



Universität Augsburg
Fakultät für Angewandte
Informatik

Erklärbare Künstliche Intelligenz – Ziele, Methoden und Herausforderungen auf dem Weg zur Menschzentrierten KI

Katharina Weitz



Menschzentrierte
Künstliche Intelligenz
Institut für Informatik

16.02.2021

Wer bin ich?

Bilder: Katharina Weitz

Bamberg



Studium der
Psychologie & Informatik

40 Std.
191 km

 **Fraunhofer**
IIS


DUMMIES *Über*


Bilder: FELI/Universität Bamberg

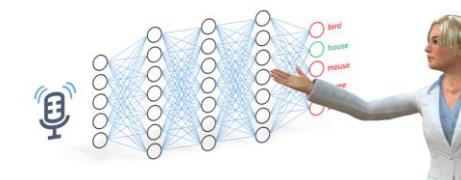

Bild: Katharina Weitz

Wissenschaftliche
Mitarbeiterin &
Doktorandin



UNIA Universität Augsburg University

Menschzentrierte Künstliche Intelligenz



Augsburg

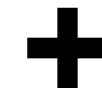


Motivation

Menschzentrierte KI

Menschen

Erklärbare Künstliche Intelligenz –
Ziele, Methoden und
Herausforderungen auf dem Weg zur
Menschzentrierten KI



Erklärbare KI

Erklärungen
über KI

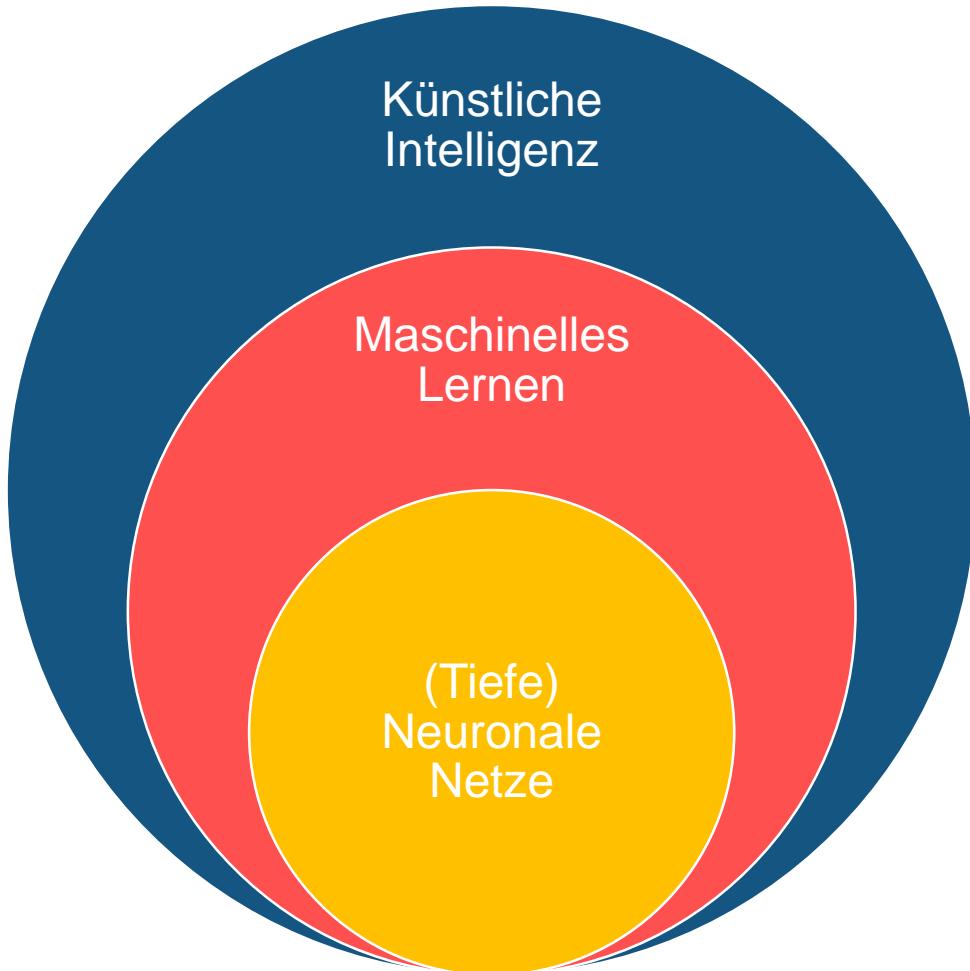
Motivation

Erklärbare Künstliche Intelligenz –
Ziele, Methoden und
Herausforderungen auf dem Weg zur
Menschzentrierten KI

Erklärbare KI

**Erklärungen
über KI**

Künstliche Intelligenz



Künstliche Intelligenz (Artificial Intelligence)

„[...] dass jeder **Aspekt des Lernens** oder jedes andere **Merkmal der Intelligenz** prinzipiell so genau beschrieben werden kann, dass eine **Maschine** dazu gebracht werden kann, ihn zu **simulieren**“

(McCarthy et al., 1955)

Maschinelles Lernen (Machine Learning)

Maschinelles Lernen bezeichnet Algorithmen, die sich durch Erfahrung automatisch verbessern.

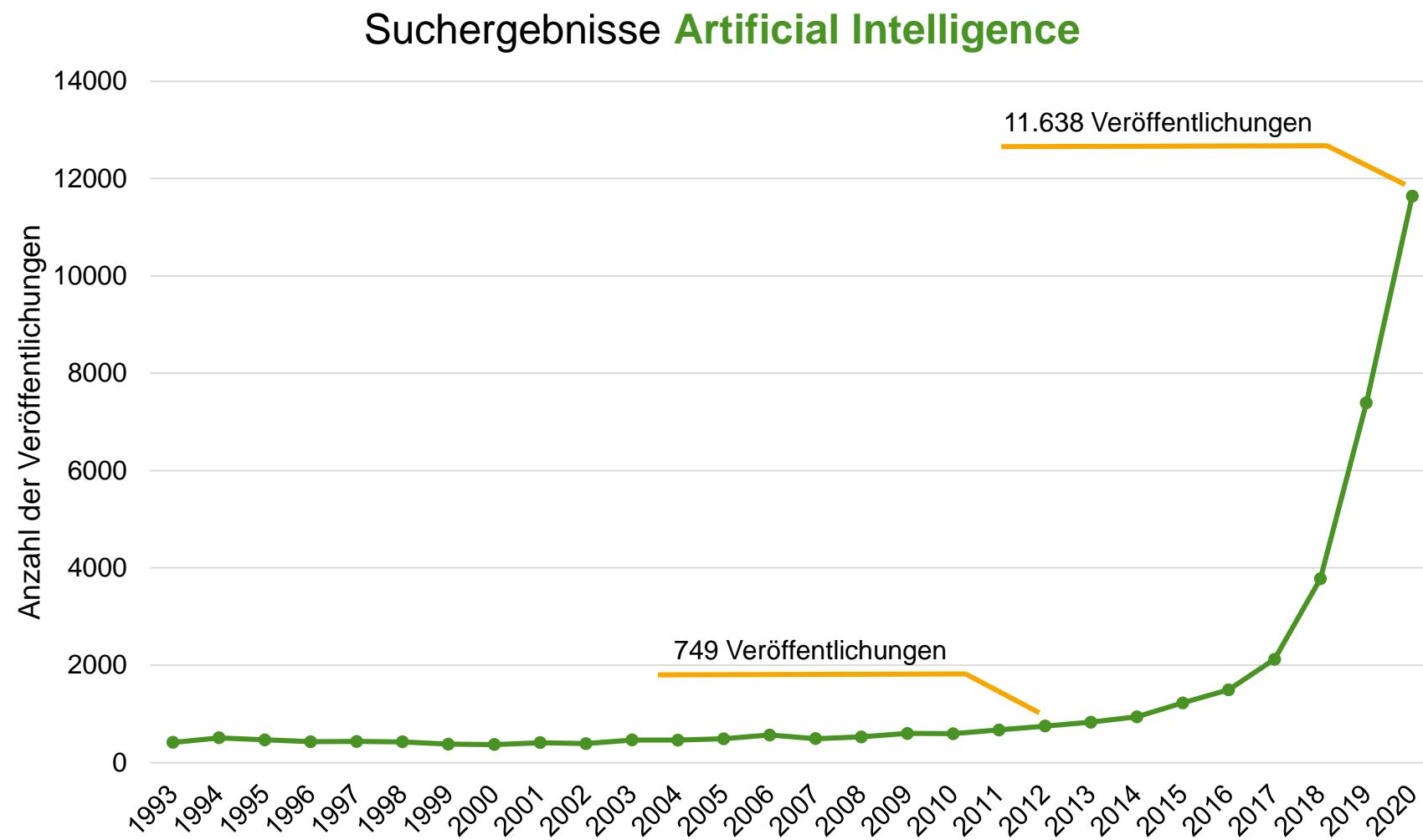
(Mitchell, 1997)

Tiefe Neuronale Netze (Deep Learning)

„Deep Learning ermöglicht es Computermodellen, die aus **mehreren Verarbeitungsschichten** bestehen, **Repräsentationen von Daten** mit mehreren Abstraktionsebenen zu lernen.“

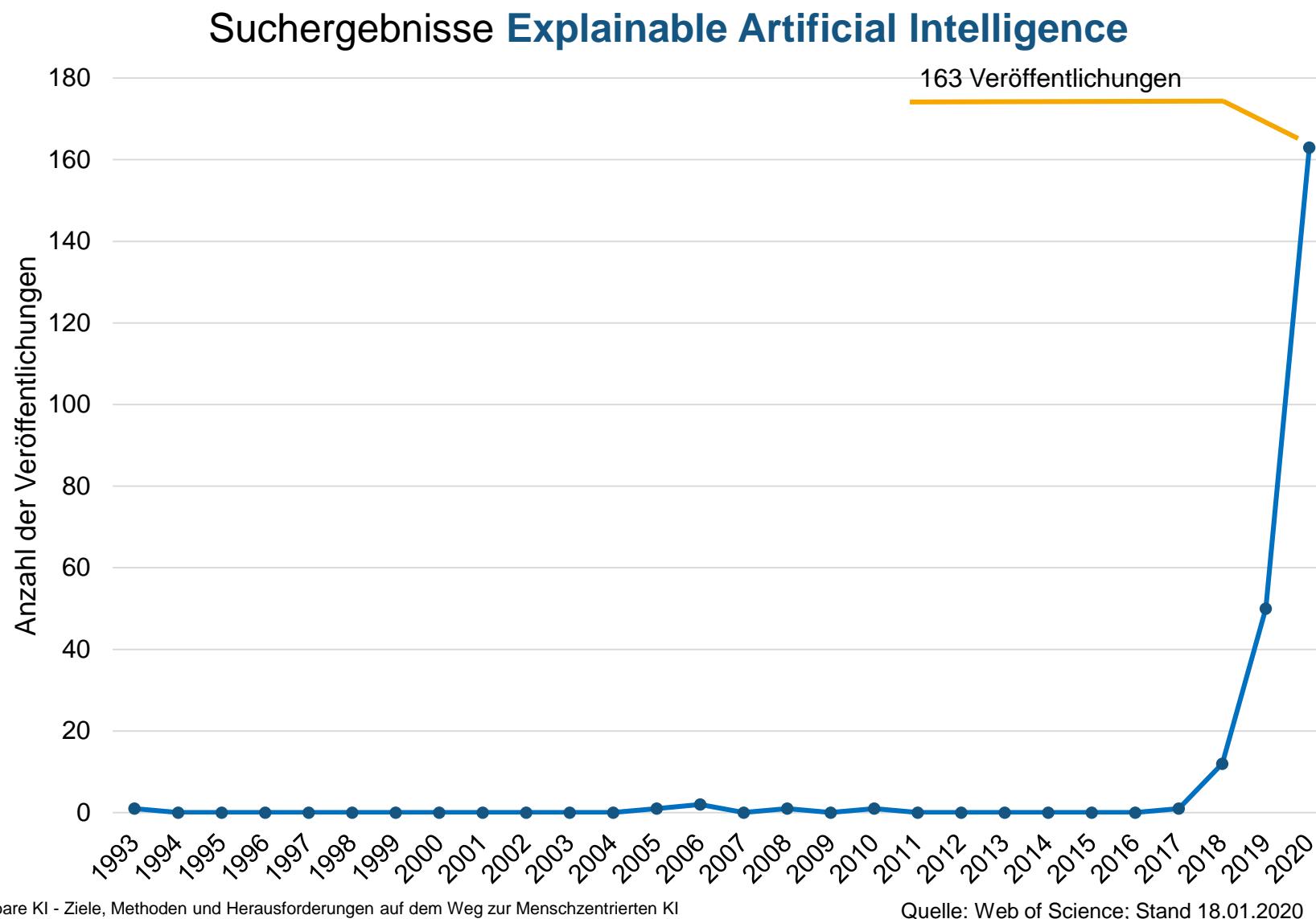
(LeCun et al., 2015)

Motivation



Quelle: Web of Science; Stand 18.01.2020

Motivation



Erklärbare KI

Definition Erklärbare KI (XAI)

“...ein Forschungsgebiet, das darauf abzielt, die **Ergebnisse von KI-Systemen** für **Menschen verständlicher** zu machen.“
(Adadi & Berrada, 2018)

“...bezeichnet jede Aktion oder Prozedur, die von einem Modell mit der Absicht durchgeführt wird, seine **internen Funktionen zu verdeutlichen** oder zu detaillieren.“

(Arrieta et al., 2020)

**Erklärungen
über KI**

Erklärbare KI

Ziele Erklärbarer KI

Interaktivität

Sicherheit

Fairness

Zugänglichkeit

Informativität

Bewusstsein für Datenschutz

uvm..

Vertrauenswürdigkeit

= Vertrauen darauf, dass sich ein Modell bei einem bestimmten Anwendungsfall **wie vorgesehen** verhält

(Arrieta et al., 2020)

Erklärungen über KI

XAI soll Nutzer:innen in die Lage versetzen zu wissen

- ob/wann/warum sie dem **KI-System vertrauen** und sich darauf verlassen können
- ob/wann/warum sie der **KI misstrauen** und sich entweder **nicht** oder **nur mit Vorsicht** auf sie verlassen sollte.

Erklärbare KI

Wofür ist XAI gedacht?

