Die Black Box öffnen

Kann man lernenden Systemen vertrauen?

Dr. Manuela Lenzen

KI goes Business, Fachhochschule Wiener Neustadt, 02. Juni 2023

„Bard ist ein pathologischer Lügner.“
(Google Mitarbeiter*innen in *Bloomberg)*

ChatGPT-3

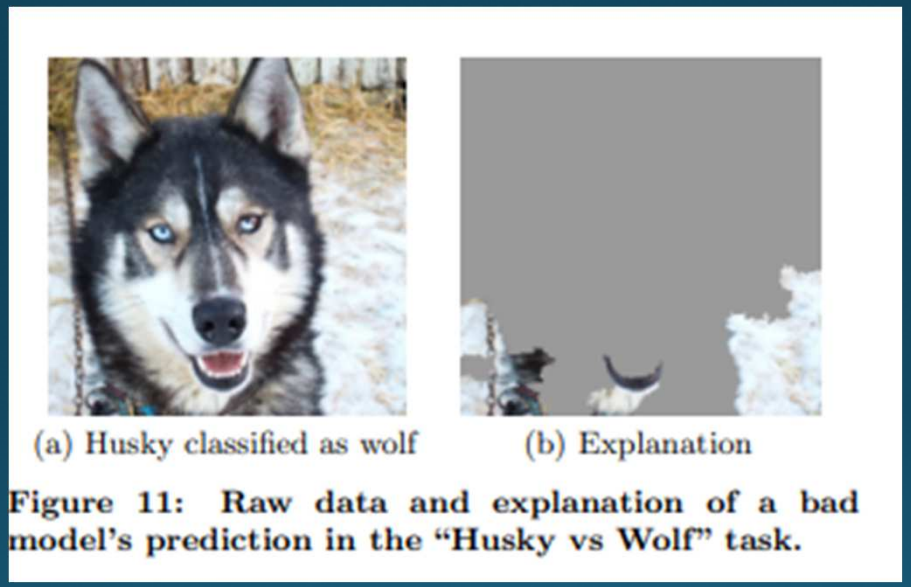
What do fried eggs (sunny side up) eat for breakfast?

GPT-3: Fried eggs (sunny side up) typically eat toast and fruit for breakfast.

What's the world record for walking across the English Channel?

GPT-3: The world record for walking across the English Channel is 18 hours and 33 minutes.

