KI-FACHKONFERENZ

Artificial Intelligence Act

Dr. Wolfgang Hildesheim
Dr. Thomas Schmid

Breakout „Bias, Robustheit & Fairness“

Berlin, 22.11.2021 | 14.15-15.00 Uhr

European Commission - Press release

Europe fit for the Digital Age: Commission proposes new rules and actions for excellence and trust in Artificial Intelligence

Brussels, 21 April 2021

The Commission proposes today new rules and actions aiming to turn Europe into the global hub for trustworthy Artificial Intelligence (AI). The combination of the first-ever legal framework on AI and a new Coordinated Plan with Member States will guarantee the safety and fundamental rights of people and businesses, while strengthening AI uptake, investment and innovation across the EU. New rules on Machinery will complement this approach by adapting safety rules to increase users' trust in the new, versatile generation of products.

Margrethe **Vestager**, Executive Vice-President for a Europe fit for the Digital Age, said: *"On Artificial Intelligence, trust is a must, not a nice to have. With these landmark rules, the EU is spearheading the development of new global norms to make sure AI can be trusted. By setting the standards, we can pave the way to ethical technology worldwide and ensure that the EU remains competitive along the way. Future-proof and innovation-friendly, our rules will intervene where strictly needed: when the safety and fundamental rights of EU citizens are at stake."*

The European approach to trustworthy AI

The new rules will be applied directly in the same way across all Member States based on a future-proof definition of AI. They follow a risk-based approach:

Unacceptable risk: AI systems considered a clear threat to the safety, livelihoods and rights of people **will be banned**. This includes AI systems or applications that manipulate human behaviour to circumvent users' free will (e.g. toys using voice assistance encouraging dangerous behaviour of minors) and systems that allow 'social scoring' by governments.

High-risk: AI systems identified as high-risk include AI technology used in:

- **Critical infrastructures** (e.g. transport), that could put the life and health of citizens at risk;
- **Educational or vocational training**, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);
- **Safety components of products** (e.g. AI application in robot-assisted surgery);
- **Employment, workers management and access to self-employment** (e.g. CV-sorting software for recruitment procedures);
- **Essential private and public services** (e.g. credit scoring denying citizens opportunity to obtain a loan);
- **Law enforcement** that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence);
- **Migration, asylum and border control management** (e.g. verification of authenticity of travel documents);
- **Administration of justice** and **democratic processes** (e.g. applying the law to a concrete

set of facts).

High-risk AI systems will be subject to **strict obligations** before they can be put on the market:

- **Adequate risk assessment and mitigation systems;**
- **High quality of the datasets** feeding the system to minimise risks and discriminatory outcomes;
- **Logging of activity to ensure traceability of results**;
- **Detailed documentation** providing all information necessary on the system and its purpose for authorities to assess its compliance;
- **Clear and adequate information** to the user;
- **Appropriate human oversight** measures to minimise risk;
- High level of **robustness**, **security** and **accuracy**.

In particular, **all remote biometric identification** systems are considered high risk and subject to strict requirements. Their live use in publicly accessible spaces for law enforcement purposes is prohibited in principle. Narrow exceptions are strictly defined and regulated (such as where strictly necessary to search for a missing child, to prevent a specific and imminent terrorist threat or to detect, locate, identify or prosecute a perpetrator or suspect of a serious criminal offence). Such use is subject to authorisation by a judicial or other independent body and to appropriate limits in time, geographic reach and the data bases searched.

Limited risk, i.e. AI systems with specific transparency obligations: When using AI systems such as chatbots, users should be aware that they are interacting with a machine so they can take an informed decision to continue or step back.

Packaging by SIG

Durchschnittliche Nährwerte

	Je 100 ml	1 Glas (250 ml)**	% (250 ml)*
Brennwert	198 kJ 47 kcal	495 kJ 118 kcal	6 %
Fett	1,5 g	3,8 g	5 %
- davon gesättigte Fettsäuren	1,1 g	2,8 g	14 %
Kohlenhydrate	5,0 g	13 g	5 %
- davon Zucker	5,0 g	13 g	14 %
Eiweiß	3,4 g	8,5 g	17 %
Salz	0,11 g	0,28 g	5 %
Calcium	120 mg (15 %***)	300 mg (38 %***)	

***NRV = Nährstoffbezugswert

*Referenzmenge für einen durchschnittlichen Erwachsenen (8.400 kJ / 2.000 kcal).
**1 Portion = 1 Glas (250 ml) Bio-H-Fettarme Milch. Die Packung enthält 4 Portionen.

75 % Karton aus nachwachsenden Rohstoffen
+
25 % pflanzenbasierte Kunststoffe mit Mengenausgleich[1]
+
Verzicht auf Alu bei gleichbleibender Produktqualität