“XAI wird [...] es menschlichen Anwendern ermöglichen, die aufkommende Generation von KI Partnern zu **verstehen**, ihnen **angemessen zu vertrauen** und sie **effektiv zu verwalten**, während diese gleichzeitig ein hohes Leistungsniveau beibehalten“

(Adadi & Berrada, 2018; Gunning, 2017)

Die Idee, Maschinelles
Lernen erklärbar zu
gestalten, ist nicht neu!

MYCIN

(Shortliffe et al., 1975;
Shortliffe & Buchanan,
1975)

WENN

- 1) die Färbung des Organismus grampositiv ist,
und
- 2) die Morphologie des Organismus kokkus ist,
und
- 3) die Wachstumsform des Organismus
kettenförmig ist

DANN

Es gibt suggestive Hinweise (.7), dass die
Identität des Organismus **Streptokokken** sind

Diagnose

Therapievorschläge

Erklärungen über KI

Tiefe Neuronale Netze

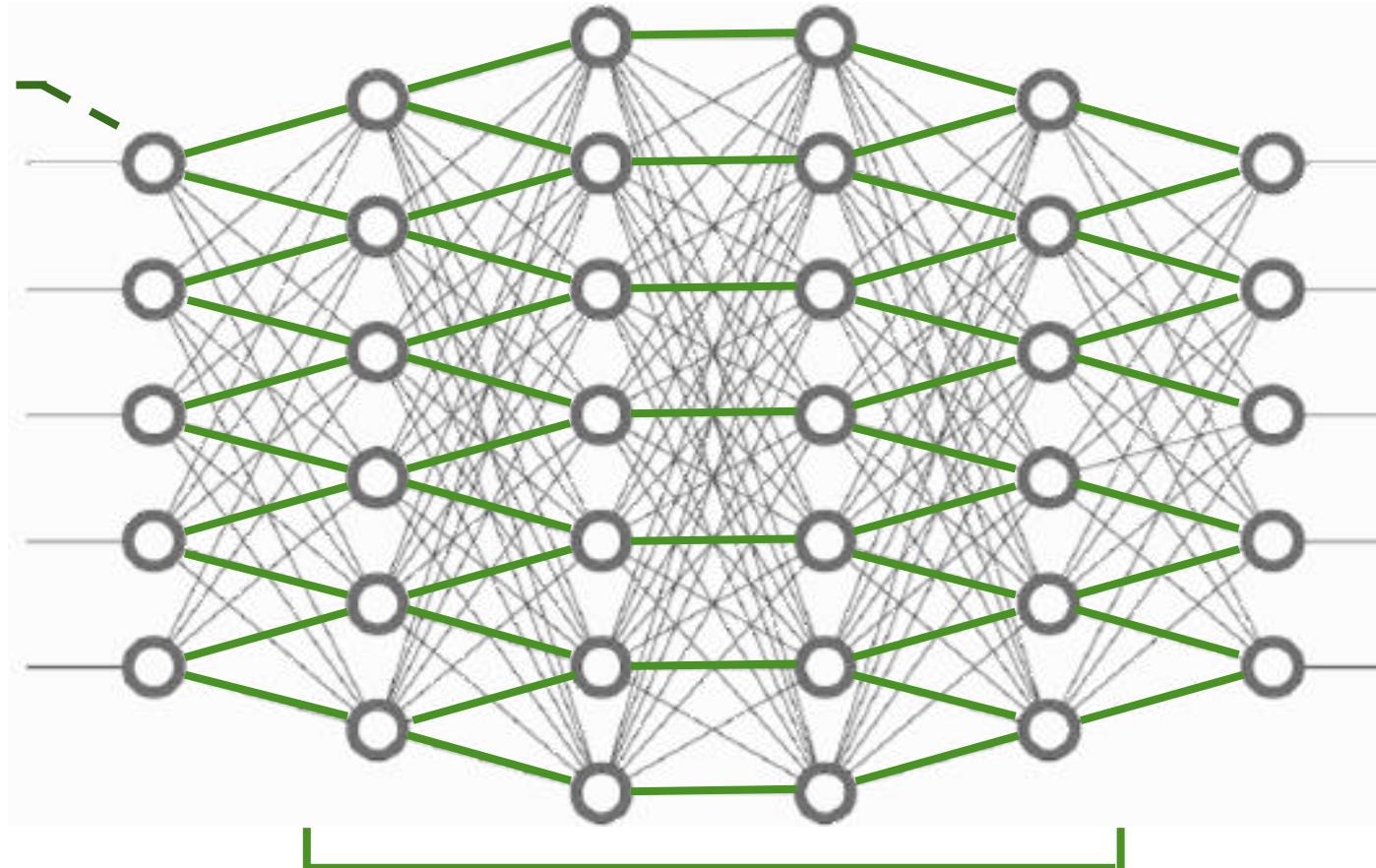
Deep Neural Networks (DNN)

Erklärbare KI



Bild: Katharina Weitz

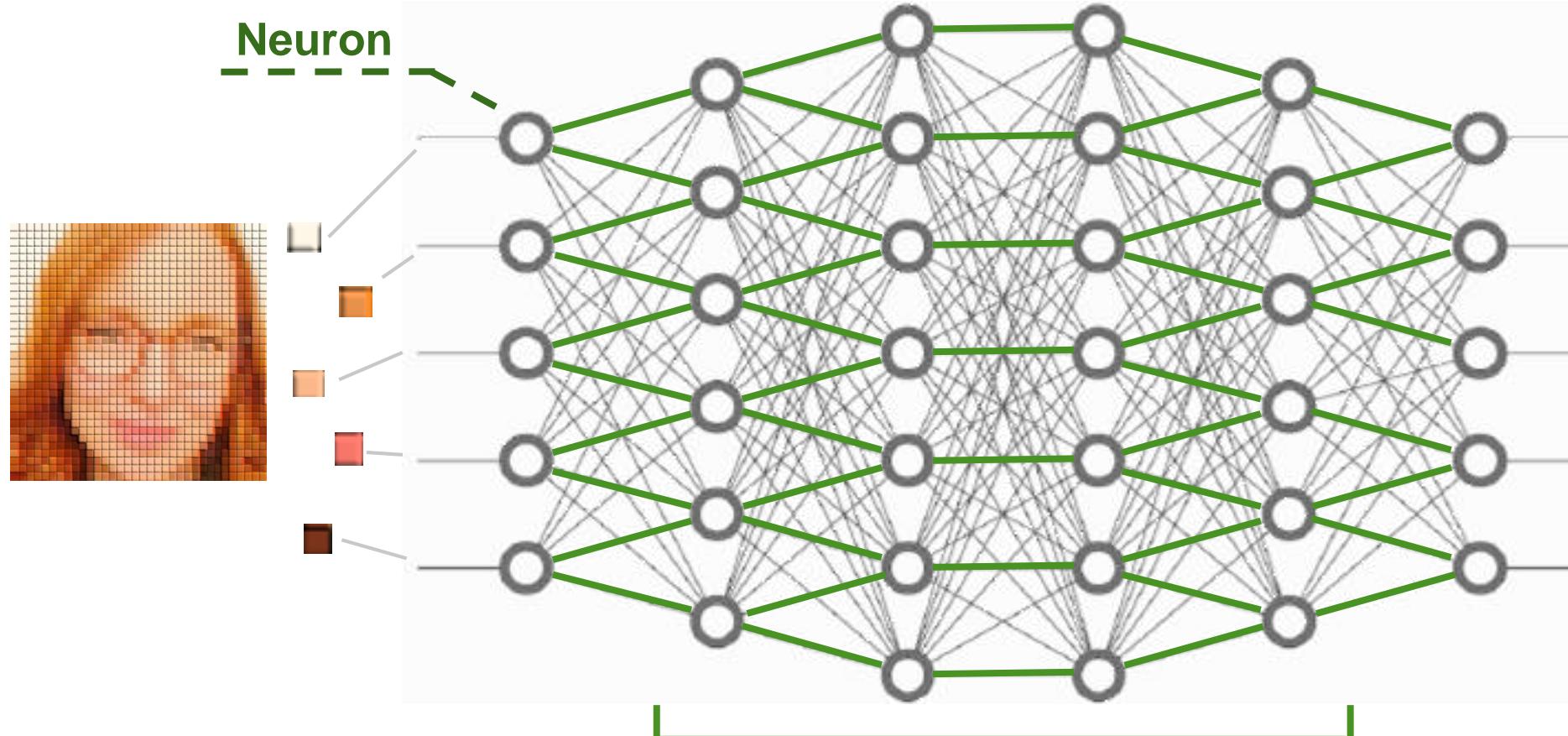
Neuron



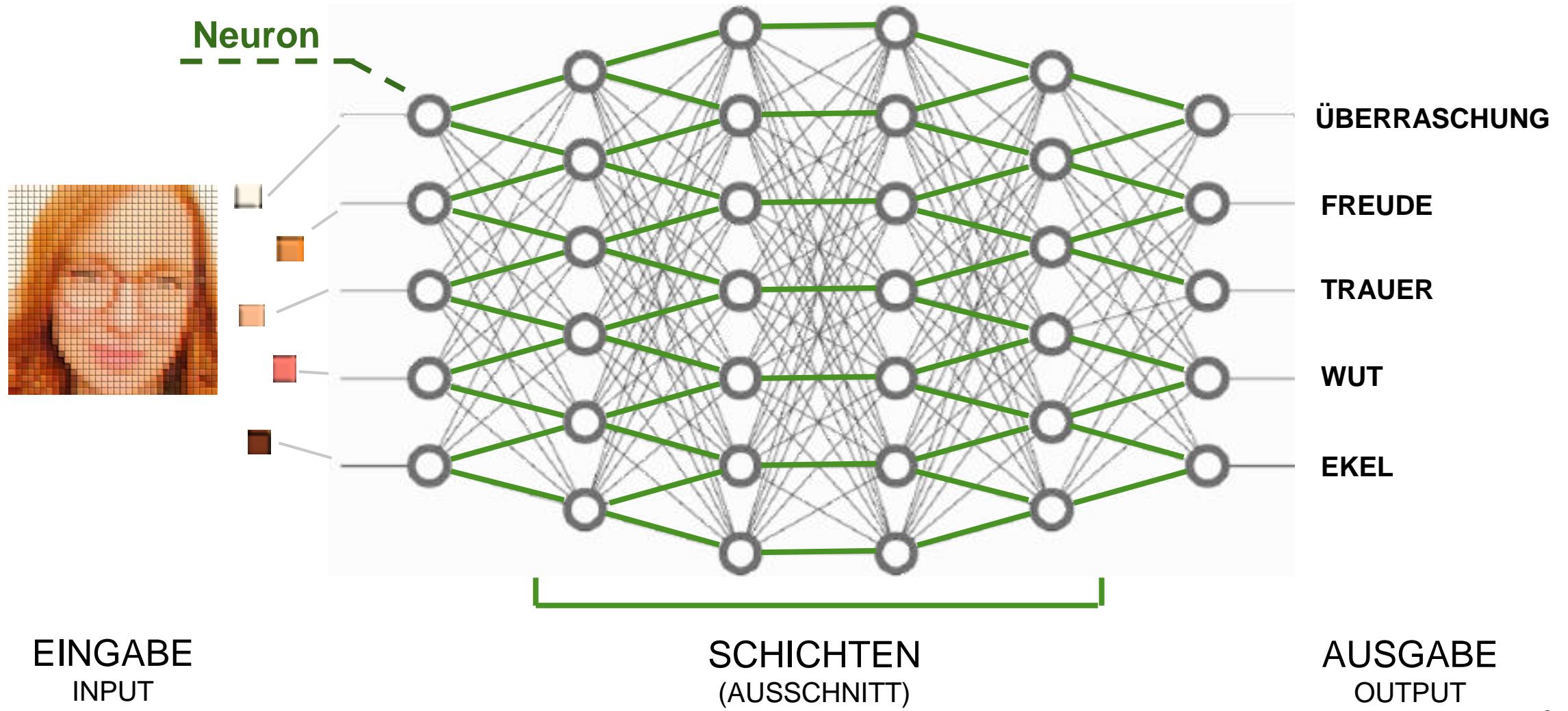
EINGABE
INPUT

SCHICHTEN
(AUSSCHNITT)

Erklärbare KI



Erklärbare KI

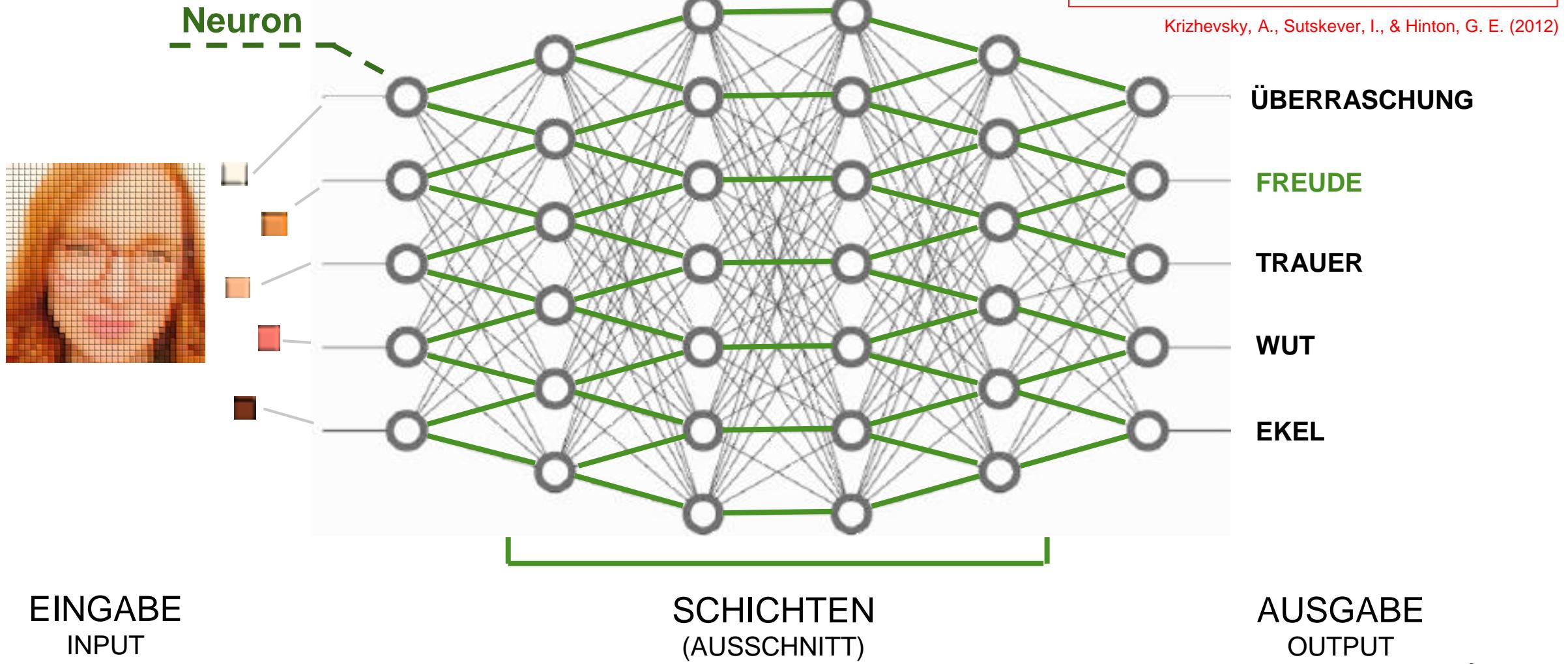


Erklärbare KI

500.000 Neurone

6.000.000 Parameter

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012)



