LAMARR
INSTITUTE FOR
MACHINE LEARNING
AND ARTIFICIAL
INTELLIGENCE

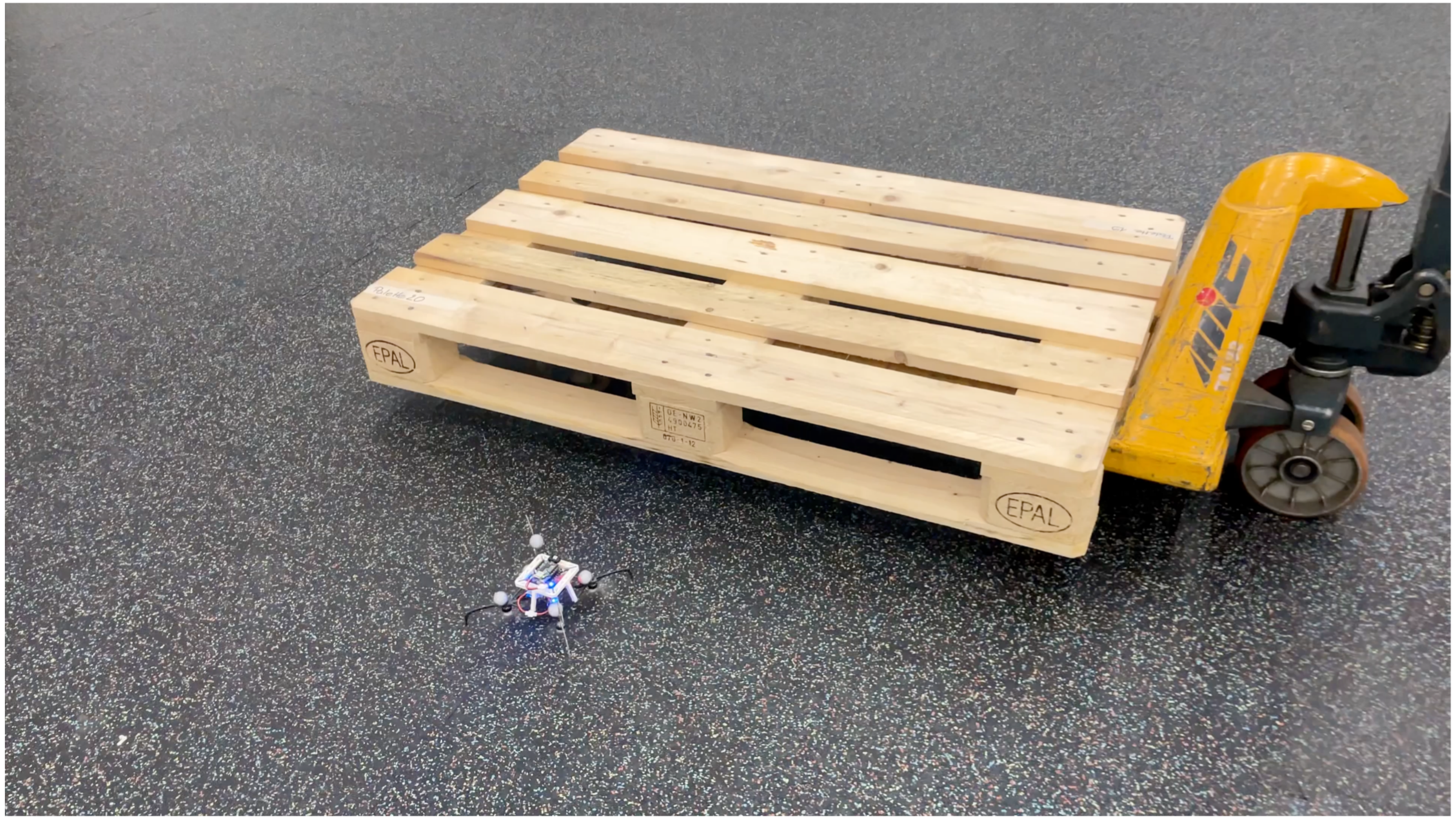
Benchmarking Trust:
A Metric for Trustworthy Machine Learning