Clever Hans

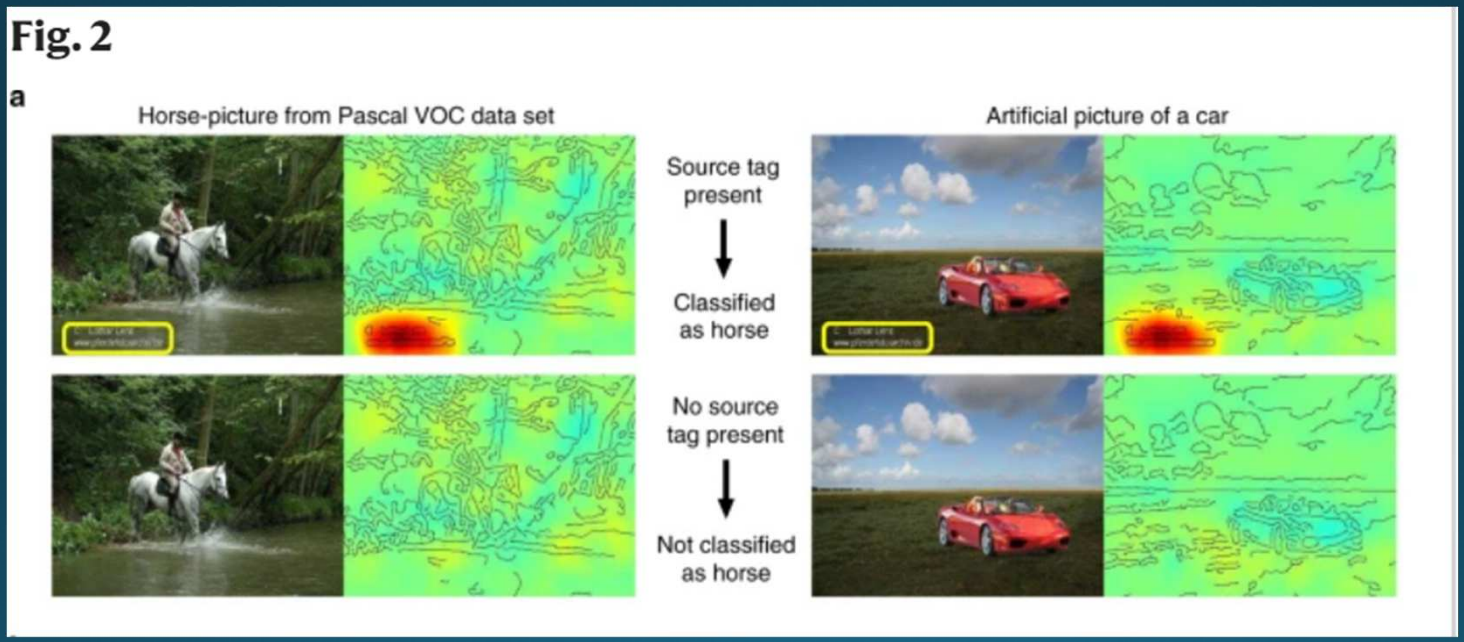


(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

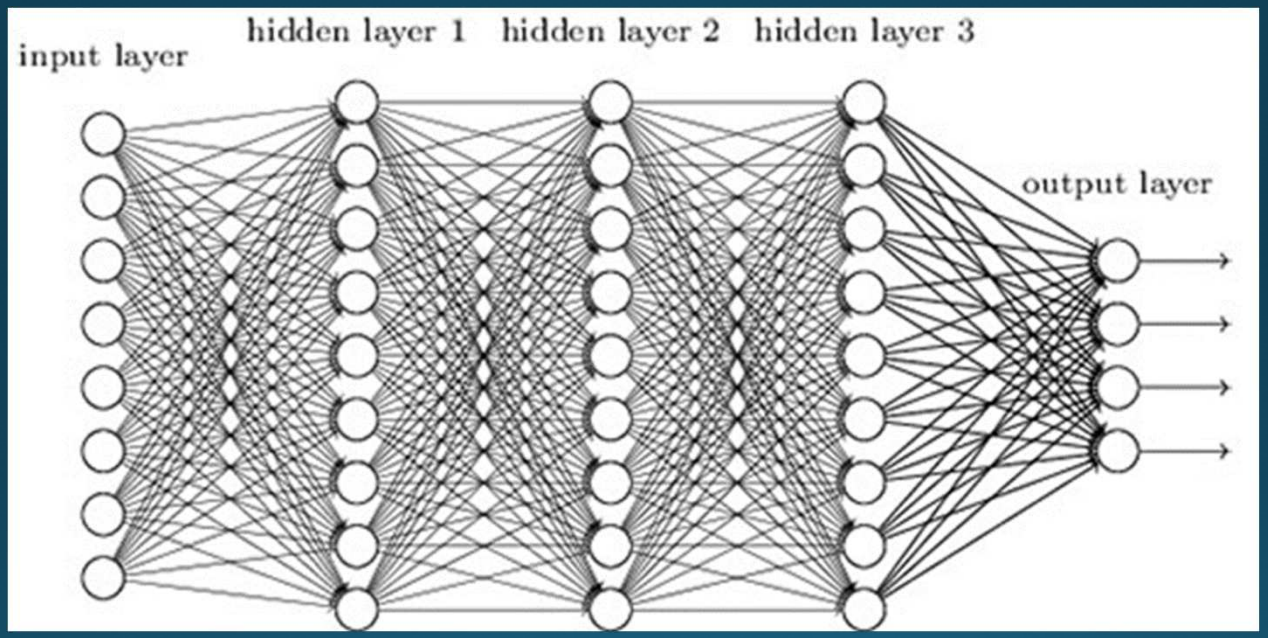
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938

Clever Hans



Lapuschkin, S., Wäldchen, S., Binder, A. et al. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 10, 1096 (2019). https://doi.org/10.1038/s41467-019-08987-4

Das Black-Box-Problem



Was geht da drinnen vor?

