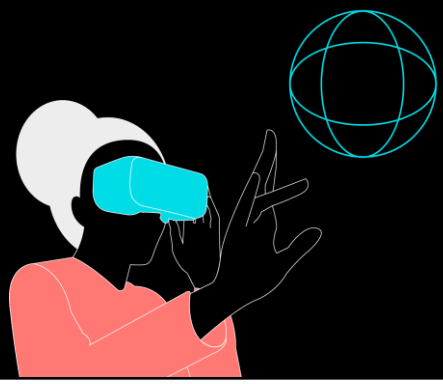
Fairness bei algorithmischen Entscheidungsprozessen



TÜV AI.LAB

Marc Hauer
Senior Solutions Architect
marc@tuev-lab.ai

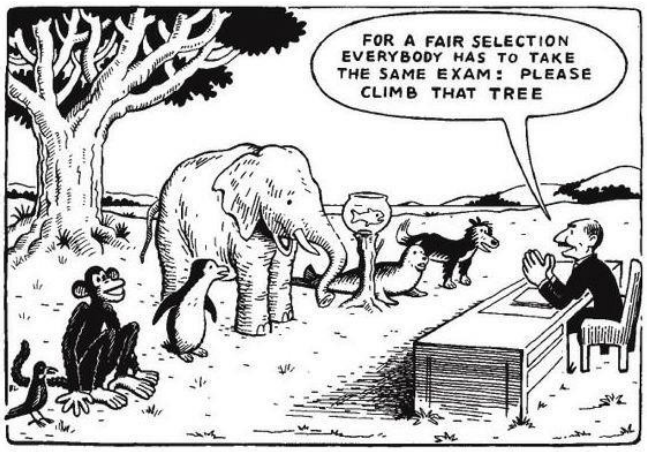
https://www.linkedin.com/in/marc-hauer-ai-lab/

https://scholar.google.de/citations?user=vsf4PHAAAAAJ&hl=de

https://www.trusted-ai.com/

https://www.tuev-lab.ai/

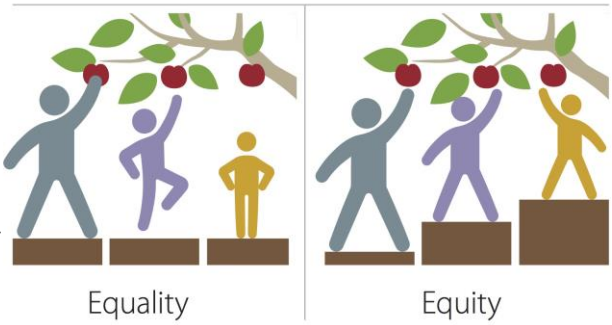
Definitionen von Fairness



2

Herausforderungen für „faire“ KI

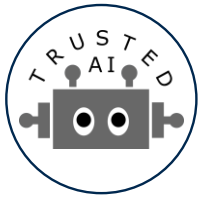
1. Was ist Fairness?
2. Wie geht man mit einander widersprechenden Fairnessvorstellungen um?
3. Wie geht man mit Gruppen ungleicher Größe um?
4. Berechnet man Maße mit künstlichen Daten einer gewünschten Verteilung oder mit echten, ungleich verteilten Daten?
5. Uvm…



TÜV AI.LAB

Wie lernt ein System von Daten?

DIY:
Sie sind heute eine
„Support Vector Machine“



4

Beispiel von der Trusted AI GmbH (mit Erlaubnis)

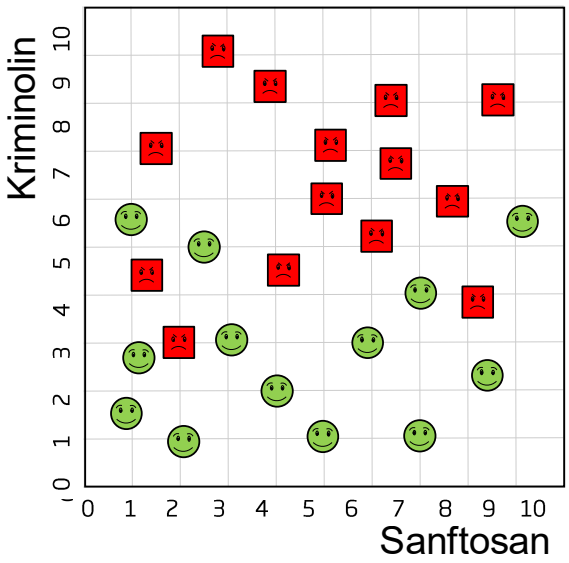
Bösartige Kriminelle

Unschuldige Bürger

Zeichnen Sie eine Linie so zwischen die Smileys, dass die roten möglichst gut von den grünen getrennt sind. Kleben Sie ihn fest.

Gratulation: Sie haben eine Support Vector Machine trainiert!

Der Holzspieß dient nun als Entscheidungsregel, ob eine Person als kriminell gilt oder unschuldig zu sein scheint.

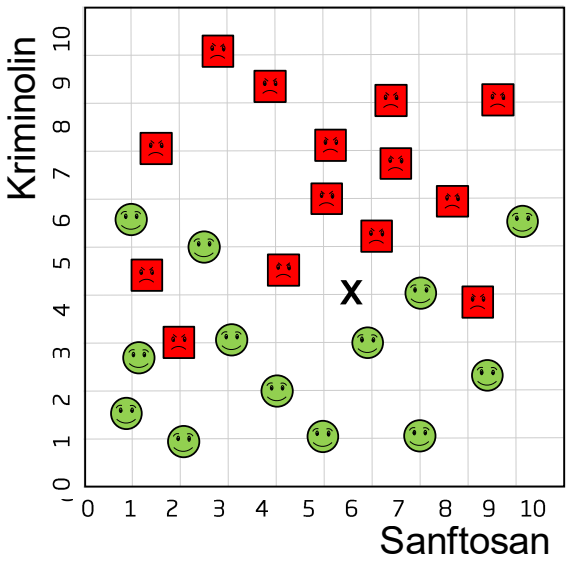


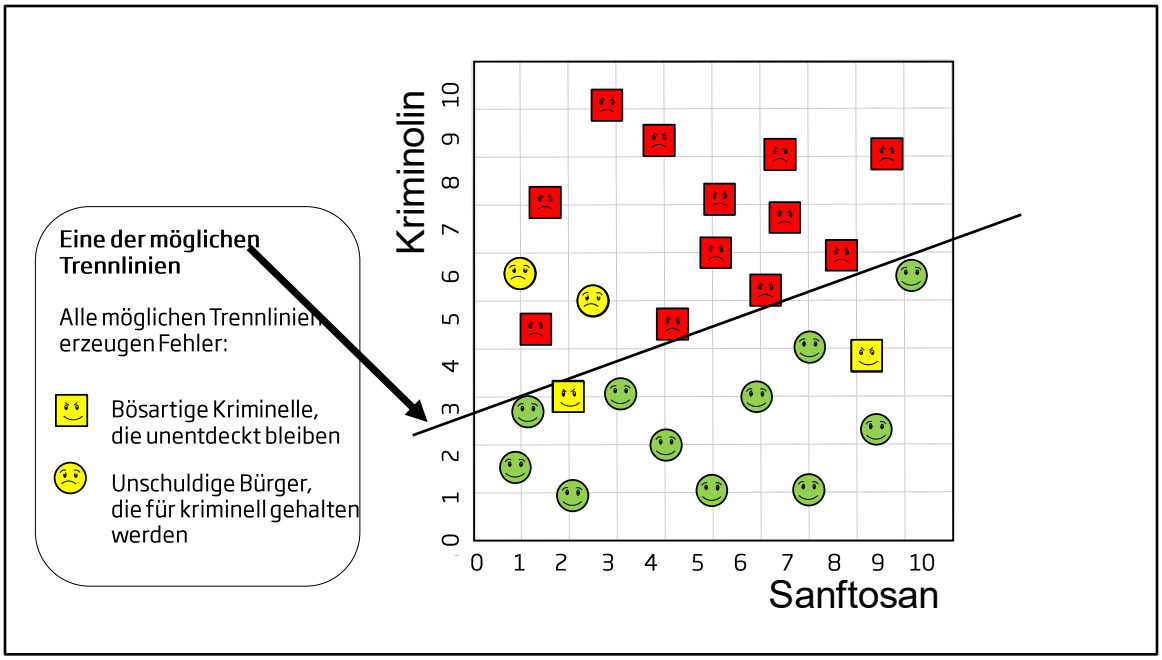
Bösartige Kriminelle

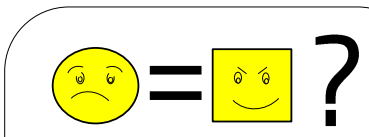
Unschuldige Bürger

Bewerten Sie Frau Müller:

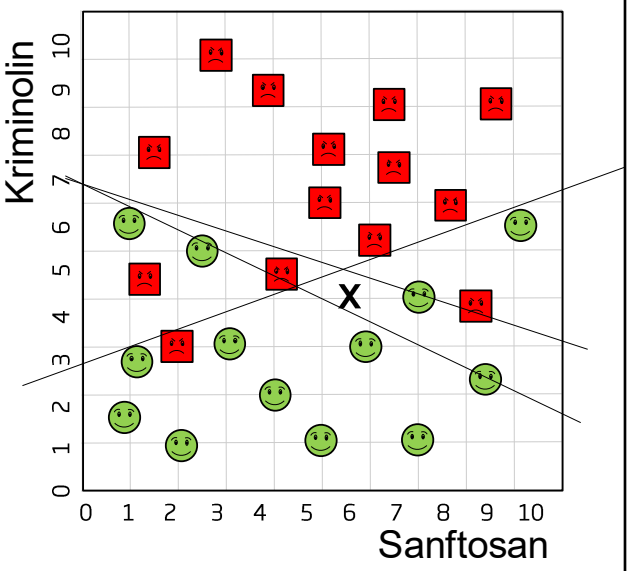
5.5 Sanftosan
4.0 Kriminolin

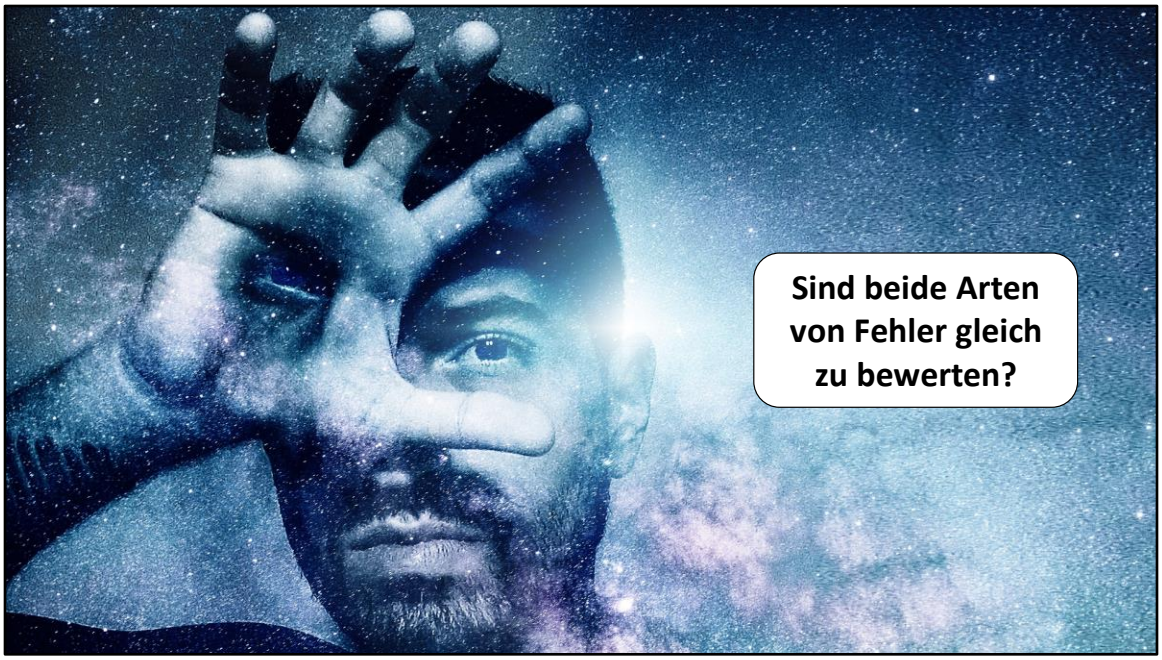


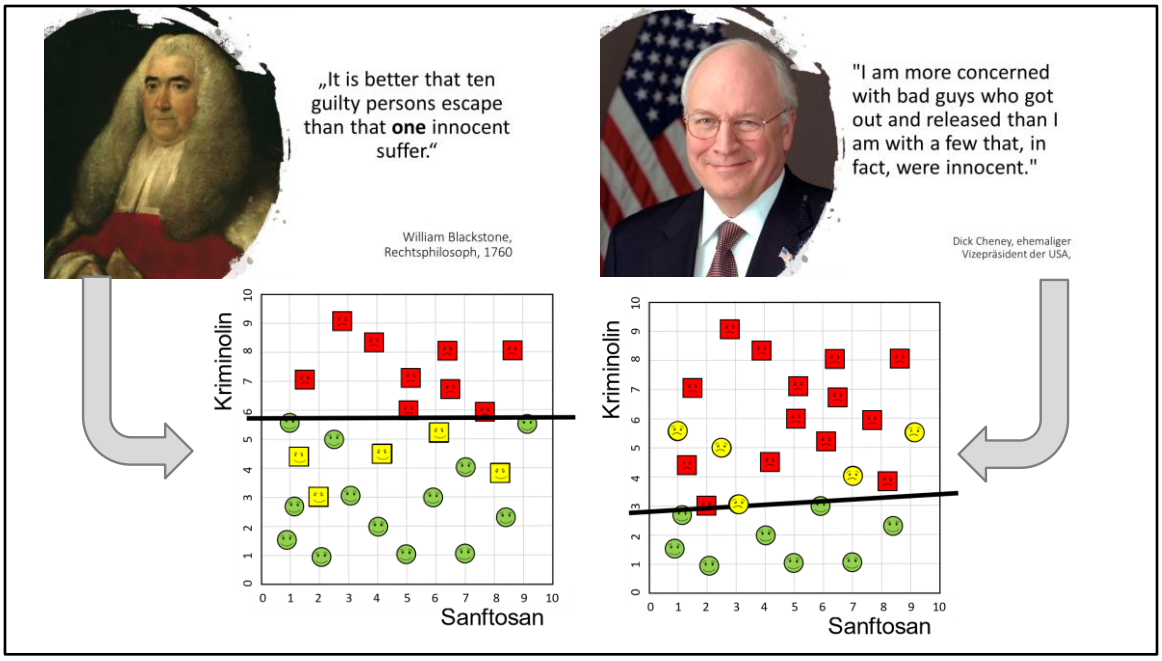

Eine der möglichen Trennlinien
Alle möglichen Trennlinien erzeugen Fehler:
Bösartige Kriminelle, die unentdeckt bleiben
Unschuldige Bürger, die für kriminell gehalten werden
Kriminolin
0 1 2 3 4 5 6 7 8 9 10
Sanftosan
0 1 2 3 4 5 6 7 8 9 10

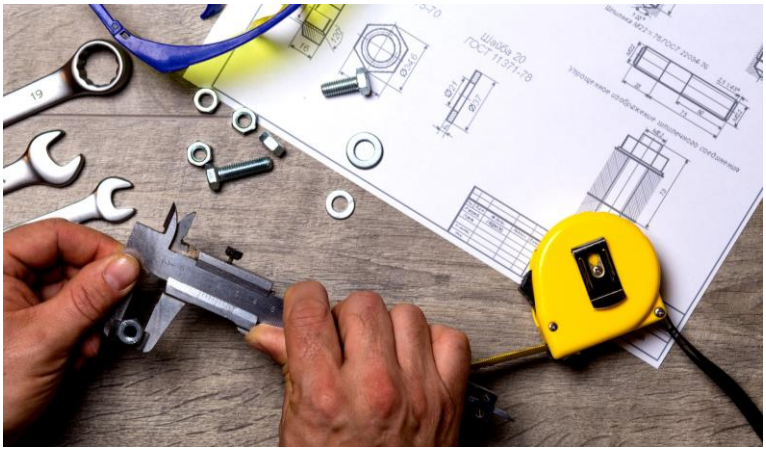


Wenn beide Fehler als gleich schlimm gelten, gibt es mehrere optimale Trennlinien mit möglichst wenigen Fehlern.




Sind beide Arten von Fehler 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,
Kriminolin
Sanftosan
Kriminolin
Sanftosan



Qualitätsmaße

- Sensitivität
- Spezifität
- Genauigkeit
  (accuracy)
- Es gibt über 25 weitere

1. Beobachtung

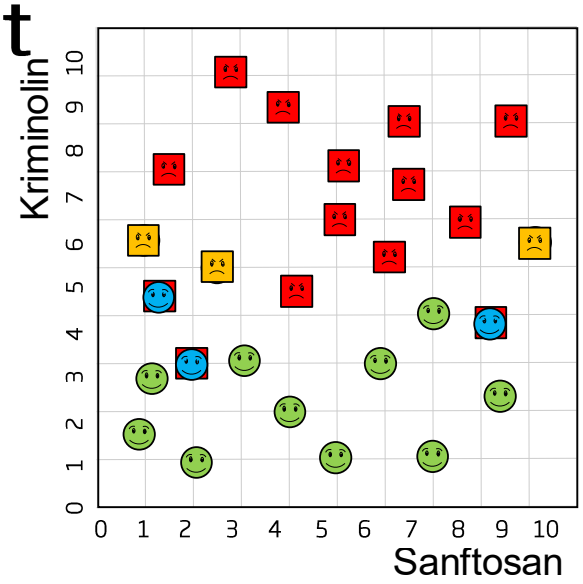
Was durch eine künstliche Intelligenz
optimiert werden soll,
ist eine gesellschaftliche Entscheidung!

Datenqualität

Noch nicht entdeckte Steuerbetrüger

Unschuldig im Gefängnis

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



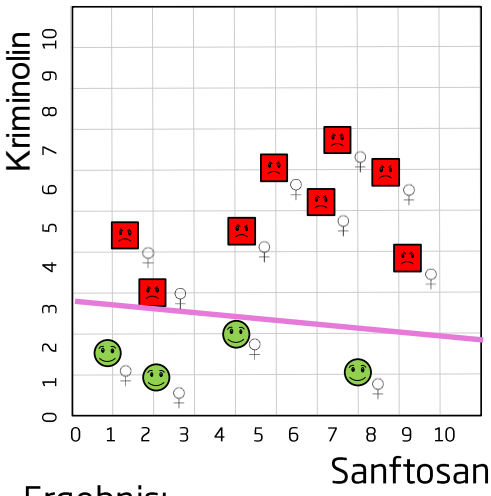
2. Beobachtung

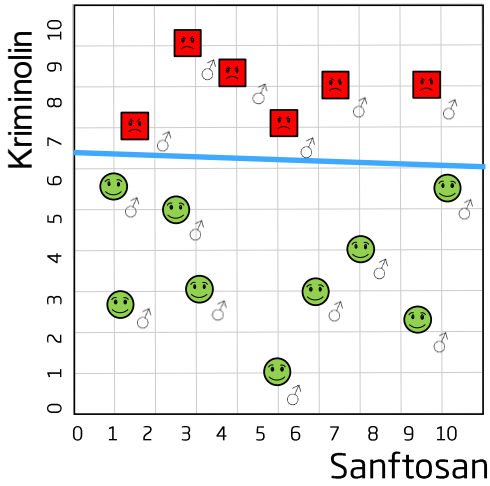
Wie gut die Maschine lernt, ist direkt abhängig von der Qualität der Daten.

Diskriminierung bei Apple Pay?

- https://twitter.com/dhh/status/1192540900393705474
- Der Autor ist ein Softwareentwickler
- Tatsächlich wurde der Fall offiziell untersucht (Neil Vigdor: „Apple Card Investigated After Gender Discrimination Complaints“, NYT, 10.11.2019, https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html (Paywall)).
- Die Ungleichbehandlung war sachlich gerechtfertigt!

Geschönt übersetzt sagt er:
Hinter der Apple Pay Karte steckt ein sexistisches Programm. Meine Frau und ich sind steuerlich gemeinsam veranlagt. Wir leben zusammen und sind schon lange verheiratet. Warum erhalte ich durch euren Algorithmus einen 20 mal höheren Kreditrahmen?

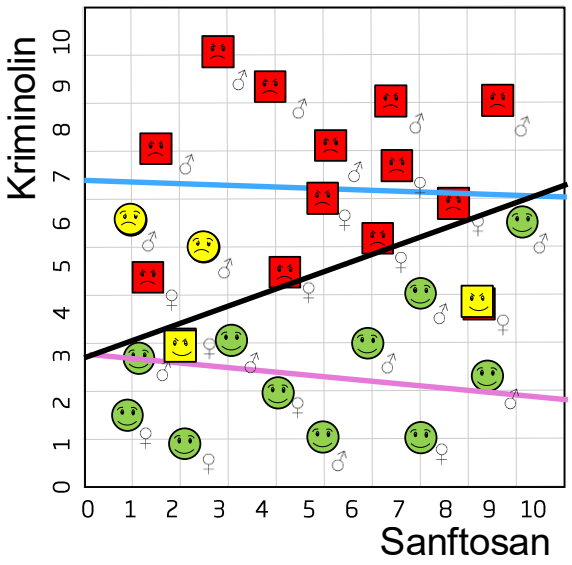




Ergebnis:
In diesem fiktiven Beispiel wird für jede Teilgruppe eine optimale Entscheidungsregel ohne Fehler gefunden.

Legt man dagegen beide Gruppen zusammen, diskriminiert die trainierte Support Vector Machine **Männer:**

Zwei weibliche Kriminelle gelten als unschuldig, zwei unschuldige männliche Bürger als kriminell.



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.

Ist es besser, die sensitive Information wegzulassen?

Diskriminierung messen

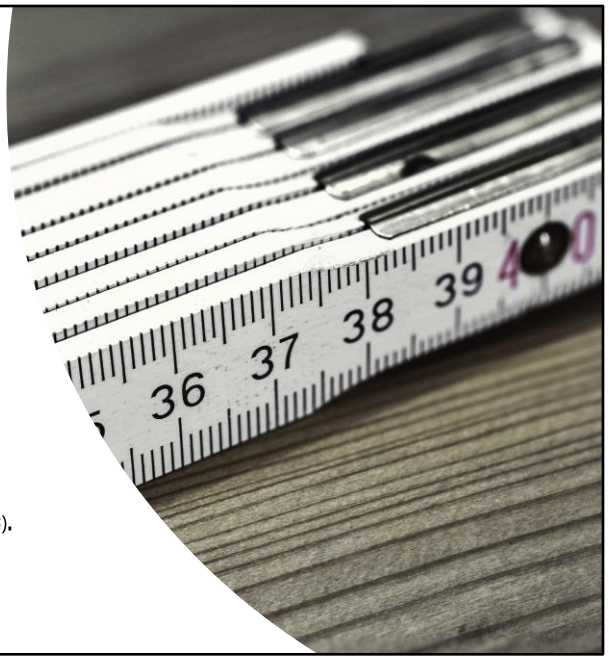
Qualitätsmaß(e) wählen, z.B. Falsch-Positiv-Rate

(Statistische) Gleichheit der Teilgruppen im Qualitätsmaß fordern.

- Buolamwini: Teilgruppe sollte mind. 80% des maximalen Wertes haben (Buolamwini, 2017, S.49).

Vorsicht: Manche Fairnessmaße widersprechen einander (Zweig & Krafft, 2018).

- Gesellschaft muss bestimmen.



Kurzfassung Leistungsmetriken

- Klassifikation
  - Auf Basis der Grundwahrheit (Konfusionsmatrix)

TÜV AI.LAB

Definition 12 (Confusion Matrix)

The confusion matrix is a table used to evaluate the quality of a decision-making system. It contains four entries: The number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) (see Figure 2.10). These entries can be retrieved from a ground truth (see Definition 3, p. 13). The values in the matrix can be used to compute various quality measures (see Table 2.2).

P+N	PP	PN	
P	TP	FN	TPR = TP/P FNR = FN/P
N	FP	TN	FPR = FP/N TNR = TN/N
	PPV = TP/PP FDR = FP/PP	FOR = FN/PN NPV = TN/PN	

Figure 2.10: *P* is the overall number of positive instances ($TP + FN$). *N* is the overall number of negative instances ($FP + TN$). *PP* is the overall number of as positive classified instances ($TP + FP$). *PN* is the overall number of as negative classified instance ($FN + TN$).

Auszug Dissertation

Kurzfassung Leistungsmetriken

- Klassifikation
  - Auf Basis der Grundwahrheit (Konfusionsmatrix)
  - Qualität

Quality measure	Formula
Accuracy (ACC)	$\frac{TP+TN}{TP+TN+FP+FN}$
True positive rate (TPR), Recall or sensitivity	$\frac{TP}{TP+FN}$
True negative rate (TNR) or specifity	$\frac{TN}{TN+FP}$
False positive rate (FPR) or fall-out	$\frac{FP}{FP+TN}$
False negative rate (FNR) or miss rate	$\frac{FN}{FN+TP}$
Positive predictive value (PPV) or precision	$\frac{TP}{TP+FP}$
False discovery rate (FDR)	$\frac{FP}{FP+TP}$
Negative Predictive Value (NPV)	$\frac{TN}{TN+FN}$
False omission rate (FOR)	$\frac{FN}{FN+TN}$
Matthews correlation coefficient (MCC)	$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
F1-Score	$2 * \frac{PPV * TPR}{PPV + TPR}$
ROC-AUC	komplizierter

TÜV AI.LAB

22

Receiver operating characteristic – area under the curve
Fläche unter der Kurve der Betriebskennlinier des Beobachters

Kurzfassung Leistungsmetriken

- Klassifikation
  - Auf Basis der Grundwahrheit (Konfusionsmatrix)
  - Qualität
  - Fairness

Fairness measure	Requires equality of
Overall accuracy equality	ACC
Separation Conditional procedure accuracy Equalized odds	TPR and FPR
Equal opportunity	TPR
Error rate balance	FPR and FNR
Sufficiency	PPV and FOR
Conditional use accuracy	PPV and NPV

TÜV AI.LAB

Problem → Lösung?

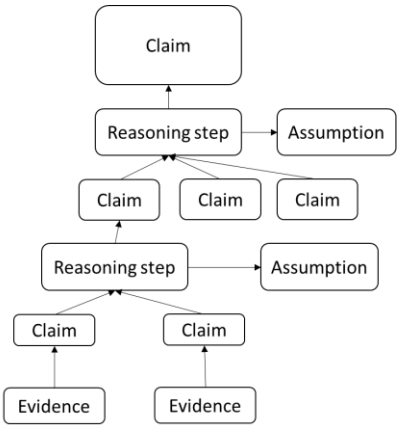
Aktuelle Lösungsansätze:

- Normung
  - DIN SPEC 91512 – Fairness von KI in Finanzdienstleistungen
- Forschung
  - „Aufräumen" des Feldes
- Accountability Mechanismen
  - Persönliche Verantwortung zuschreiben/übernehmen
- Monitoring
- Assurance Cases

TÜV AI.LAB

Was sind Assurance Cases?

- Begründete Argumentation, dass eine Behauptung (Claim) wahr ist, unterstützt durch Beweise (Evidence)
- Unterteilung durch erklärte Argumente (Reasoning)
- Es können kontextuelle Informationen angehängt werden, z.B. Annahmen (Assumptions)



TÜV AI.LAB

25

Ganz abstrakt ist ein Assurance Case eine begründete Argumentation die durch eine Reihe von Beweisen gestützt wird und aussagt, dass ein System für eine bestimmte Anwendung in einer bestimmten Umgebung funktioniert wie vorgesehen.

Sie werden bereits stark und weiter zunehmend zur Absicherung sicherheitskritischer ADM-Systemen, wie z. B. in autonomen Fahrzeugen oder der Avionik eingesetzt, aber die Technik ist auch in der Philosophie gängig um komplexe Überlegungen zu strukturieren. Die Hauptaufgabe eines Assurance Case ist es, zu begründen, warum und unter welchen Annahmen ein Beweis eine Behauptung impliziert. Dazu wird die Hauptbehauptung (main-claim) in Teilbehauptungen (sub-claims) zerlegt, die entweder ebenfalls auf der Erfüllung von hierarchisch strukturierten Teilbehauptungen beruhen oder direkt aus Beweisen abgeleitet werden können.

Jede Zerlegung einer Behauptung wird durch eine Argumentation explizit gemacht, die die Idee hinter einer Zerlegung erklärt.

Außerdem werden alle relevanten Annahmen für die Schlussfolgerung, dass die Sub-Claims den Main-Claim implizieren, explizit gemacht und mit dem Argument verbunden. Um das Verständnis eines Arguments zu erleichtern, können weitere

kontextuelle Informationen ergänzt werden.

Also zusammenfasst: das Ziel eines Assurance Case ist es, ein Argumentationsframework bereitzustellen, in dem die Aussage, dass die Beweise die Behauptung unter den gegebenen Annahmen unterstützen, begründet, verstanden und dokumentiert werden kann.

Rotable technologies hat eine Plattform zur Verwaltung von Rotationen von Medizinstudenten entwickelt.

Das bedeutet, dass jeder Medizinstudent seine praktische Ausbildung mindestens für einen festgelegten Zeitraum in mehreren medizinischen Bereichen absolvieren muss und daher für bestimmte medizinische Abteilungen in mehreren Krankenhäusern eingeplant wird. rotable technologies bietet eine softwarebasierte Lösung zur Verbesserung der Erstellung und Verwaltung eines solchen Rotationsplans.

Um die Lösung weiter zu verbessern, haben sie eine KI-basierte Methode entwickelt, um die Rotationsplanung zu automatisieren. Da sie einen starken Stakeholder-Fokus aufweisen müssen, um die notwendige Akzeptanz zu gewährleisten, haben sie beschlossen, einen Assurance Case zur Fairness ihres Produkts zu erstellen.

Dafür haben sie im Rahmen des Industrieprojektes fAIr by design mit den Firmen winnovation consulting und Rania Wazir kooperiert. Wir vom Algorithm Accountability Lab der RPTU Kaiserslautern wurden als externe Experten zur Mitarbeit eingeladen, weil wir die prinzipielle Idee, ACs für nicht-funktionale Anforderungen wie fairness zu verwenden, zwei Jahre zuvor veröffentlicht haben.

Als wir mit dem AC angefangen haben, befand sich das Upgrade bereits in der Entwicklung, daher war der AC primär für einen Continuous Deployment Prozess und die Stakeholderkommunikation gedacht.

Da das Upgrade nach dem bevorstehenden AI-Act höchstwahrscheinlich als Hochrisiko-Anwendung eingestuft wird, wird der AC potenziell aber auch für künftige Audit- oder Zertifizierungsprozesse verwendet werden.

Ergebnisse und Feedback

~20 Argumentationsschritte

~70 Sub-Claims

~80 geforderte Evidenzen, davon ~30 einzigartige

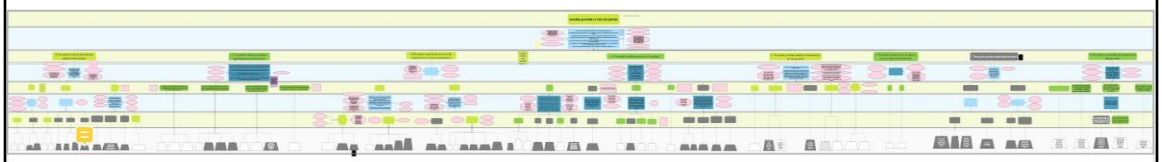
~40-45 Stunden Aufwand pro Teilnehmer (~300h in Summe)

Half bei der Identifizierung mehrerer fehlender Funktionen und wichtiger Tests: z.B: Software ermöglicht Vergleich von KPIs mehrerer Pläne.

Half einem Rechtsexperten bei der Reflektion über potentielle Probleme

Hoher Zeitkonsum (es gab zuvor weder ein Verfahren noch praktische Erfahrungen)

→ Signifikante Verbesserungen möglich



Unten sehen Sie den endgültigen Assurance Case zum Zeitpunkt der Einreichung der Veröffentlichung. Aus Gründen der Vertraulichkeit darf ich nicht viel mehr ins Detail gehen, was wohl der größte Nachteil unserer Arbeit ist, aber es lässt sich nicht ändern.

Insgesamt umfasste der AC:
~20 Argumentationsschritte
~70 Sub-Claims
~80 erforderliche Nachweise, ~30 davon einzigartig
~40-45 Stunden Arbeit pro Teilnehmer

Es stellte sich auch heraus, dass er viel zu viel Zeit in Anspruch nahm. Allerdings haben wir bei null angefangen, ohne jeglichen Prozess, was uns ebenfalls viel Zeit gekostet hat.

Zusammenfassung

Assurance Cases können die prinzipiellen Probleme in der Fairness/Ethik-Debatte nicht lösen, aber sie…

→… verlangen, dass alle Stakeholderperspektiven berücksichtigt und diskutiert werden.

→… stellen eine Dokumentation der Argumentation bereit (z.B., für externe Reviews).

→… skizzieren automatisierbare Tests und händisch überprüfbare Artefakte.

→… bieten Langzeitschutz gegen ungewollte Änderungen und Fehler (z.B. durch kontinuierliches Lernen oder Updates).

→… bieten eine wiederverwertbare Argumentationsgrundlage für zukünftige Produkte.

TÜV AI.LAB

28

Der vorgeschlagene Ansatz ist kein deterministischer Prozess und wird daher nicht zu einem eindeutigen Ergebnis führen. Er kann auch nicht das Haupt- und Grundsatzproblem lösen, wie man Fairness oder andere ethische Anforderungen in einer quantifizierten und allgemein akzeptierten Weise definiert. Nichtsdestotrotz wird ein pragmatischer Ansatz beschrieben, um zu einer gut dokumentierten Argumentation darüber zu gelangen, wann und unter welchen Annahmen ein System als "fair genug" angesehen wird, um es zu nutzen.

Indem die Argumentation als Assurance Case modelliert wird, ist sie nachvollziehbar dokumentiert und kann für eine externe Überprüfung oder ein Audit, z.B. im Falle eines Rechtsstreits oder für einen Zertifizierungsprozess, leicht offengelegt werden.

Durch die Dokumentation der Argumentation und die automatisierte Bereitstellung der Nachweise bietet dieser Prozess das Potential, einen langfristigen Schutz vor ungewollten Änderungen, z.B. durch einen kontinuierlichen KI-Lernprozess oder Fehler bei der Änderung des Codes, zu bieten.

Zudem kann dieselbe Argumentationsbasis auf ähnliche Anwendungen übertragen werden und bietet somit einen guten Ausgangspunkt für die Sicherstellung der

Erfüllung einer bestimmten Anforderung. Außerdem können im Laufe der Zeit Best Practices entwickelt werden, um die Qualität einer Gruppe von KI-basierten Anwendungen zu verbessern.

Zumindest ist das in manchen Safety Kontexten so passiert, wo die Anwendung dieses Frameworks inzwischen ein De-facto-Standard ist, zum Beispiel in der Luft- und Raumfahrt, so dass ich wirklich hoffe, dass das auch bei den ACs für ethische Anforderungen passieren wird. Das setzt aber voraus, das mehr mit dem Framework experimentiert wird und die Erfahrungen damit geteilt werden. Und da kommen vielleicht auch Sie ins Spiel. Vielen Dank!


DISKUSSIONSRUNDE