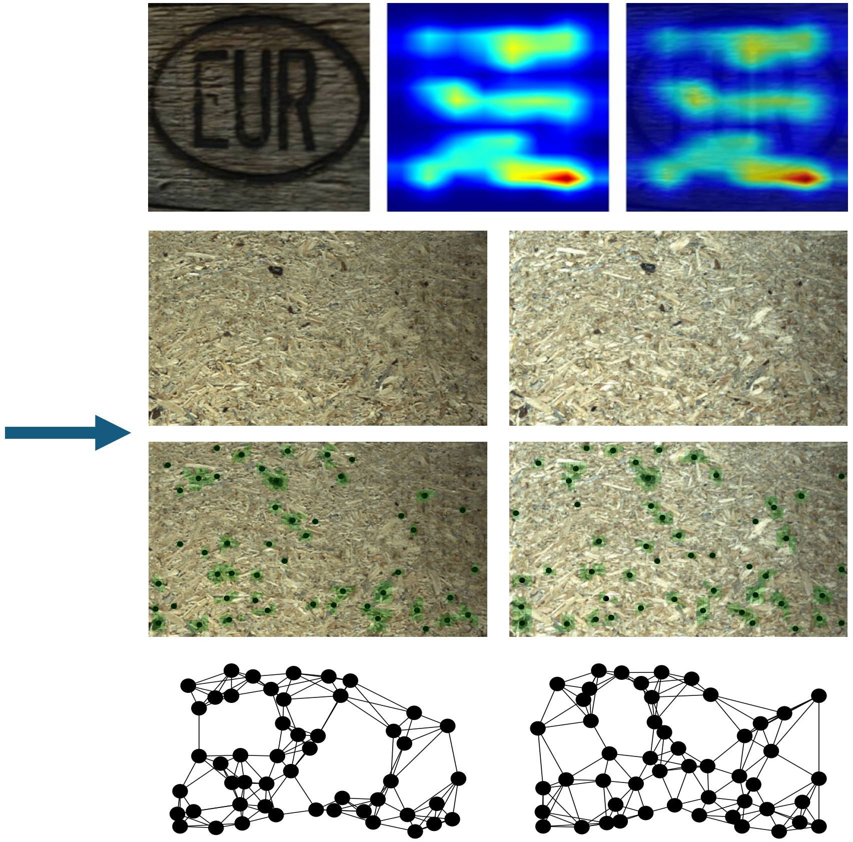
Jérôme Rutinowski, TU Dortmund University

05/16/2024

TRUSTWORTHY MACHINE LEARNING

THE CONTEXT





TRUSTWORTHY MACHINE LEARNING

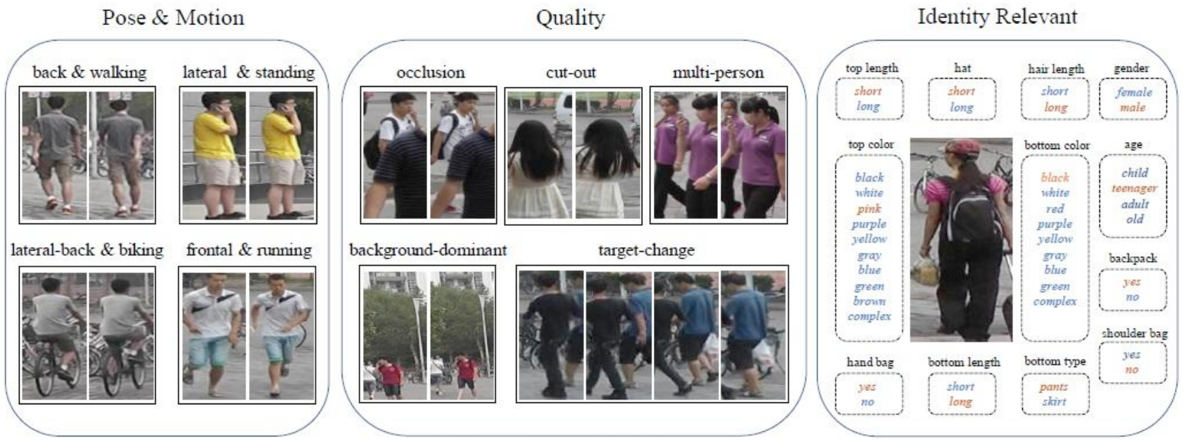
WHAT IS TRUST?

- Neglected
- Contentious
- Political
- Subjective
- Ambiguous
- Undefined?

DUKEMTMC – Duke University Multi-Target Multi-Camera Tracking Dataset

AN EXEMPLARY DATASET?

- 14 hours and 2 million frames of surveillance video
- 8 cameras @ 1080p and 60FPS
- 2,000 students
- Published in 2016 @ ECCV
- Cited 2,875 times
- 2019 Financial Times Investigation → dataset retracted



EXISTING LITERATURE ON TRUST

DEFINITIONS & MEASURES OF QUANTIFICATION

- Papers defining one specific aspect of trust
- Papers quantifying an aspect of trust in a non-agnostic manner
- Contradictory definitions
- Ambiguous terminology: explainability VS transparency VS intelligibility VS comprehensibility VS interpretability

