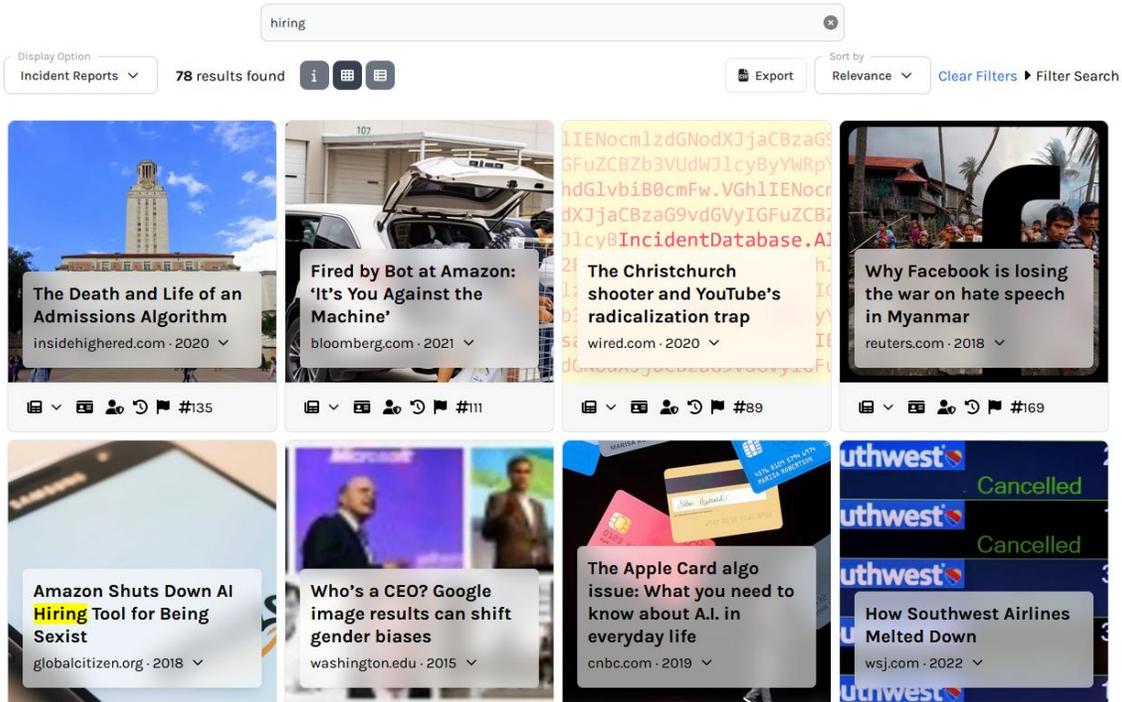


TRUSTWORTHY AI

Priv.-Doz. Dipl.Ing. Dr.techn. Dominik Kowald, BSc.
Research Area Manager – Fair AI

Dipl.Ing. Patrick Ratheiser, MSc.
CEO - LeftShiftOne



[<https://incidentdatabase.ai/>]

Negative Auswirkungen durch AI

AI Incidents Database

> 2000 Vorfälle

~ 80 für Arbeitsmarkt

Wieso?

Biases in Daten

Unklare Fairness Definitionen

Arbeitsmarkt ist Hochrisikoanwendung
laut EU AI Act

TRUSTWORTHY AI

Evaluierung und Zertifizierung von Trustworthy AI

Negative Auswirkungen von AI führen zu schlechter Reputation
bzw. im schlimmsten Fall zu menschlichem Leid

Vor allem in Hochrisikobereichen laut dem EU AI Act:

Arbeitsmarkt

Gesundheitssektor

Finanz- und Kreditwesen

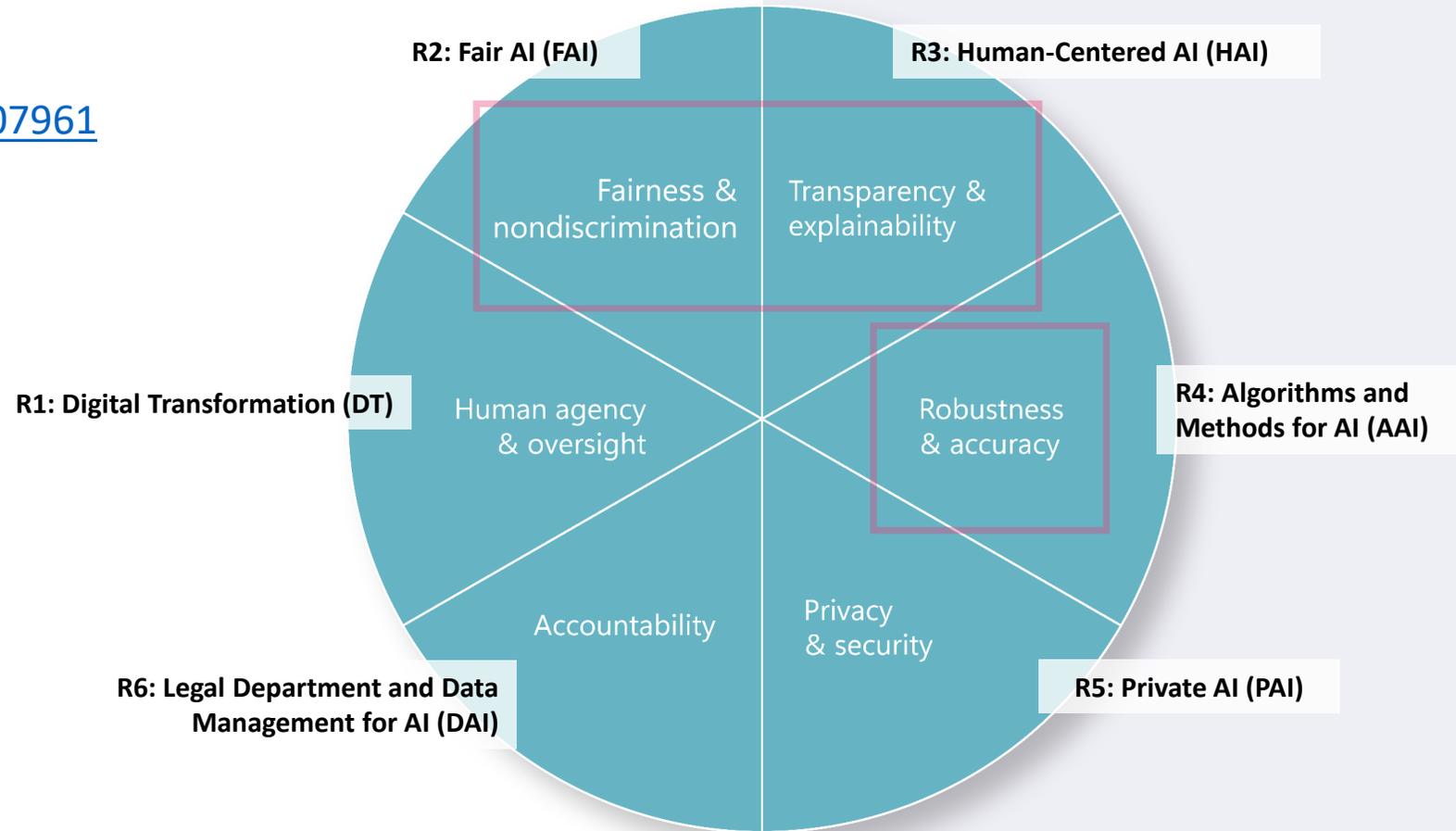
Die Zertifizierung von Trustworthy AI ist ein Weg um die
Vertrauenswürdigkeit von AI Modellen zu zeigen
→ Evaluierung ist eine Grundvoraussetzung dafür

Frage ist wie man die Vertrauenswürdigkeit von AI Modellen zeigen kann?

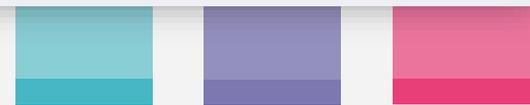
White Papers:

<https://zenodo.org/records/11207961>

1. Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*
2. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys, 55(9), IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*
3. OECD AI Principles (Focuses on inclusivity, transparency, and accountability)
4. UNESCO Recommendation on the Ethics of Artificial Intelligence (Framework for ethical AI development & governance, human rights)
5. Google AI Principles (fairness, accountability, privacy, and safety in AI)
6. Microsoft AI Principles (fairness, reliability, safety, privacy, inclusiveness, transparency, and accountability)
7. European Commission: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>



TRUSTWORTHY AI



TRUSTWORTHY AI

Trustworthy AI Metriken

Genauigkeit

Wie viele Entscheidungen des AI Modells sind richtig?

