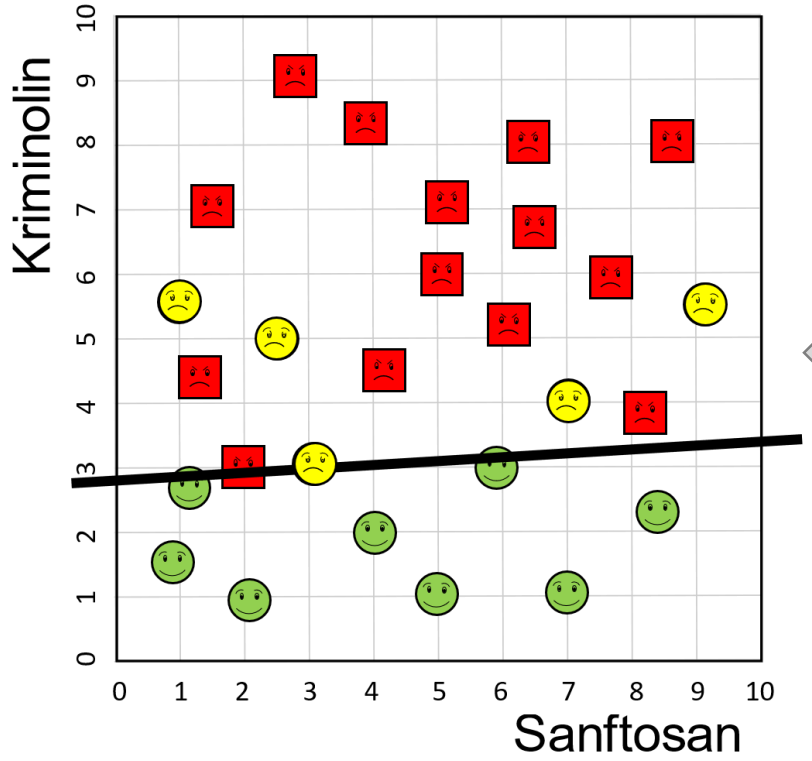
William Blackstone,
Rechtsphilosoph, 1760



"I am more concerned with bad guys who got out and released than I am with a few that, in fact, were innocent."

Dick Cheney, ehemaliger
Vizepräsident der USA,





1. Beobachtung

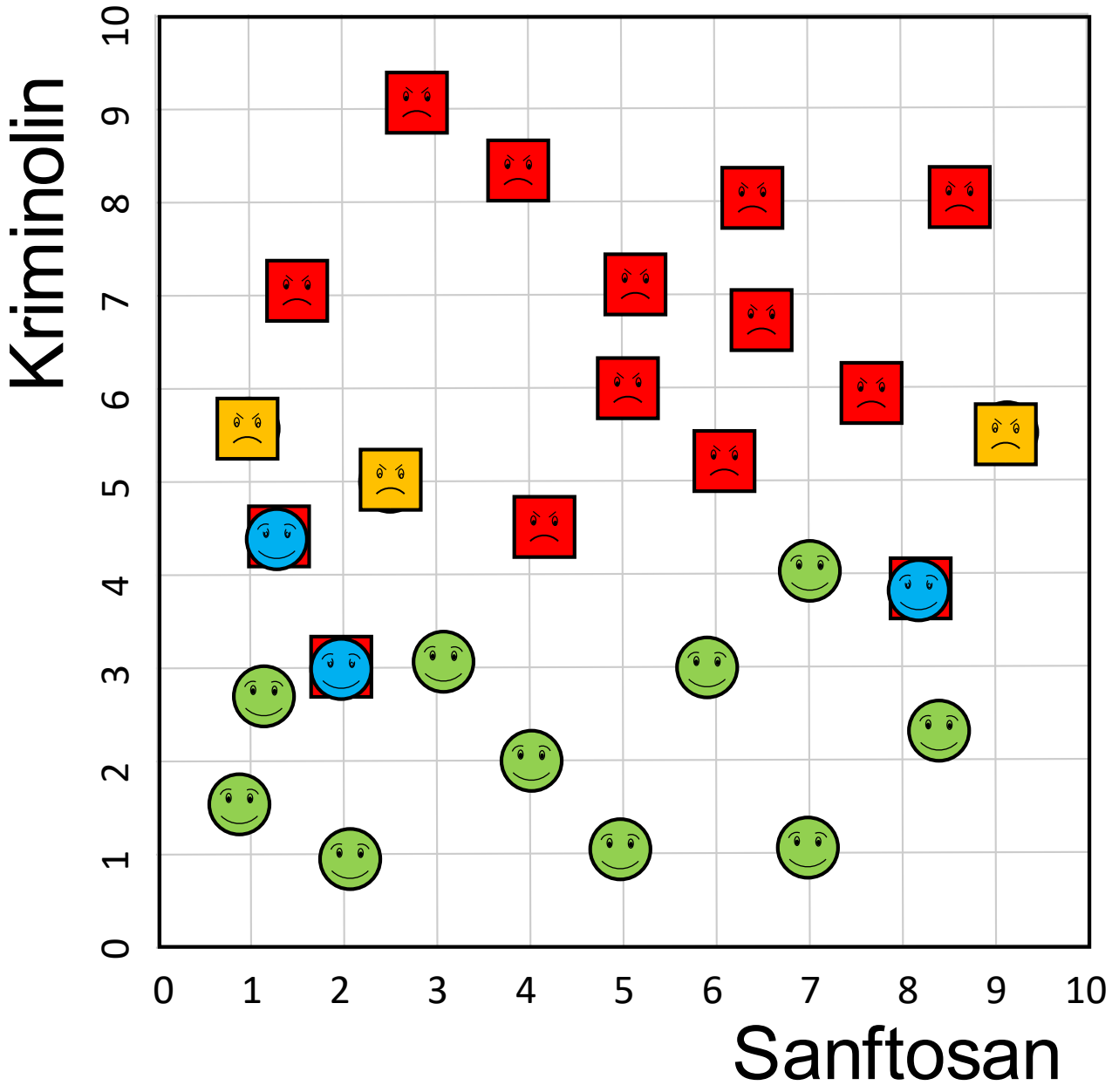
Was durch eine künstliche Intelligenz
optimiert werden soll,
ist eine gesellschaftliche Entscheidung!