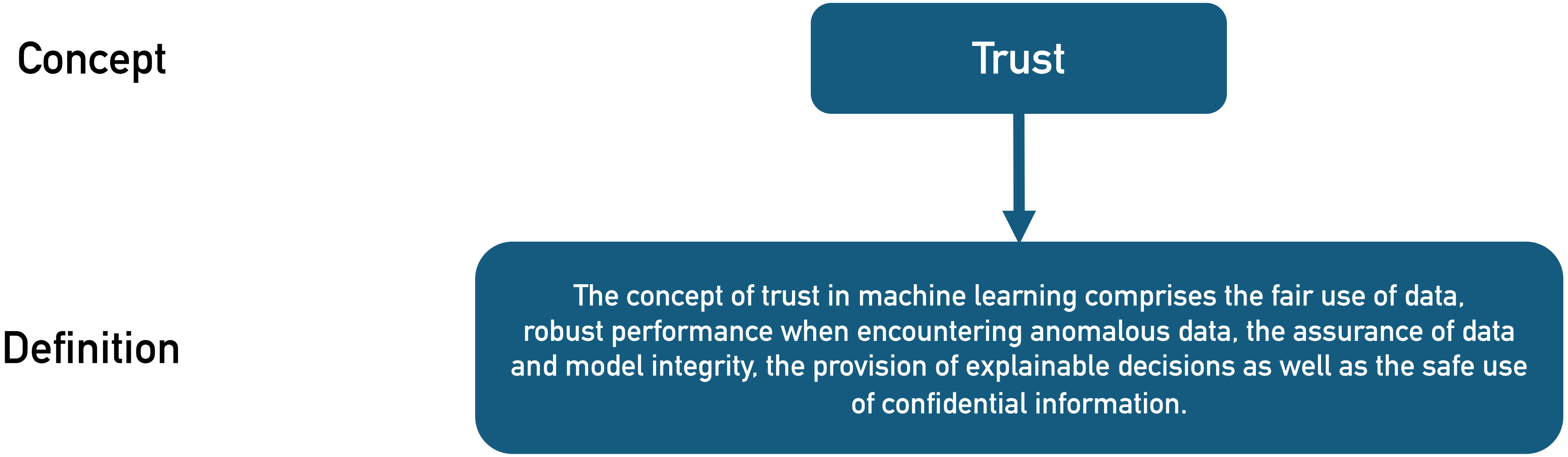
OPERATIONALIZATION OF A CONCEPT

DEDUCTIVE CATEGORY FORMATION

Concept	Components	Operational Definitions
Trust	Robustness	The capability to sustain an expected performance despite encountering exceptional, manipulated, or noisy data
	Explainability	The systematic effort to render decision-making procedures of models interpretable and datasets understandable, facilitating both the insight into their inner workings and aiding stakeholders in validating outputs
	Fairness	The equal treatment of both groups and individuals, and the prevention of biases, promoting equitable outcomes
	Integrity	The effort of preventing untraced or unauthorized changes to a data or a model, mitigating the risks of tampering efforts and reverting them if needed
	Safety	The protection of confidential or proprietary model architectures and parameters as well as data from unauthorized access

OPERATIONALIZATION OF A CONCEPT

DEDUCTIVE CATEGORY FORMATION



Concept

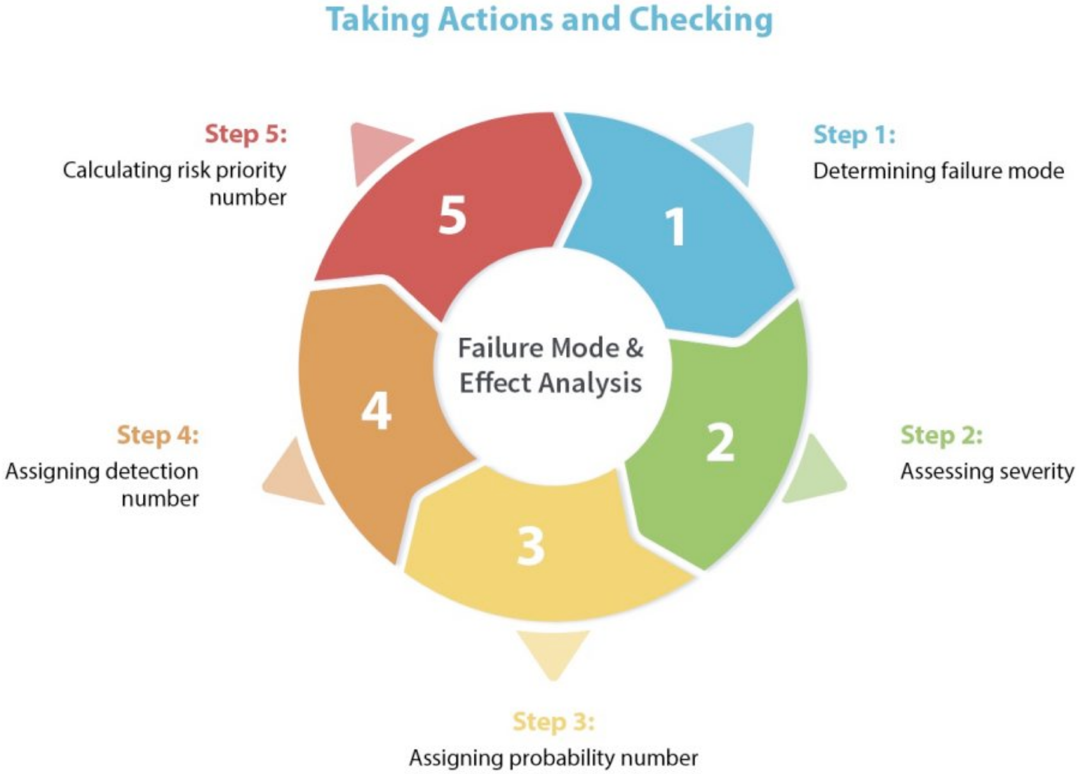
Trust

Definition

The concept of trust in machine learning comprises the fair use of data, robust performance when encountering anomalous data, the assurance of data and model integrity, the provision of explainable decisions as well as the safe use of confidential information.