Erklärbare KI



Bilder aus dem Vortrag: Explaining Black-Box Machine Learning Predictions - Sameer Singh

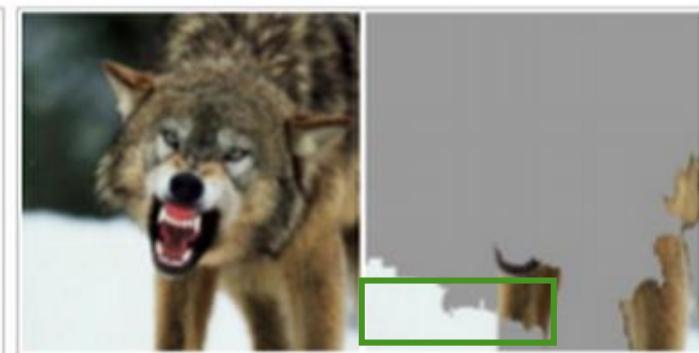
Erklärbare KI



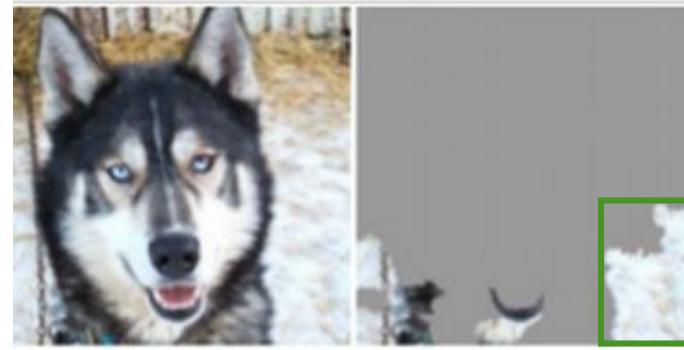
Predicted: **wolf**
True: **wolf**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**



Predicted: **wolf**
True: **husky**



Predicted: **husky**
True: **husky**



Predicted: **wolf**
True: **wolf**

Bilder aus dem
Vortrag: Explaining
Black-Box Machine
Learning Predictions -
Sameer Singh

Erklärbare KI

Beispiele



Eine Vielzahl von XAI Methoden

Intrinsische Erklärbarkeit

Erklärbarkeit, die sich aus der Gestaltung des ML Modells ergibt

Weniger komplexe Modelle verwenden, die interpretierbar bleiben

Post-Hoc Erklärbarkeit

Erklärbarkeit wird nach der Verwendung des ML Modells geschaffen, indem das Modell analysiert wird

Komplexe Modelle erklärbar machen

(Molnar, 2019)

Erklärbare KI

Beispiele

Eine Vielzahl von XAI Methoden

Post-Hoc Erklärbarkeit

Featurebasierte Erklärungen

- Layerwise Relevance Propagation (LRP)
- LIME
- SHAP

Beispielbasierte Erklärungen

- Counterfactuals

Regelbasierte Erklärungen

- Relationen

Konzeptuelle Erklärungen

- Konzeptaktivierungsvektoren (TCAV)

Strategieerklärungen

- HIGHLIGHTS Algorithmus

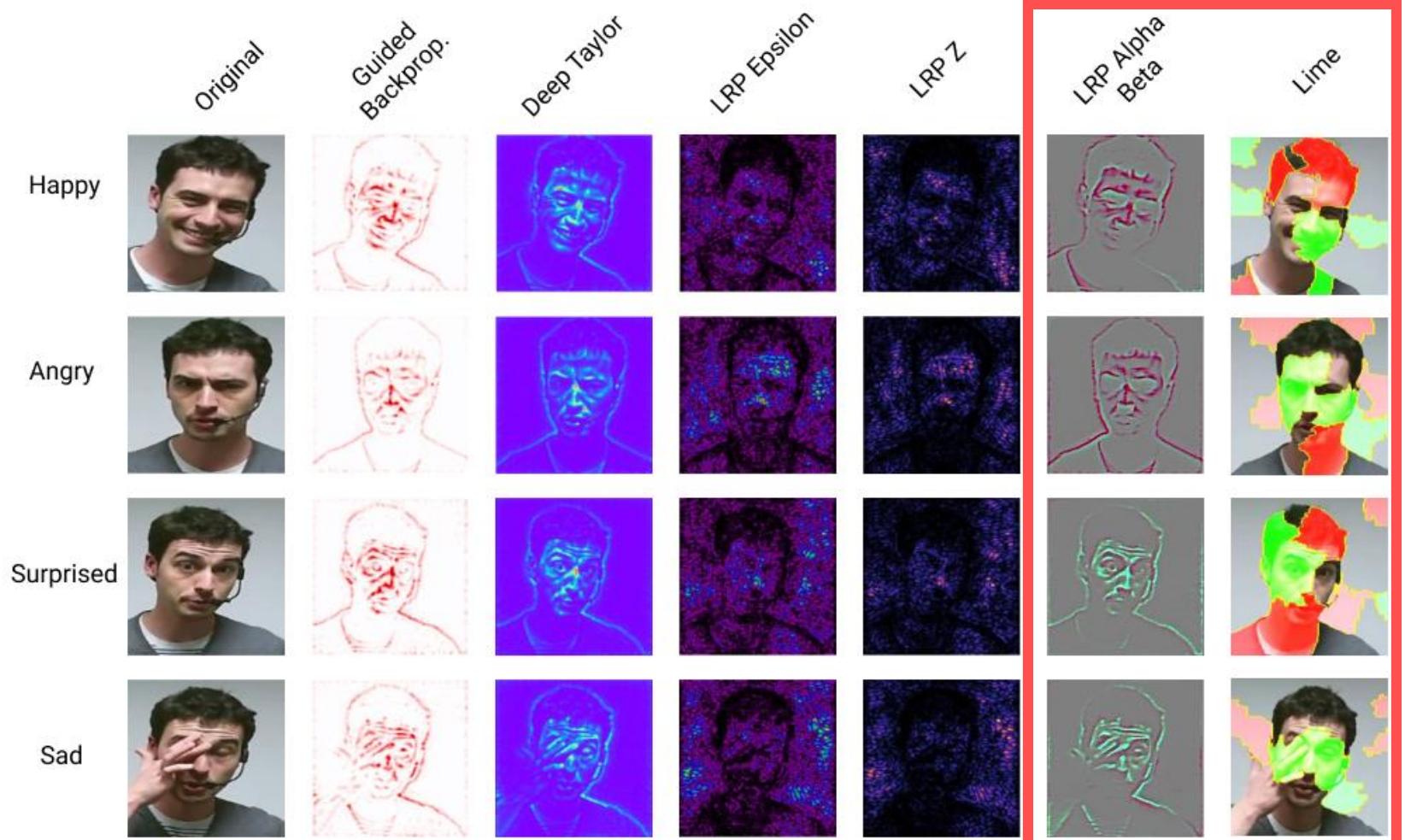
Erklärbarkeit wird nach der Verwendung des ML Modells geschaffen, indem das Modell analysiert wird

Komplexe Modelle erklärbar machen

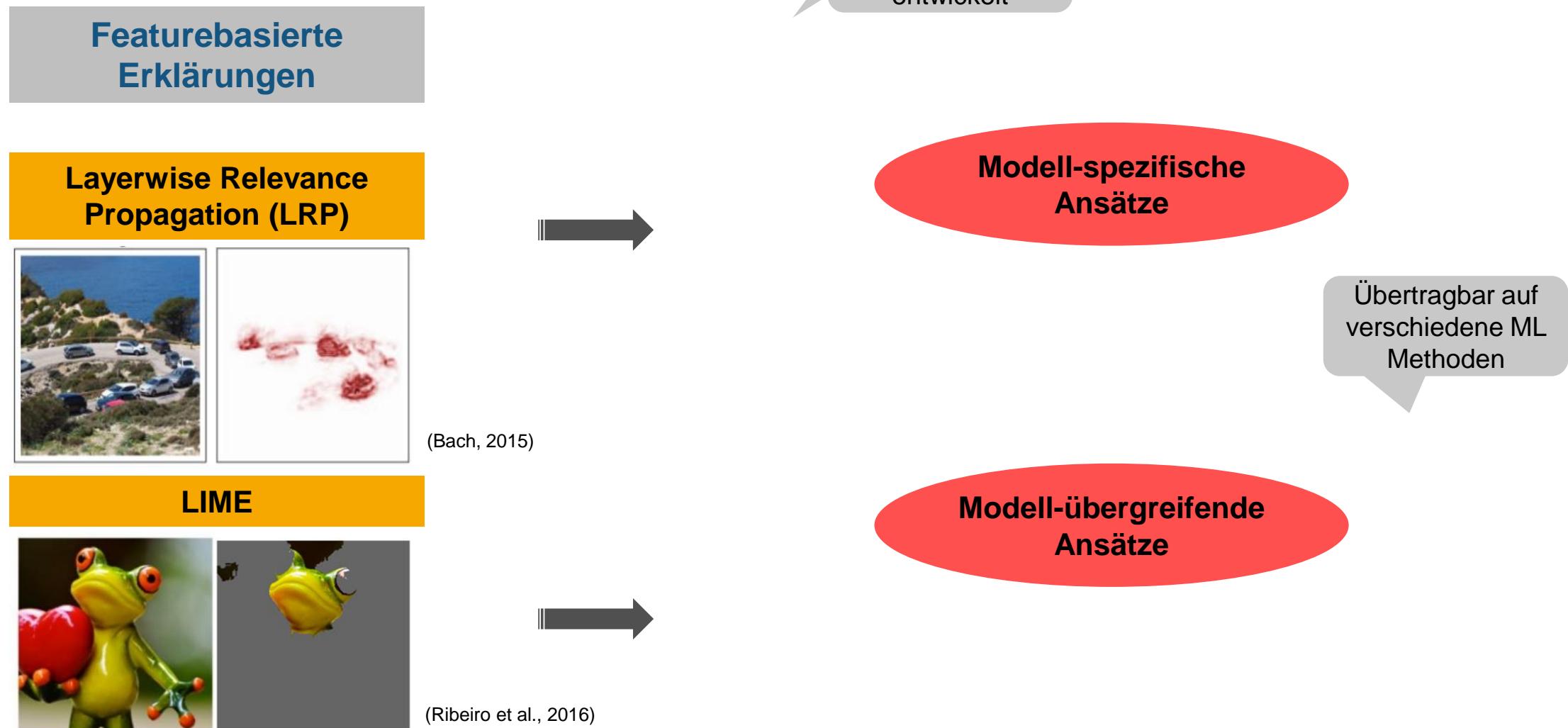
Erklärbare KI

Featurebasierte Erklärungen

Methoden, die die
**Wichtigkeit von
Features**
(z.B. Pixeln) hervorheben



Erklärbare KI

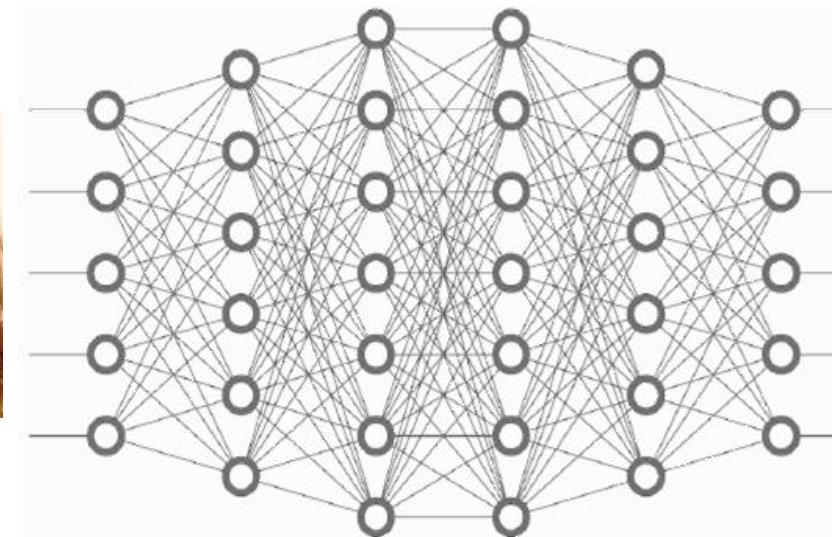


Erklärbare KI

Featurebasierte Erklärungen

Layerwise Relevance Propagation (LRP)

(Bach et al., 2015; Kohlbrenner, 2017;
Lapuschkin et al., 2017)