Datenqualität

Noch nicht entdeckte Finanzbetrüger

Unschuldig im Gefängnis

Falsche Datenpunkt-zuordnungen haben Einfluss auf das Training der Support Vector Machine und damit auf die nachfolgenden Entscheidungen.



2. Beobachtung

Wie gut die Maschine lernt, ist direkt abhängig von der Qualität der Daten.

3. Beobachtung

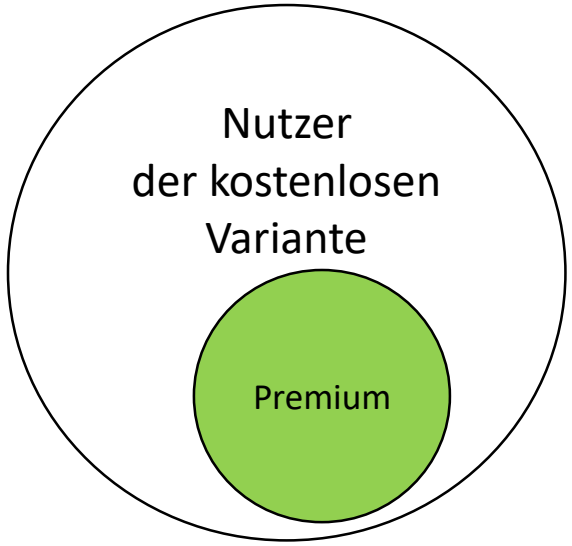
Eine geschützte Information kann wichtig sein, um bessere Entscheidungen zu treffen. Diskriminierung wird nicht per se dadurch vermieden, dass die Information vorenthalten wird.

Ethische Entscheidungen im maschinellen Lernen

- Was genau „gelernt“ (optimiert) werden soll, ist eine gesellschaftliche Frage, wenn es um Entscheidungen über Menschen und gesellschaftliche Teilhabe geht.
- Ob Daten als Grundlage für eine soziale Fragestellung geeignet sind, muss dann auch die Gesellschaft entscheiden.
- Auch wahrhaftige Daten stellen immer nur einen Ausschnitt aus der Wirklichkeit dar – sie bedürfen der Einordnung und Interpretation.
- Die Frage nach Diskriminierung, ihrer Entdeckung und ihres Ausgleichs bedarf der gesellschaftlichen Diskussion – unabhängig davon, wer die Entscheidung trifft – Mensch oder Maschine.

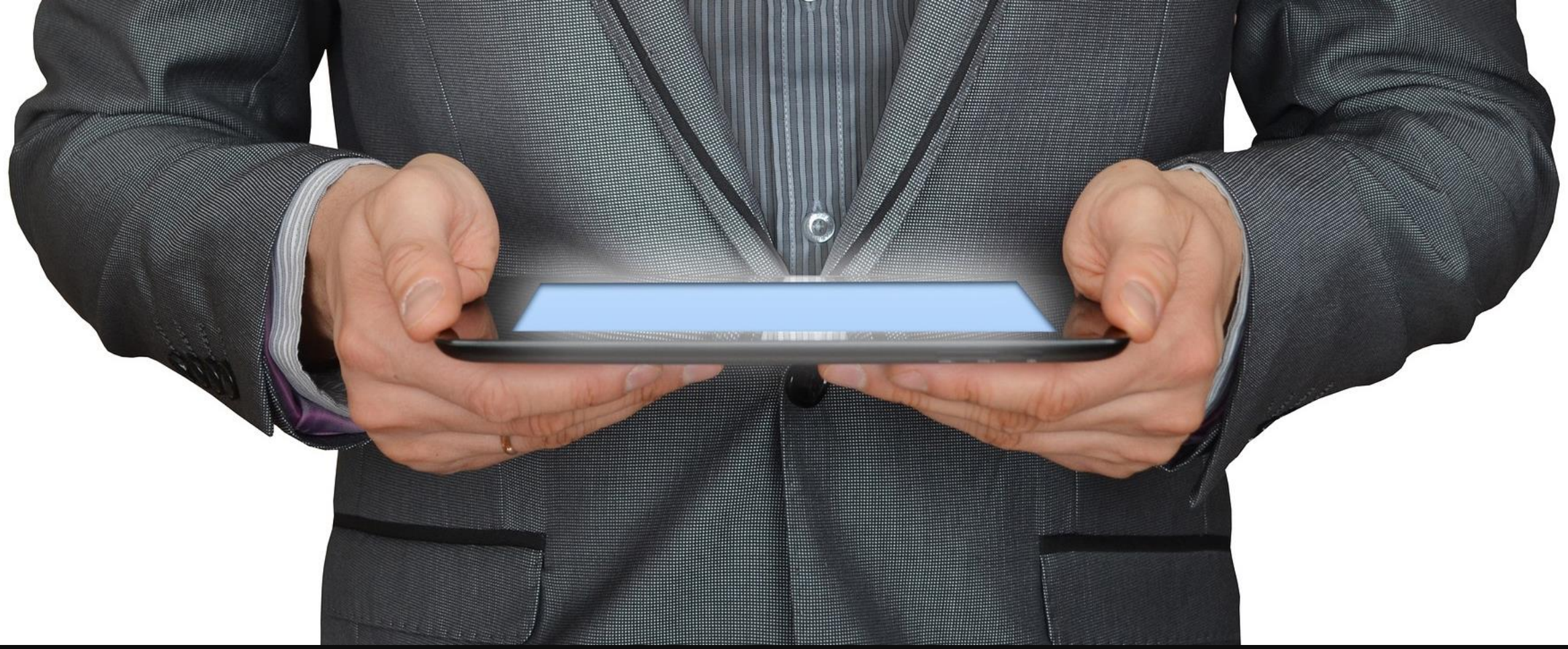
Empfehlungssysteme und Nudging

Wie Apps Geld verdienen



Monetarisierung im
Freemium-Modell

In der Aufmerksamkeitsökonomie gibt es den Anreiz, uns so lange wie möglich auf der Plattform zu halten.