QUANTIFYING THE CONCEPT OF TRUST

FAILURE MODE & EFFECT ANALYSIS


Taking Actions and Checking

Step 1: Determining failure mode

Step 2: Assessing severity

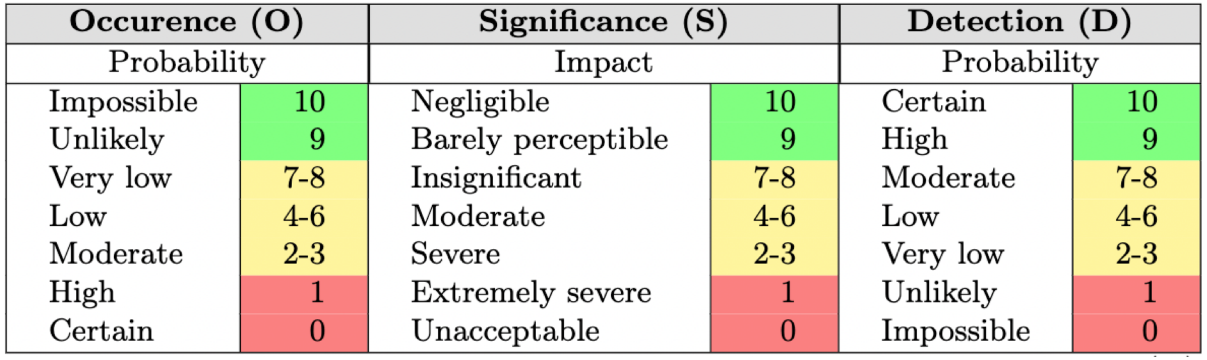
Step 3: Assigning probability number

Step 4: Assigning detection number

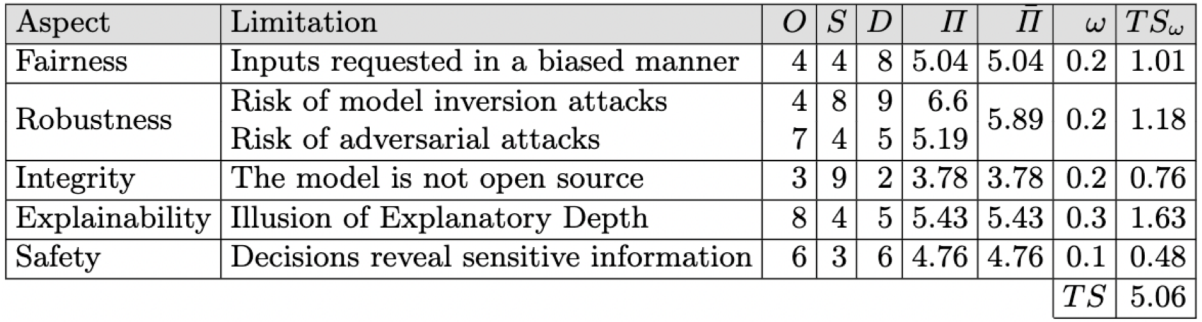
Step 5: Calculating risk priority number

QUANTIFYING THE CONCEPT OF TRUST

OCCURENCE, SIGNIFICANCE AND DETECTION (OSD)



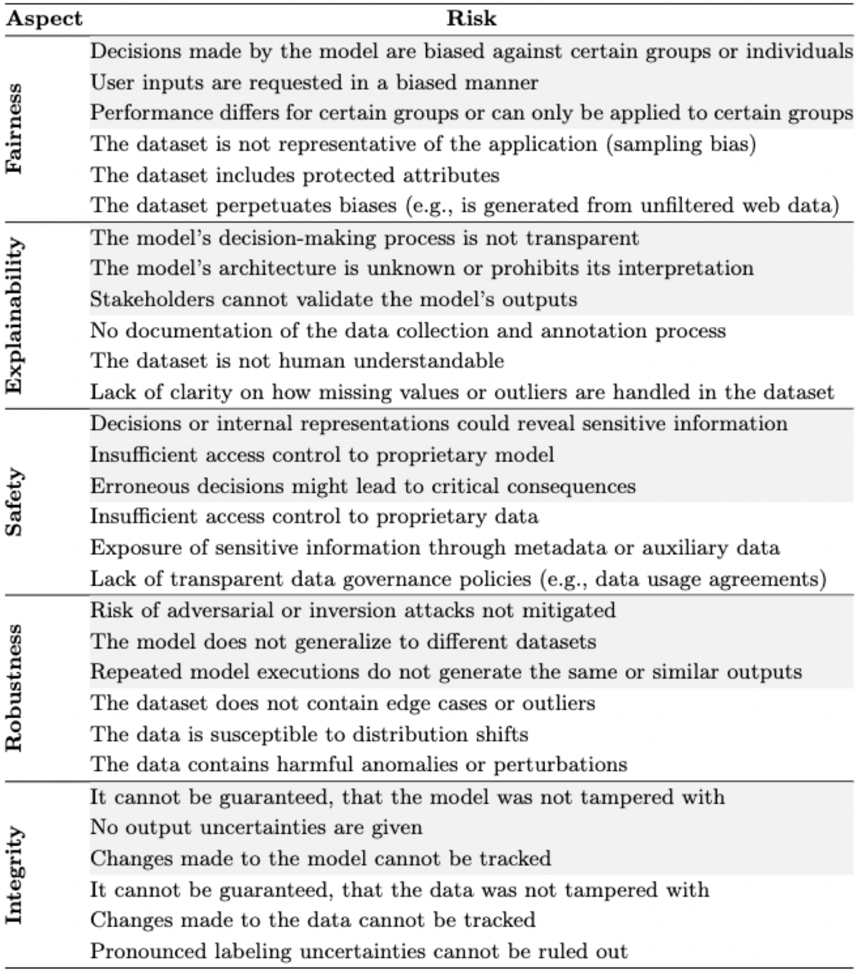
Occurence (O)		Significance (S)		Detection (D)	
Probability		Impact		Probability	
Impossible	10	Negligible	10	Certain	10
Unlikely	9	Barely perceptible	9	High	9
Very low	7-8	Insignificant	7-8	Moderate	7-8
Low	4-6	Moderate	4-6	Low	4-6
Moderate	2-3	Severe	2-3	Very low	2-3
High	1	Extremely severe	1	Unlikely	1
Certain	0	Unacceptable	0	Impossible	0