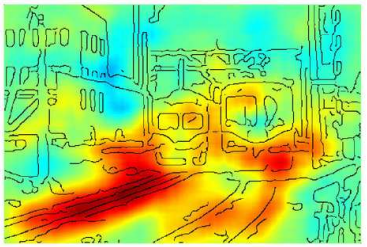
eXplainable AI (XAI)

Black-box Algorithmus

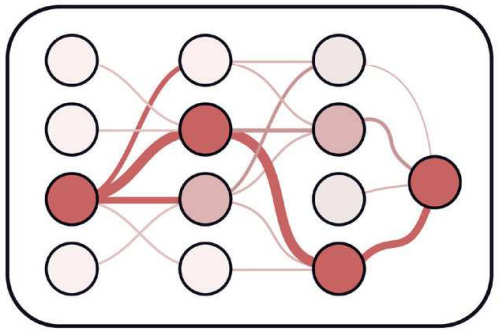
"Zug"

Aber, worauf schaut der Algorithmus?

Können wir sicher sein, dass der Zug korrekt erkannt wird?



Hier schaut der Algorithmus NICHT auf den Zug, sondern auf die Schienen!



Mit der Methode des Fraunhofer HHI und der TU Berlin kann visualisiert werden, worauf der black-box Algorythmus schaut

eXplainable AI (XAI)

- LIME

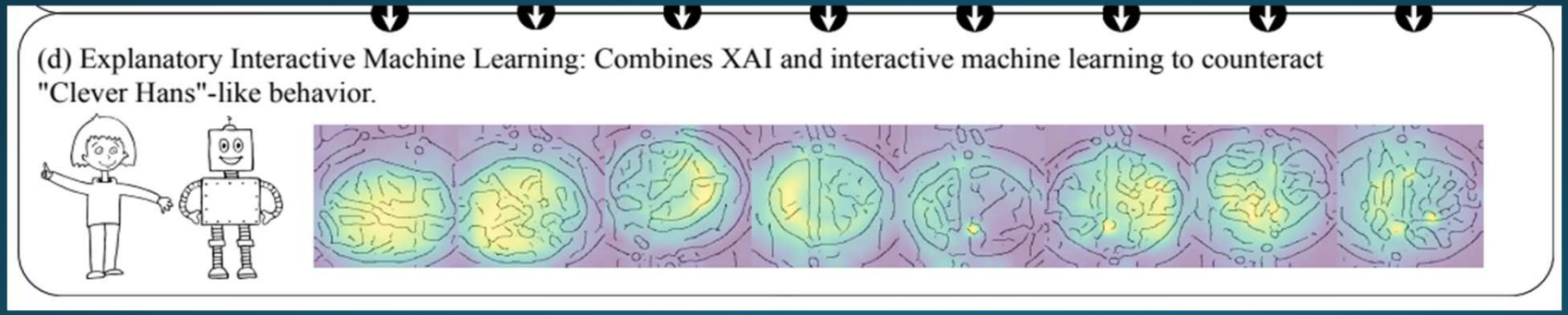
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$)**, "Acoustic guitar"** ($p = 0.24$) **and "Labrador"** ($p = 0.21$)

Marco Tulio Ribeiro: LIME - Local Interpretable Model-Agnostic Explanations

eXplainable AI (XAI)

- Auskunft über Sicherheit
- Lösungsweg
- Datenbasis
- Z.B.: Explanatory Interactive Learning



Patrick Schramowski et al (2020): Making deep neural networks right for the right scientific reasons by interacting with their explanations. arXiv:2001.05371 .

Herausforderungen

- Je besser die Ergebnisse, desto weniger kritisch sind die Nutzer*innen.
- Für viele Entscheidungen gibt es keine klaren Kriterien.
- Wie ist eine Überprüfung möglich?
- Wie wird KI in Entscheidungsprozesse eingebunden?