A/B-Testing als radikales Instrument zur digitalen Produktentwicklung


Langnese
EISKREM


LANGNESE

Produktänderungen in der realen Welt

ALDI ändert sein Supermarkt-Layout

- Kosten in den USA: $ 1.6 Milliarden[1]
- Kosten in Australien: $ 1 Milliarde[2]
- Kosten ALDI Nord: 5.6 Milliarden €[3]

1 http://www.businessinsider.de/aldis-new-store-design-mimics-whole-foods-2017-2?r=US&IR=T
2 http://www.dailymail.co.uk/news/article-4418206/Aldi-spends-1-billion-change-store-layout-Australia.html
3 http://www.faz.net/aktuell/wirtschaft/unternehmen/aldi-projekt-zu-groesstem-umbau-der-unternehmensgeschichte-15124932.html

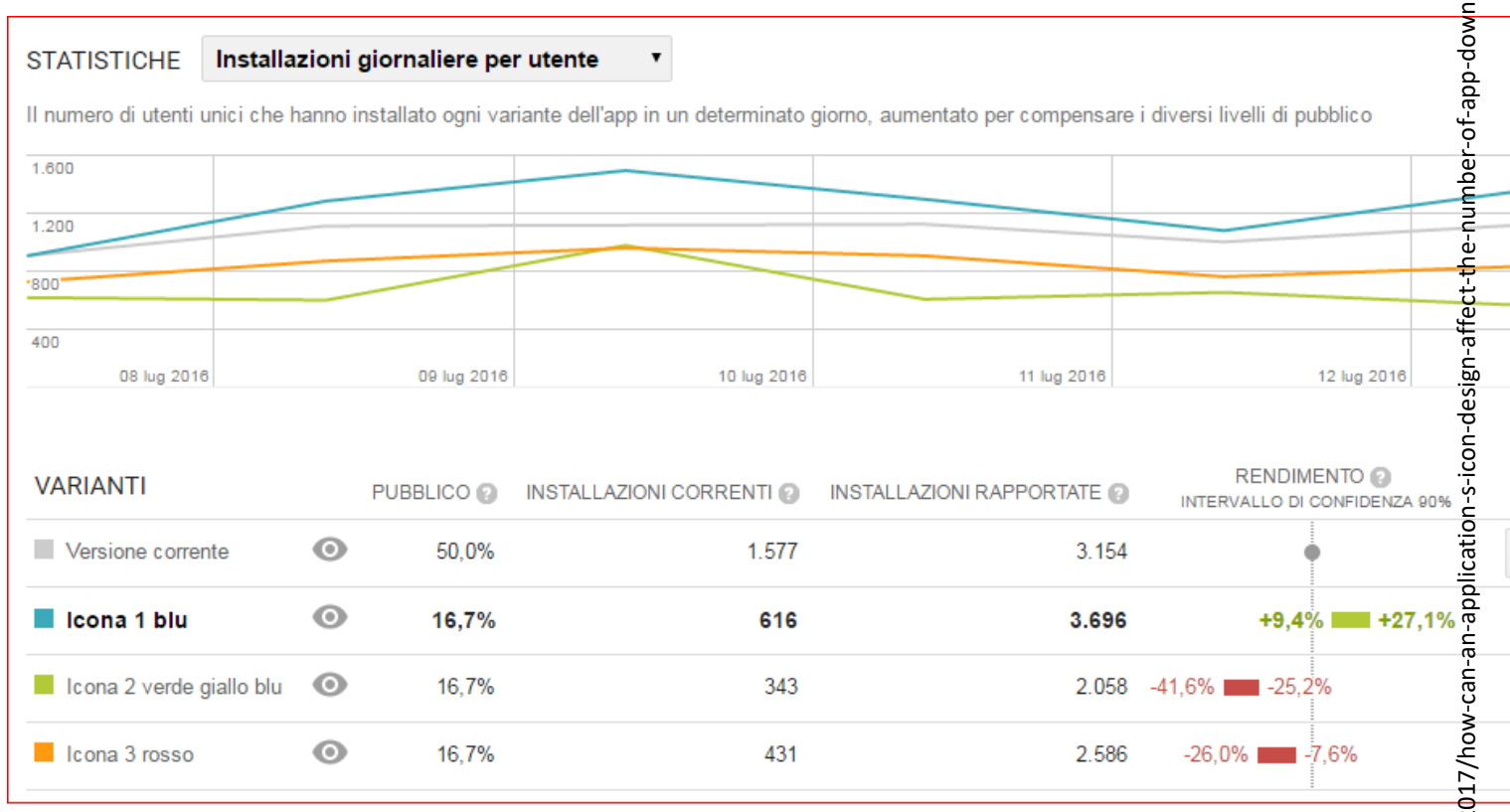
27% mehr als das alte Design (links)

Logo Design

http://blog.icondesignlab.com/en/2017/how-can-an-application-s-icon-design-affect-the-number-of-app-downloads/

Digitale „Game“-Changer

- Günstige Entwicklung von Produktvarianten.
- Günstiges Ausbringen der Produktvarianten.
- Randomisierte, kontrollierte Studie mit allen (!) Nutzern – **ohne Notwendigkeit der Zustimmung**.
- Unmittelbares Konsumentenfeedback.
- Oftmals fertige experimentelle Umgebungen (z.B. Google Play Store).
- Zusammenfassend: Kurzer und günstiger Entscheidungszyklus.