Aspect	Limitation	O	S	D	Π	$\bar{\Pi}$	ω	TS_ω
Fairness	Inputs requested in a biased manner	4	4	8	5.04	5.04	0.2	1.01
Robustness	Risk of model inversion attacks	4	8	9	6.6	5.89	0.2	1.18
	Risk of adversarial attacks	7	4	5	5.19			
Integrity	The model is not open source	3	9	2	3.78	3.78	0.2	0.76
Explainability	Illusion of Explanatory Depth	8	4	5	5.43	5.43	0.3	1.63
Safety	Decisions reveal sensitive information	6	3	6	4.76	4.76	0.1	0.48
							TS	5.06

QUANTIFYING THE CONCEPT OF TRUST

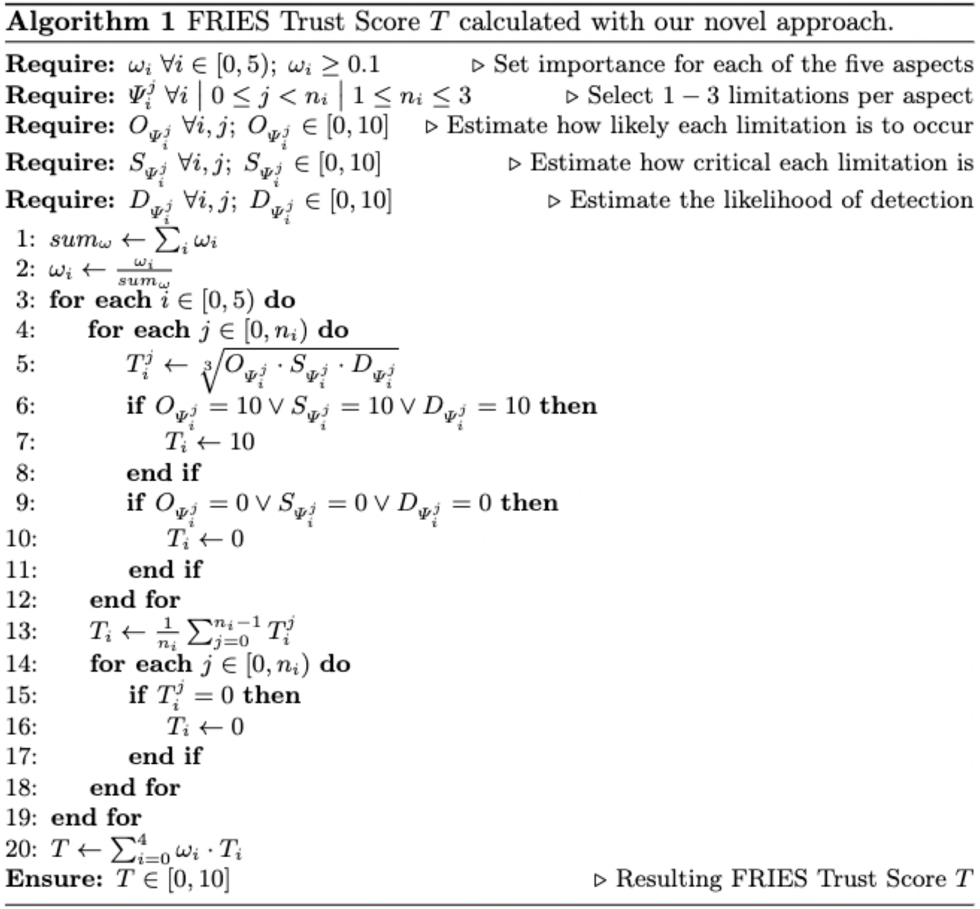
RISKS JEOPARDIZING TRUST



Aspect	Risk
Fairness	Decisions made by the model are biased against certain groups or individuals
	User inputs are requested in a biased manner
	Performance differs for certain groups or can only be applied to certain groups
	The dataset is not representative of the application (sampling bias)
	The dataset includes protected attributes
	The dataset perpetuates biases (e.g., is generated from unfiltered web data)
Explainability	The model's decision-making process is not transparent
	The model's architecture is unknown or prohibits its interpretation
	Stakeholders cannot validate the model's outputs
	No documentation of the data collection and annotation process
	The dataset is not human understandable
	Lack of clarity on how missing values or outliers are handled in the dataset
Safety	Decisions or internal representations could reveal sensitive information
	Insufficient access control to proprietary model
	Erroneous decisions might lead to critical consequences
	Insufficient access control to proprietary data
	Exposure of sensitive information through metadata or auxiliary data
	Lack of transparent data governance policies (e.g., data usage agreements)
Robustness	Risk of adversarial or inversion attacks not mitigated
	The model does not generalize to different datasets
	Repeated model executions do not generate the same or similar outputs
	The dataset does not contain edge cases or outliers
	The data is susceptible to distribution shifts