ÜBERRASCHUNG

FREUDE

TRAUER

WUT

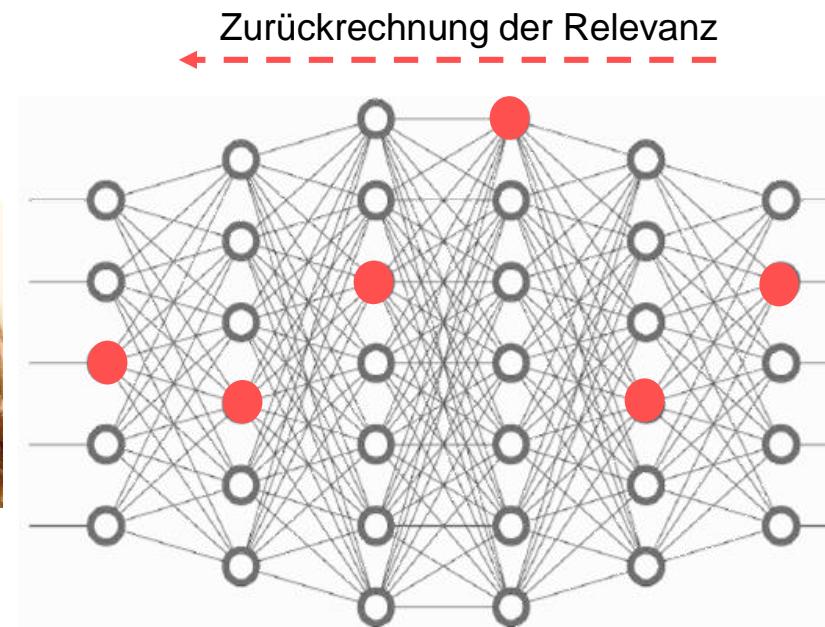
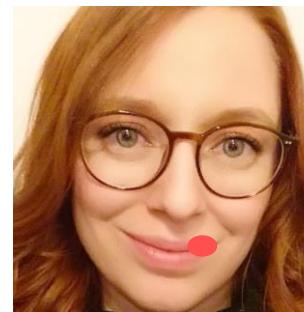
EKEL

Erklärbare KI

Featurebasierte Erklärungen

Layerwise Relevance Propagation (LRP)

(Bach et al., 2015; Kohlbrenner, 2017;
Lapuschkin et al., 2017)



ÜBERRASCHUNG

FREUDE

TRAUER

WUT

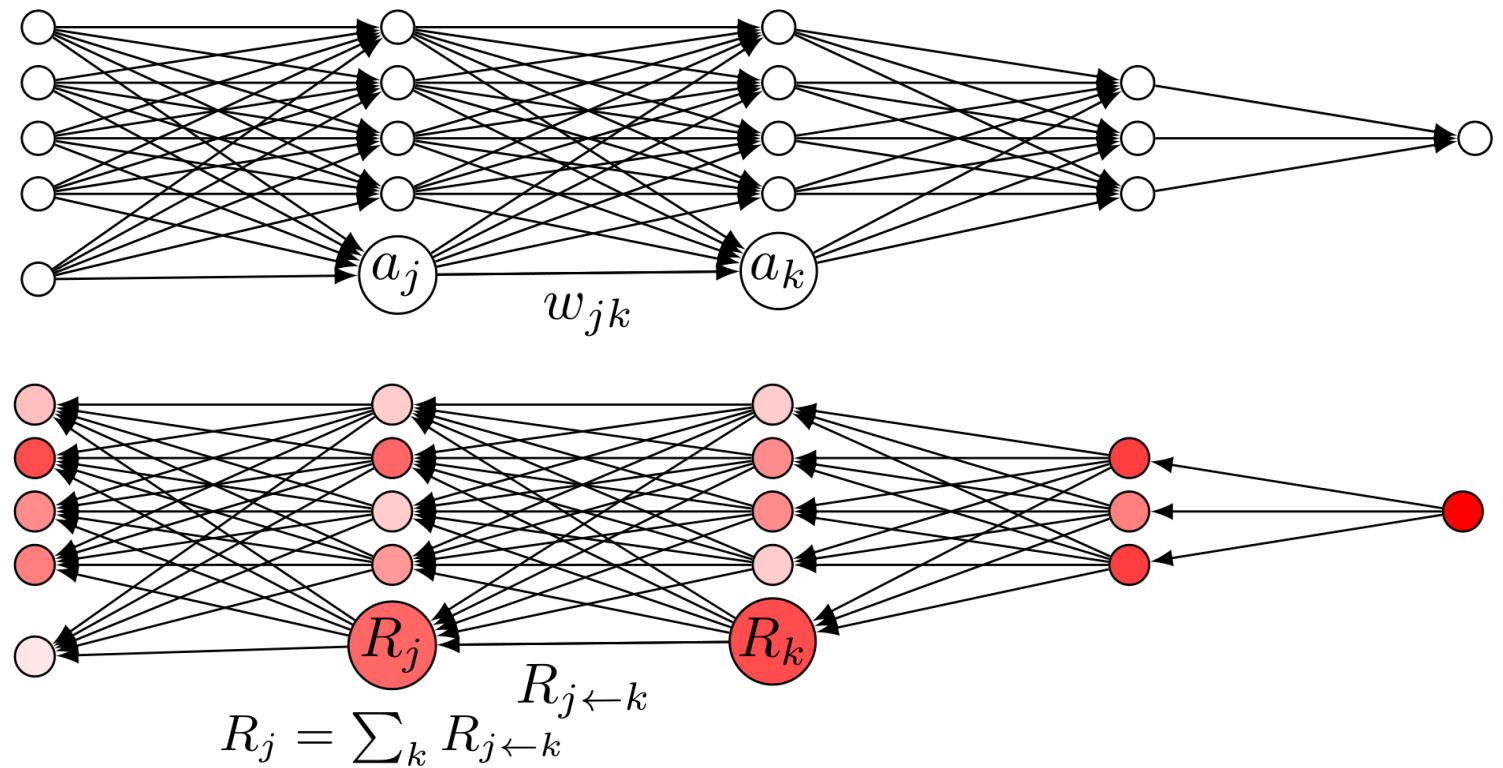
EKEL

Erklärbare KI

Featurebasierte Erklärungen

Layerwise Relevance Propagation (LRP)

(Bach et al., 2015; Kohlbrenner, 2017;
Lapuschkin et al., 2017)



Bilder: Huber et al. (2019)

Erklärbare KI

Featurebasierte Erklärungen

Layerwise Relevance Propagation (LRP)

(Bach et al., 2015; Kohlbrenner, 2017;
Lapuschkin et al., 2017)

AlphaBeta Regel (ohne Bias) ist definiert als

$$R_{j \leftarrow k} = \left(\alpha \frac{(a_j w_{jk})^+}{\sum_j (a_j w_{jk})^+} + \beta \frac{(a_j w_{jk})^-}{\sum_j (a_j w_{jk})^-} \right) R_k$$

mit

$$(a_j w_{jk})^+ := \begin{cases} a_j w_{jk}, & a_j w_{jk} > 0 \\ 0, & a_j w_{jk} \leq 0 \end{cases}$$

$(a_j w_{jk})^-$ wird analog dazu berechnet.

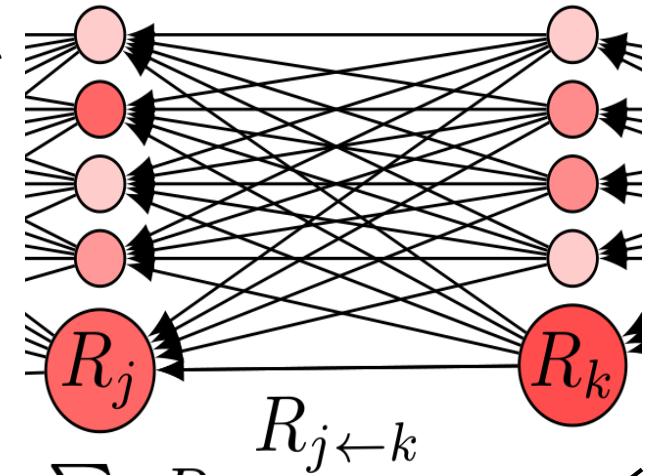


Bild: Huber et al. (2019)

Erklärbare KI

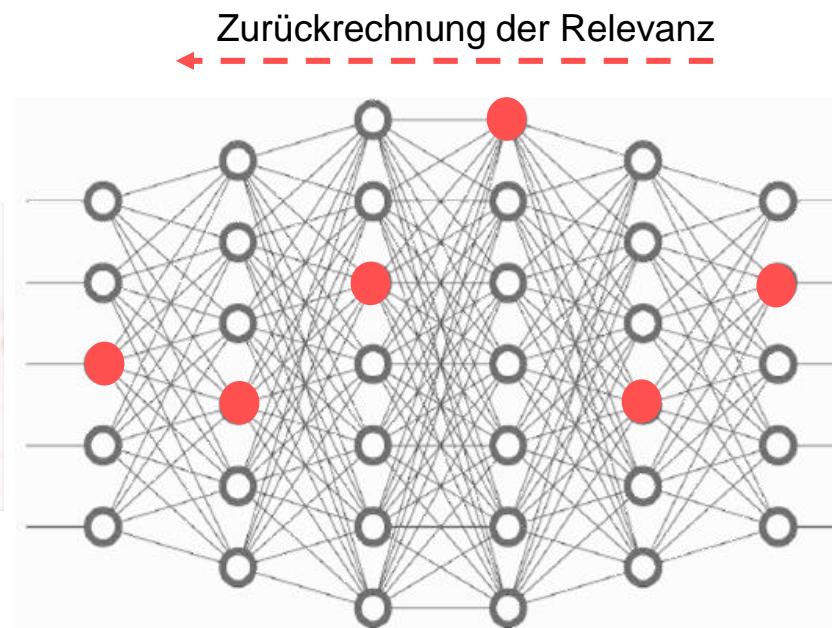
Featurebasierte Erklärungen

Layerwise Relevance Propagation (LRP)

(Bach et al., 2015; Kohlbrenner, 2017;
Lapuschkin et al., 2017)



Bild: Katharina Weitz



ÜBERRASCHUNG

FREUDE

TRAUER

WUT

EKEL

Erklärbare KI

Featurebasierte Erklärungen

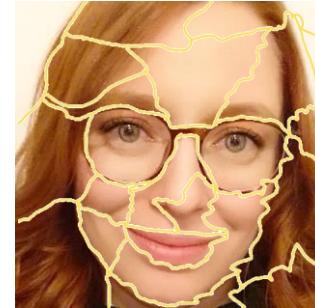
LIME

(Ribeiro et al., 2016)



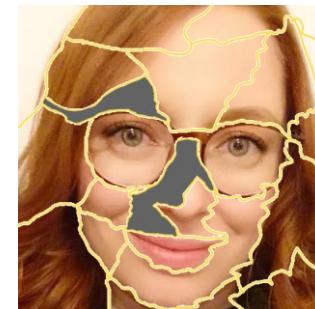
Originalbild

Klassifikation:
Freude
90%

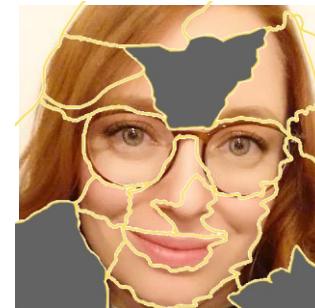


Segmentierung in
Superpixel

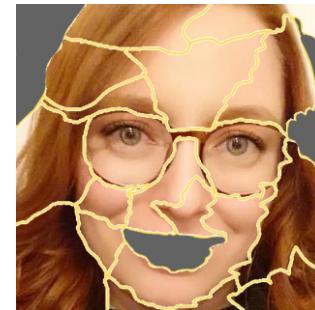
„Störungs“-Bilder (Perturbed Images)



Klassifikation:
Freude
20%



Klassifikation:
Freude
80%



Klassifikation:
Freude
50%

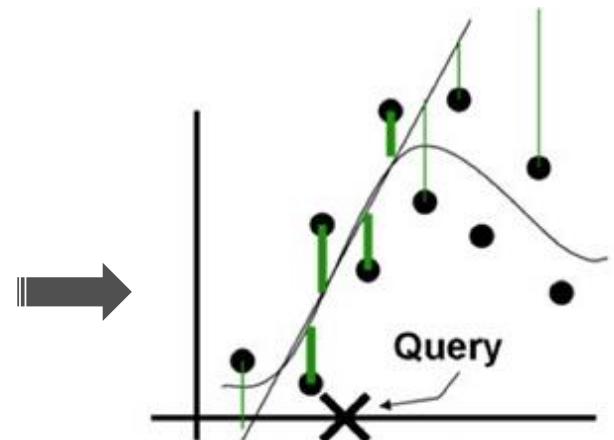


Bild: Ribeiro et al. (2016)

Bilder: Katharina Weitz; Darstellung orientiert an Ribeiro et al. (2016)

Erklärbare KI

Featurebasierte Erklärungen

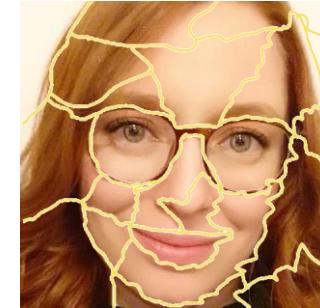
LIME

(Ribeiro et al., 2016)



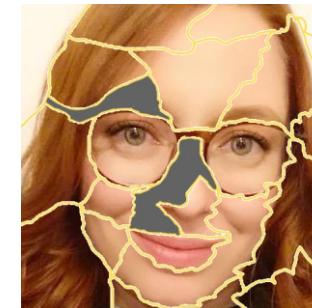
Originalbild

Klassifikation:
Freude
90%

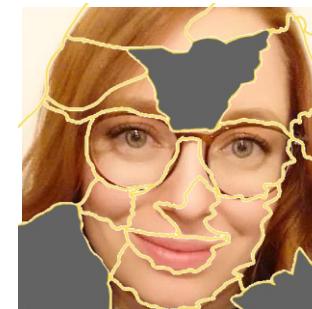


Segmentierung in
Superpixel

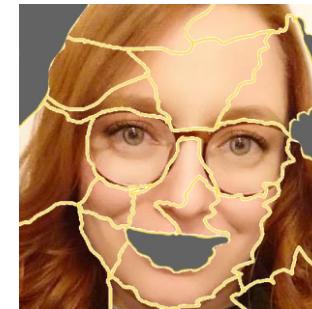
„Störungs“-Bilder (Perturbed Images)



Klassifikation:
Freude
20%



Klassifikation:
Freude
80%



Klassifikation:
Freude
50%



Bilder: Katharina Weitz; Darstellung orientiert an Ribeiro et al. (2016)

Erklärbare KI

Beispielbasierte Erklärungen
(z.B. Counterfactuals)

Regelbasierte Erklärungen
(z.B. zum Beschreiben von Relationen)

Original



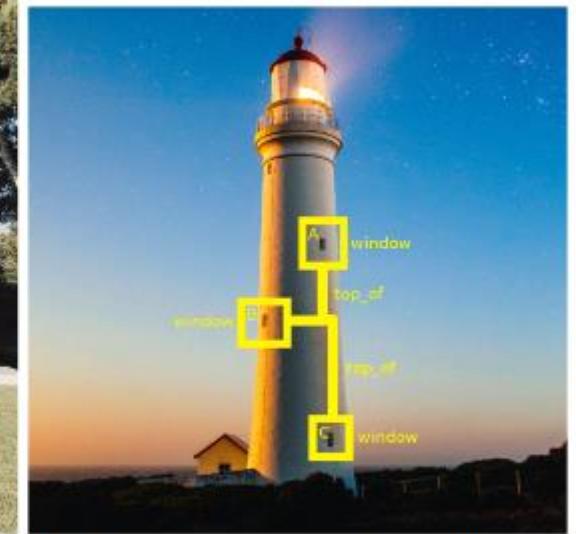
Umgewandelt



Bild: Zhu et al. (2017)



(a) A house, because three windows left
of each other.