Versione corrente

Icona 1 blu


Icona 2 verde giallo blu

Icona 3 rosso

http://blog.icondesignlab.com/en/2017/how-can-an-application-s-icon-design-affect-the-number-of-app-downloads/

Aufmerksamkeits- und monetarisierungssteigernde Maßnahmen

A/B Testing in Games

- Revenue fördernde und damit aufmerksamkeitsverbrauchende Optimierung
- Basierend auf psychologischen Tricks und Kniffen[1]
  - Sozialer Druck (Bestenliste, Austausch von Gütern)
  - Indirekte Bezahlung über Symbolwährungen statt Bezahlung mit Geld
  - Intermittierende Belohnungen
  - Zeitdruck für Entscheidungen
  - Intransparenz
  - …

Das Risiko

- Kinder sind besonders gefährdet
  - Z.T. zu hoher Druck: Spiele lassen Tiere sterben, wenn man sich nicht genug kümmert.
  - Den Kindern fehlt wertvolle Entwicklungszeit.
- In VR könnte dies zu besonders aufmerksamkeitsheischenden Produkten führen (totale Immersion).

Es sind **personalisierte Anpassungen** per A/B-Testing plus Nutzung von **Machine Learning** denkbar:

Jede(r) wird an seiner oder ihrer Schwachstelle oder zu einem Schwachpunkt erwischt.

Was ist zu tun?

Steuerungsmöglichkeiten

- Etikette/Selbstverpflichtung
- Deklaration von Tests und Zustimmung durch TeilnehmerInnen
- Transparenz durch Zugang zu Daten
  - Individuelle Ebene
    - Summe des investierten Geldes
    - Zeit gespielt/verbracht
  - Gesamtpopulation
    - Wieviel mehr Geld erzeugt das Spiel durch Evolution?
    - Wieviel mehr Lebenszeit pro Nutzer?

Forschungsbedarf

- Kategorisierung von Monetarisierungs- und aufmerksamkeitsheischenden Maßnahmen
  - Einordnung digitaler Produkte anhand einer Skala
  - Verbote einiger Maßnahmen für Kinderspiele
- Nützlichkeitsanalysen der Steuerungsmöglichkeiten.
- Was kommt bei VR dazu?
- Wie klärt man am besten auf?

Positive Aspekte

- Können Strategien für ‚Serious Games' genutzt werden?
- Wie können die Strategien für andere Produkte genutzt werden?

Wie gut sind die Robo-Richter?

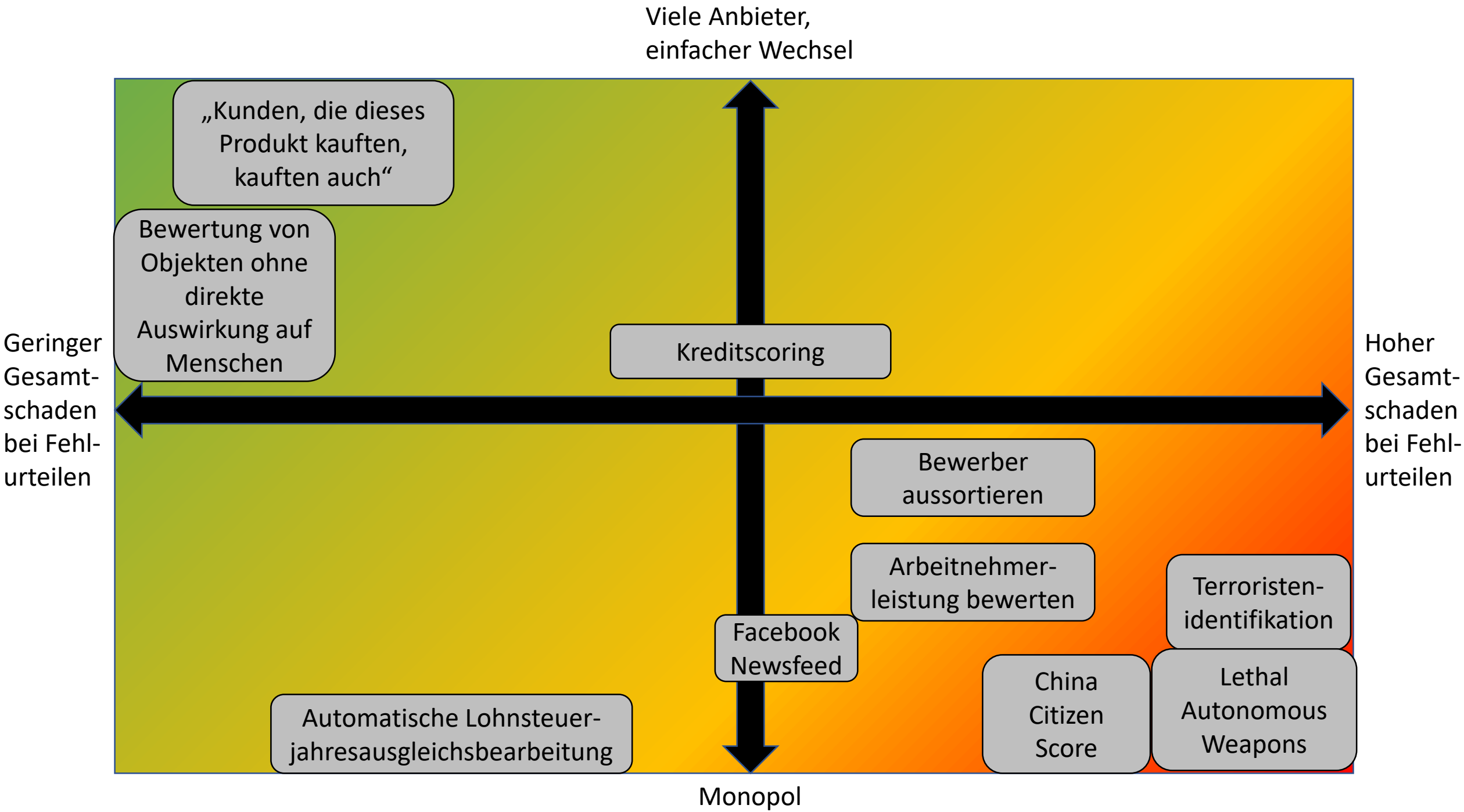
- Ganz schön schlecht: COMPAS
  - Hochrisiko-Kategorie:
    - Gewöhnliche Kriminaltaten: nur zu 70% richtig!
    - Schwere Straftaten: nur zu 20% richtig!
- Ein amerikanisches Terroristenidentifikationssystem tönt:
  - „Nur 0.008% falsch Positive!“
  - Bei 55 Millionen Einwohner sind das 4.400 Unschuldige, um wenige Hundert zu identifizieren.
  - Von den „Hochrisikopersonen“ also vermutlich unter 20%!
- Im medizinischen Bereich teilweise besser als Doktoren!