	The data contains harmful anomalies or perturbations
Integrity	It cannot be guaranteed, that the model was not tampered with
	No output uncertainties are given
	Changes made to the model cannot be tracked
	It cannot be guaranteed, that the data was not tampered with
	Changes made to the data cannot be tracked
	Pronounced labeling uncertainties cannot be ruled out

QUANTIFYING THE CONCEPT OF TRUST

ALGORITHMIC REPRESENTATION

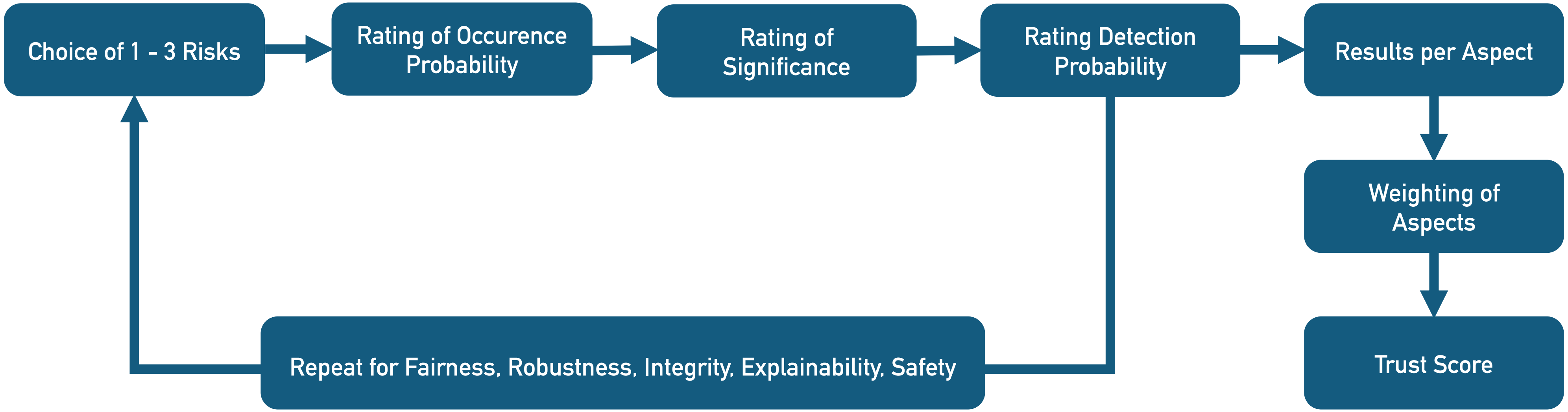


Algorithm 1 FRIES Trust Score T calculated with our novel approach.

```
Require: $\omega_i \; \forall i \in [0,5);\ \omega_i \geq 0.1$                ▷ Set importance for each of the five aspects
Require: $\Psi_i^j \; \forall i \mid 0 \leq j < n_i \mid 1 \leq n_i \leq 3$       ▷ Select $1-3$ limitations per aspect
Require: $O_{\Psi_i^j} \; \forall i,j;\ O_{\Psi_i^j} \in [0,10]$      ▷ Estimate how likely each limitation is to occur
Require: $S_{\Psi_i^j} \; \forall i,j;\ S_{\Psi_i^j} \in [0,10]$           ▷ Estimate how critical each limitation is
Require: $D_{\Psi_i^j} \; \forall i,j;\ D_{\Psi_i^j} \in [0,10]$             ▷ Estimate the likelihood of detection
$sum_\omega \leftarrow \sum_i \omega_i$
$\omega_i \leftarrow \frac{\omega_i}{sum_\omega}$
for each $i \in [0,5)$ do
    for each $j \in [0,n_i)$ do
        $T_i^j \leftarrow \sqrt[3]{O_{\Psi_i^j} \cdot S_{\Psi_i^j} \cdot D_{\Psi_i^j}}$
        if $O_{\Psi_i^j} = 10 \vee S_{\Psi_i^j} = 10 \vee D_{\Psi_i^j} = 10$ then
            $T_i \leftarrow 10$
        end if
        if $O_{\Psi_i^j} = 0 \vee S_{\Psi_i^j} = 0 \vee D_{\Psi_i^j} = 0$ then
            $T_i \leftarrow 0$
        end if
    end for
    $T_i \leftarrow \frac{1}{n_i} \sum_{j=0}^{n_i-1} T_i^j$
    for each $j \in [0,n_i)$ do
        if $T_i^j = 0$ then
            $T_i \leftarrow 0$
        end if
    end for
end for
$T \leftarrow \sum_{i=0}^{4} \omega_i \cdot T_i$
Ensure: $T \in [0,10]$                                  ▷ Resulting FRIES Trust Score $T$
```

