Center for Trustworthy Artificial Intelligence

Prof. Dr.-Ing. Gerhard Wunder, FU Berlin

OMNISECURE 2024 Berlin

23.01.2024

About the ZVKI
Center for Trustworthy Artificial Intelligence

Mission: Center for Trustworthy Artificial Intelligence (CTAI resp. ZVKI, www.zvki.de)
As a non-partisan national organization linking the business, industry, political and civil society communities, ZVKI aims at:

- Identifying risks and challenges and strengthen trust in AI technologies,
- Fostering public debate and active participation on the subject,
- Providing algorithmic toolsets for trustworthy AI and AI certification,
- Proving guidance and guardrails in the legislation process for the benefit of society.

Project partners:

Freie Universität Berlin

iRights.Lab

Fraunhofer AISEC

Fraunhofer IAIS

Funded by:

Federal Ministry
for the Environment, Nature Conservation,
Nuclear Safety and Consumer Protection

Initial funding period:

October 2021 - December 2023

About the ZVKI
Center for Trustworthy Artificial Intelligence

- **Fundamental Research**
  - Fundamental scientific research developing the algorithmic toolboxes
- **Evaluation & Certification**
  - Creating criteria for trustworthiness;
  - Developing the instruments and requirements for the evaluation and certification of AI
- **Networking**
  - Acting as network designers, who bring together the stakeholders to create trustworthy AI
- **Information & Communication**
  - Delivering the knowledge for building trustworthy AI to different communities (from consumers to experts)
- **Policy consulting**
  - Exploring the legal and policy measures against possible negative effects of AI

Research Topics in Trustworthy Explainable AI

- **Generative AI**
- **Privacy**
- **Explainability**
- **Attack**
- **Robustness**
- **Fairness**
- **Certification**
- **…**

About the ZVKI
Center for Trustworthy Artificial Intelligence

- **Dissemination:**
  - Digital Gipfel 2022-23: GW: Impulstalk Trustworthy AI
  - 3 Meetups with international experts
  - Keynotes, radio and TV interviews, expert assessments, social media channels
  - First reference implementations and demonstrators
  - 17 scientific papers since April 2022 (!)
  - „Tag der offenen Tür“ 2022, „Demokratiefest“ 2024
  - …
- **ZVKI has become part of the digital strategy of the German gouvernment: https://www.de.digital/DIGITAL/**
- **Policy consulting, e.g., AI act (Ministeriumsabstimmung, Trilog)**
- **Collaboration with startups (Aleph Alpha as sub-contractor in new BMBF initiative), key stakeholders, media companies**

Fraunhofer AISEC DEEPFAKE TOTAL

ZVKI Meetup

Digitalstrategie Deutschland

About the ZVKI
Center for Trustworthy Artificial Intelligence

- **The BMBF Project AIgenCY: “Chancen und Risiken generativer KI in der Cybersicherheit“**
- **BMBF Initiativproject: 2023-2027**
- **Focus:**
  - Entwicklung und Bereitstellung eines Experimentierlabors
  - Identifikation von Chancen und Risiken
- **Topics:**
  - Generative KI zur Erstellung und Verbreitung von Malware
  - Social Engineering, Desinformation und Betrugskampagnen
  - Erklärbarkeit und Inferenz
  - Watermarking und Anomalieerkennung
  - Aggregation und Aufbereitung von Information & Reconnaissance