(b) A tower, because three windows on
top of each other.

Bild: Rabold et al. (2019)

Erklärbare Künstliche Intelligenz –
Ziele, Methoden und
Herausforderungen auf dem Weg zur
Menschzentrierten KI

Erklärbare KI

**Erklärungen
über KI**

Menschzentrierte KI

Menschzentrierte KI

Erklärbare Künstliche Intelligenz –
Ziele, Methoden und
Herausforderungen auf dem Weg zur
Menschzentrierten KI



Menschen

Menschzentrierte KI

Menschzentrierte KI



Menschzentrierte KI

= Perspektive auf Künstliche Intelligenz und Maschinelles Lernen, die besagt, dass intelligente Systeme mit dem **Bewusstsein entworfen** werden müssen, dass sie **Teil eines größeren Systems** sind.

(Riedl, 2019)

Menschliche Akteur:innen



Menschzentrierte KI

Menschen



Quelle: <https://www.iwd.de/artikel/kuenstliche-intelligenz-angst-vor-dem-unbekannten-448593/>

Technologischer Fortschritt

Kein Vertrauen in KI? Das steckt hinter der Angst vor Künstlicher Intelligenz

Quelle: <https://www.maschinenmarkt.vogel.de/kein-vertrauen-in-ki-das-steckt-hinter-der-angst-vor-kuenstlicher-intelligenz-a-876226/>

INTERVIEW MIT KI-EXPERTEN

03.06.2020, 11:20 Uhr

Google, Alexa und Co.: Beherrschen wir noch die Maschine – oder hat uns die KI schon im Griff?

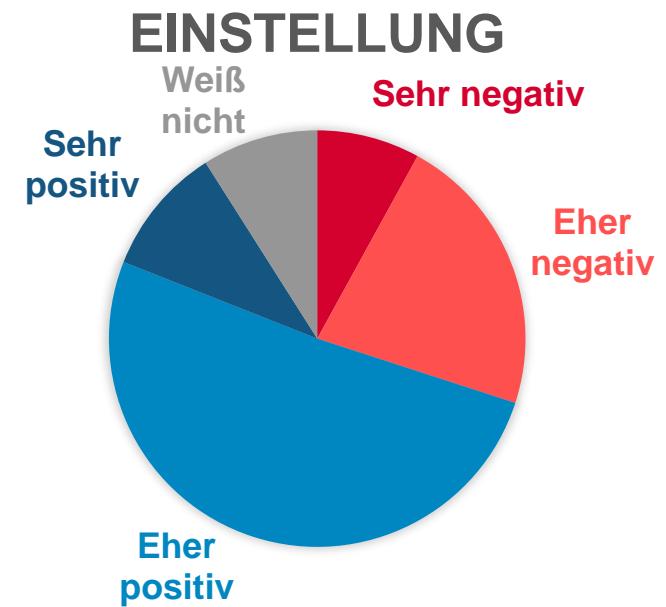
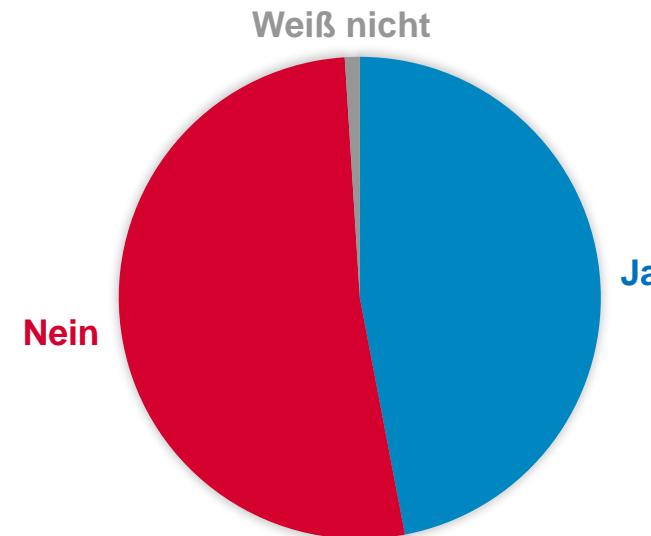
Quelle: <https://www.ingenieur.de/technik/fachbereiche/itk/google-alexand-co-warum-wir-ki-assistenten-nicht-immer-trauen-koennen-und-was-sich-in-zukunft-aendern-muss/>

Menschzentrierte KI

Einstellungen gegenüber KI (Eurobarometer, 2017)



ETWAS ÜBER KI GEHÖRT



Menschzentrierte KI

Einstellungen gegenüber KI (Eurobarometer, 2017)



ETWAS ÜBER KI GEHÖRT

Weiß nicht

Die **Meinung zu KI** hängt stark von der Auseinandersetzung mit Informationen/Wissen ab.

Befragte, die in den letzten 12 Monaten etwas über KI gehört/gelesen/gesehen haben, haben mit höherer Wahrscheinlichkeit eine **positive Einstellung** zu künstlicher Intelligenz und Robotern
(75% vs. 49%, die keine Berührungspunkte mit KI hatten)

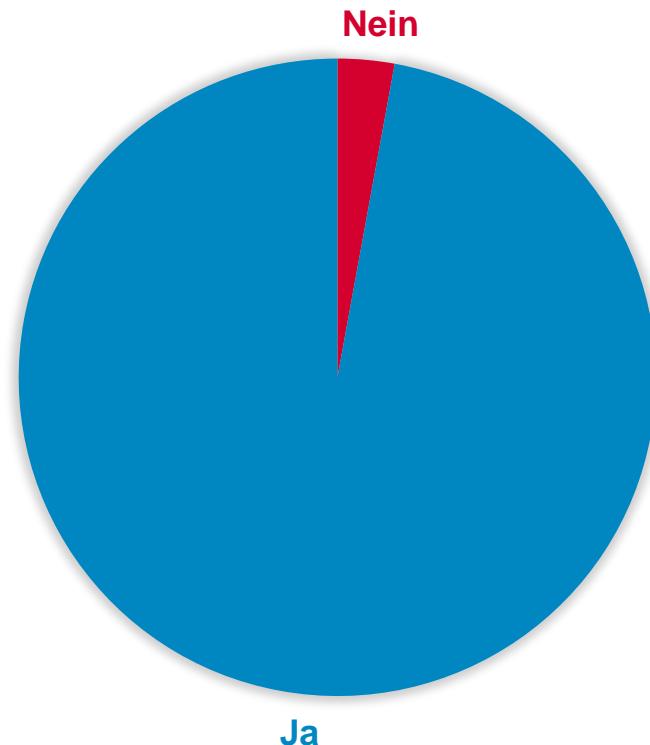
Eher positiv

Menschzentrierte KI

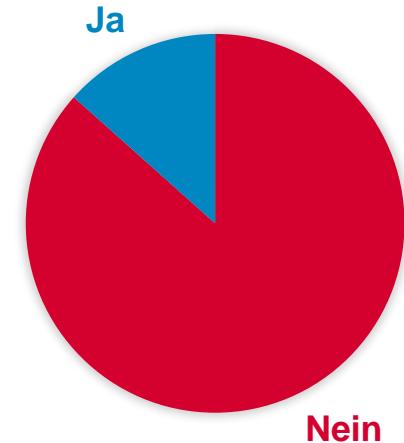
Deutschland: Einstellung gegenüber KI (Unveröffentlichte Daten; Weitz, 2020)



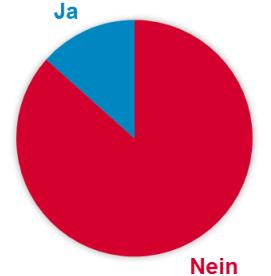
ETWAS ÜBER KI GEHÖRT



ETWAS ÜBER XAI
GEHÖRT



ETWAS ÜBER XAI
GEHÖRT



Was wir wissen

“Die Meinung zu KI hängt **stark von der Auseinandersetzung mit Informationen/Wissen ab.**”

Erklärungen für Menschen (Miller, 2019)

- Erklärungen sind vergleichend
- Erklärungen sind ausgewählt
- Wahrscheinlichkeiten sind nicht so wichtig
- Erklärungen sind sozial

Menschzentrierte KI



Was wir nicht wissen

Selbstwirksamkeits-
erwartung

Mentale
Modelle

Vertrauen

Leistung

Kognitive
Belastung

Erklärbare KI



Menschzentrierte KI



Mentale Modelle über KI

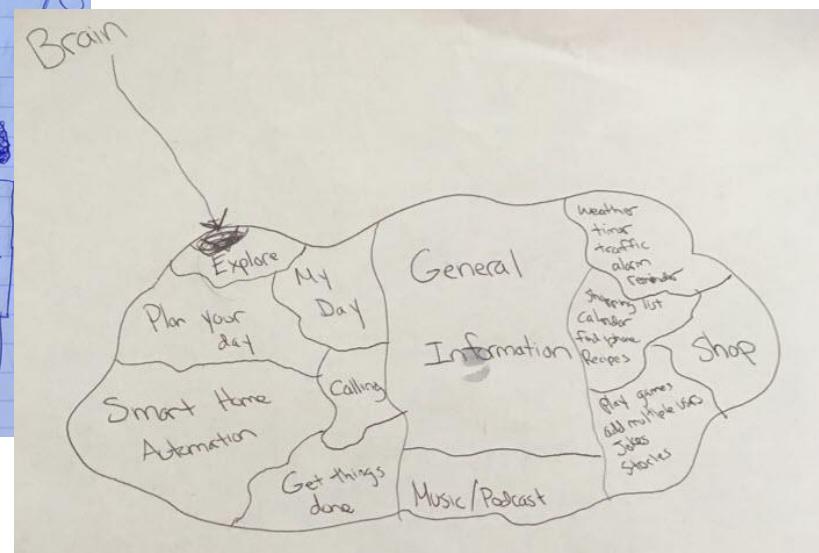
Mentales Modell

= ist die **kognitive Repräsentation**, die ein Benutzer über ein komplexes Modell hat
(Halasz & Moran, 1983; Norman, 2014)



Mentales Modell über Siri

Mentales Modell über Google Home



Bilder: Budidu (2019)

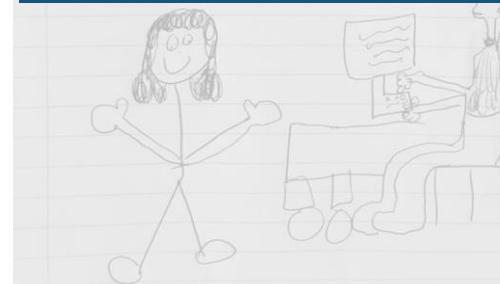
Menschzentrierte KI

Menschen

Mentale Modelle über KI

Mental Model

Korrekte Mentale Modelle sind die Grundlage,
um **angemessenes Vertrauen** zu entwickeln!



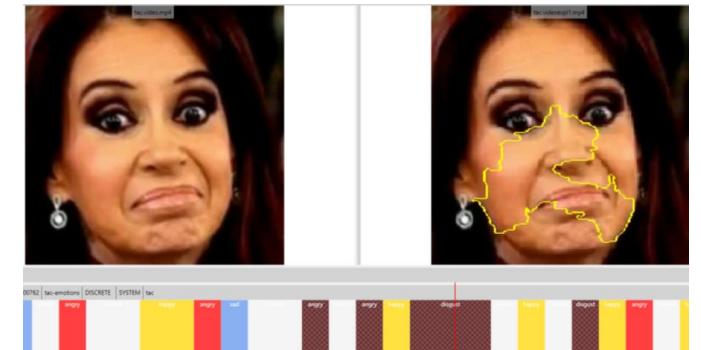
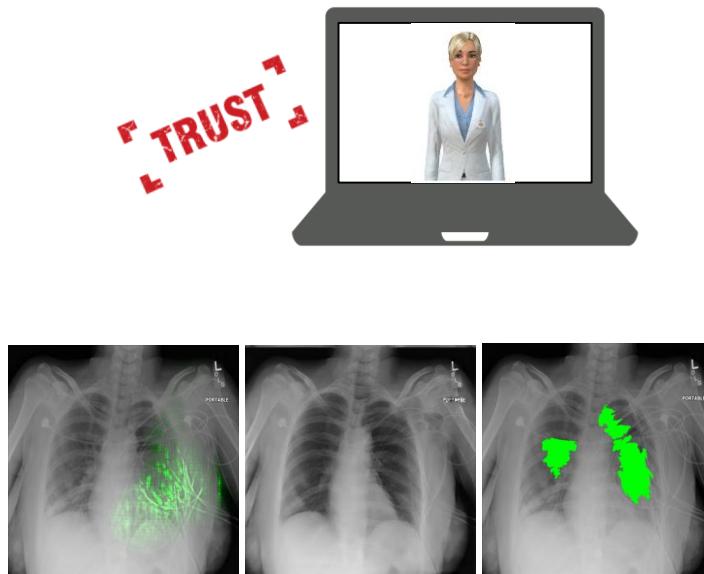
Mentales Model über
Siri



Bilder: Budidu (2019)