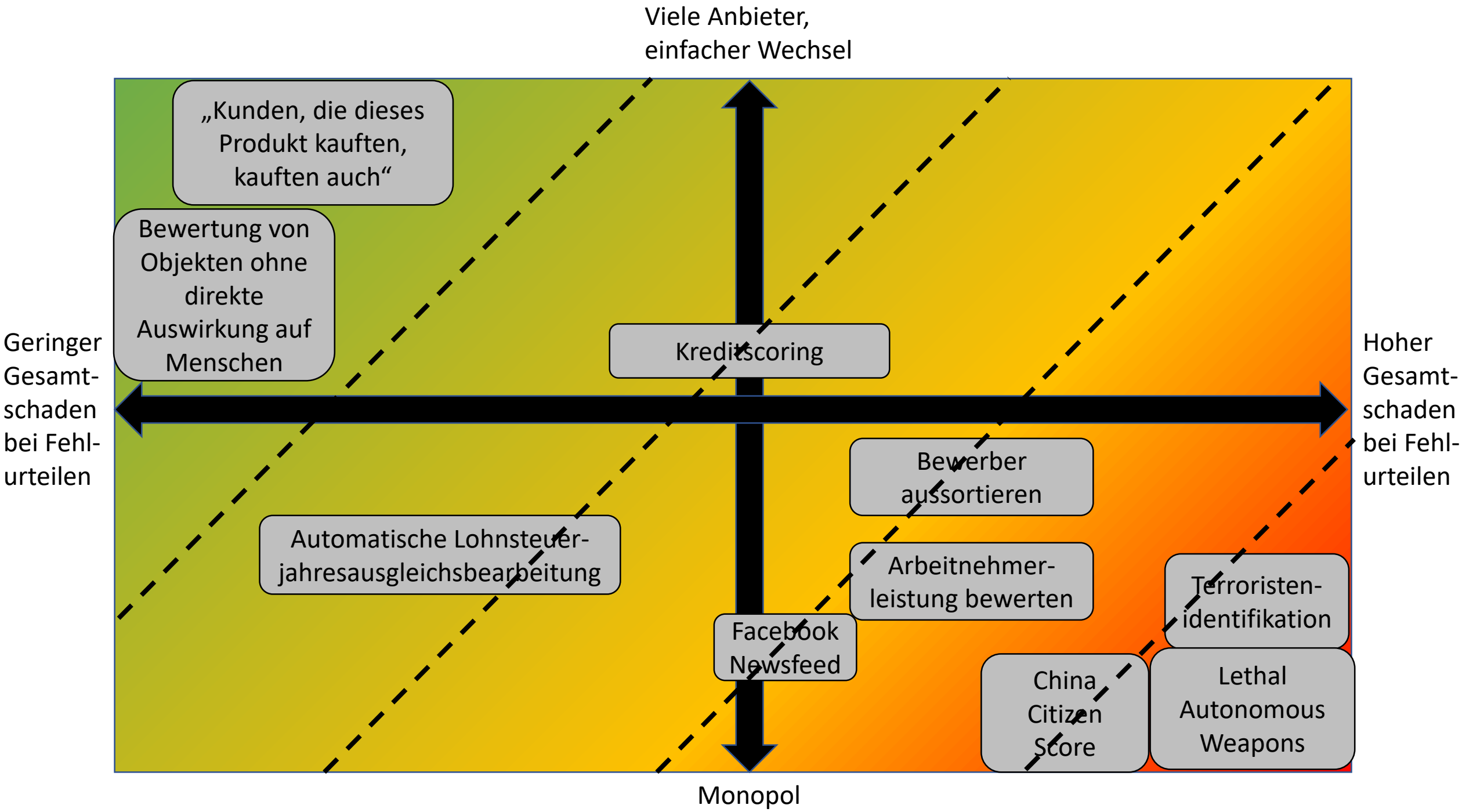
Wie bewerten bezüglich der Regulierungsnotwendigkeit?