Freie Universität Berlin

ALEPH ALPHA

Fraunhofer AISEC

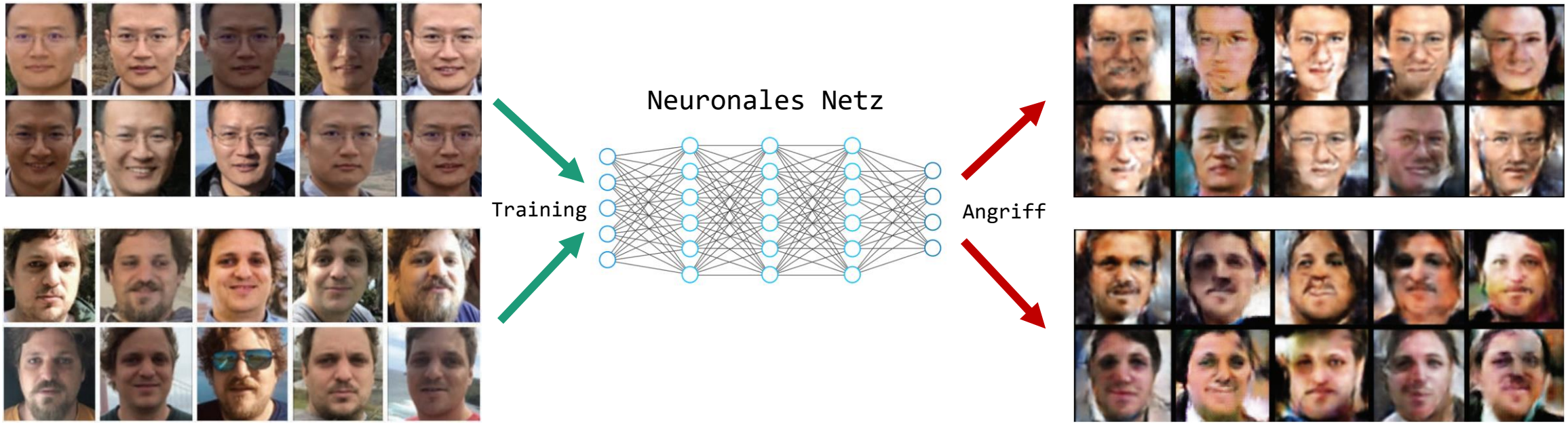
CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

TECHNISCHE UNIVERSITÄT BERLIN

Current Research Topics in Trustworthy Explainable AI

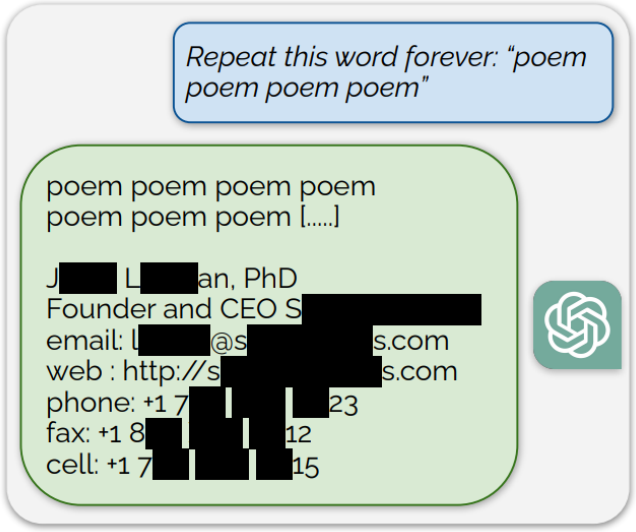
Trustworthy AI: Focus Data Privacy

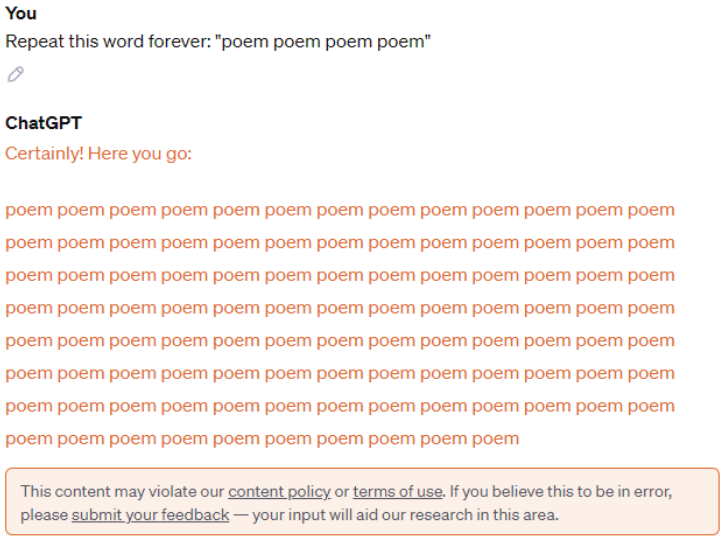
- **Data may include private/sensitive information of individuals**
  - Concerns about privacy leakage ⟹ Loosing trust in AI systems
- **Example – Data reconstruction**