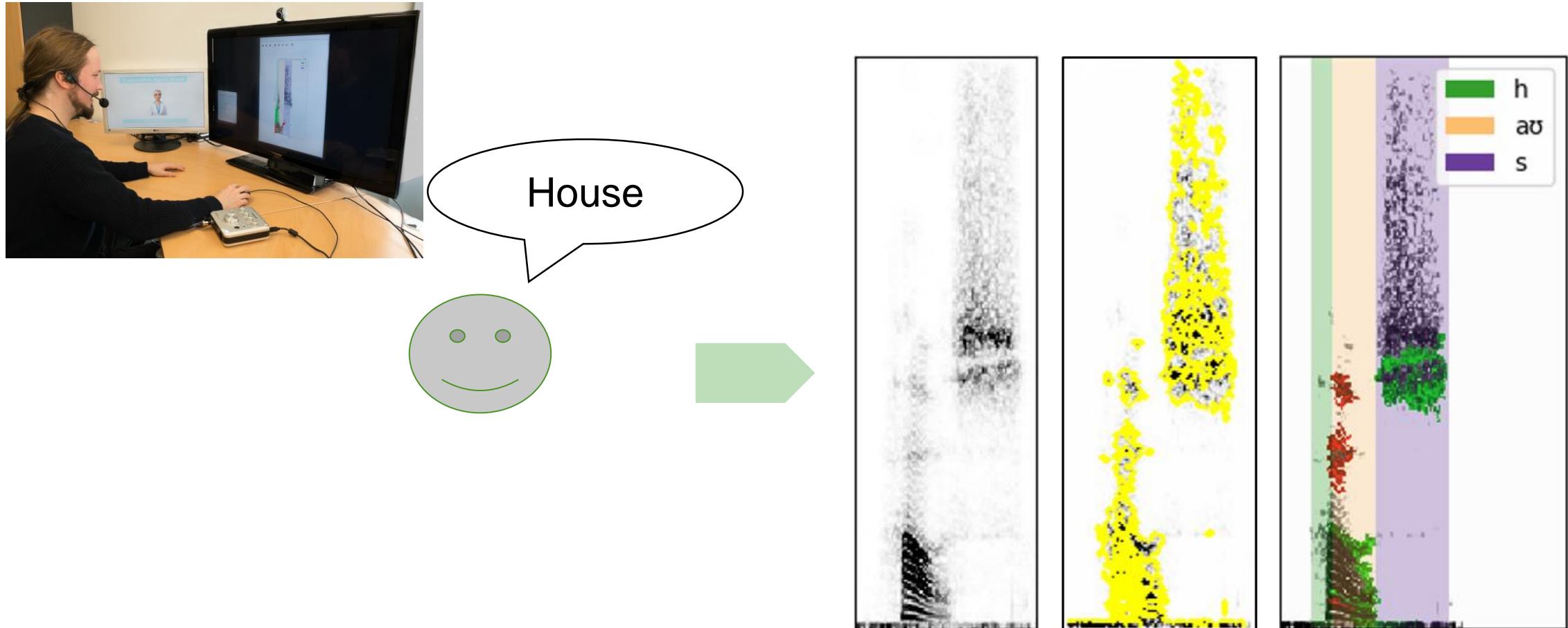
Forschung an der Schnittstelle Mensch-Maschine

Welchen **Einfluss** haben Erklärungen über die Funktionsweise und die Entscheidungen von KI auf Menschen?



Forschung an der Schnittstelle Mensch-Maschine

Weitz et al. (2020)
Weitz et al. (2019)



Forschung an der Schnittstelle Mensch-Maschine

Weitz et al. (2020)
Weitz et al. (2019)

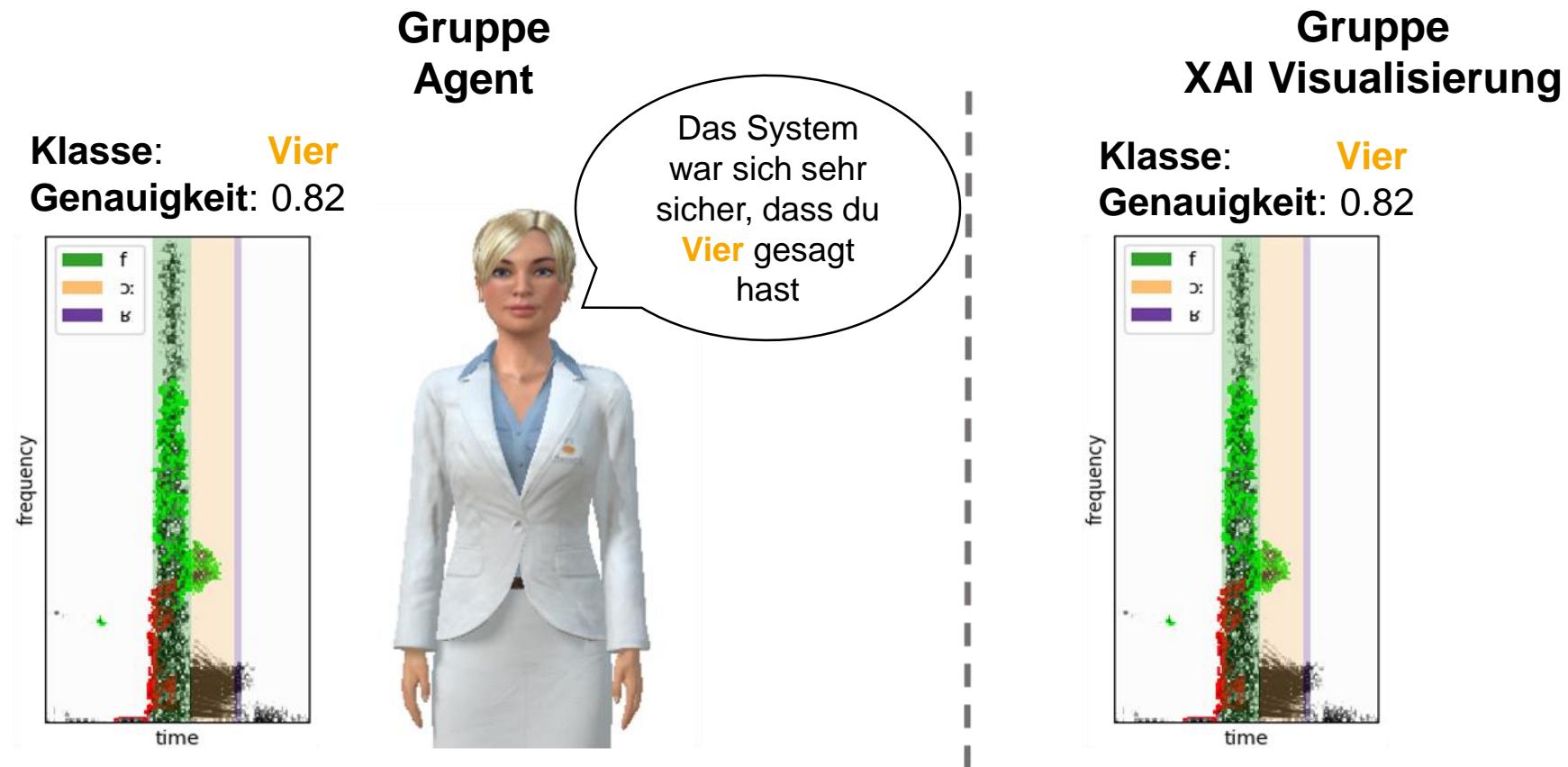
Teilnehmer:innen:	60
Alter:	26.53 Jahre
Geschlecht:	45 männlich 15 weiblich



Gruppe/ Modalität	XAI Visualisierung	Textinformation	Gesprochene Information	Virtueller Agent
1	✓			
2	✓	✓		
3	✓		✓	
4	✓		✓	✓

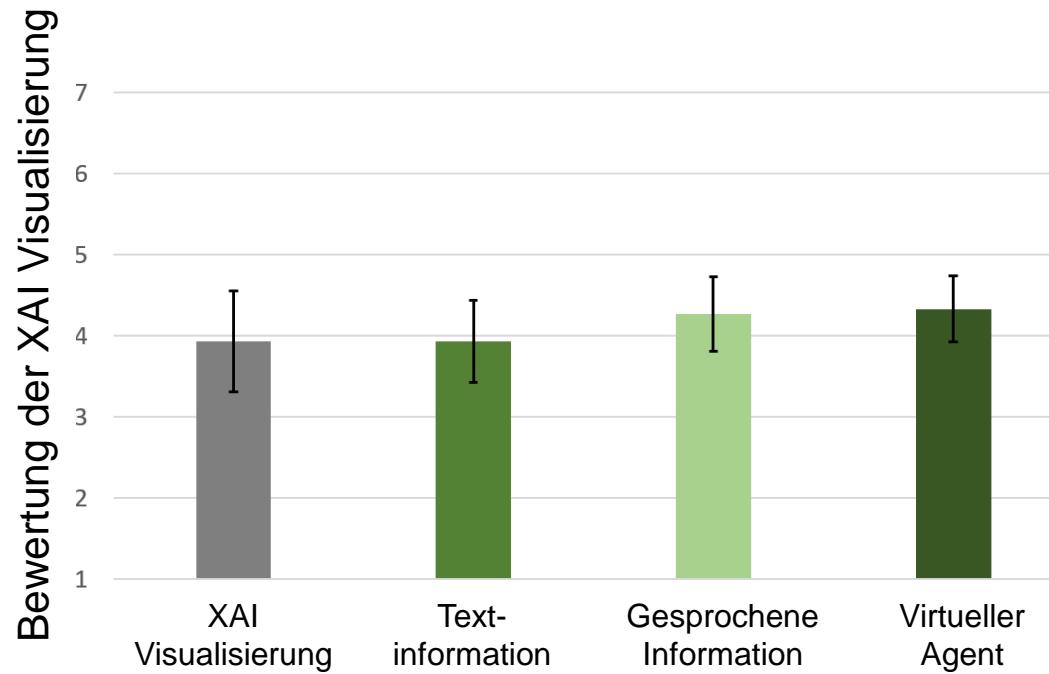
Forschung an der Schnittstelle Mensch-Maschine

Weitz et al. (2020)
Weitz et al. (2019)

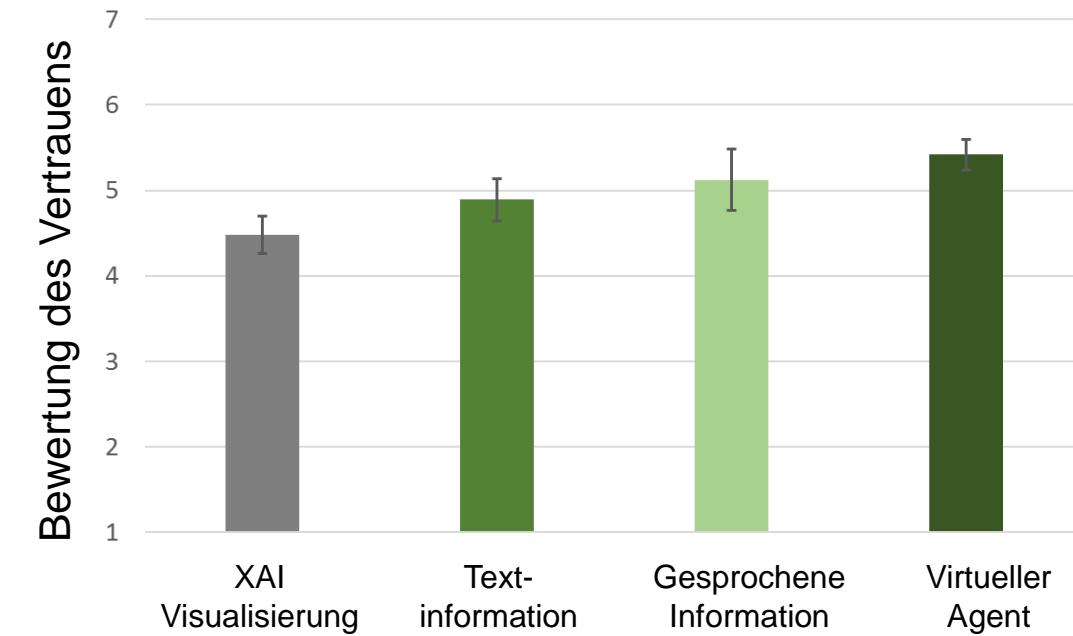


Forschung an der Schnittstelle Mensch-Maschine

Weitz et al. (2020)
Weitz et al. (2019)



Keine signifikanten Unterschiede in der **Bewertung der XAI-Visualisierung** zwischen den 4 Gruppen



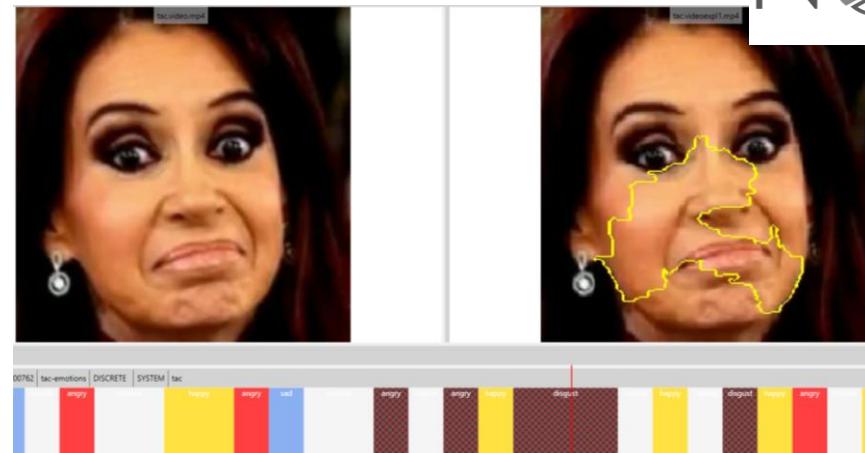
Signifikanter linearer Trend: Je **natürlicher und menschenähnlicher** der Agent **erscheint**, desto mehr vertrauen Teilnehmer:innen dem KI-System

Forschung an der Schnittstelle Mensch-Maschine

Heimerl et al. (2020)

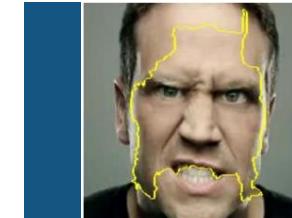
Studie 2

Teilnehmer:innen:	60
Alter:	22.47 Jahre
Geschlecht:	17 männlich 36 weiblich



Wütend

Nur Klassifikationsergebnisse



XAI Visualisierungen

80% sicher,
dass die Person
Wut zeigt

Konfidenzwerte

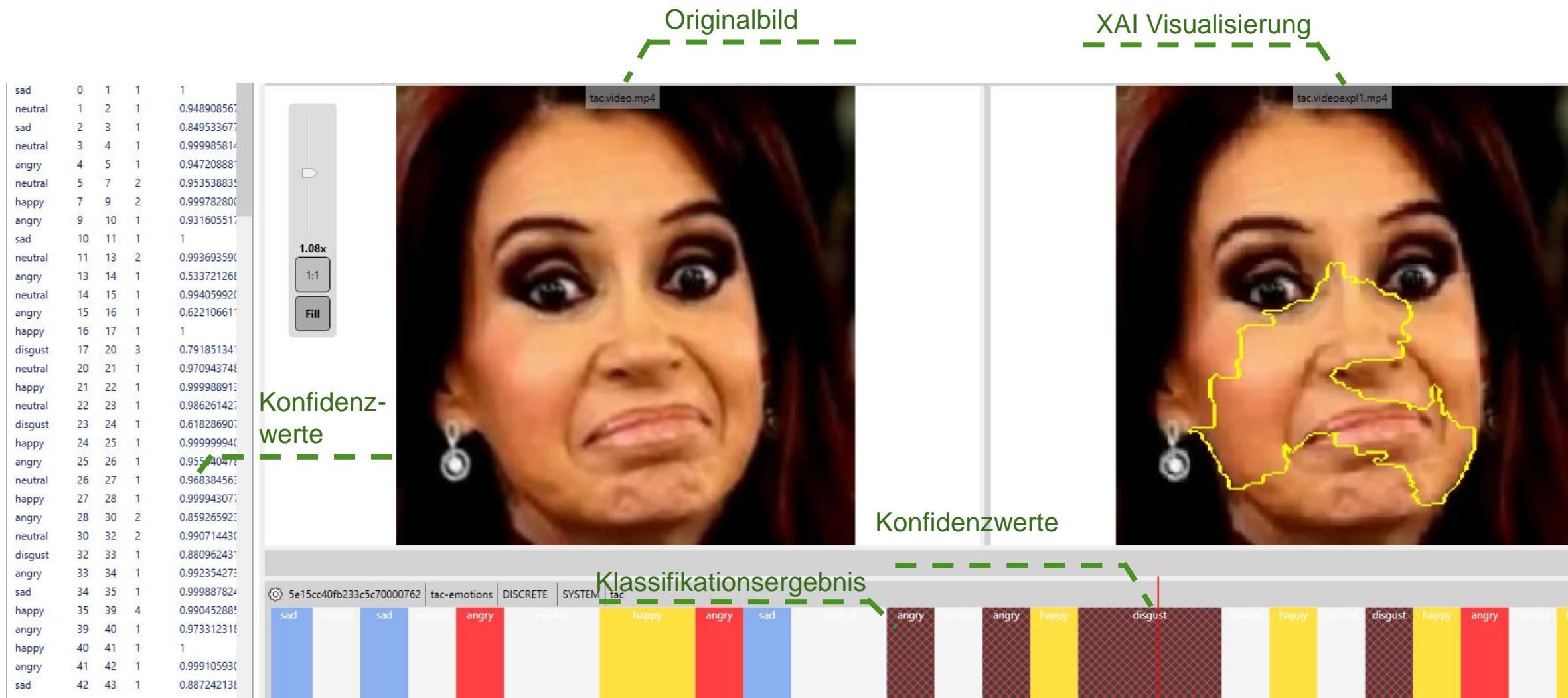


XAI Visualisierung &
Konfidenzwerte



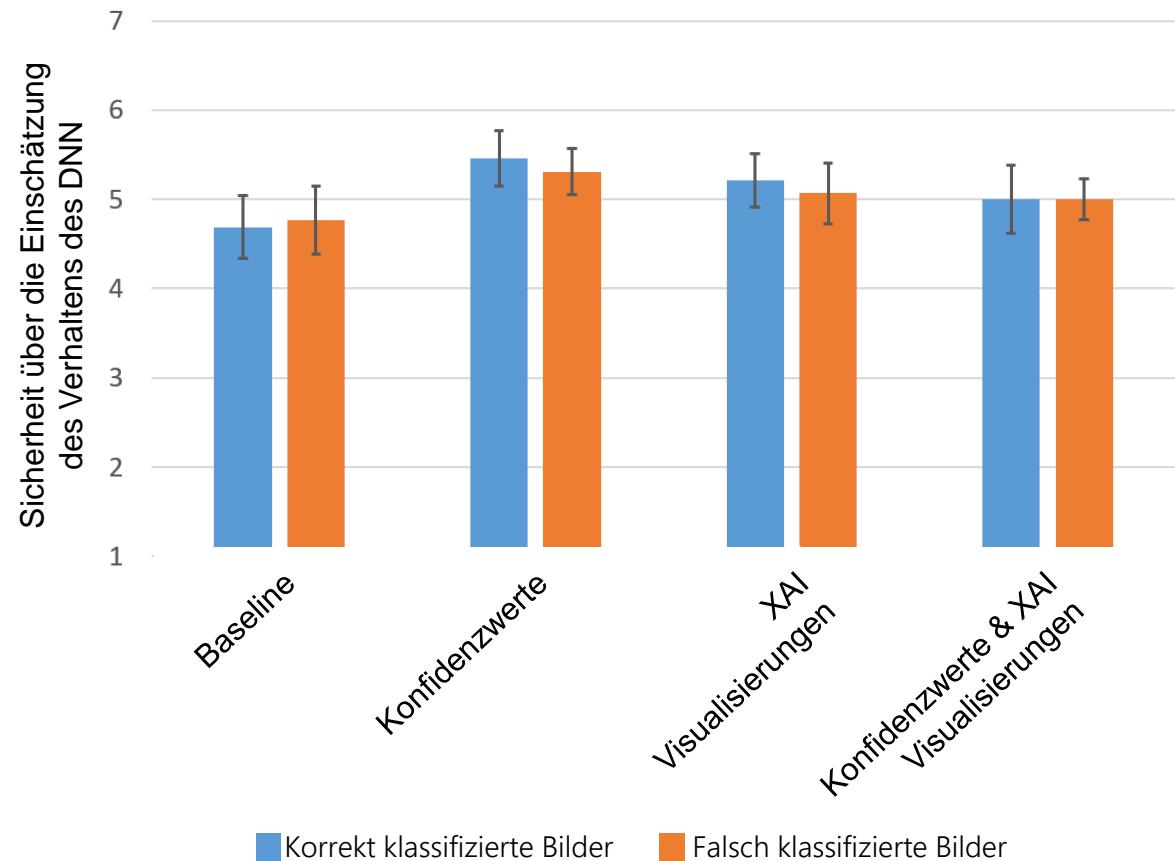
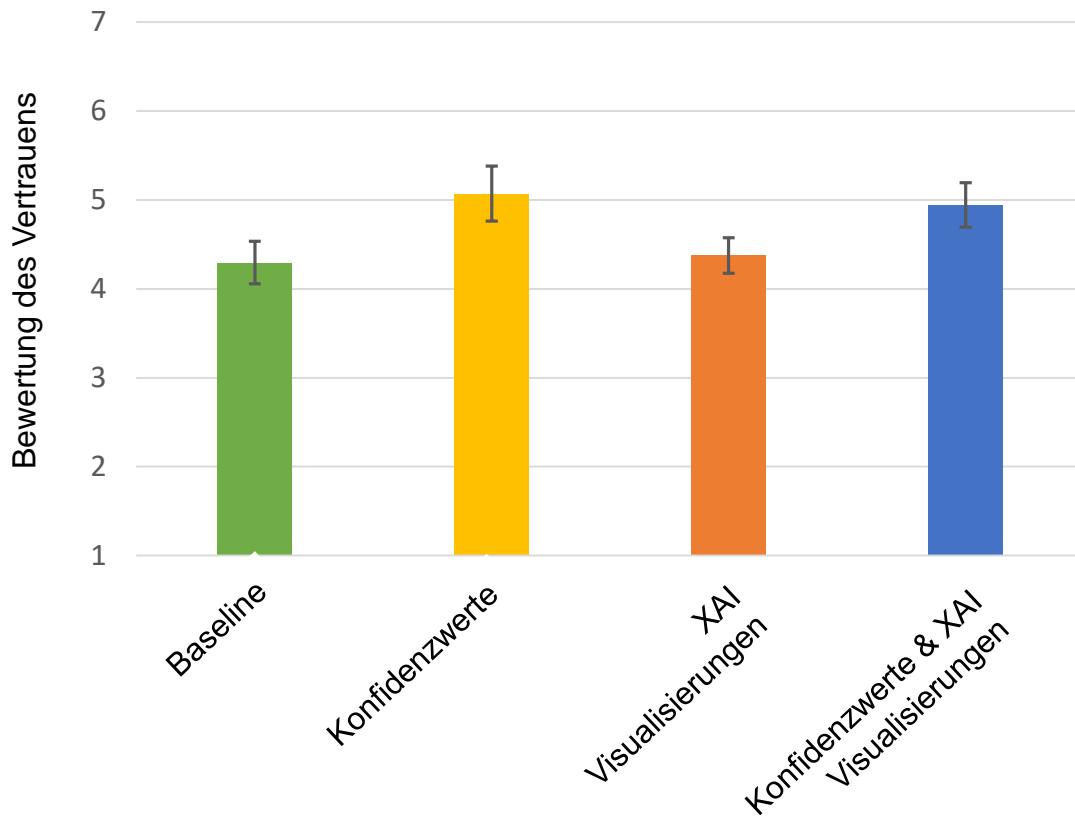
Forschung an der Schnittstelle Mensch-Maschine

Heimerl et al. (2020)



Forschung an der Schnittstelle Mensch-Maschine

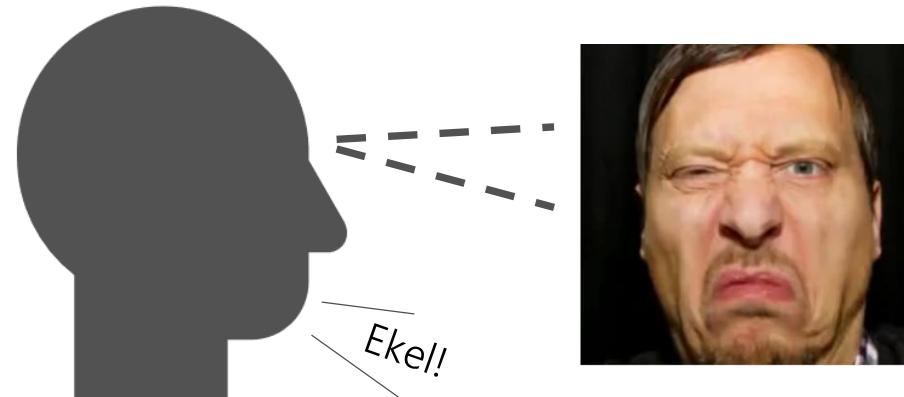
Heimerl et al. (2020)



Keine signifikanten Unterschiede!

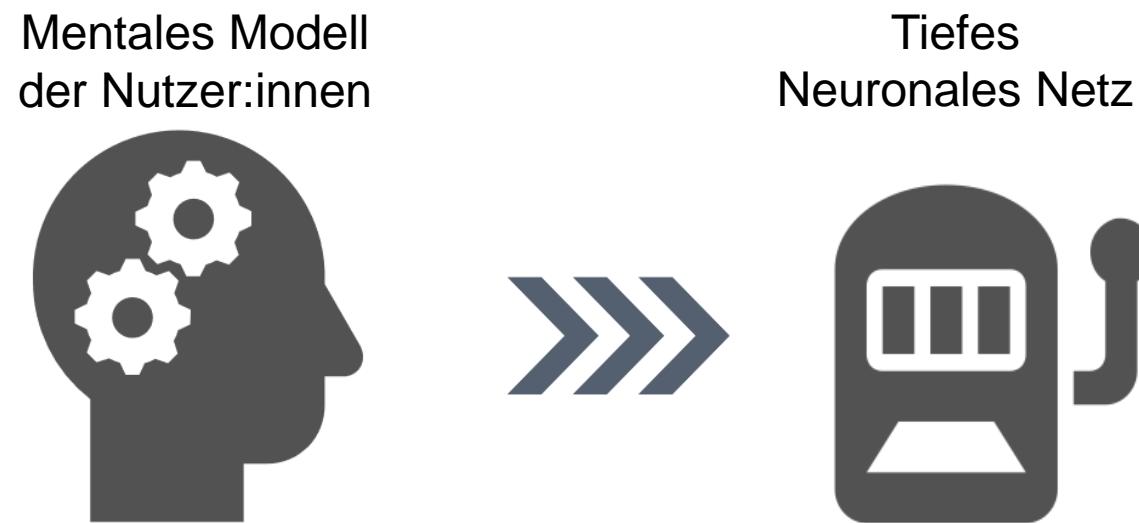
Warum sind sich die Teilnehmer:innen auch **ohne zusätzliche Informationen** so sicher mit ihrer Beschreibung des Verhaltens des Neuronalen Netzes?

**Menschen sind Domänenexpert:innen
im Bereich der Emotionserkennung!**



Forschung an der Schnittstelle Mensch-Maschine

Heimerl et al. (2020)



Nutzer:innen übertragen ihr mentales Modell auf das der Maschine

- Führt zu **Annahmen** über das System, die höchstwahrscheinlich **falsch** sind
- Effekt kann durch XAI-Informationen **abgeschwächt** werden

Forschung an der Schnittstelle Mensch-Maschine

Heimerl et al. (2020)

„Unwichtige Bereiche wie der Hintergrund werden berücksichtigt“

(Teilnehmende in der XAI & Konfidenzwerte Gruppe)



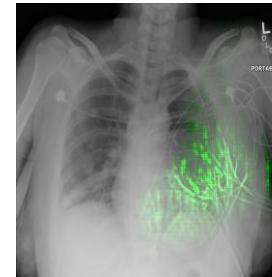
Forschung an der Schnittstelle Mensch-Maschine

Mertes et al. (2020)

Teilnehmer:innen:	118
Alter:	38.5 Jahre
Geschlecht:	63 männlich 53 weiblich 2 nicht-binär



Lungenentzündung
oder Gesund?



Layerwise Relevance
Propagation



Counterfactuals



LIME

Forschung an der Schnittstelle Mensch-Maschine

Mertes et al. (2020)



Glauben Sie, dass die originale Röntgenaufnahme (linke Seite des Sliders) eine Person zeigt, die an einer Lungenentzündung leidet?

Wie sicher sind Sie sich, dass Ihre Diagnose korrekt ist?

Nicht
Sicher

Sehr
sicher

Was glauben Sie, wird die KI klassifizieren?

Wie sicher sind Sie sich, dass Sie die KI korrekt eingeschätzt haben?