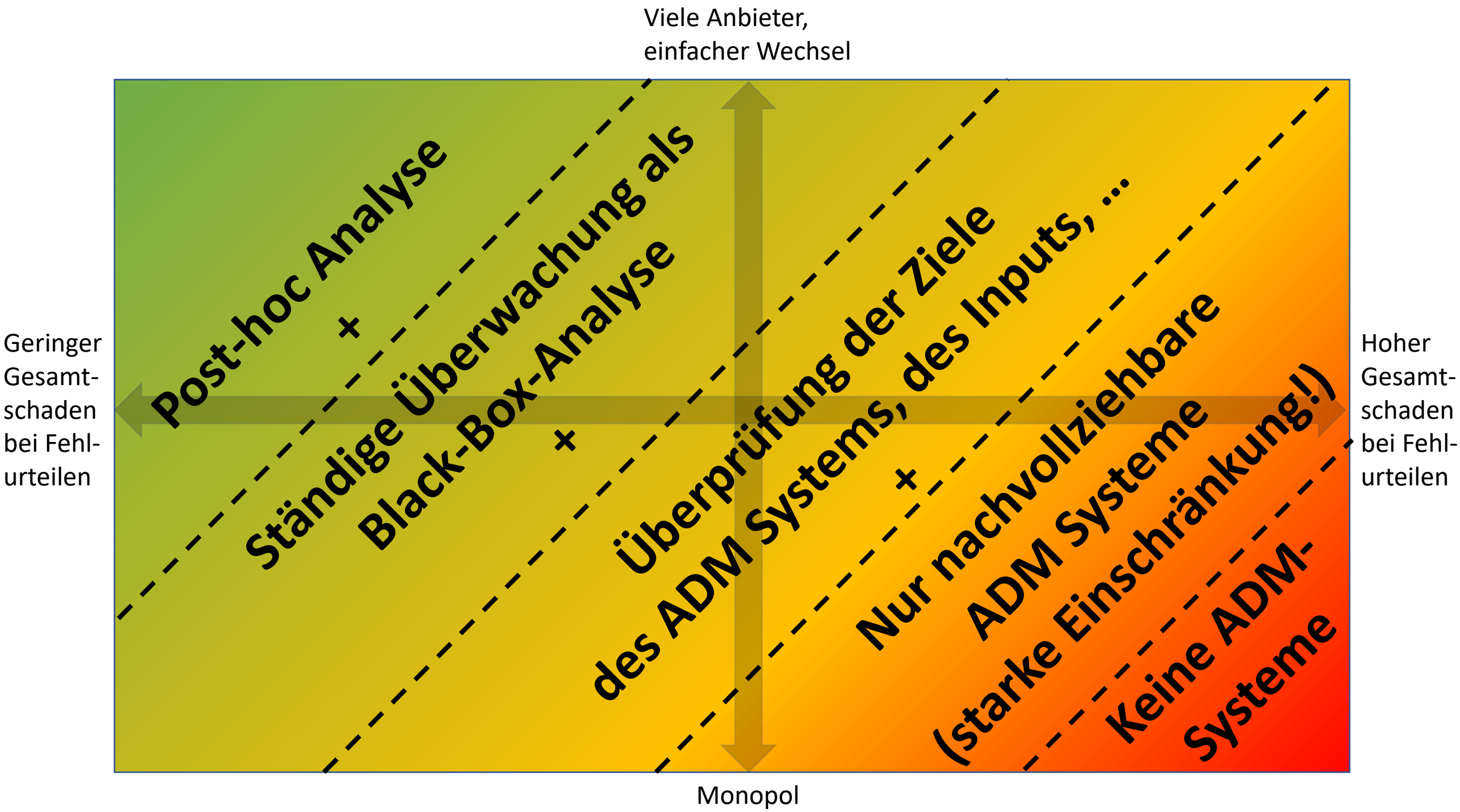
1. Schadenstiefe

$$\sum \begin{array}{c} \text{Schaden für Individuum(Fehlurteil)} \\ +\text{Schaden für Gesellschaft(Fehlurteil)} \end{array}$$

2. Anbietervielzahl und Wechselmöglichkeit


Viele Anbieter,
einfacher Wechsel
„Kunden, die dieses Produkt kauften, kauften auch“
Bewertung von Objekten ohne direkte Auswirkung auf Menschen
Kreditscoring
Geringer Gesamt-schaden bei Fehl-urteilen
Hoher Gesamt-schaden bei Fehl-urteilen
Bewerber aussortieren
Arbeitnehmer-leistung bewerten
Terroristen-identifikation
Facebook Newsfeed
China Citizen Score
Lethal Autonomous Weapons
Automatische Lohnsteuer-jahresausgleichsbearbeitung
Monopol


Viele Anbieter,
einfacher Wechsel
„Kunden, die dieses Produkt kauften, kauften auch“
Bewertung von Objekten ohne direkte Auswirkung auf Menschen
Kreditscoring
Geringer Gesamt-schaden bei Fehl-urteilen
Hoher Gesamt-schaden bei Fehl-urteilen
Bewerber aussortieren
Automatische Lohnsteuer-jahresausgleichsbearbeitung
Arbeitnehmer-leistung bewerten
Terroristen-identifikation
Facebook Newsfeed
China Citizen Score
Lethal Autonomous Weapons
Monopol


Viele Anbieter,
einfacher Wechsel
Geringer Gesamt-schaden bei Fehl-urteilen
Hoher Gesamt-schaden bei Fehl-urteilen
Monopol
Post-hoc Analyse
+
Ständige Überwachung als
Black-Box-Analyse
+
Überprüfung der Ziele
des ADM Systems, des Inputs, …
+
Nur nachvollziehbare
ADM Systeme
(starke Einschränkung!)
Keine ADM-
Systeme


Viele Anbieter,
einfacher Wechsel
Staatliche KI hat oft ein höheres Schadenspotenzial (superlineare Effekte)
Monopol
Geringer Gesamt-schaden bei Fehl-urteilen
Hoher Gesamt-schaden bei Fehl-urteilen
Klasse 4
Keine ADM-Systeme
Monopol

Fallbeispiel: AMS-Algorithmus

>66% zugewiesene Wahrschein-lichkeit, Zielkriterium 1 zu erfüllen.	ALLE ANDEREN	< 25% zugewiesene Wahr-scheinlichkeit Zielkriterium 3 zu erfüllen

- Teilt Arbeitslose in 3 Klassen ein:
  - Hohe Integrationschancen – keine weiteren Maßnahmen nötig.
  - Mittlere Integrationschancen – mit Maßnahmen
  - Niedrige Integrationschancen – Maßnahmen nicht sinnvoll.
- Ist noch im Teststadium
- Mitarbeiter können Entscheidung hochstufen.

Fallbeispiel: AMS

- Firma: Synthesis Forschung GmbH
- Methode: Logistische Regression
- Inputdaten sind bekannt
- Groundtruth:
  - Zielfunktion 1: Erfolgreiche Integration heißt innerhalb von 7 Monaten mind. 90 ungeförderte Arbeitstage
  - Zielfunktion 3 (sic!): Innerhalb von 24 Monaten mind. 180 ungeförderte Beschäftigungstage.
- Accuracy: 80% aller Entscheidungen sind korrekte Entscheidungen (in Klasse A und C – B bleibt „übrig“).

http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

Sonstige Beobachtungen

- Die Beschreibung ist schon fast vorbildlich (noch nicht ganz nachvollziehbar).
- Hier wird die Grenze des Black-Box-Ansatzes deutlich:
  - Trotz Transparenz können nur künstliche Datensätze ersonnen werden,
  - Die echte Verteilung der Datensätze ist unbekannt.
  - Daher kann Verteilung der verschiedenen Gruppen (Geschlecht, Alter, Herkunft) über die drei Klassen nicht nachvollzogen werden.

http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

Was fehlt?

- Wie kam Ausschreibung zustande?
- **Was genau soll optimiert werden im sozialen Gesamtprozess?**
- Wie genau wird das gemessen?
- Bekommen Bürger ihre Einteilung auch ohne Gespräch? Wichtig für Black-Box-Analyse

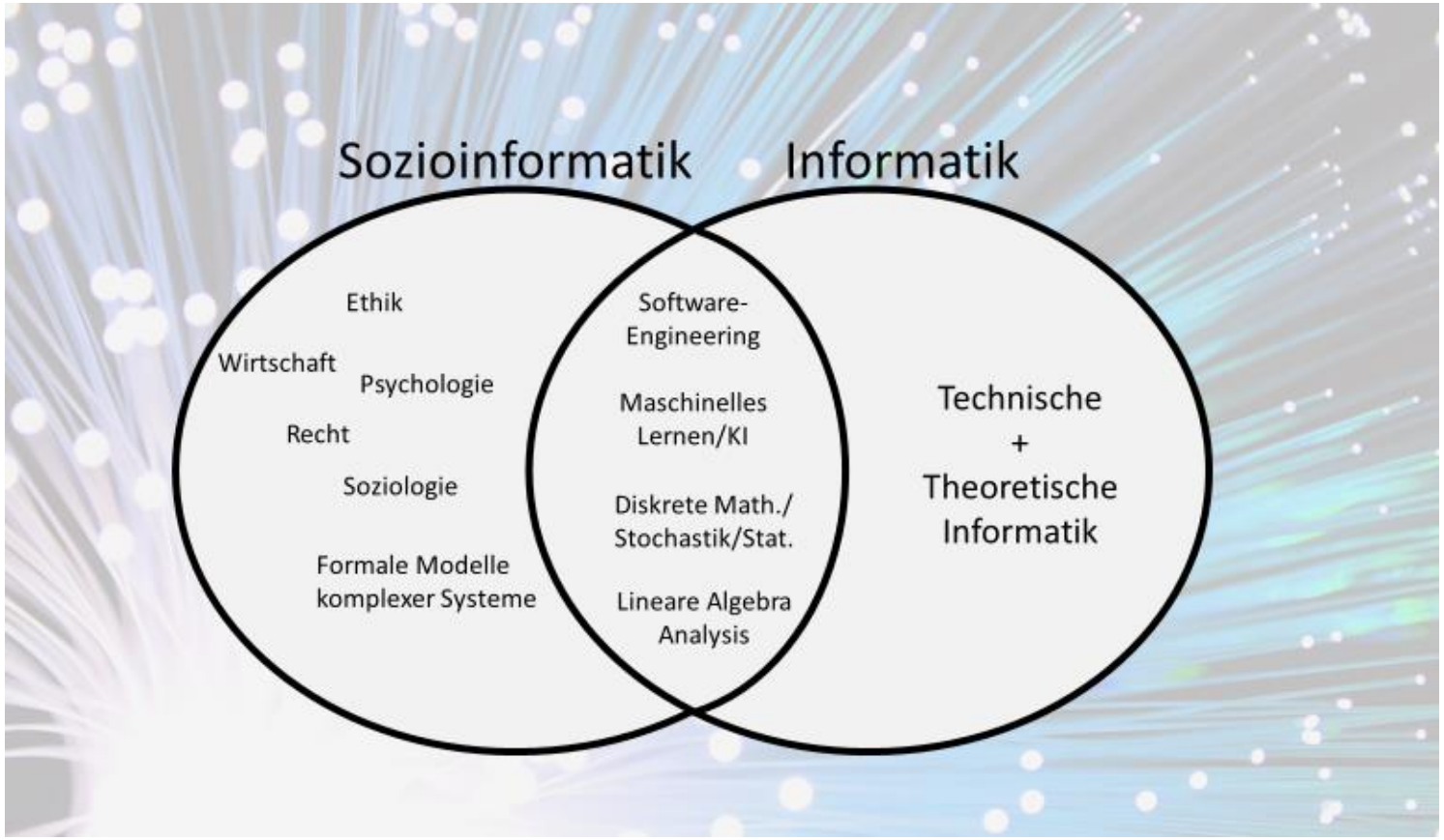
AMS kontert Kritikern: Förderbudget beim Arbeitsmarktservice kommt Frauen überproportional zu Gute. Der neue Algorithmus diskriminiere nicht.

Wichtig: Sozialverträglichkeitsregeln (Ausschnitt)

- Klassifizierung muss mit Bürger:in im Dialog besprochen werden.
- Darf nur unterstützend sein.
- Daten dürfen nicht älter als 4 Jahre sein.
- Modell muss kontinuierlich neu gelernt werden.

Wer kann und sollte überwachen?

- Viele Institutionen schon da:
  - Verbraucherschutz
  - Landesmedienanstalten
  - Betriebsräte
  - NGOs
  - …
- Müssen digital „aufgenordet“ werden:
  - brauchen eigene Data Scientists und Sozioinformatiker.



Studiengang Sozioinformatik

- **Immer noch bundesweit einzigartig (!).**
- **Modellierung, Analyse und Gestaltung sozio-informatischer Systeme.**
- **Bundespolitik sehr interessiert an Absolventen.**
- **Nachfrage von anderen Universitäten zur Beratung.**

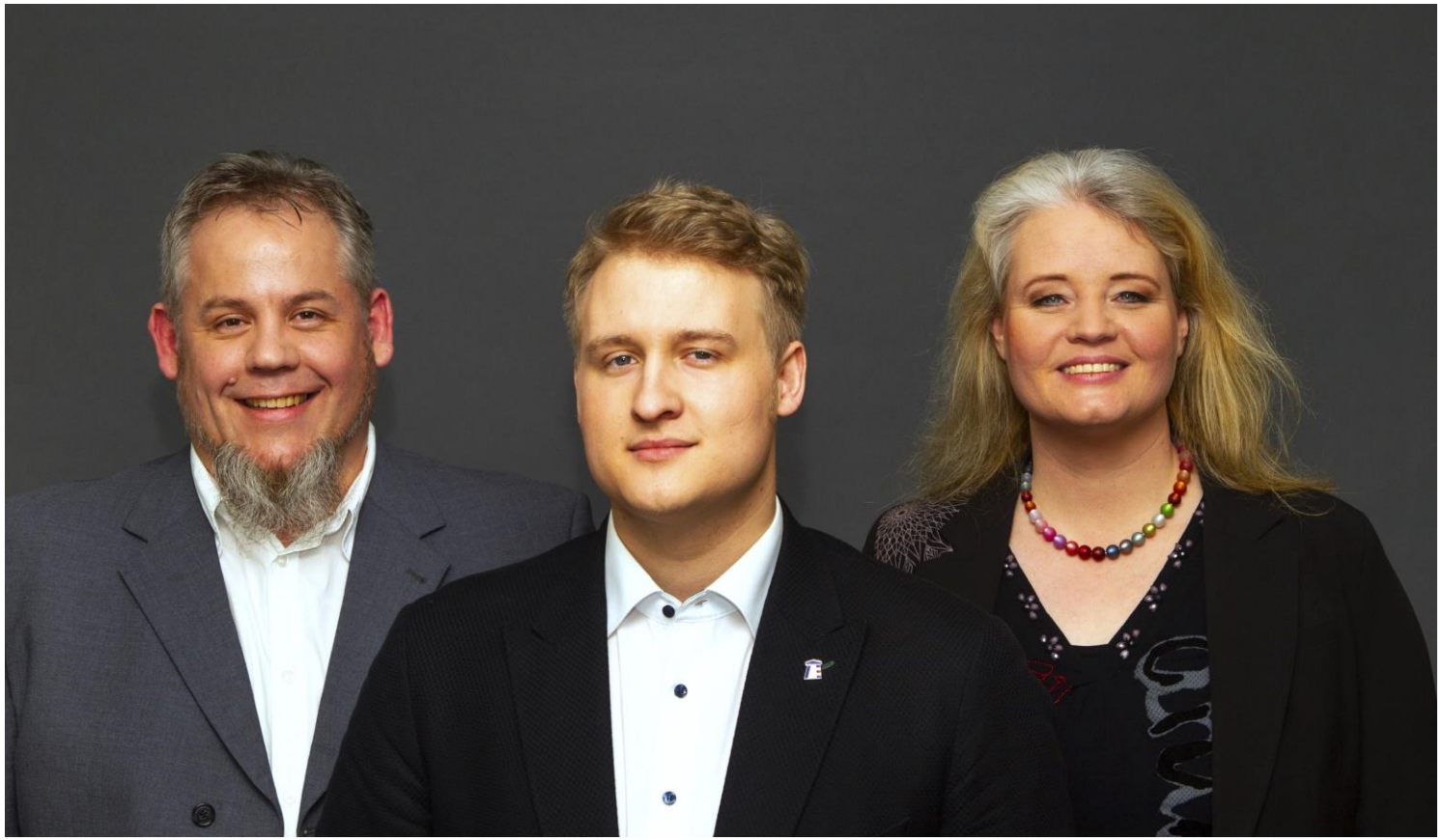
Sie designen sozioinformatisch relevante IT-Systeme

Sie programmieren diese Systeme

Sie bewerten solche Systeme

Sie beraten die Gesellschaft

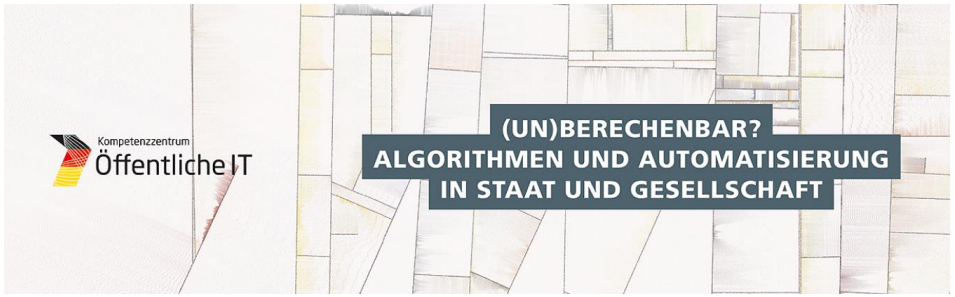
Berufsbild Sozioinformatik



Gründung: Trusted AI GmbH

- Gründung 2019
  - Ehepaar Zweig, Tobias Krafft
- Einzelcoachings zum Thema KI und allen verwandten Fragen
- Workshops für Betriebsräte, Lehrer:innen, Kirche, Landesmedienanstalten, Manager, Ministerien, Kanzleien, etc.
- Begleitung bei der Ausschreibung, beim Kauf und bei der Inbetriebnahme von KI zur Sicherstellung ethischer Entscheidungen

Weitere Informationen



1. Studie für die Bertelsmann-Stiftung:
   Zweig, Fischer & Lischka: „Wo Maschinen irren können“ (Serie AlgoEthik, No. 4, 2018)
2. Zwei Kapitel im Sammelband (Un)Berechenbar? des Fraunhofer FOKUS, Kompetenzzentrum ÖFIT, 2018
   1. Zweig & Krafft: „Fairness und Qualität algorithmischer Entscheidungen“
   2. Krafft & Zweig: „Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann“
3. Studie für die Konrad-Adenauer-Stiftung „Algorithmische Entscheidungen: Transparenz und Kontrolle“ (Zweig, erscheint 2019)
4. Studie vom Fraunhofer FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT): Opiela, Mohabbat Kar, Thapa & Weber: Exekutive KI 2030 – Vier Zukunftsszenarien für Künstliche Intelligenz in der öffentlichen Verwaltung, 2018)

Ab Herbst:

Buch für die breite Öffentlichkeit zum Thema

Kontakt

Prof. Dr. Katharina A. Zweig
Algorithm Accountability Lab
Gottlieb-Daimler-Str. 48
67663 Kaiserslautern

aalab.informatik.uni-kl.de

zweig@cs.uni-kl.de
@nettwerkerin bei Twitter