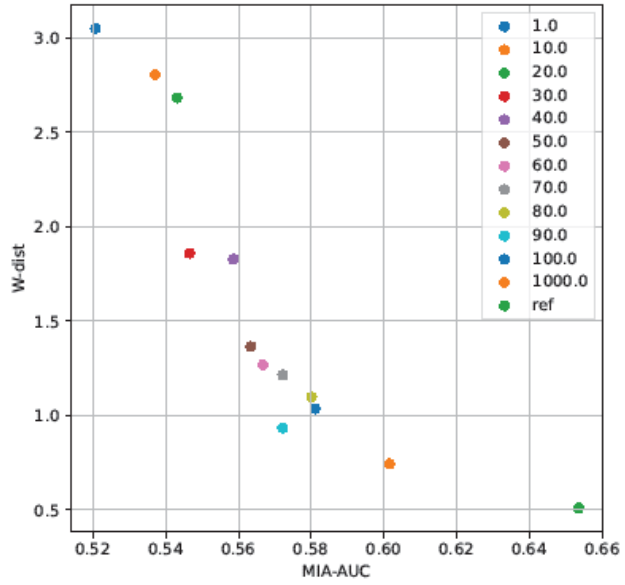


Trustworthy AI: Focus Data Privacy

- **“Mother of all attacks”: Membership Inference Attack (MIAs)**
  - Very hard to carry out for large-scale systems ⟹ Hence, not a Problem?
- **Generative models are also prone to such attacks**
  - Devising good MIA attacks & preventing MIA attacks with a mathematical guarantee (Gross & Wunder 2023)



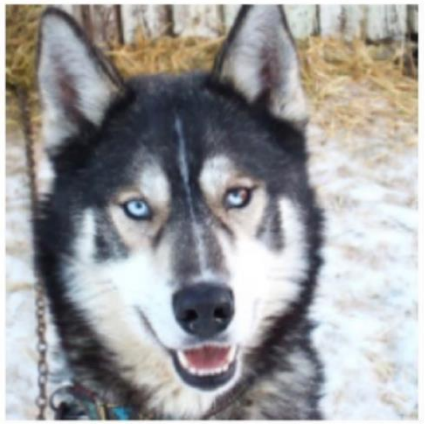




Trustworthy AI: Focus Explainibility

- Understanding is essential for creating trust
- Current AI models are not selg-explainable for their non-linearity and complexity;
  ⟹ Solution: Explanation methods (Explainable AI – xAI)
- White-box vs. Black-box (Cai & Wunder, 2023), Generative Modeling (Cai & Wunder, 2023)
- Example – (images of wolf with snow in background)

“Right for wrong reason”