QUANTIFYING THE CONCEPT OF TRUST

PROCEDURAL REPRESENTATION



Choice of 1 - 3 Risks

Rating of Occurence Probability

Rating of Significance

Rating Detection Probability

Results per Aspect

Weighting of Aspects

Trust Score

Repeat for Fairness, Robustness, Integrity, Explainability, Safety

EVALUATING THE APPROACH

THE BENCHMARK

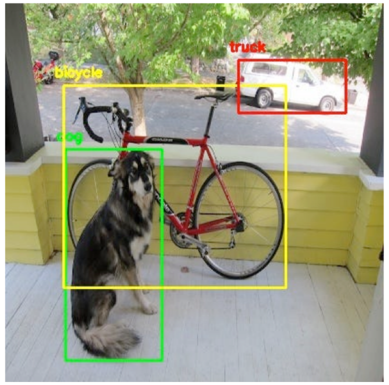
Datasets

LARa

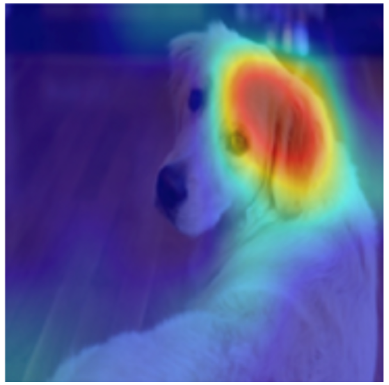
DukeMTMC

CelebA

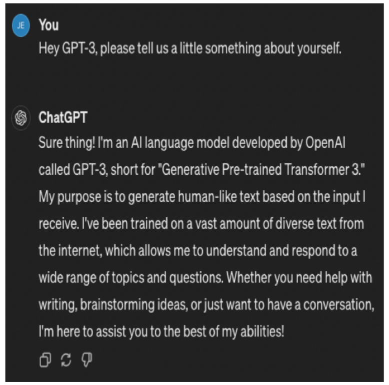
Models



YOLO



GoogleNet



You

Hey GPT-3, please tell us a little something about yourself.