Fairness

Wie hoch ist die Gefahr, dass Nutzer auf Grund von sensitiven Attributen (zB Geschlecht, Alter) durch Entscheidungen des AI Modells diskriminiert werden?

Transparenz

Wie einfach ist es Entscheidungen des AI Modells zu erklären?

Robustheit

Wie einfach ist es das AI Modell auszutricksen?

Weitere Metriken

Reproduzierbarkeit (zB Grad der Randomisierung) und Effizienz (zB Zeit/Stromverbrauch)



Trustworthy AI Metriken

- Dashboard

AI MODEL TRUSTWORTHINESS EVALUATION




TRUSTWORTHINESS: 0.6023

Trustworthiness is given for 1 out of 4 metrics.

DATE: 4/8/2024, 5:35:26 PM

SERVER CONFIG

CPU Prozessor Intel(R) Core(TM) i7-1260P, 12 Core(s)

RAM 64GiB

GPU NVIDIA T550

MODEL INFO

Select a model
LogisticRegression

MODEL LogisticRegression
LIBRARY sklearn
LIBRARY VERSION 1.4.1.post1

TRUSTWORTHINESS SCORES

ACCURACY

0.83

TRANSPARENCY

0.655

FAIRNESS

0.729

ROBUSTNESS

0.332

DATA INFO

DATA SET German Credit Data

LINK [Link](#)

RESULT DOWNLOAD

TRAIN SET [Download](#)

TEST SET [Download](#)

PREDICTIONS [Download](#)

Computed as the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group. The ideal value of this metric is 1.

Fairness with respect to:

Age: 0.729

Gender: 0.857

C: 0.21000000000000002	class_weight:	1	l1_ratio:	12	max_iter: 100	fit_intercept: true
intercept_scaling: 1	penalty:	0.0001	verbose:	0	random_state: 7	multi_class: auto
n_jobs:	verbose:				warm_start: false	solver: liblinear

Trustworthy AI Metriken

– Glassbox (LeftShiftOne)

The screenshot displays a 'glassbox model card' for a 'RandomForestClassifier' from 'knowcenter'. The model is identified as 'RANDOMFORESTCLASSIFIER' with a checksum of 'df9d00b3f84d3b9f232ca3ca9ab17ec9' and a size of '1718361770.7343333'. The 'Trustworthiness' is highlighted as 'STABLE'. A navigation bar includes 'SHOW DESCRIPTION', 'RANDOM FOREST', 'ENSEMBLE', 'CLASSIFIER', 'BINARY CLASSIFICATION', 'REGULARIZED', and 'TREE BASED'. Below this, there are sections for 'Benchmarks', 'Ethic Scores', and 'Code & Data'. The 'Ethic Scores' section includes 'Reproducibility', 'Fairness', 'Transparency', 'Robustness', and 'Efficiency', each with a progress indicator. The 'Code & Data' section shows 'DATA' for 'statlog+german+credit+data' with 'train', 'test', and 'evaluate' buttons.

Zusammenfassung

Problem:

KI Vorfälle aufgrund von Daten Biases, Diskriminierung durch AI-basierte Entscheidungen – besonders in sensiblen Bereichen wie Arbeitsmarkt, Gesundheitssektor, Finanz- und Kreditwesen

Lösungsansätze:

Zertifizierung von Trustworthy AI auf Basis von a priori bestimmten Metriken: Genauigkeit und Robustheit (Wahrscheinlichkeit, dass AI-Entscheidungen richtig sind), Fairness (keine Diskriminierung), Transparenz (einfache Erklärbarkeit von AI-Entscheidungen), Reproduzierbarkeit & Effizienz

Vorteile und Nutzen:

Fairness (Diskriminierung aufgrund von Biases erkennen), Transparenz und Erklärbarkeit, Effizienzsteigerung, Entscheidungsunterstützung für AI Auditoren und AI Entwickler

Branchen:

Alle Branchen, in denen AI eingesetzt wird, vor allem Hochrisikoanwendungen wie Human Resources/Arbeitsmarkt, Finanz- und Kreditwesen, Gesundheitssektor

TRUSTWORTHY AI



Know Center Research GmbH

Sandgasse 34/2
A-8010 Graz
+43 316 873 30801
info@know-center.at



TRUSTWORTHY AI