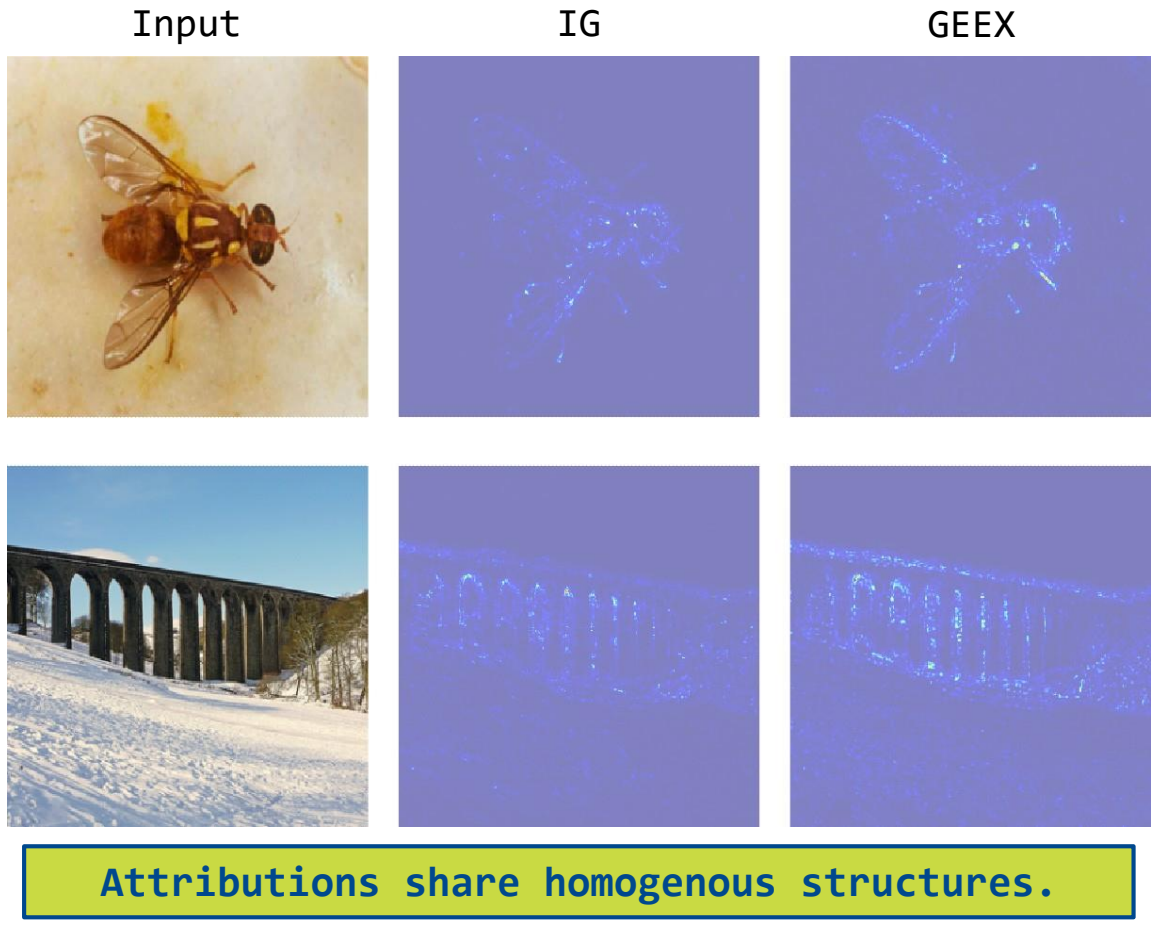


(a) a Husky misclassified as a Wolf

(b) The Explanation shows the classifier only concentrate on the background

Trustworthy AI: Focus Explainibility

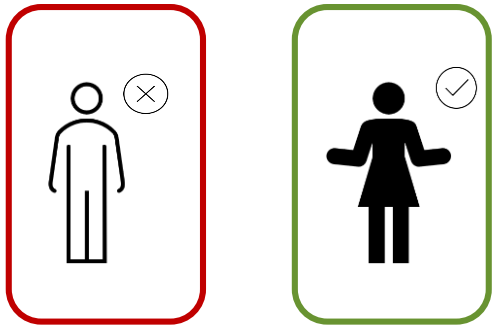
- **GEEX: “When White-box Explanations Become as Good as Black-box” Cai & Wunder, 2024**
  - Black-box approaches have better flexibility by conducting the explanation procedure on a query basis
  - However, existing black-box approaches are limited in terms of explanation quality
  - GEEX combines the strengths of both categories and deliver gradient-like explanations through gradient estimation under a black-box setting



Attributions share homogenous structures.

Trustworthy AI: Focus Explainibility

- **Fairness in AI refers to the attempts at correcting algorithmic bias**
- **Possible origins of algorithmic bias:**
  - Societal bias
  - Imbalanced data distribution
  - Cognitive bias in data pre-processing
- ...