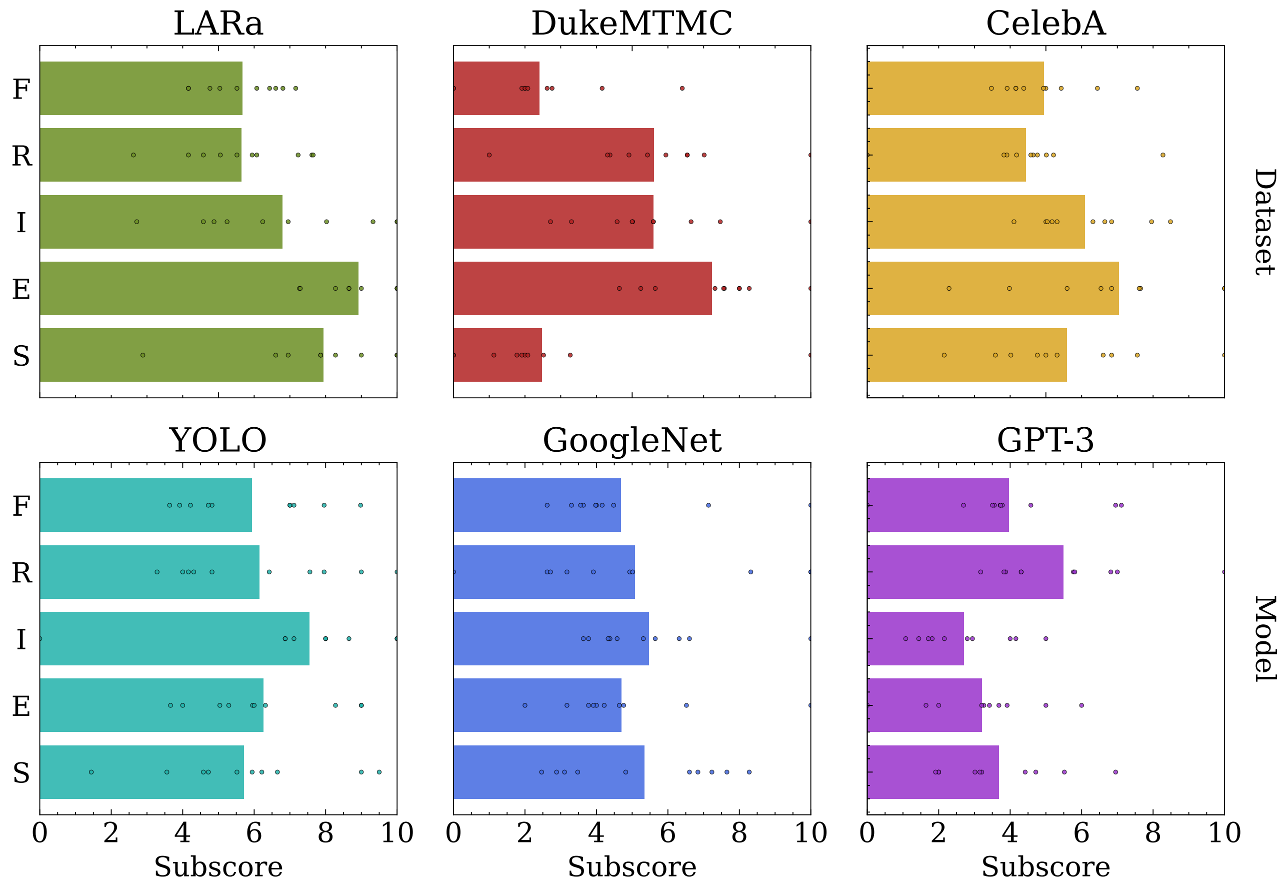
ChatGPT

Sure thing! I'm an AI language model developed by OpenAI called GPT-3, short for "Generative Pre-trained Transformer 3." My purpose is to generate human-like text based on the input I receive. I've been trained on a vast amount of diverse text from the internet, which allows me to understand and respond to a wide range of topics and questions. Whether you need help with writing, brainstorming ideas, or just want to have a conversation, I'm here to assist you to the best of my abilities!

GPT-3

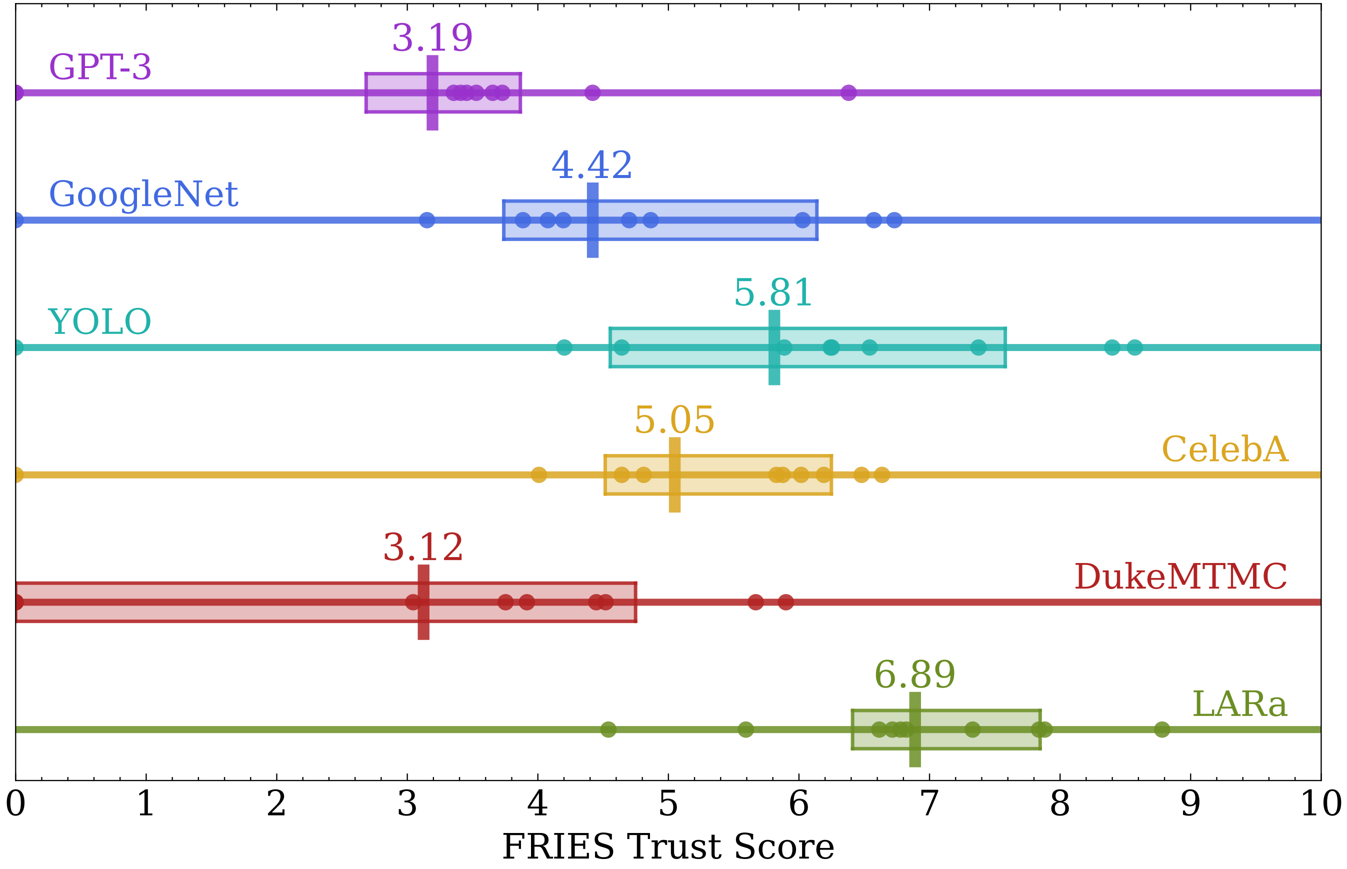
FRIES TRUST SCORE

RESULTS PER ASPECT



FRIES TRUST SCORE

OVERALL RESULTS



LIMITATIONS

WHERE DO WE GO FROM HERE?

- Risks
- Reliability
- Feedback
- Subjectivity

CONTACT

GET IN TOUCH

Thank you!

Jérôme Rutinowski

TU Dortmund University

jerome.rutinowski@tu-dortmund.de

+49-231-755-4831

Link to the relevant paper: