KI als Dichter und Richter?

Prof. Dr. K.A. Zweig
TU Kaiserslautern
Algorithm Accountability Lab
@nettwerkerin

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 …“

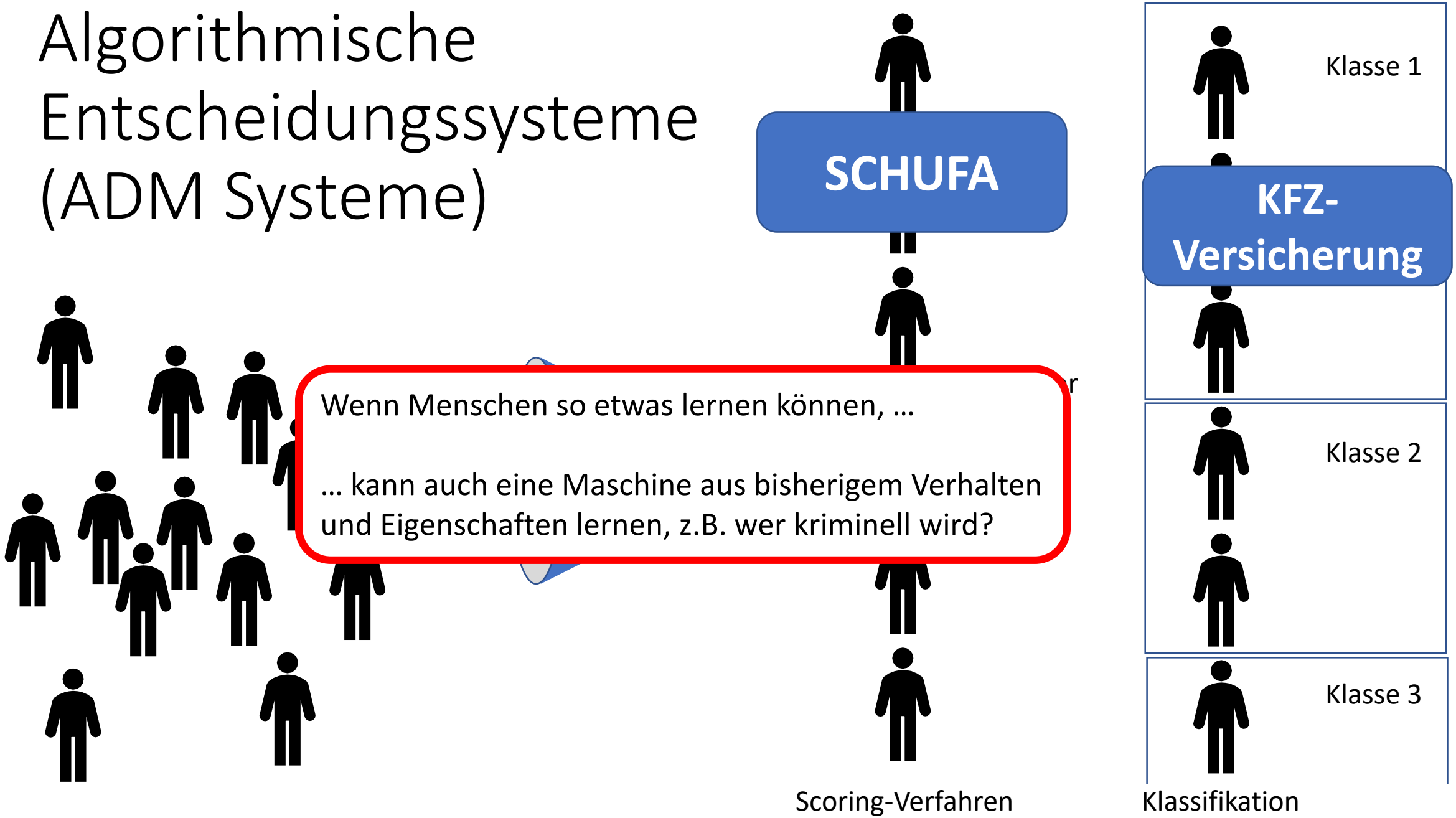
https://www.bundestag.de/gebaerdensprache/mediathek?videoid=7272804#url=bWVkaWF0aGVrb3ZlcmxheT92aWRlb2lkPTcyNzI4MDQ=&mod=mediathek

Die zwei Ängste

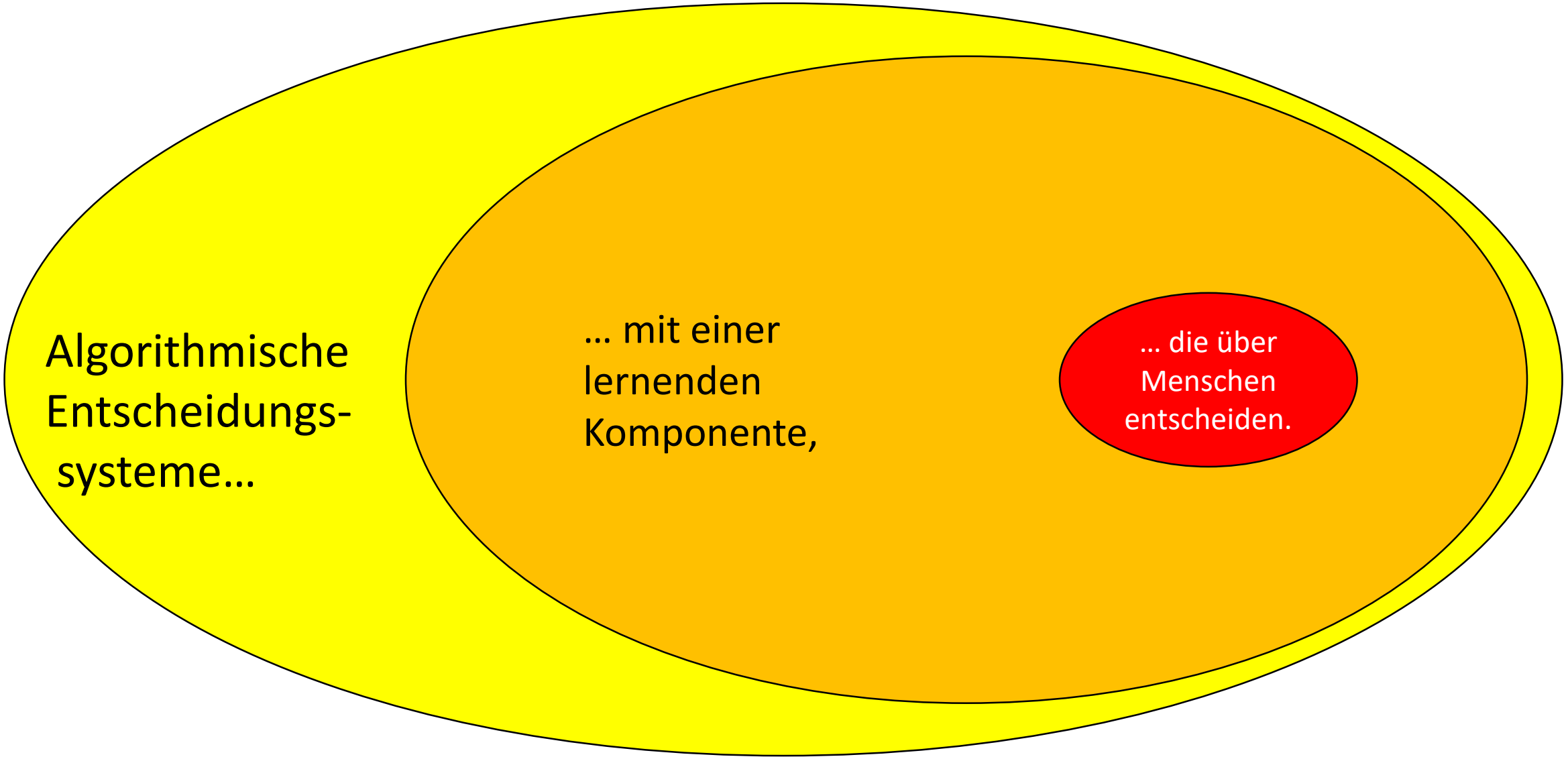
Sie wird richten

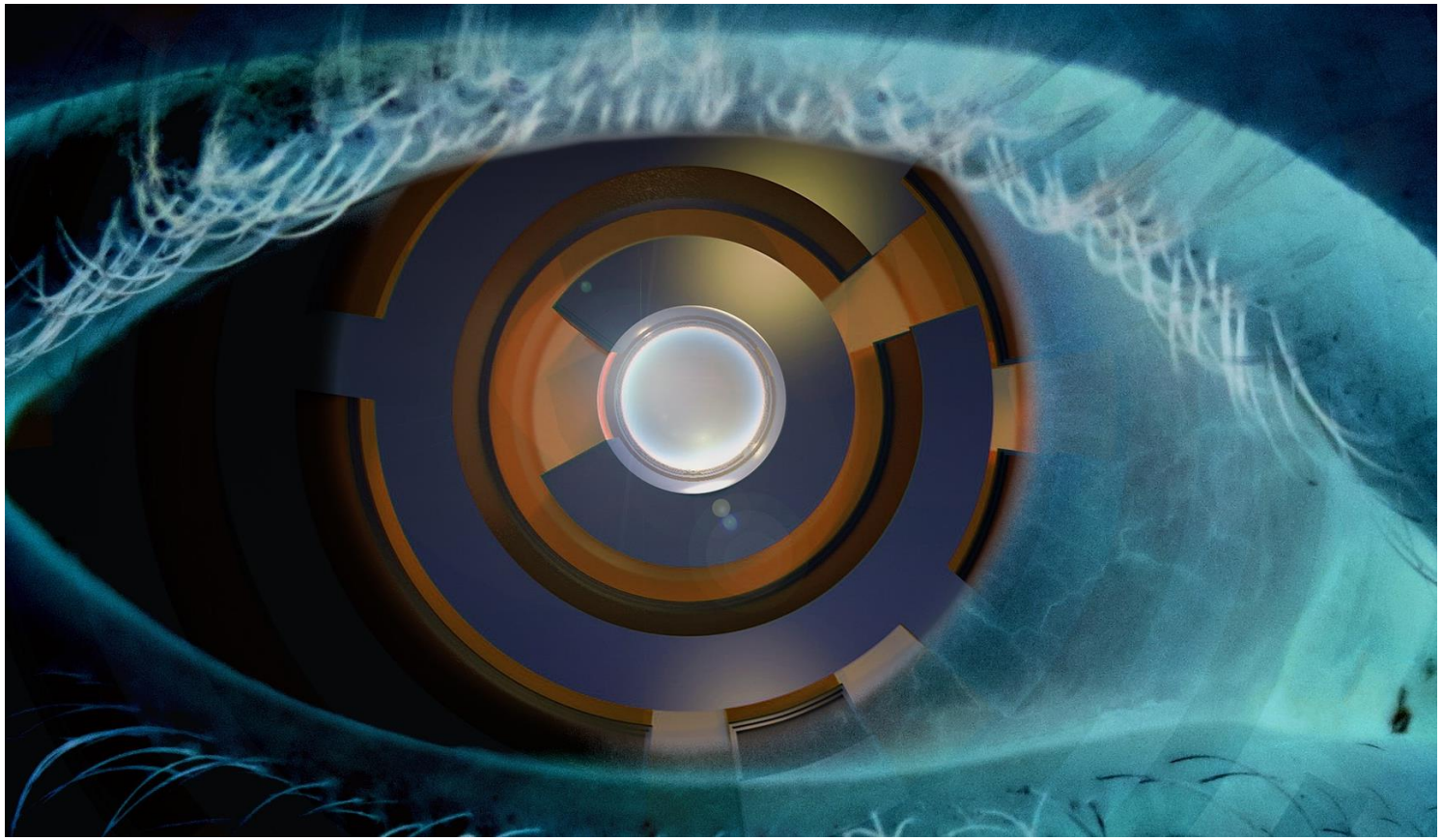
Sie wird dichten

Algorithmische Entscheidungssysteme (ADM Systeme)



Welche ADM-Systeme sind problematisch?





Begriffe

- Algorithmus
- Künstliche Intelligenz (schwache, starke)
- Maschinelles Lernen / selbstlernende Systeme

- Nachvollziehbarkeit
- Transparenz

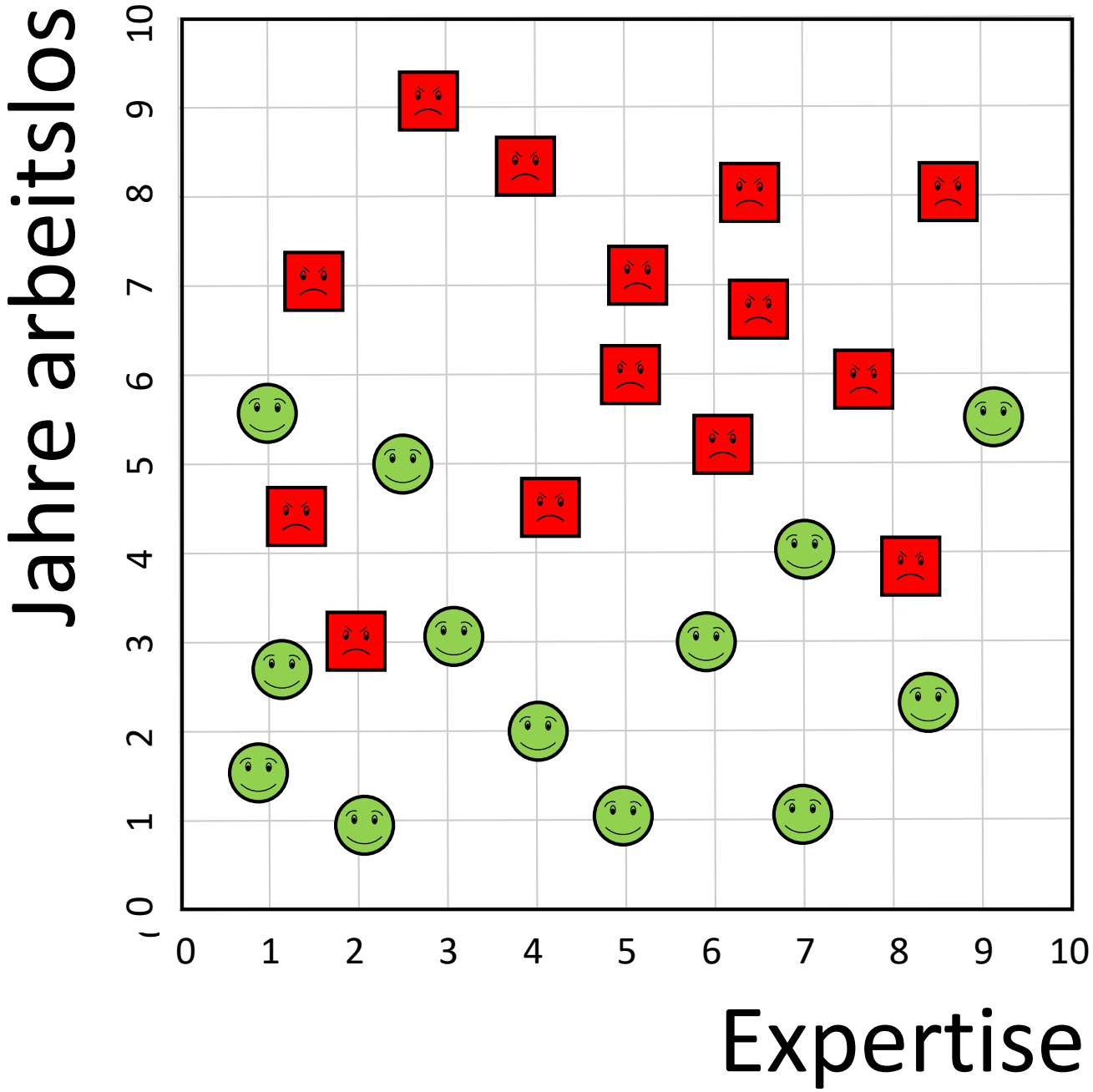
Wie „lernt“ das System von Daten?

DIY:

Sie sind heute meine „Support Vector Machine“

Weniger erfolgreiche Arbeitnehmer:innen

Erfolgreiche Arbeit-nehmer:innen



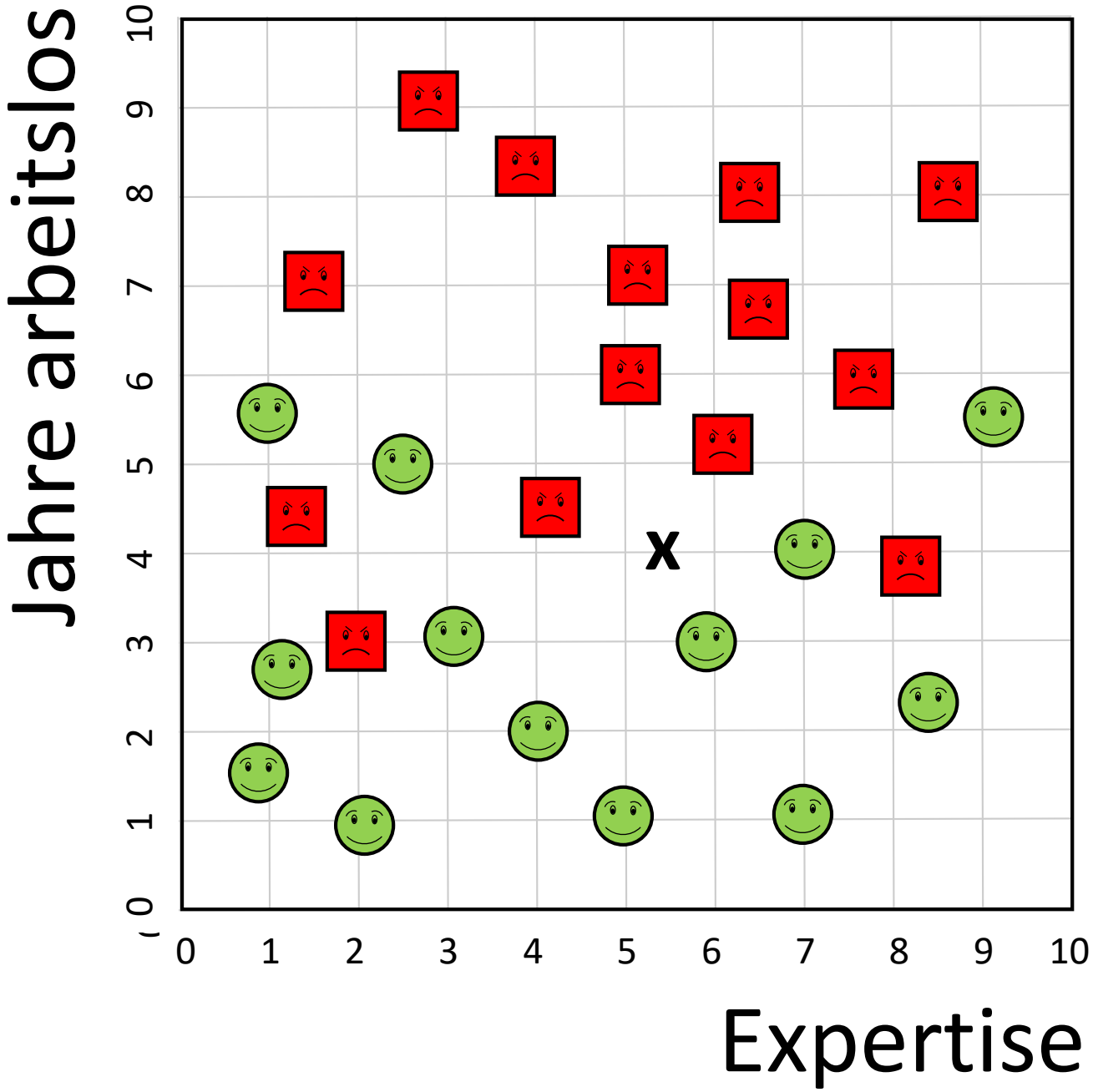
Weniger erfolgreiche Arbeitnehmer:innen

Erfolgreiche Arbeit-nehmer:innen

Bewerten Sie Frau Müller:

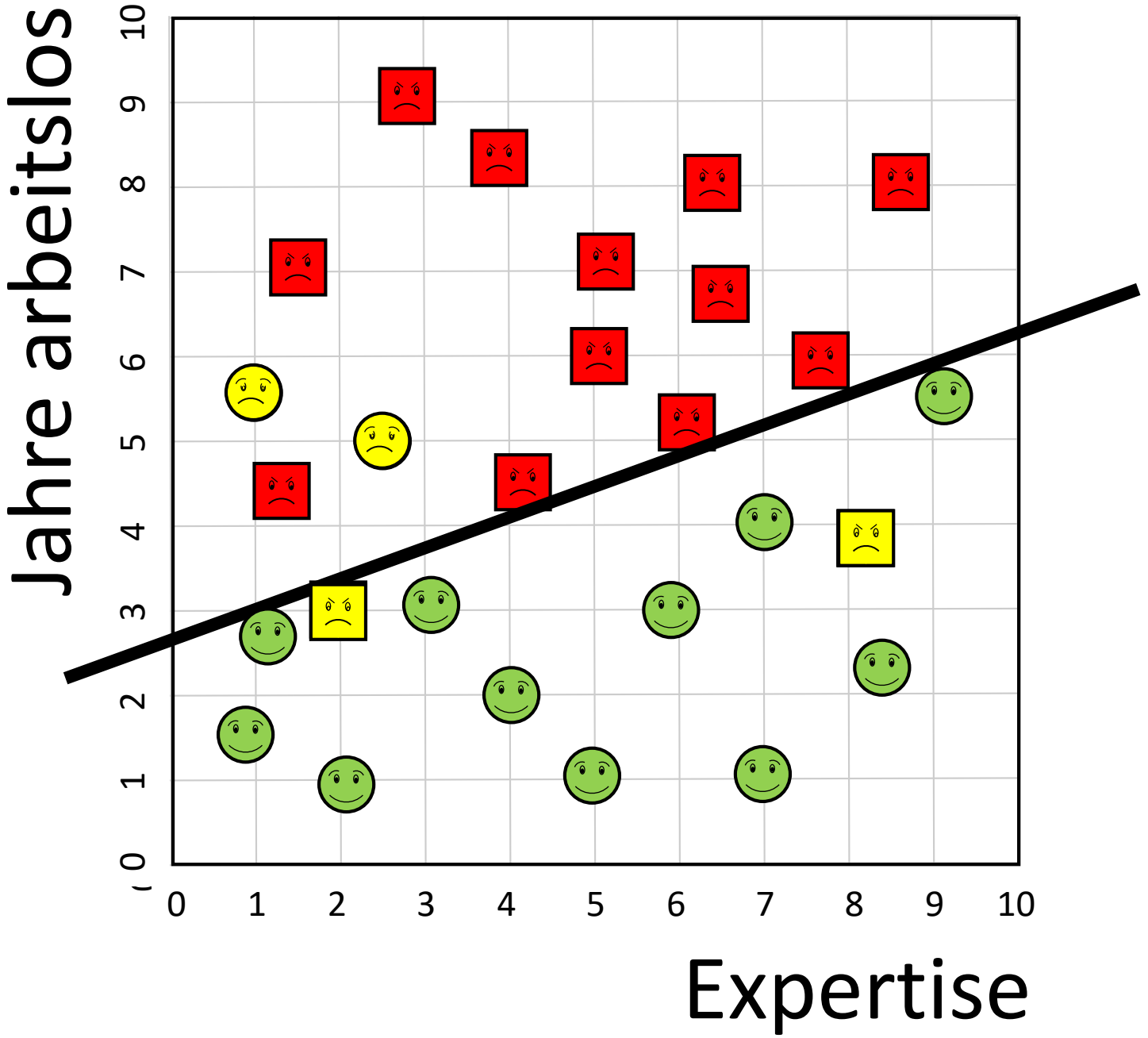
5.5 Jahre Erfahrung

4 Jahre arbeitslos



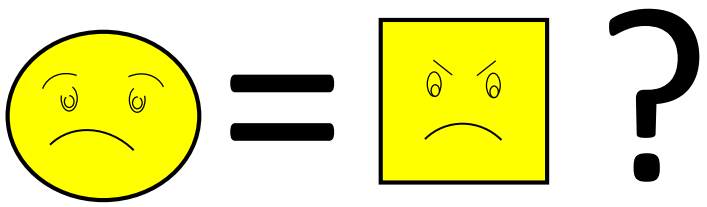
Weniger erfolgreiche Arbeitnehmer:innen

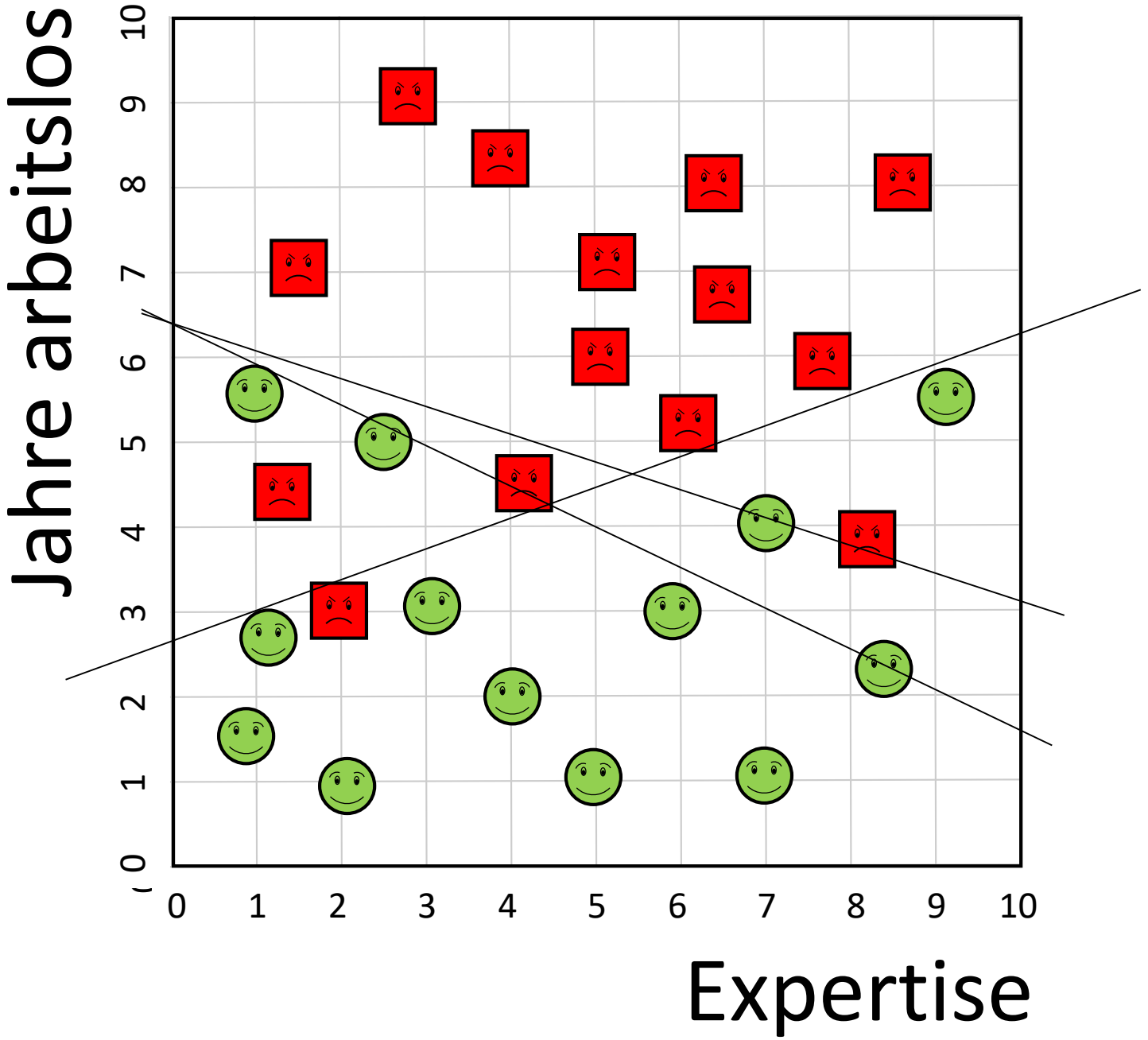
Erfolgreiche Arbeit-nehmer:innen



Weniger erfolgreiche Arbeitnehmer:innen

Erfolgreiche Arbeit-nehmer:innen



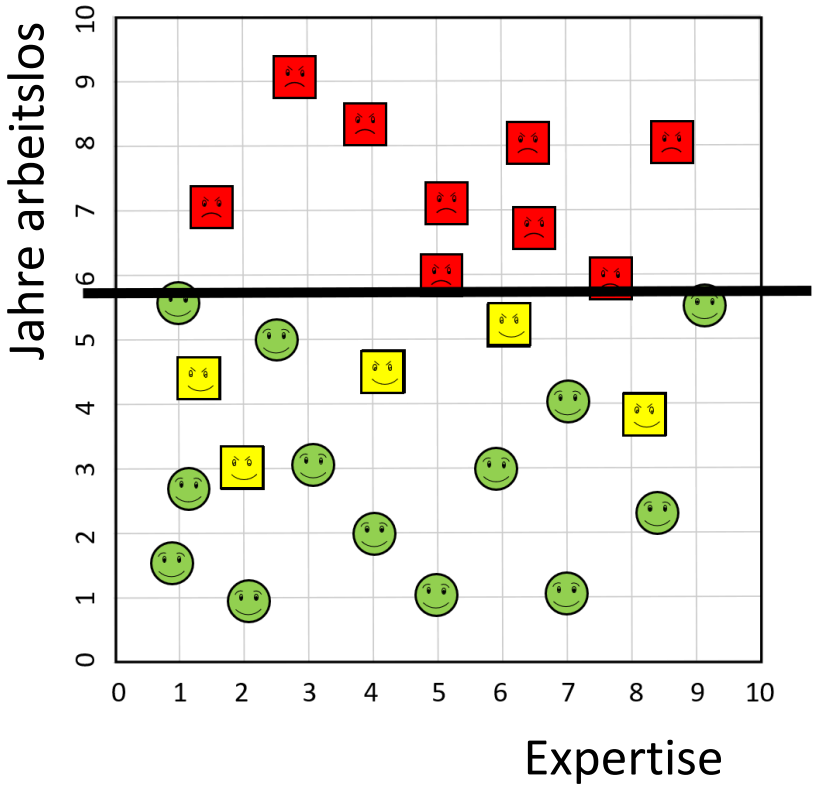


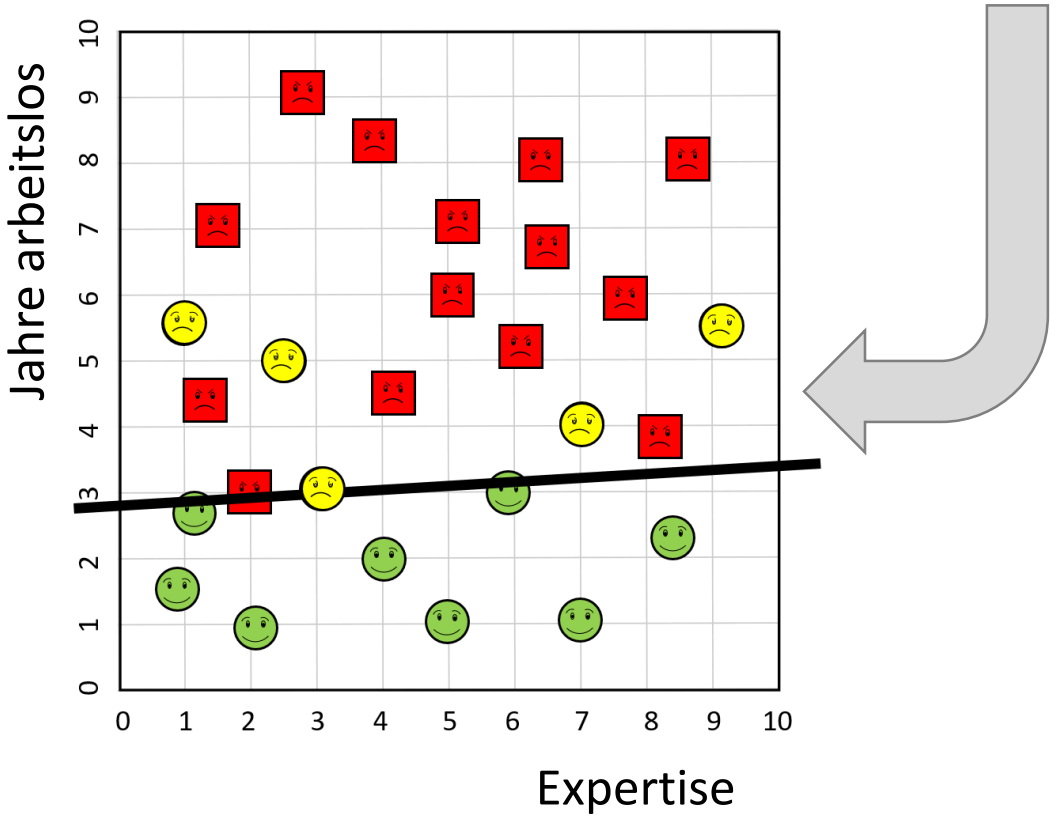
„Wir können es uns nicht leisten
erfolgreiche
Arbeitnehmer:innen
zu übersehen!“

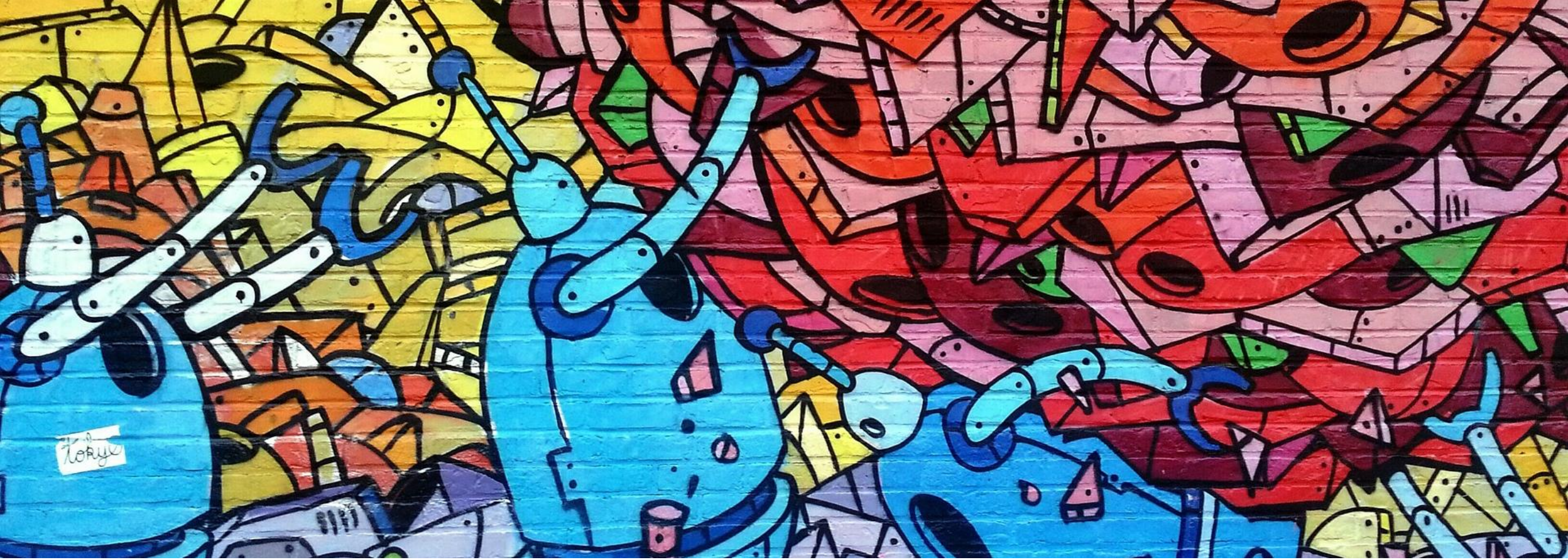
Weniger erfolgreiche Arbeitnehmer:innen

Erfolgreiche Arbeit-nehmer:innen

„Wir können es uns nicht leisten,
weniger erfolgreiche
Arbeitnehmer:innen
einzustellen!“







Warum Betriebsräte hier mitbestimmen müssen!

Qualität von ADM Systemen

1. **Wer entscheidet, wann ein ADM System „gut“ ist? Wer, wann es „fair“ ist?**

Algorithmen…

- … basieren auf Korrelationen von Eigenschaften mit gewünschtem Verhalten.
- Quasi **algorithmisch legitimierte Vorurteile**:
  - Zu 70% erfolgreich heißt:
  - Von 100 Personen, die „genau so sind wie dieser Mensch“, sind 70 nachher erfolgreich.

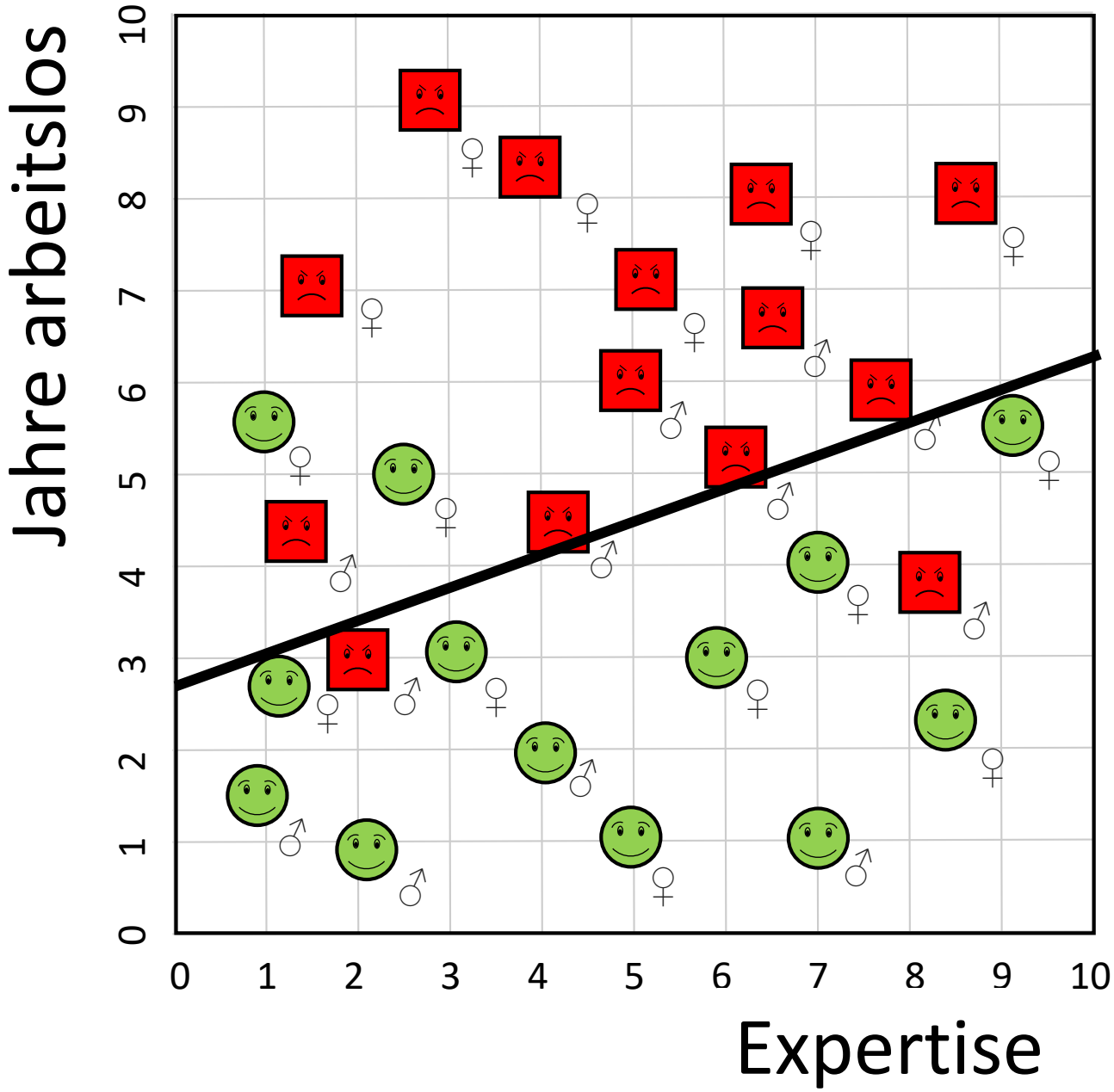
Qualität von ADM Systemen

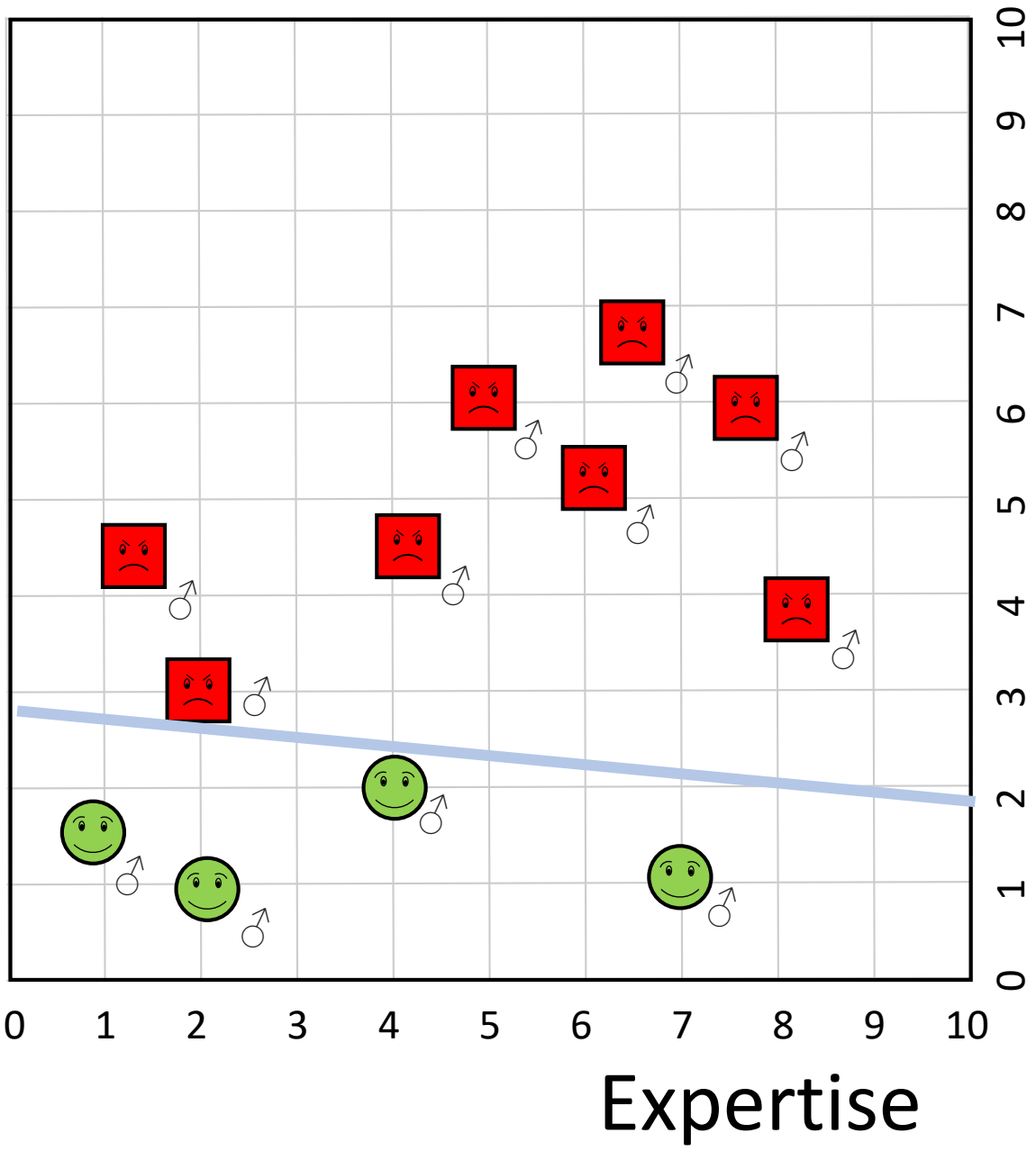
1. Wer entscheidet, wann ein ADM System „gut“ ist? Wer, wann es „fair“ ist?
2. **ADM Systeme ergeben nur Wahrscheinlichkeiten, keine Wahrheiten.**

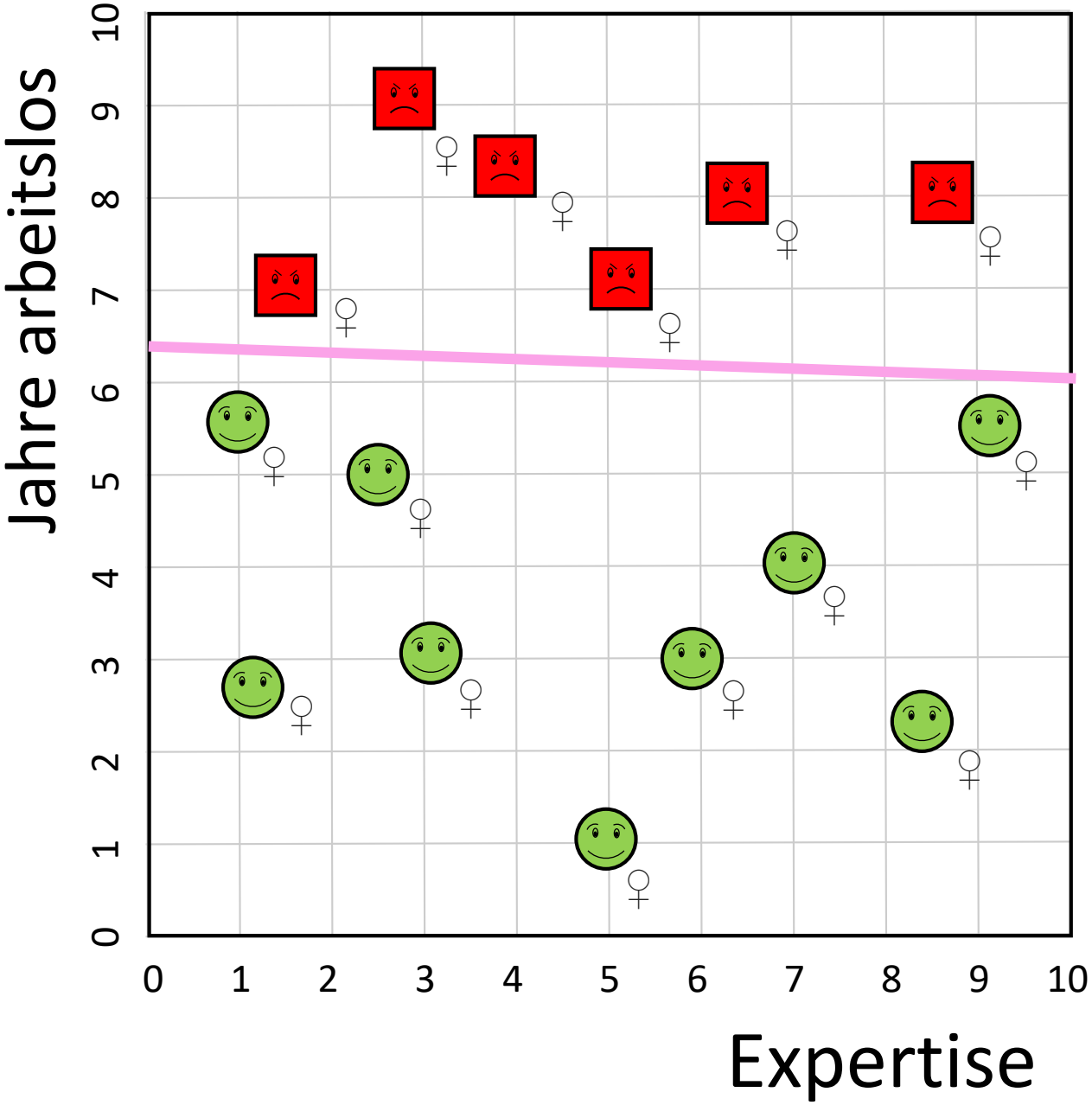
Weniger erfolgreiche Arbeitnehmer:innen

Erfolgreiche Arbeit-nehmer:innen

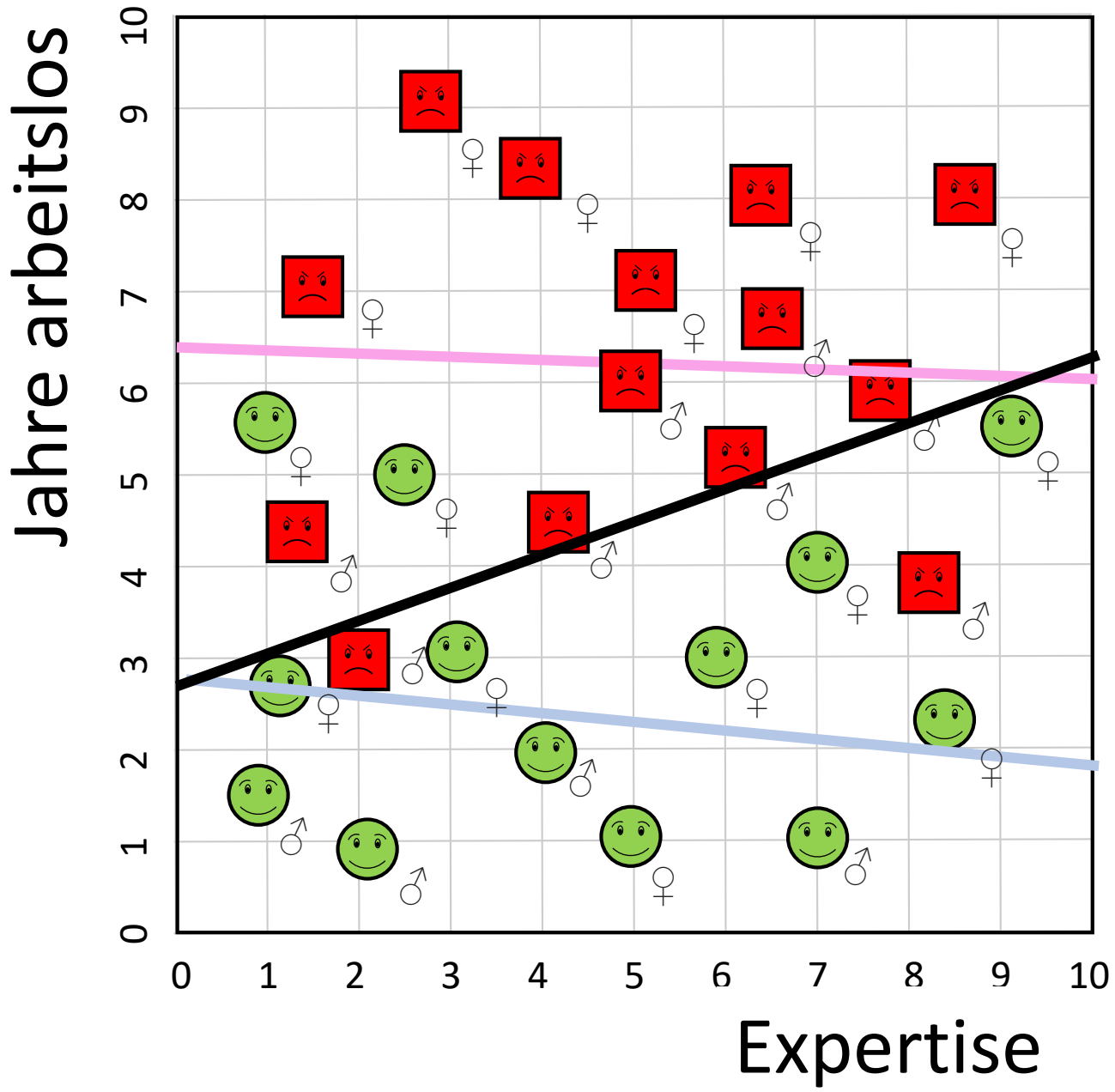
Daten-grundlage




0 1 2 3 4 5 6 7 8 9 10
Expertise


Jahre arbeitslos
0 1 2 3 4 5 6 7 8 9 10
Expertise

Effekt:
Wir diskriminieren!



Beobachtung

Eine geschützte Information kann wichtig sein, um bessere Entscheidungen zu treffen.

Diskriminierung wird nicht per se dadurch vermieden, dass die sensitive Information vorenthalten wird.

Qualität von ADM Systemen

1. Wer entscheidet, wann ein ADM System „gut“ ist? Wer, wann es „fair“ ist?
2. ADM Systeme ergeben nur Wahrscheinlichkeiten, keine Wahrheiten.
3. **ADM Systeme können diskriminieren.**

Sozio-informatische Gesamtanalyse

Sozio-informatische Betrachtungsweise

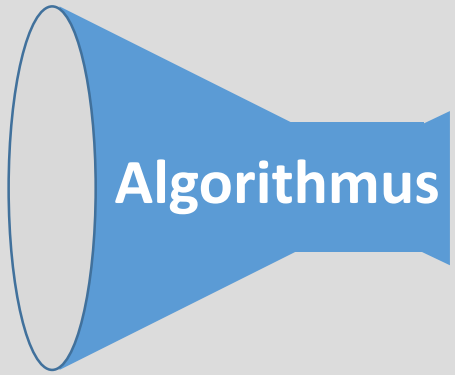
- Software schafft neue Anreize für menschliche Akteure.
- Diese reagieren auf die Anreize und wirken auf die Software ein.

Probleme der Einbettung des ADM in den sozialen Prozess

- **Aufmerksamkeitsökonomie** von Entscheiderinnen und Entscheidern.
- **„Best practice“** erfordert Nutzung der Software.
- **Delegation von Verantwortung!**
- Manchmal kann ein falsch Beurteilter **die Vorhersage prinzipiell nicht entkräften!**
  - Z.B. abgelehnte Bewerberin

Algorithmische Entscheidungs**systeme** (ADM Systeme)

Daten



Bewertete

Nutzer des ADM Systems

oder

Soziales System

Klasse 1

Klasse 2

Klasse 3

Scoring-Verfahren

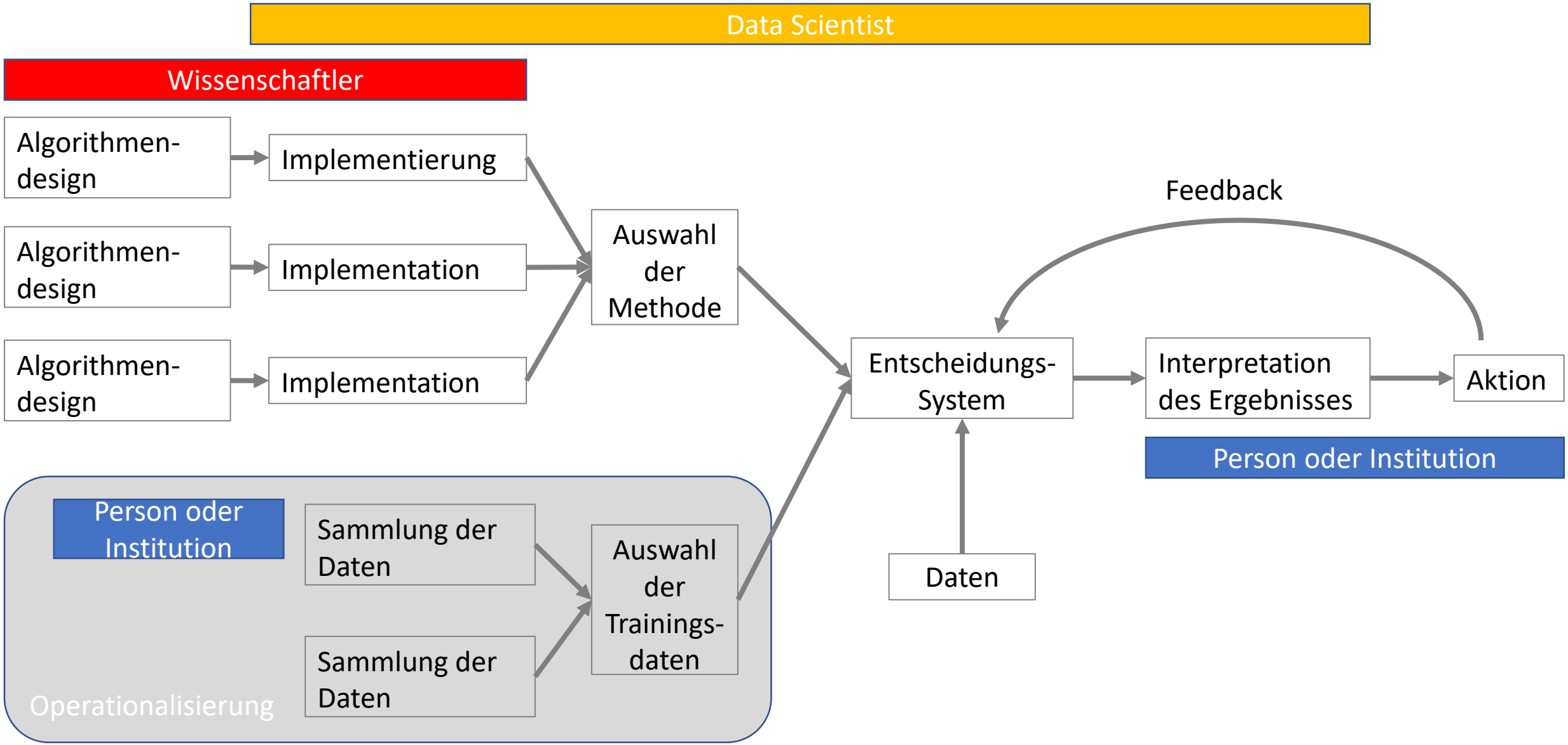
Qualität von ADM Systemen

1. Wer entscheidet, wann ein ADM System „gut“ ist? Wer, wann es „fair“ ist?
2. ADM Systeme ergeben nur Wahrscheinlichkeiten, keine Wahrheiten.
3. ADM Systeme können diskriminieren.
4. **ADM Systeme bedürfen einer sozio-informatischen Gesamtanalyse.**

Wie gut sind die Robo-Richter?

- Ganz schön schlecht: COMPAS
  - Hochrisiko-Kategorie:
    - Gewöhnliche Kriminaltaten: nur zu 50% 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!

Lange Kette der Verantwortlichkeiten

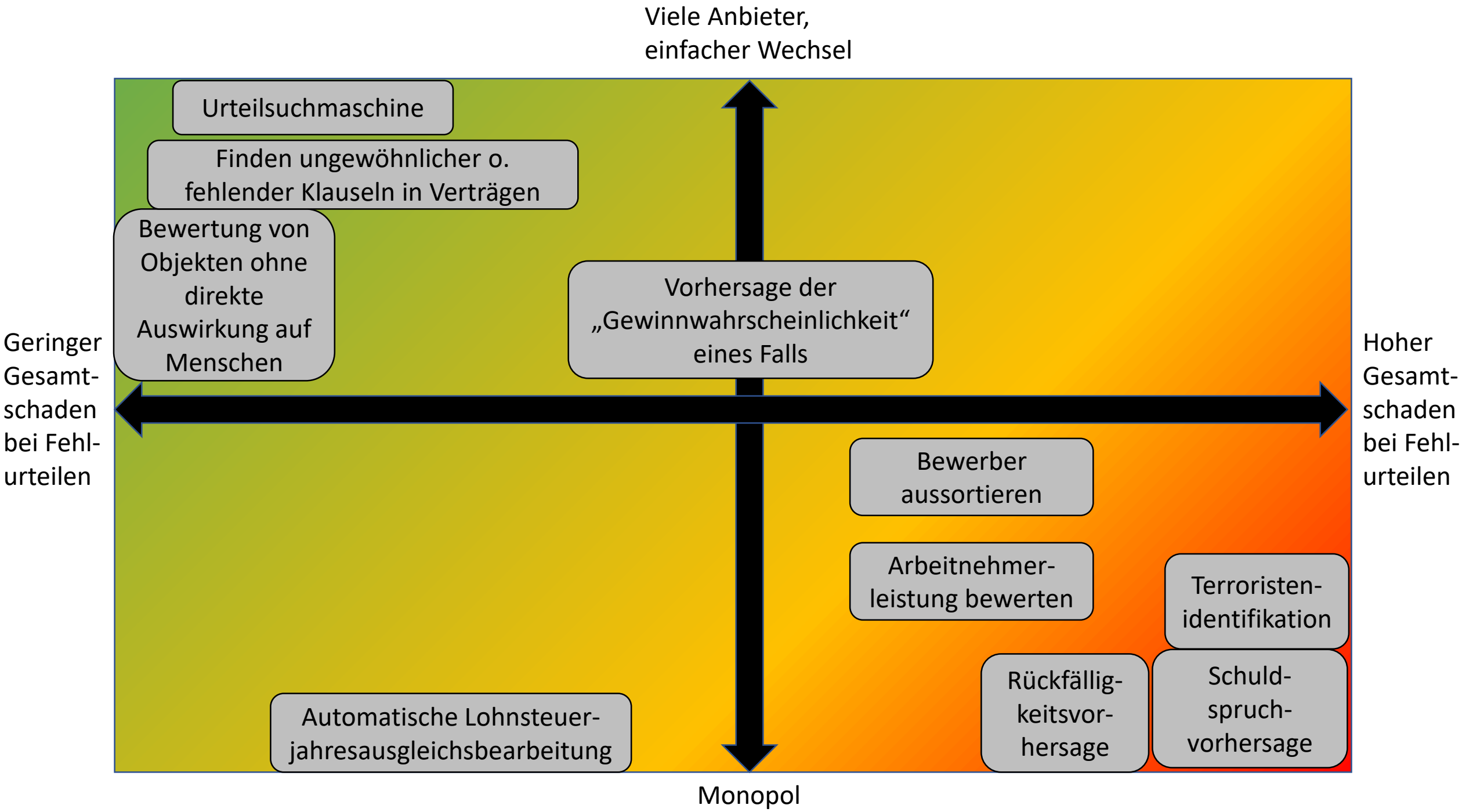


Wie bewerten bezüglich der Regulierungsnotwendigkeit?

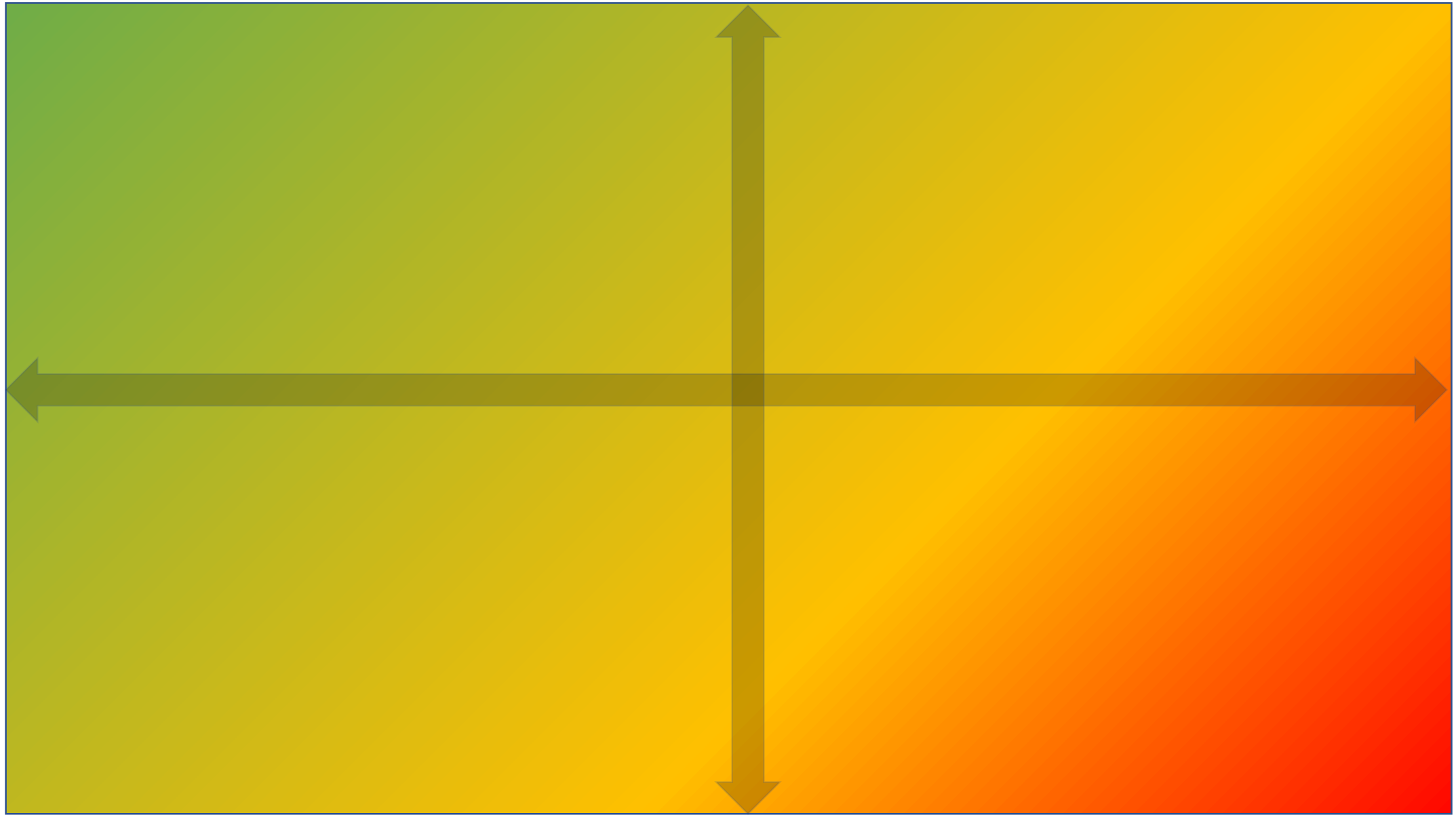
1. Schadenstiefe

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

2. Anbietervielzahl und Wechselmöglichkeit


Viele Anbieter,
einfacher Wechsel
Urteilsuchmaschine
Finden ungewöhnlicher o.
fehlender Klauseln in Verträgen
Bewertung von
Objekten ohne
direkte
Auswirkung auf
Menschen
Vorhersage der
„Gewinnwahrscheinlichkeit“
eines Falls
Geringer
Gesamt-
schaden
bei Fehl-
urteilen
Hoher
Gesamt-
schaden
bei Fehl-
urteilen
Bewerber
aussortieren
Arbeitnehmer-
leistung bewerten
Terroristen-
identifikation
Rückfällig-
keitsvor-
hersage
Schuld-
spruch-
vorhersage
Automatische Lohnsteuer-
jahresausgleichsbearbeitung
Monopol

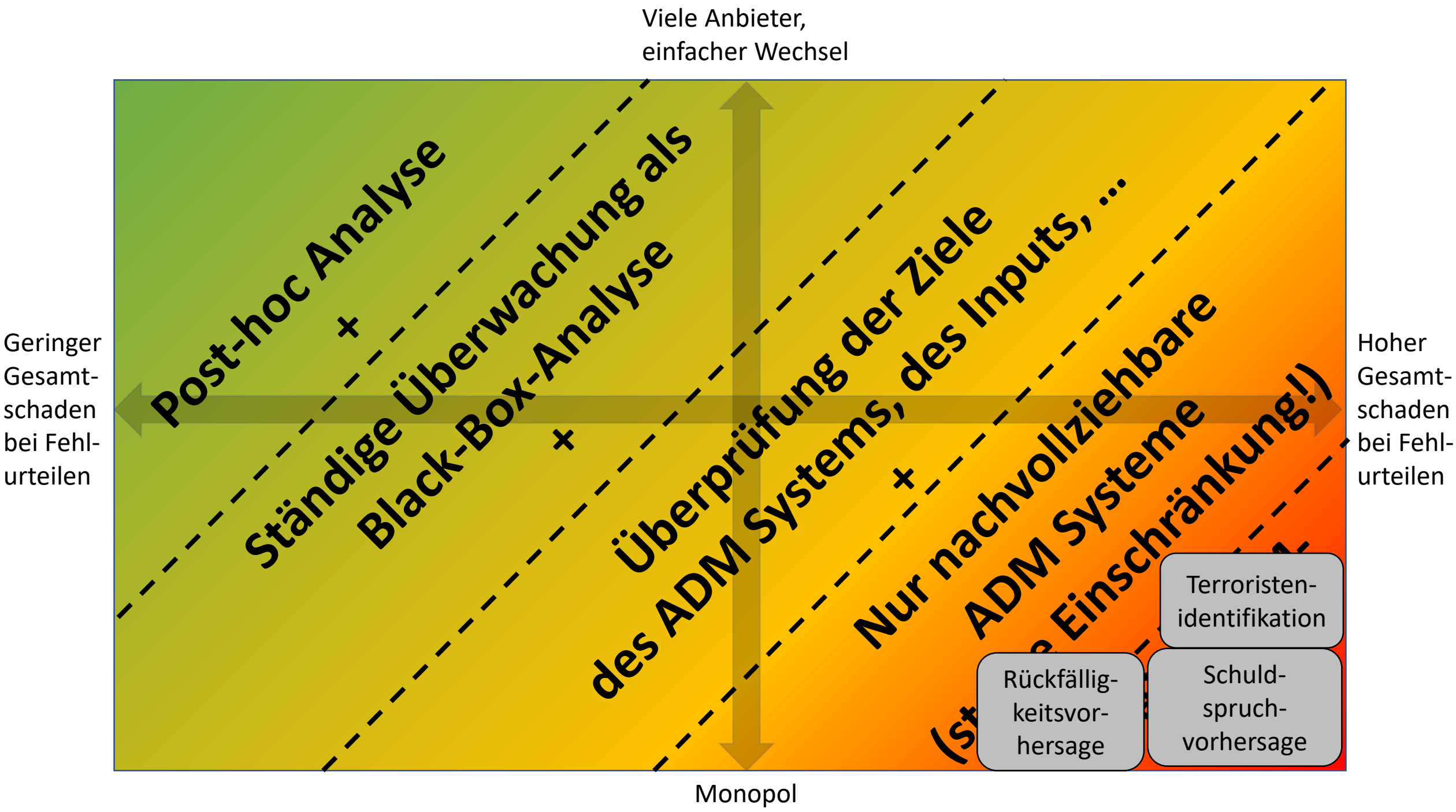
Viele Anbieter,
einfacher Wechsel

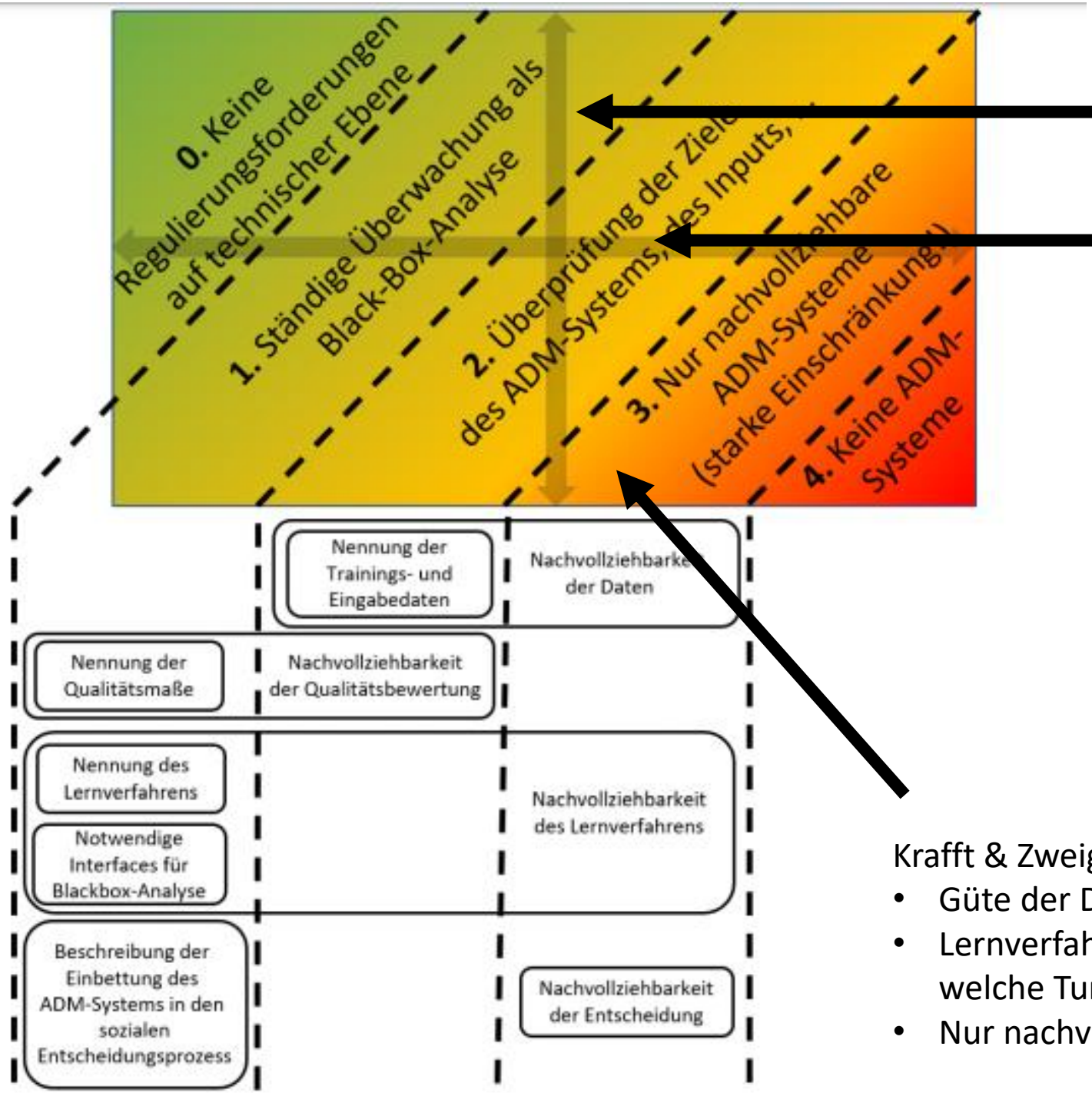


Geringer Gesamt-schaden bei Fehl-urteilen

Hoher Gesamt-schaden bei Fehl-urteilen

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
(st… Einschränkung!)
Terroristen-identifikation
Rückfällig-keitsvor-hersage
Schuld-spruch-vorhersage



Analyse auf Diskriminierung

Krafft & Zweig (2019) fordern:

- Nennung der Art der Inputdaten
- Nachvollziehbarkeit Qualität (Fairness)

Holl, Kernbeiß & Wagner-Pinter fordern [9]:

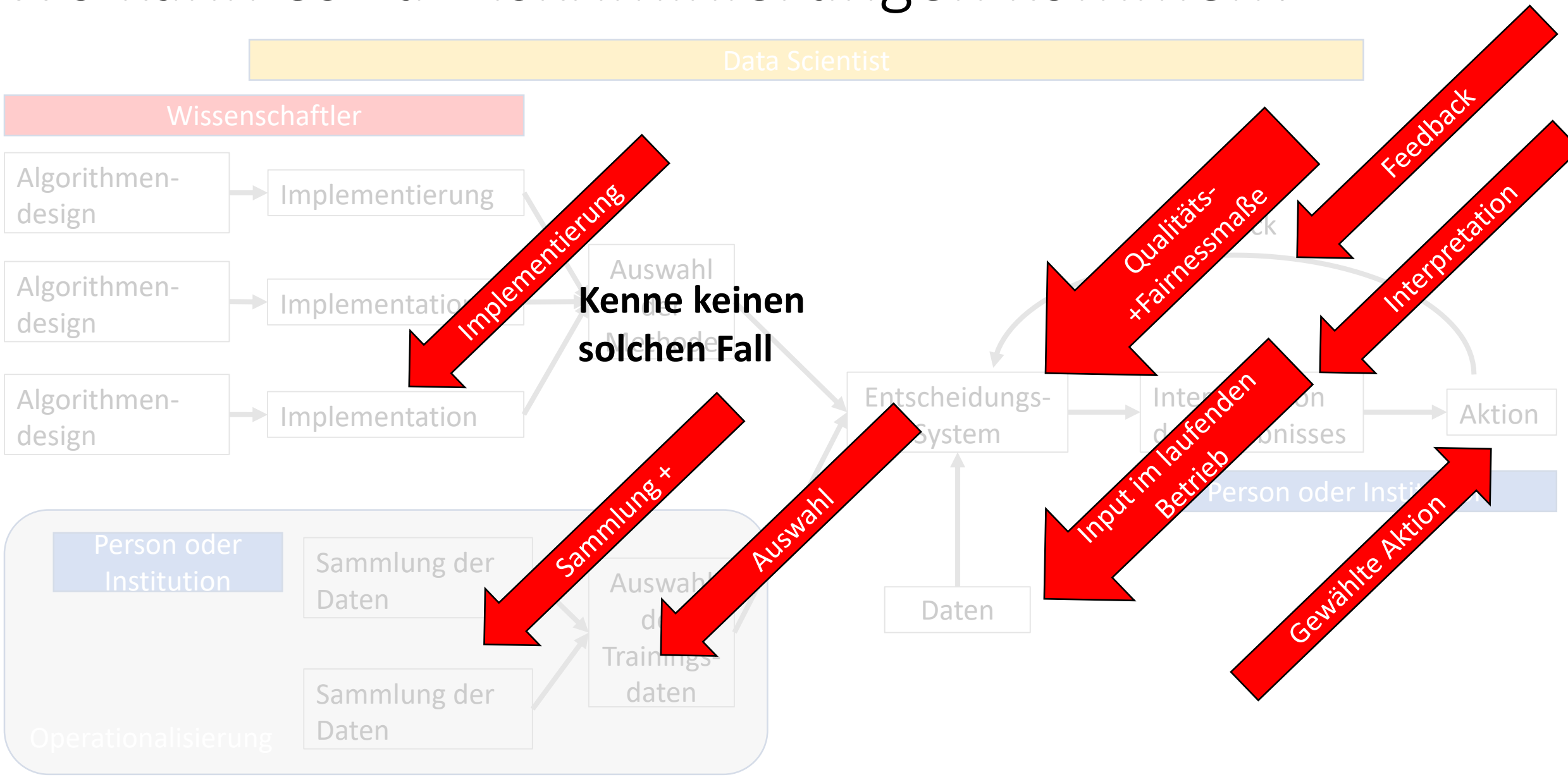
- Nur Daten mit Kausalzusammenhang zu gewünschter Vorhersage.
- Daten müssen für Betroffene(n) leicht änderbar und löschbar sein.
- Daten sind höchstens ein paar Jahre alt.

Krafft & Zweig (2019) fordern:

- Güte der Daten von außen nachvollziehbar
- Lernverfahren nachvollziehbar (Welche Methode, welche Tuningparameter?)
- Nur nachvollziehbare statistische Modelle (z.B. kein NN)

Diskriminierung durch algorithmische Entscheidungssysteme

Wo kann es zu Diskriminierungen kommen?



Daten: Diskrimierung durch …

- Unvollständigkeit: betrifft oft Minderheiten, Frauen, Frauen aus Minderheiten [1,2,3].
- Vorhandener Bias in den Daten [6].
- Imbalanciertes Datenset bezüglich Volksgruppen [3,5].
- Diskriminierung durch **Eliminierung von** sensitiven Eigenschaften vor dem Lernen.

Datengrundlage:
Was kann man tun?

- Auf vorhandenen Bias untersuchen.
- Auf Vollständigkeit und Balanciertheit untersuchen.
  - Vervollständigen und Ausgleichen.
- Lernen mit und ohne sensitive Eigenschaften im Vergleich.

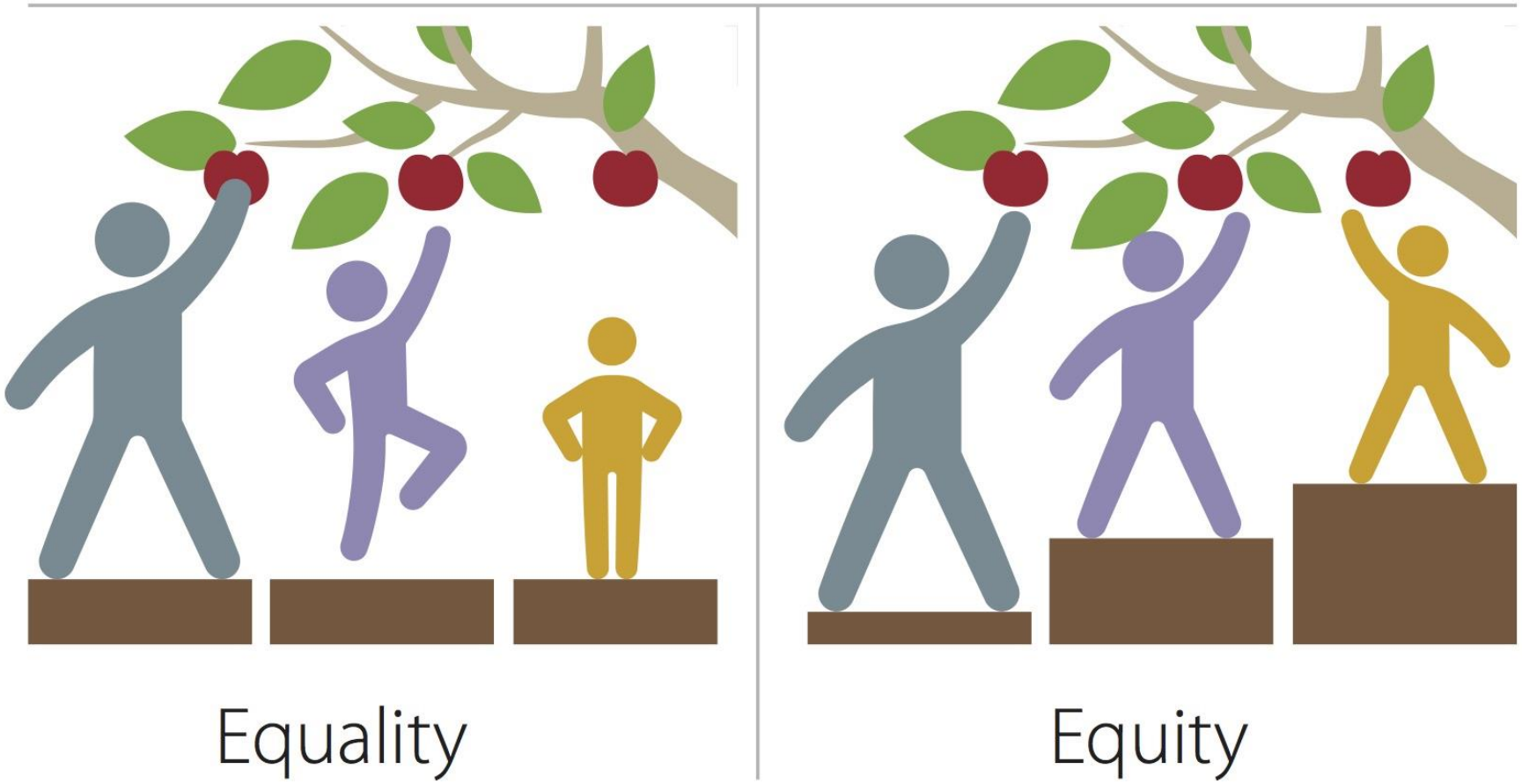
Diskriminierung durch Interpretation + Aktion: Ergebnis von Mensch und Maschine

- Vorherigen Bias durch Menschen untersuchen.
- Maschinelles Resultat auf Bias untersuchen.
- Mensch + Maschine auf Bias untersuchen.
- Differenzierte Abwägung leisten, wenn Maschine oder Mensch + Maschine bessere Urteile fällt.

Hinweis 1: American Civil Liberty Union

- ACLU fordert 2011 Risk Assessment in allen Phasen [7].
- Hofft auf objektivere Entscheidungen.
- Unterschreibt 2018 Forderung, dass pre-trial kein Risk Assessment mehr eingesetzt wird [8]:
  - Sieht mehr Nachteile als Vorteile.

Hinweis 2: Eine(r) verliert immer



Hinweis 3: Diskriminierung hängt von der genauen Umsetzung ab

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

- AMS-Algorithmus
- Teilt Arbeitslose in 3 Klassen ein:
  - Hohe Integrationschancen – keine weiteren Maßnahmen nötig.
  - Mittlere Integrationschancen – mit Maßnahmen
  - Niedrige Integrationschancen – Maßnahmen nicht sinnvoll.
- Weist Älteren (> 50), Frauen, Pflegenden höheres Risiko zu.
- **Diskriminierung?**
- **Kommt auf Umsetzung an!**

Transparenz

- Dokumentation des AMS-Algorithmus ist Blaupause für Transparenz:
- Daten
- Methode
- Qualitätstests:
  - Ca. 80% der Entscheidungen über die 1. und ca. 80% der Entscheidungen über die 3. Gruppe sind korrekt.

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

Sozialverträglichkeitsregeln hinter dem AMS-System

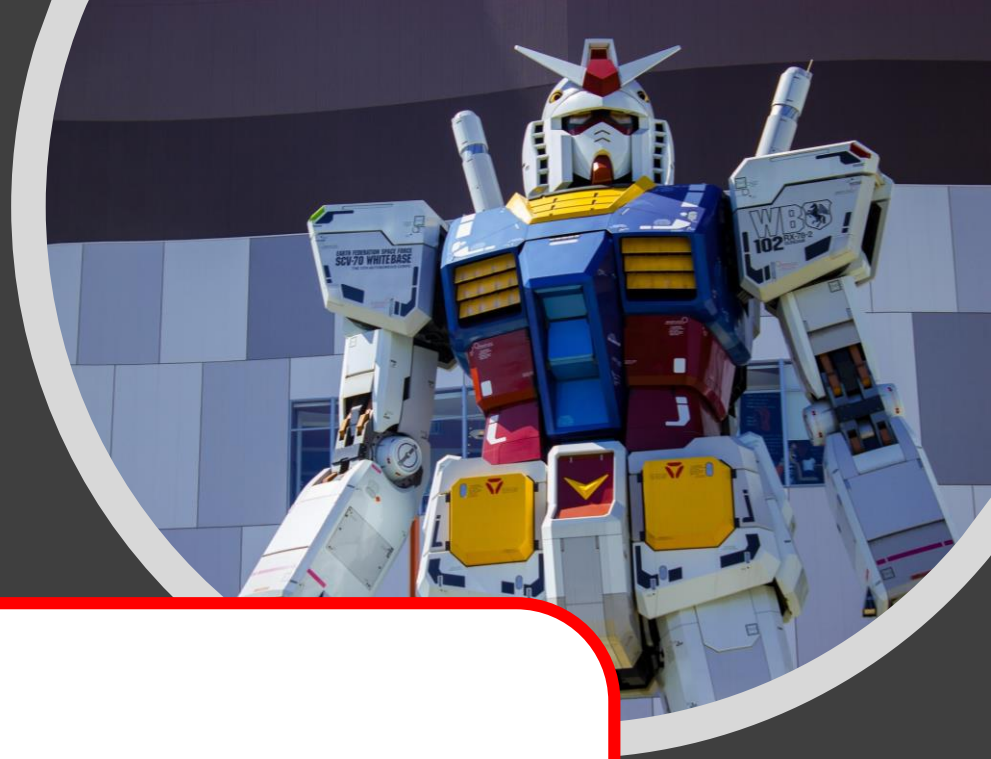
- Ergebnis des Algorithmus nur unterstützend einsetzen.
- Ergebnis muss im Dialog mit Betroffenen besprochen werden.
- Trainingsdaten nicht älter als 4 Jahre – das System muss vergessen können.

Wagner-Pinter: „Personenbezogene Wahrscheinlichkeitsaussagen (»Algorithmen«) - Stichworte zur Sozialverträglichkeit“, Technischer Report der SynthesisForschung, 2019

Qualität von ADM Systemen

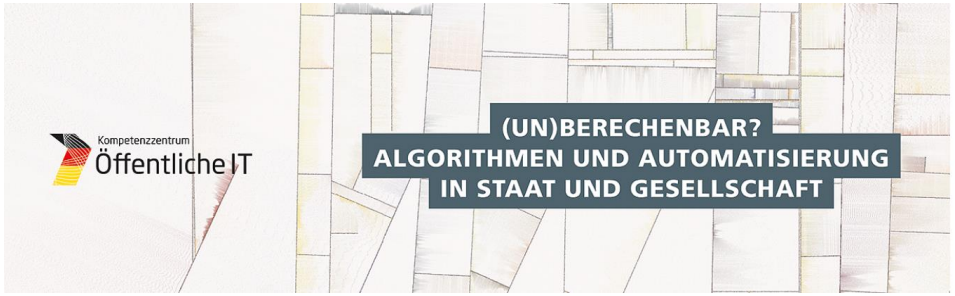
1. Wer entscheidet, wann ein ADM System „gut“ ist? Wer, wann es „fair“ ist?
2. ADM Systeme ergeben nur Wahrscheinlichkeiten, keine Wahrheiten.
3. ADM Systeme können diskriminieren.
4. ADM Systeme bedürfen einer sozio-informatischen Gesamtanalyse.

Die zwei Ängste



Keine Macht den dichtenden Robo-Richtern!

Eigene Arbeiten



[Zweig, Fischer & Lischka, 2018] Studie für die Bertelsmann-Stiftung: Zweig, Fischer & Lischka: „Wo Maschinen irren können“ (Serie AlgoEthik, No. 4, 2018)

Zwei Kapitel im Sammelband (Un)Berechenbar? des Fraunhofer FOKUS, Kompetenzzentrum ÖFIT, 2018

[Zweig & Krafft, 2018a] Zweig & Krafft: „Fairness und Qualität algorithmischer Entscheidungen“

[Krafft & Zweig, 2018b] Krafft & Zweig: „Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann“

[Zweig, 2019] Studie für die Konrad-Adenauer-Stiftung „Algorithmische Entscheidungen, Zweig: „Transparenz und Kontrolle von algorithmischen Entscheidungssystemen“, 2019, https://www.kas.de/analysen-und-argumente/detail/-/content/algorithmische-entscheidungen-transparenz-und-kontrolle

[Krafft & Zweig, 2019] Krafft & Zweig: „Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse“, Studie für die VZBV, 2019, https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-05-02_vzbv_positionspapier_algorithmenkontrolle.pdf

Ab Herbst:


Katharina
Zweig
EIN
ALGORITHMUS
HAT KEIN
TAKTGEFÜHL
Wo künstliche Intelligenz sich irrt,
warum uns das betrifft und
was wir dagegen tun können
HEYNE

Referenzen

(1) Criado-Perez: „Invisible Women“, Chatto & Windus, 2019

(2) Joy Buolamwini & Timnit Gebru: „Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification“, Proceedings of Machine Learning Research 81:1-15, 2018, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

(3) Project Gendershades: https://www.media.mit.edu/projects/gender-shades/overview/

(4) Joy Buolamwini: „Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers”, Master Thesis, MIT, 2017

(5) Evaluierung eines binären Klassifikators, https://de.wikipedia.org/wiki/Beurteilung_eines_bin%C3%A4ren_Klassifikators#Wahrheitsmatrix:_Richtige_und_falsche_Klassifikationen

(6) Jeffrey Dastin: „Amazon scraps secret AI recruiting tool that showed bias against women”, Reuters, 10.10.2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

(7) American Civil Liberty Union: "Smart Reform is Possible", August 2011, https://www.aclu.org/files/assets/smartreformispossible.pdf, zuletzt abgerufen am 30.5.2019

(8) https://civilrights.org/2018/07/30/more-than-100-civil-rights-digital-justice-and-community-based-organizations-raise-concerns-about-pretrial-risk-assessment/, zuletzt abgerufen am 20.4.2019.

(9) Holl, Kernbeiß, Wagner-Pinter: „Personenbezogene Wahrscheinlichkeitsaussagen (»Algorithmen«) - Stichworte zur Sozialverträglichkeit, 2019

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