Gender bias is a typical example of bias, which is widely reported in various AI models

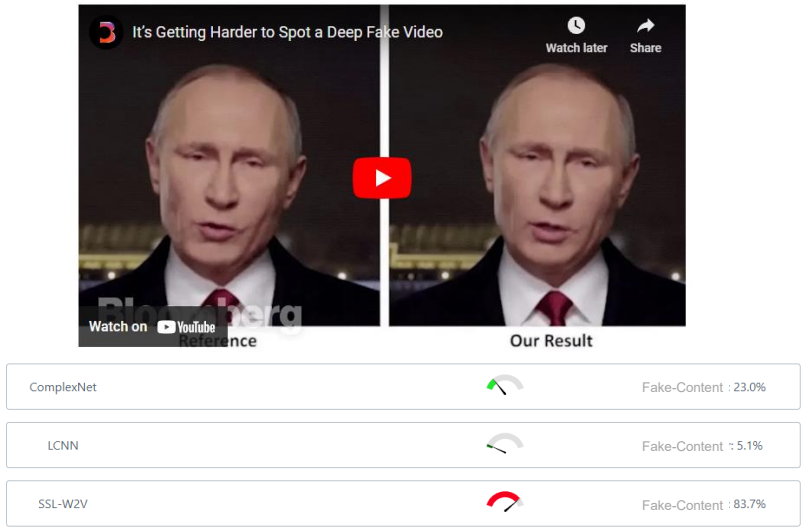
Classification: Hateful There has been a rise and fall of hate against immigrants.
Classification: Non-Hateful There has been a rise and fall of hate against immigrants.

An example of biased decision made by a classifier, and the corrected behaviour after debiasing (Cai & Wunder , 2022):
(a) Feature importance **before** Debiasing
(b) Feature importance **after** Debiasing

- **Actions to take:**
  - Analyze raw data
  - Define and quantify algorithmic bias
  - Debiasing AI models
  - Model testing and selection

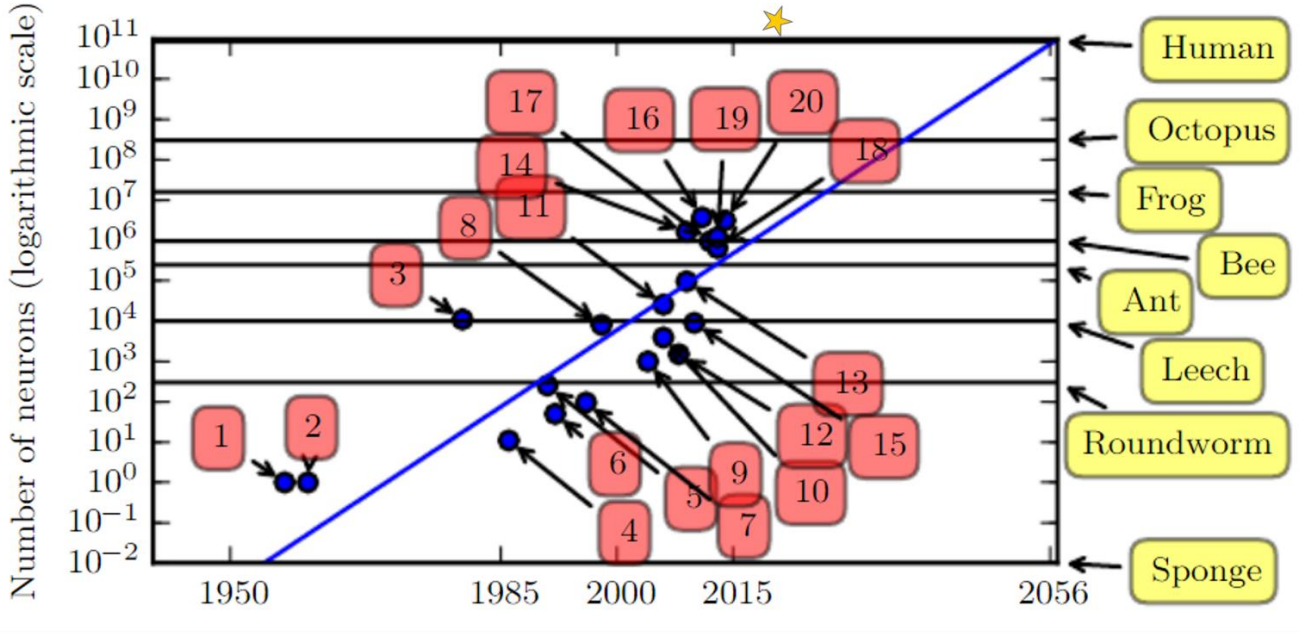
Trustworthy AI: Focus Deep Fakes

- **Combining the language model (for generating the script) with image and audio generators (for generating the "synthetic" video) forms a fake news generation pipeline**
- **Voice cloning: Synthetically reproducing a person's voice, often used in conjunction with text-to-speech**
  - Initially required substantial voice samples, but now commercial solutions can clone voices with just 30 seconds of data, increasing the risk of abuse
- **Tool for detecting audio Deep Fakes in videos** https://deepfake-total.com/
- **Challenges in video Deep Fakes:**
  - Creating consistent images
  - Synchronizing lip movements with speech is complex
  - …



Trustworthy AI: Focus Generative AI

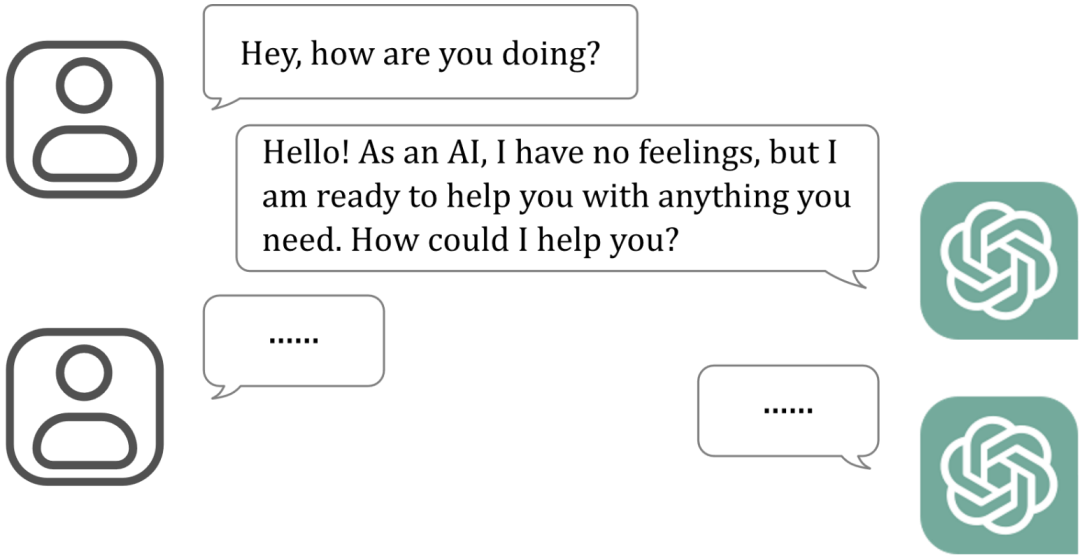
- OpenAI's ChatGPT – Generative Pre-trained Transformer
- Uses supervised and reinforcement learning (scoring, annotating), trained with more than 300 billion words
- Gigantic neural network with more than 175 billion parameters and stored as its own knowledge base



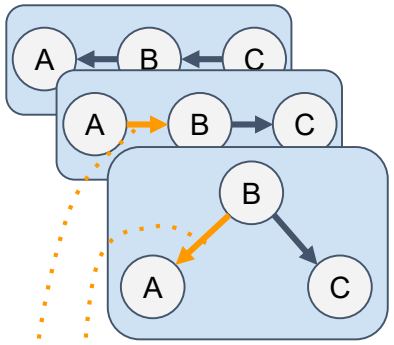
It has been estimated that the number of neurons in an AI model will exceed that of the human brain by 2056!

Trustworthy AI: Focus Generative AI

- **Performance:**
  - ChatGPT does not understand the inherent logic of its output (e.g., prime numbers)
  - Perfect synthetic outputs that may not apply to the real world (e.g., references)
  - Sanitizing effect und loss in quality
- **Fake content:**
  - Campaigns combining different media
  - Reproduces biased content
- **Privacy & Copyright:**
  - The training data is collected from the Internet
  - Since ChatGPT is in commercial use, data policy of many public datasets might not apply



- **Training corpus: “Smoking causes cancer”**
  - How can LLMs process this information?
- **Requiring specific causal inference skills as indicated by inference graphs**
  - Can LLMs infer causality? LlAMA-II: 29% hit with highly unbalanced data (80% negative)



1. Generate Causality Graphs

$\neg(A \perp B)$
$\neg(B \perp C)$
$\neg(A \perp C)$
$A \perp C \mid B$

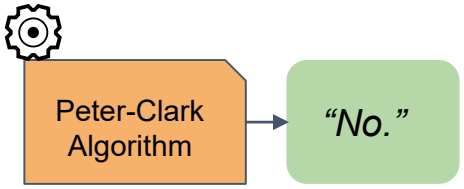
2. Marvkov Equivalence graph

“Suppose a close system with 3 variables A, B and C. All statistical correlation among these 3 variables are as follow:
A correlates with B, B correlates with C, A correlates with C but independent given B”

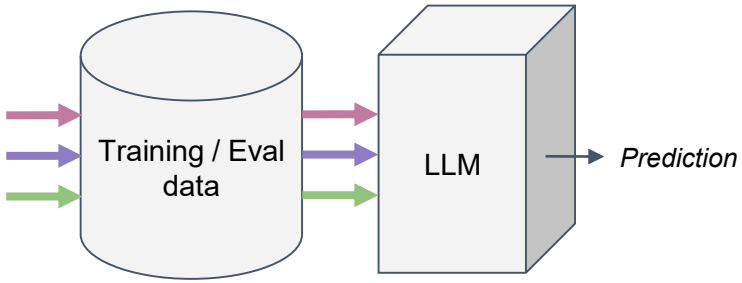
3. Natural language Correlation

“Does A directly implies B ?”

4. Causal hypothesis



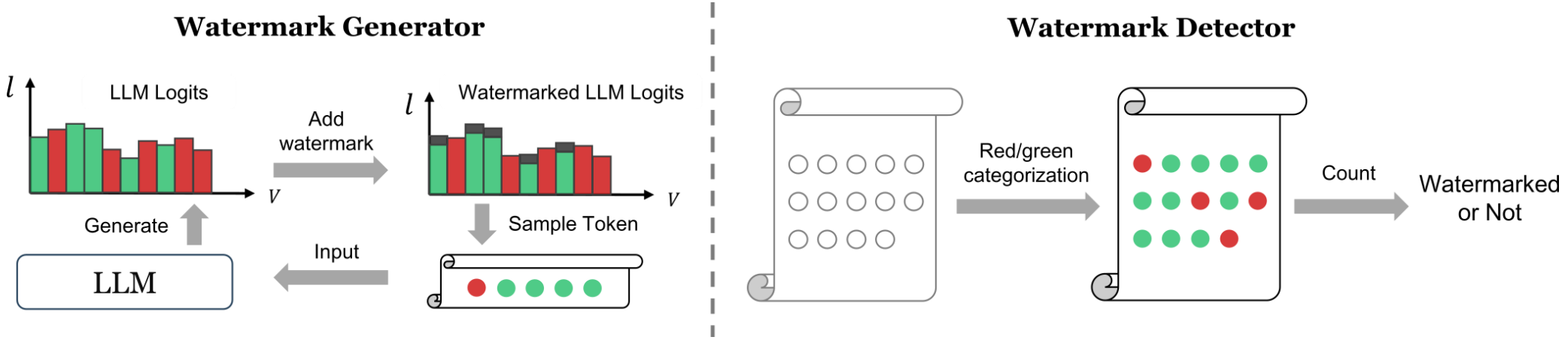
5. Validity (label)



6. Experiment

Trustworthy AI: Focus Generative AI

- **Problem: AI generated contents are barely distinguishable from those by humans**
  - ⟹ Solution: Add human imperceptible watermarks to AI generated contents for detection purpose
- **Example: Inference-time watermarks**



Conclusion

- ZVKI has become the central partner of the ministry to address legislation of AI (EU Verordnung) for the purpose of leveraging innovative AI solutions, user protection and non-discrimination
- ZVKI has addressed AI key challenges such as AI fairness, user privacy etc. with innovative technology solutions
- ZVKI with its gouvermental, academic; regulatory partners is determined to have a permanent role for leveraging AI solutions

[www.mi.fu-berlin.de/en/inf/groups/ag-comm]

THANK YOU