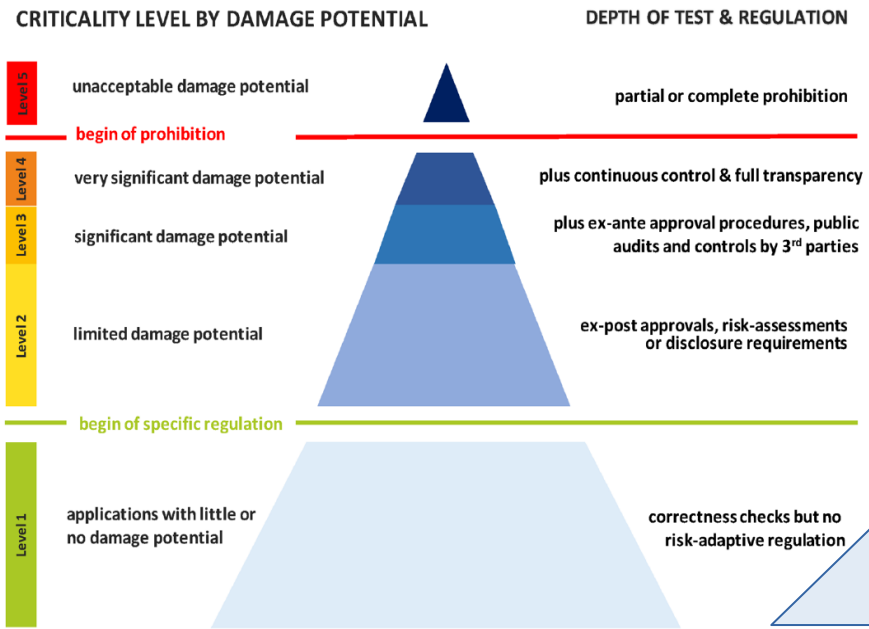
[1]Im Rahmen der Verpackungsherstellung kommen Kunststoffe zum Einsatz, bei deren Produktion im Rahmen von zertifizierten Massenbilanzverfahren fossile Ressourcen durch pflanzliche Rohstoffe ersetzt wurden. Diese Maßnahme trägt maßgeblich zu einer verbesserten Ökobilanz (CB-100732C vom

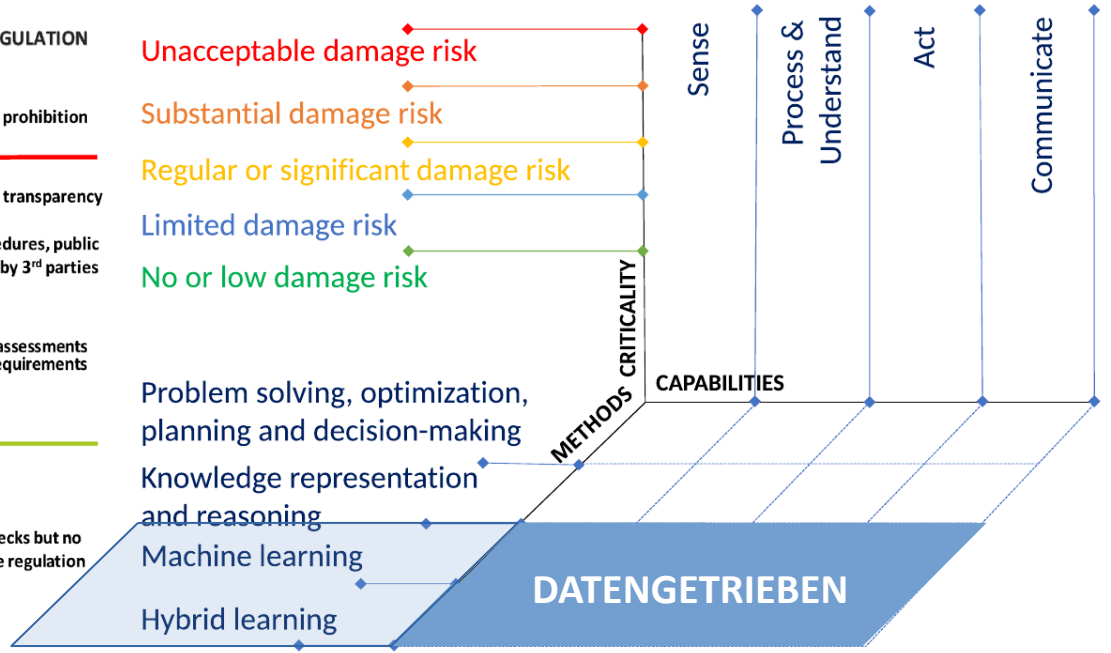
Durchschnittliche Nährwerte	pro 100 ml	RM* pro 100 ml
Energie	547 kJ 129 kcal	7 %
Fett	0,1 g	<1 %
- davon gesättigte Fettsäuren	<0,1 g	<1 %
Kohlenhydrate	30 g	12 %
- davon Zucker	25 g	28 %
Ballaststoffe	1,0 g	
Eiweiß	0,9 g	2 %
Salz	2,6 g	43 %

*RM: Referenzmenge für einen durchschnittlichen Erwachsenen (8400 kJ / 2000 kcal)

LAKTOSE FREI
GLUTEN FREI

High-Risk & Kritikalität – Concept Paper: https://link.springer.com/article/10.1007/s13218-021-00736-4

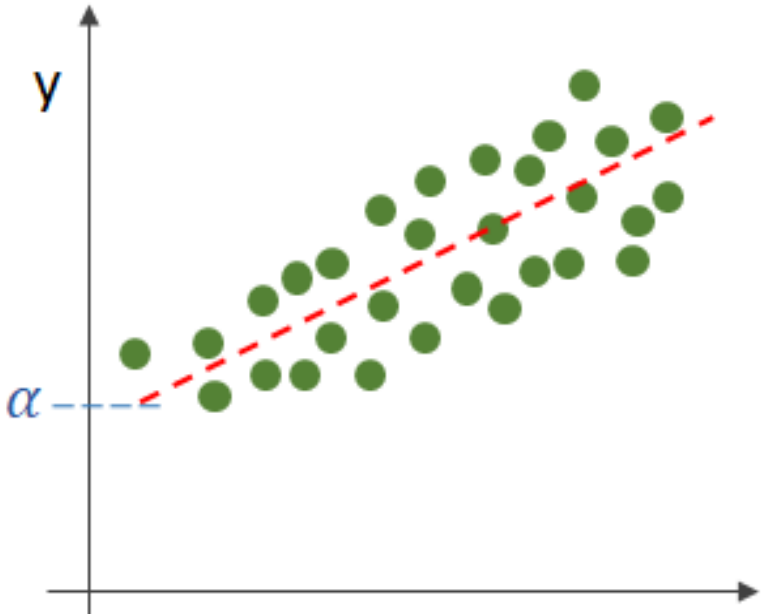


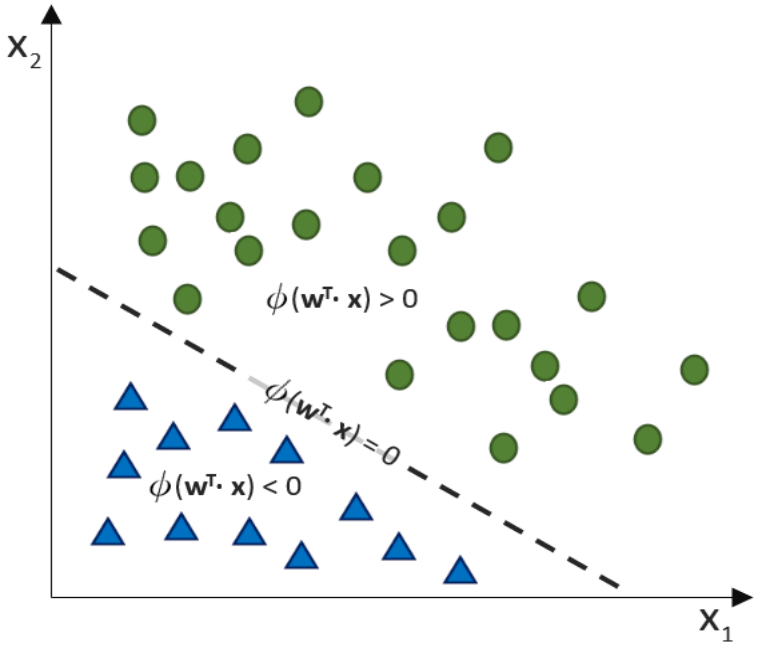


Klassifikation & Bias

Quelle: ...

Regression und Klassifikation

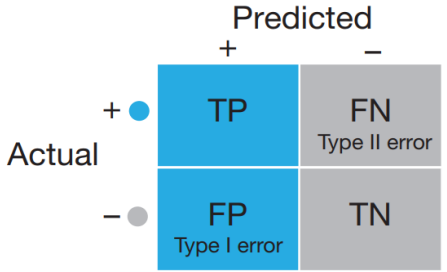




Metriken für Klassifikation | Accuracy

- Anteil korrekt kategorisierter Elemente an Grundgesamtheit
- → eindimensionales Fehlermaß (Performance of classifier)
- Vorteile
  - einfache Interpretierbarkeit
  - für Multi-Class-Probleme erweiterbar
- Nachteile
  - keine Aussage über Verhältnis von TP, TN, FN, FP
  - irreführend bei unbalancierten Daten
  - Accuracy-Paradoxon

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$



Metriken für Klassifikation | Sensitivität & Spezifität

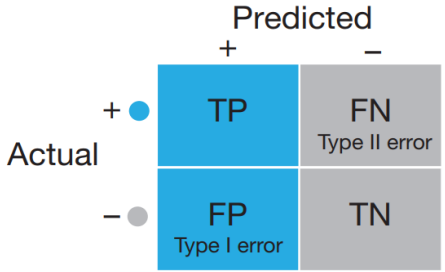
- Sensitivity = True positive rate (TPR), Recall

$$Sensitivity = \frac{TP}{TP + FN}$$

- Specificity = True negative rate (TNR)

$$Specificity = \frac{TN}{TN + FP}$$

- Vorteile
  - umfassenderes Bild als eindimensionale Fehlermaße
- Nachteile
  - nur auf binäre Klassifikatoren anwendbar



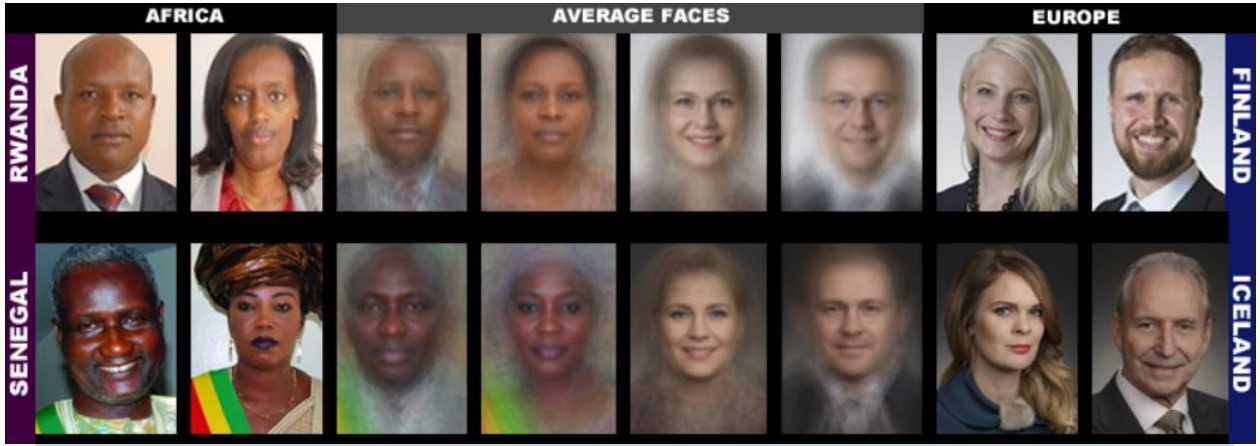
Bias & Fairness

Quelle: ...

Benachteiligende Gesichtserkennung

- Ungleichverteilte Trainingsdaten (Bias) → unfaire Funktionsweise des Systems
- MIT-Studie zu Gesichtserkennung (2018)
  - Gesichter von Frauen werden schlechter erkannt als von Männern (8-20%)
- kein Einzelfall, sondern Grundsatzproblem
- mehrere etablierte Erkennungssysteme betroffen (u.a Microsoft, Face++)

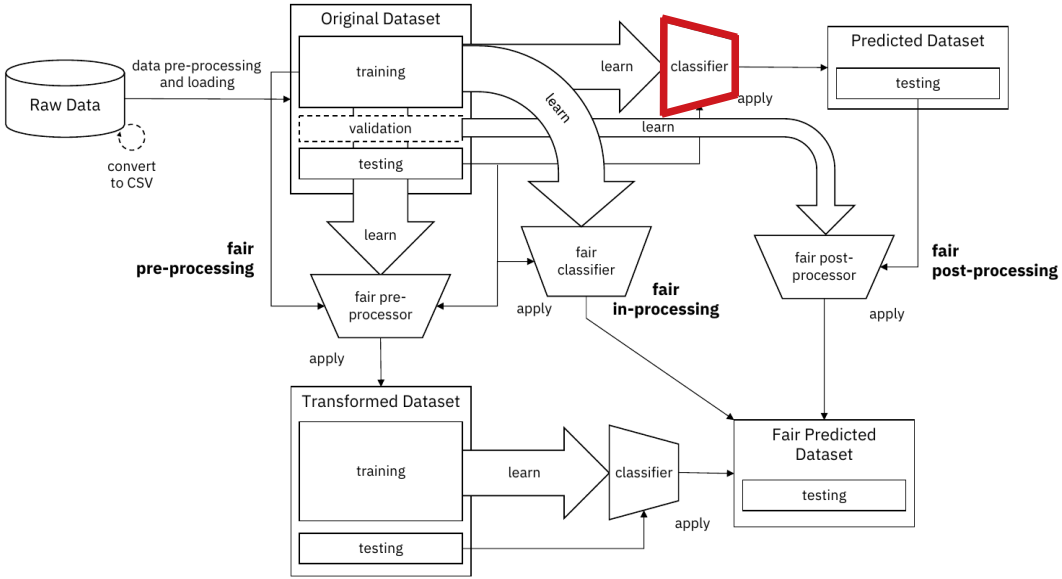
Benachteiligende Gesichtserkennung



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%

Anti-Bias Systeme

- "Einfachste" Lösung: Daten ohne Bias
  - Vorverarbeitung mit Upsampling, Downsampling
- Nicht immer möglich
  - Erdbeben
  - Seltene Krankheiten
- Algorithmische Gegenmaßnahmen
  - Beispiel: AI Fairness 360 (IBM)



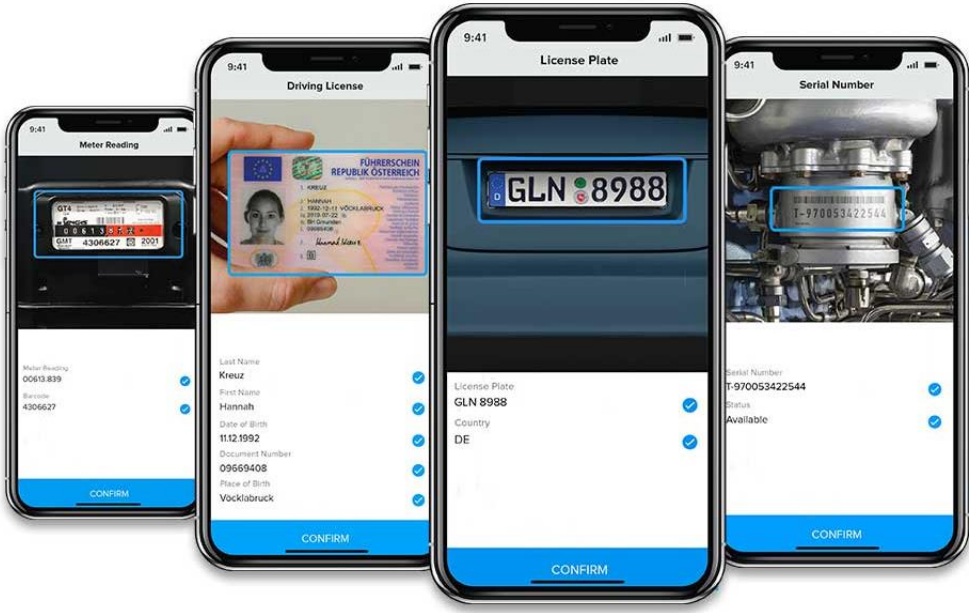
Quelle: Bellamy RKE et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943 [cs.AI], 2018

Lokalisation & Detektion

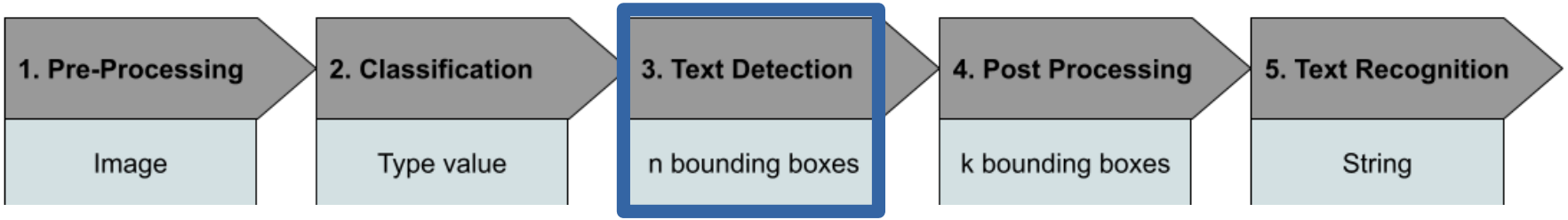
Quelle: ...

Beispiel | Mobile Texterkennung

- Tausende Nutzer
- Millionen Bilder pro Woche
- Viele Beteiligte
- Skalierbarkeit und hohe Qualität erforderlich
- Wichtiges Features: Lokalisierung

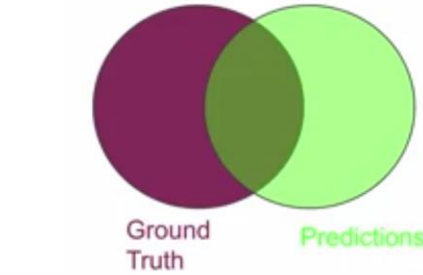


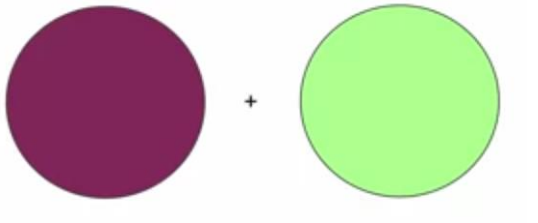
Beispiel | Mobile Texterkennung

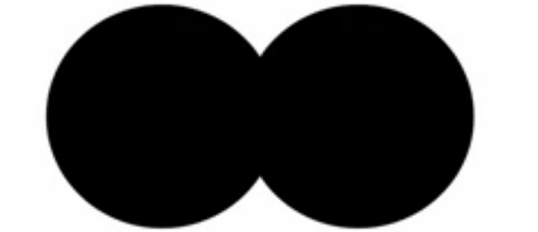


Metriken für Lokalisation & Detektion

- Area of Overlap
  - Summe(True Positives)
- Combined Area
  - Gesamt-Pixel
- Area of Union
  - Combined Area – Area of Overlap



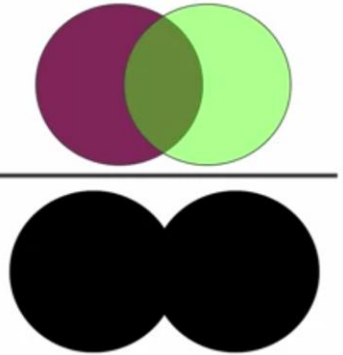




Quelle: https://www.coursera.org/lecture/advanced-computer-vision-with-tensorflow/evaluation-with-iou-and-dice-score-zbusx

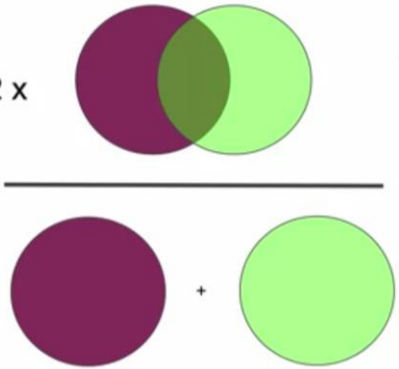
Metriken für Lokalisation & Detektion | IOU vs. Dice Score

$$\text{IOU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$



$$\text{Dice Score} = 2 \times \frac{\text{Area of Overlap}}{\text{Combined Area}}$$

$$2 \times \frac{\quad}{\quad + \quad}$$



Quelle: https://www.coursera.org/lecture/advanced-computer-vision-with-tensorflow/evaluation-with-iou-and-dice-score-zbusx


LEFT REARWARD VEHICLE CAMERA
MEDIUM RANGE VEHICLE CAMERA
RIGHT REARWARD VEHICLE CAMERA
LANE LINES
LANE LINES
ROAD FLOW
IN-PATH OBJECTS
ROAD LIGHTS
OBJECTS
ROAD SIGNS

Erklärbarkeit von KI – Explainable AI Examples

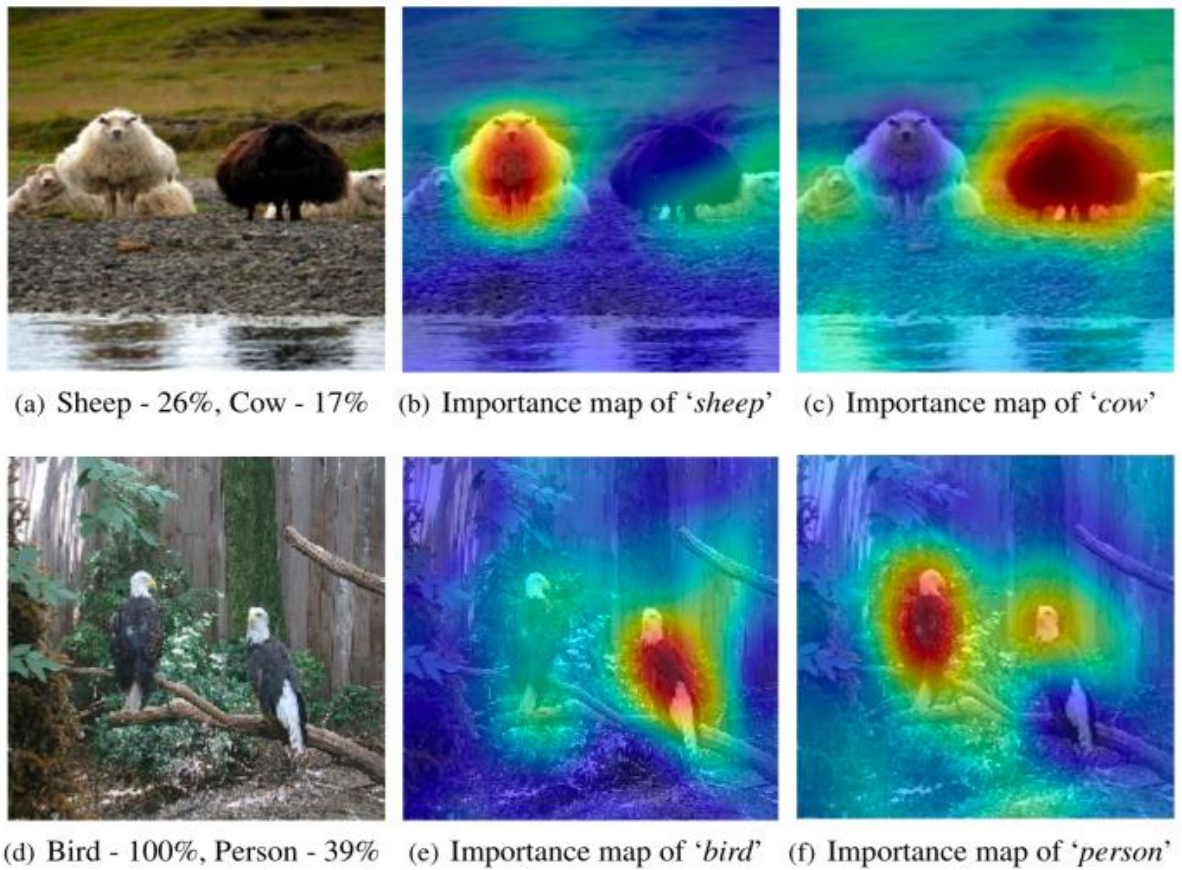


Image retrieved from:
Vitali Petsiuk, Abir Das, & Kate Saenko. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models.

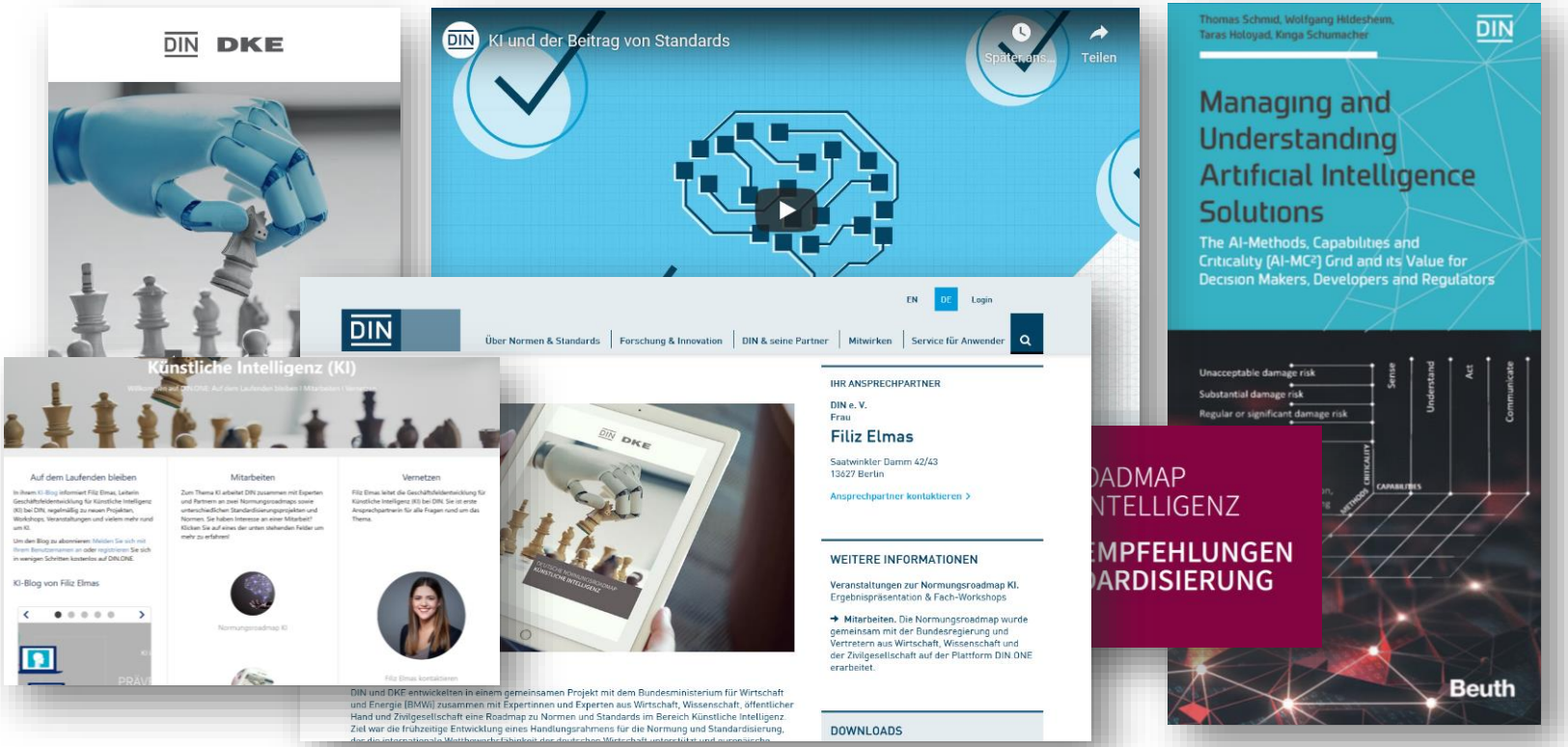
Fazit

Quelle: ...

Normungsroadmap Künstliche Intelligenz v2 startet – Machen Sie mit!

Informationsmaterial

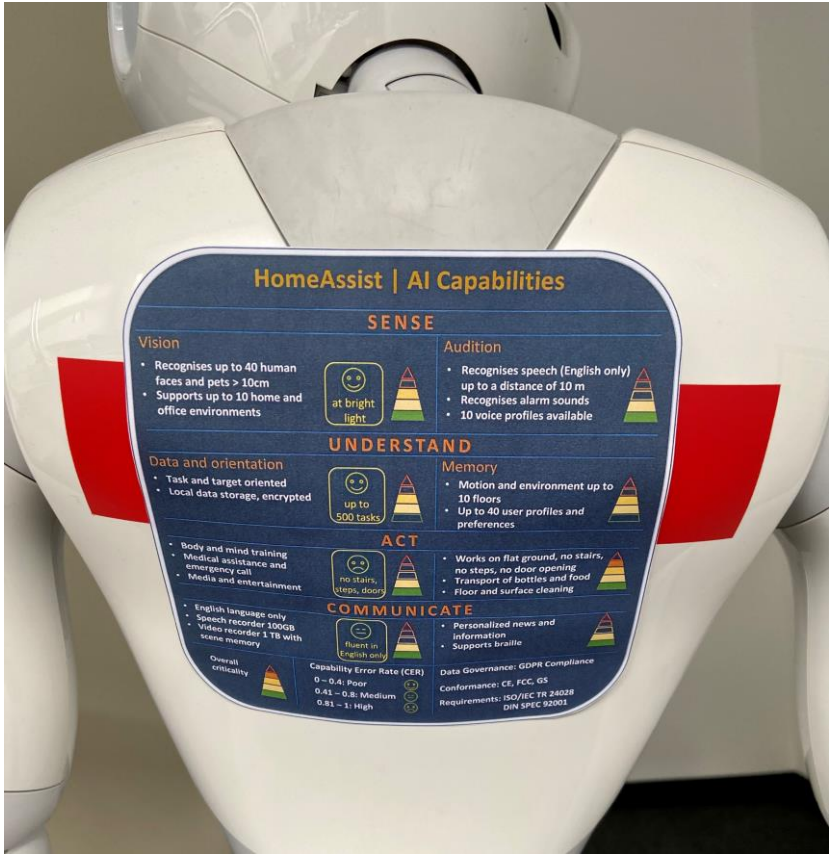
https://link.springer.com/article/10.1007%2Fs13218-021-00736-4



Fazit

- Eine KI-Anwendung = viele KI-Prozesse
  - Jeder Prozess muss wohldefiniert sein
  - Für jeden einzelnen Prozess muss Bias/Robustheit/Fairness kontrolliert sein
  - Qualität

- Ende-zu-Ende-Evaluation unbedingt nötig
  - Risiko und Fehlerbetrachtung müssen für die gesamte Anwendung durchgeführt werden (“AI-factory”)
  - Verständnis von Fehler/ Risken sollte im Ökosystem (Anbieter, Kunde, ...) normiert und standardisiert werden



1. Runde Fragen aus dem Chat…
…vor der geplanten Umfrage ☺

Moderation Frau Dr. Reinel

Warum interessiert Sie dieser Breakout?

https://pingo.coactum.de/719298

Warum interessiert Sie dieser Breakout?

→ 31 Antworten (Auswahl)

- Gesellschaftliche Relevanz
- Wie kann man Bias, Robustheit und Fairness überprüfbar machen, bzw. dokumentieren
- Minmierung von Bias durch diverse Teams, Verbot von Personalauswahl- und Biometrieanwendngen prüfen, AGG/DSGVO um Prüf-und Informationspflichten für algorithm. Systeme erweitern
- Wie erhöht man die Akzeptanz für KI?
- Dieses Thema ist eine der vielleicht verstecktesten Fallen der KI
- Anwendung in industriellen Systemen, die robust sein müssen
- Das Thema ist in regulierten Industrien oft elementar, bevor dort KI in Produktion gebracht werden kann.
- Weil wir den historischen Moment nutzen sollten, Ungleichheiten aus der analogen Welt nicht mit in die Entwicklung neuer Technologien zu nehmen
- Engagement in der deutschen Fachgesellschaft für Geschlechterstudien zum Thema AI, Bias, Diskriminierung
- Wir machen datengetriebene Entscheidungsunterstützung

Für welchen Anwendungsfall würden Sie Bias/Robustheit/Fairness messen?

https://pingo.coactum.de/719298

Für welchen Anwendungsfall würden Sie Bias/Robustheit/Fairness messen?

→ 26 Antworten (Auswahl)

- Regulierung Biometriesysteme
- Personalauswahlysteme (4x)
- Industrieübergreifend bzw. alle Anwendungsbereiche, wo Daten eine Rolle spielen (4x)
- Anwendung in der Medizintechnik (6x), z.B. Diagnostik oder medizinische Bildverarbeitung
- Für KI-Einsatz in der öffentlichen verwaltung
- Dosimeter, Bildqualität in der Medizin
- Industrie 4.0 (2x), z.B. predictive maintanance
- Automatisierungstechnische systeme / industrielle produktion / industrie 4.0
- Sicherheitstechnik
- Einsatz von Robotik mit KI
- Computer Vision Use-cases im Smart Home
- Geschlechtergerechtigkeit und Intersektionalität (2x)

Welche Ihrer Anwendungen fälltin die High Risk-Klasse nach EU-Vorschlag?

- **Critical infrastructures** (e.g. transport), that could put the life and health of citizens at risk;
- **Educational or vocational training**, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);
- **Safety components of products** (e.g. AI application in robot-assisted surgery);
- **Employment, workers management and access to self-employment** (e.g. CV-sorting software for recruitment procedures);
- **Essential private and public services** (e.g. credit scoring denying citizens opportunity to obtain a loan);
- **Law enforcement** that may interfere with people's fundamental rights (e.g. evaluation of the reliability of evidence);
- **Migration, asylum and border control management** (e.g. verification of authenticity of travel documents);
- **Administration of justice** and **democratic processes** (e.g. applying the law to a concrete

Welche Ihrer Anwendungen fällt in die High Risk-Klasse nach EU-Vorschlag?

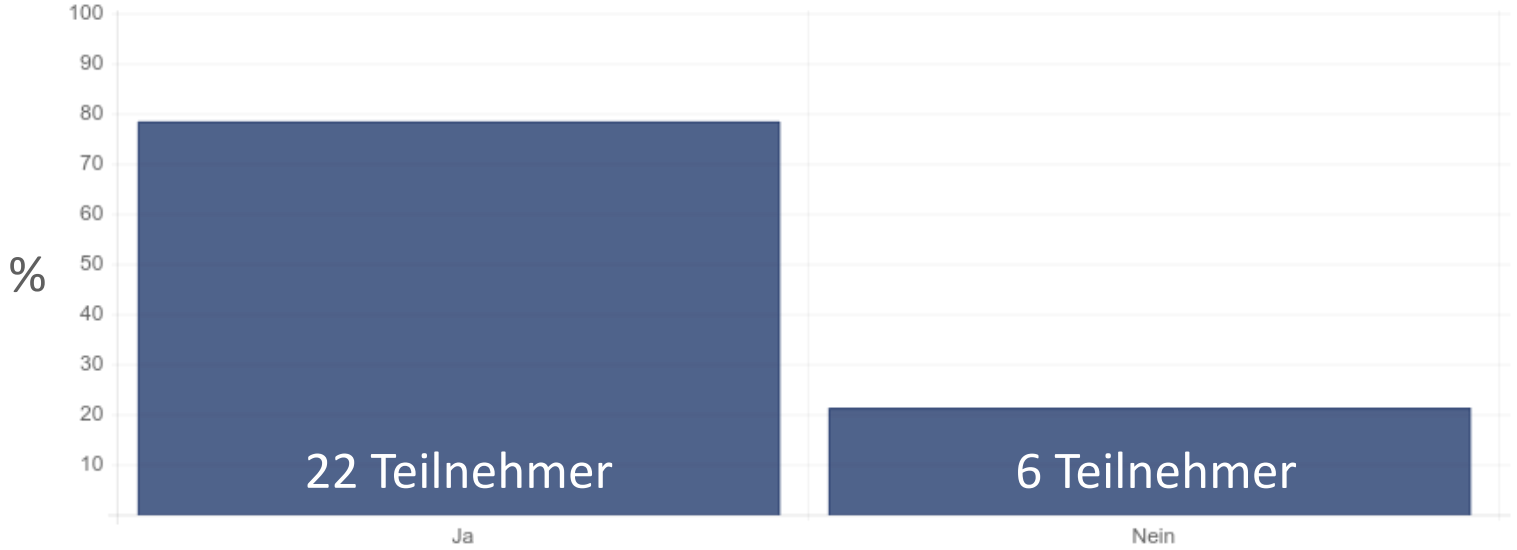
→ 26 Antworten (Auswahl)

- Autonomes Fahren/ Fliegen (3x)
- Medizinprodukte (3x)
- Medizinische Bildqualität
- Robotic surgeries
- Kritische infrastrukturen z.b. energieversorgung, umwelt, katastrophenmanagement, krisenmanagement
- Education, Online-Prüfungsbeaufsichtigung, Erkennung von Prüfungsbetrug
- Gesichtserkennung (3x), z.B. Identifizierung bei Prüfungen oder zivile Sicherheit
- Objekterkennung
- Empfehlunsalgorithmen auf Plattformen (z.B. Youtube-Videos)
- Vergabe von sozialleistungen
- Jobmatching
- keine (4x)

Haben Sie Interesse, im Rahmen der deutschen Normungsroadmap ein Ökosytem für dieses Thema aufzubauen?

https://pingo.coactum.de/719298

Haben Sie Interesse, im Rahmen der deutschen Normungsroadmap ein Ökosytem für dieses Thema aufzubauen? → 28 Antworten



Wenn sie Interesse haben das Thema gemeinsam weiter zu bearbeiten können sie uns gern eine Mail schicken

https://pingo.coactum.de/719298

2. Runde Fragen aus dem Chat…
…dann Ergebnisse der Umfrage!

Moderation Frau Dr. Reinel

Ergebnis der Umfrage

Vielen Dank für Ihre Aufmerksamkeit!

Wir planen eine Arbeitsgruppe zum Thema aufzusetzen. Wenn sie Interesse haben das Thema gemeinsam weiter zu bearbeiten können sie uns gern eine Mail schicken und wir laden sie ein.

Ihr Ansprechpartner:

Dr. Wolfgang Hildesheim
IBM Deutschland

Tel.: +49 160 5509 598
E-Mail: hildeshe@de.ibm.com

Ihr Ansprechpartner:

Dr. Thomas Schmid
Universität Leipzig

Tel.: +49 341 97 32136
E-Mail: schmid@informatik.uni-leipzig.de

Text hinzufügen