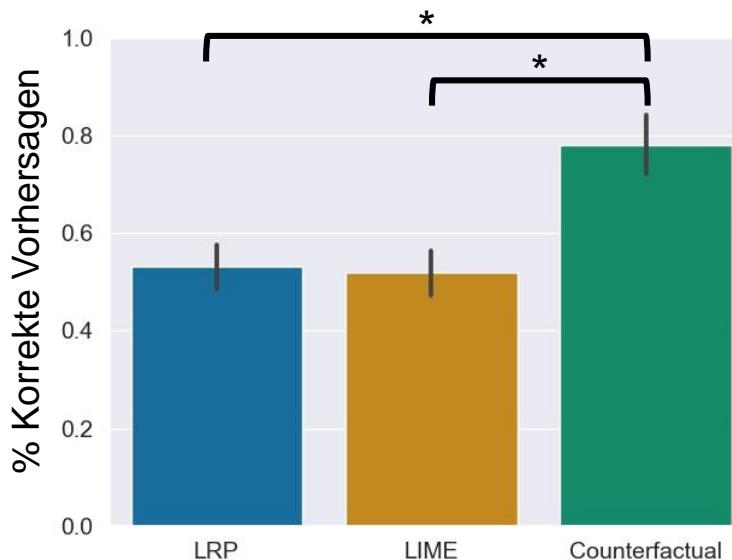
Nicht
sicher

Sehr
sicher

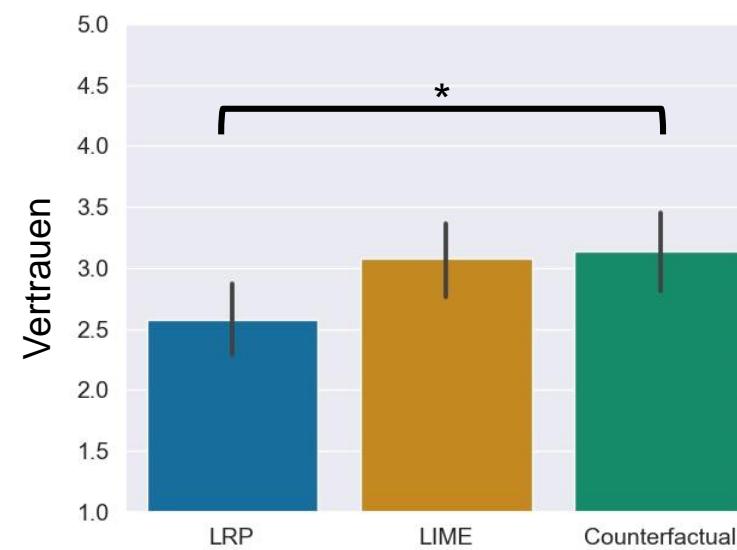
Bitte erläutern Sie kurz Ihre Einschätzung:

Forschung an der Schnittstelle Mensch-Maschine

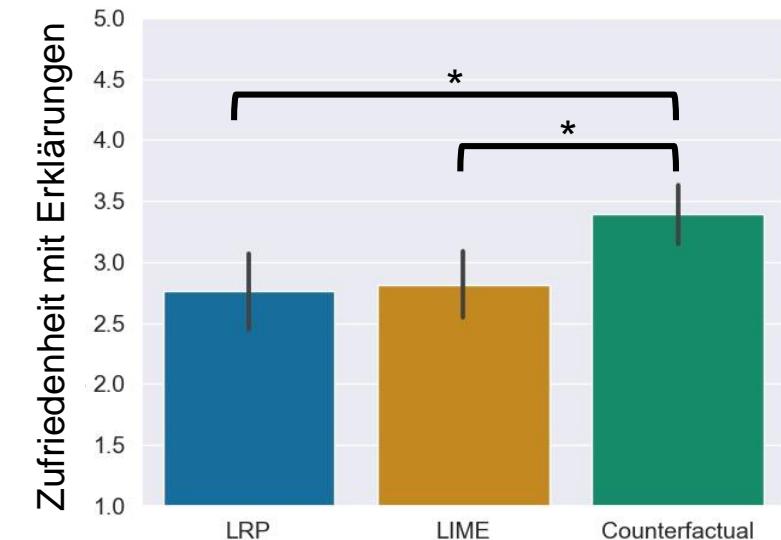
Mertes et al. (2020)



Counterfactuals halfen den Teilnehmer:innen, die Klassifikation des Modells korrekt vorherzusagen.



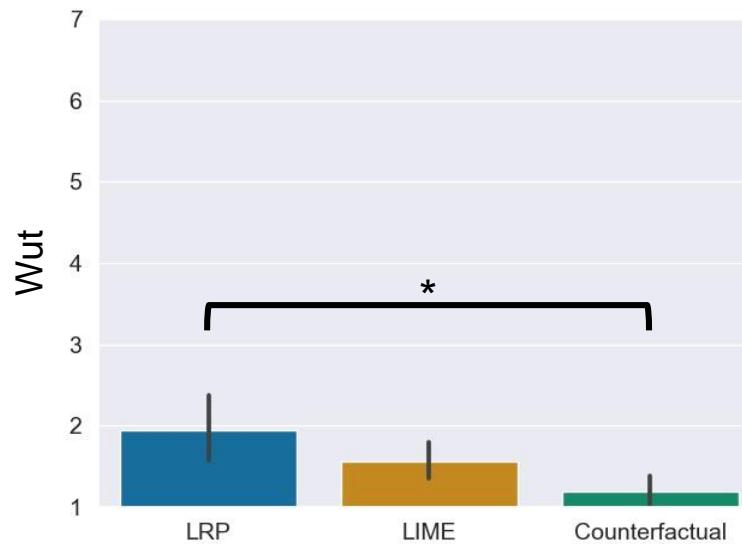
Teilnehmer:innen vertrauten dem Modell, das Counterfactual Erklärungen lieferte, mehr.



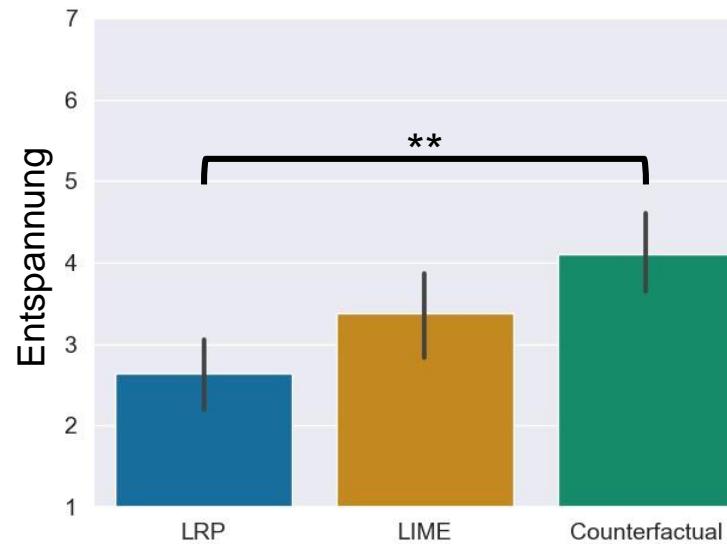
Teilnehmer:innen waren zufriedener mit den Counterfactual Erklärungen.

Forschung an der Schnittstelle Mensch-Maschine

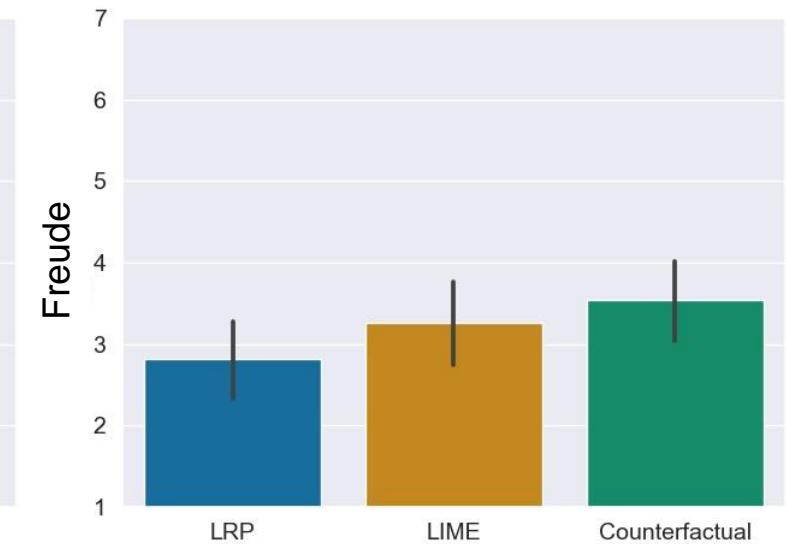
Mertes et al. (2020)



Teilnehmer:innen waren weniger verärgert, wenn sie Counterfactual Erklärungen erhielten.



Teilnehmer:innen waren entspannter, wenn sie Counterfactual Erklärungen erhielten.



Keine Unterschiede!

* $p < .05$ ** $p < .001$

Zusammenfassung

Erklärungen beeinflussen unsere Wahrnehmung von KI!
→ Nicht immer positiv!

Nicht jede Erklärung ist eine hilfreiche Erklärung!



Wenn wir im Fokus Endanwender:innen haben, müssen die gewonnenen Forschungsergebnisse diese auch erreichen!

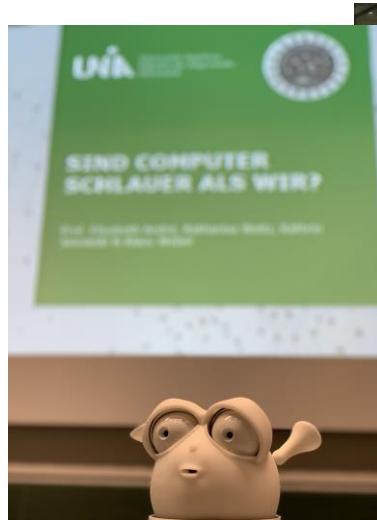


Bild: Katharina Weitz



Bild: StMWK



Bild: Science Slam Z.DB

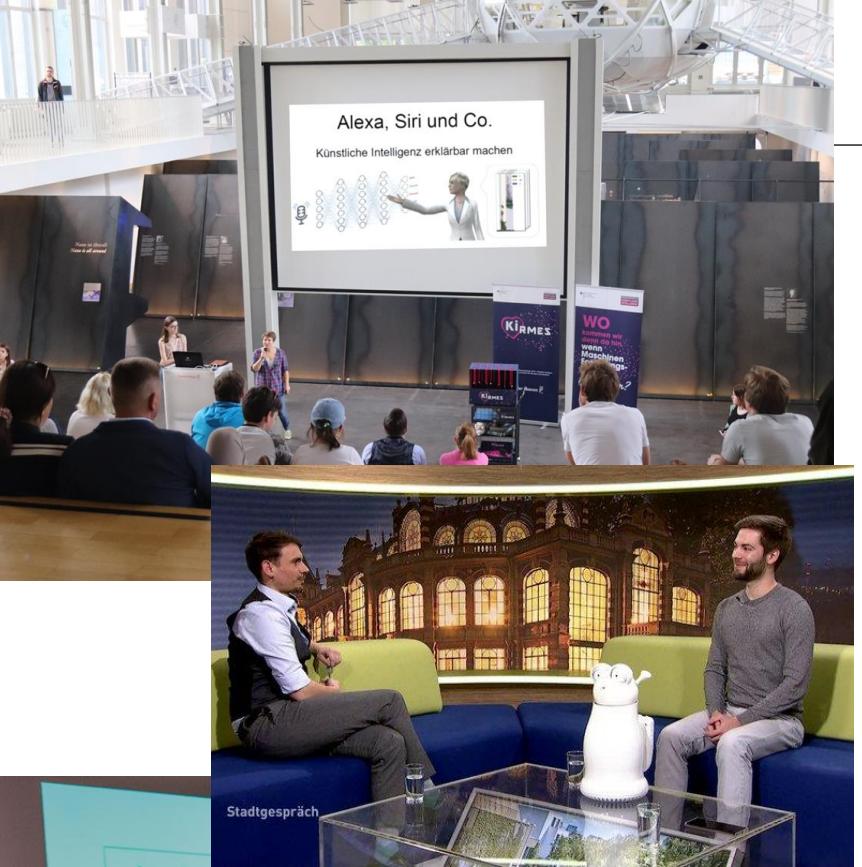


Bild: Augsburg TV (a.tv)

Referenzen

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138-52160.
- Amir, D., & Amir, O. (2018). Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 1168-1176).
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140.
- Budidu, R. (2019). Mental Models for Intelligent Assistants. <https://www.nngroup.com/articles/mental-model-ai-assistants/>
- Halasz, F. G., & Moran, T. P. (1983). Mental models and problem solving in using a calculator. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 212-216).
- Eurobarometer 2017:
<https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjvn6zRhPPAhUCy6YKHbToAM8QFjACegQIAhAB&url=https%3A%2F%2Fec.europa.eu%2Fcommfrontoffice%2Fpublicopinion%2Findex.cfm%2FResults%2Fdownload%2FDocumentKy%2F78998&usg=AOvVaw19eYS0AiR4cJJ9C-Xqlod>
- Heimerl, A., Weitz, K., Baur, T., & André, E. (2020). Unraveling ML Models of Emotion with NOVA: Multi-Level Explainable AI for Non-Experts. *IEEE Transactions on Affective Computing*.
- Huber, T., Schiller, D., & André, E. (2019). Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)* (pp. 188-202). Springer, Cham.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (pp. 2668-2677). PMLR.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks
- Kohlbrenner, M. H. (2017). On the stability of neural network explanations. Bachelor's Thesis.
- Lapuschkin, S., Binder, A., Müller, K.-R., & Samek, W. (2017). Understanding and comparing deep neural networks for age and gender classification. In *Proceedings of the international conference on computer vision* (pp. 1629–1638)
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444.
- Lehrstuhl Menschzentrierte KI: www.hcm-lab.de
- McCarthy, J., Minsky, M. L., & Rochester, N. (1955). A proposal for the dartmouth summer research project on artificial intelligence. Zu finden in: McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4), 12-12.
- Mertes, S., Huber, T., Weitz, K., Heimerl, A., & André, E. (2020). This is not the Texture you are looking for! Introducing Novel Counterfactual Explanations for Non-Experts using Generative Adversarial Learning. *arXiv preprint arXiv:2012.11905*.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38.
- Mitchell, T. (1997). Machine learning. McGraw-Hill international editions – computer science series. McGraw-Hill Education.
- Molnar, Christoph. (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable, <https://christophm.github.io/interpretable-ml-book/>.
- Norman, D. A. (2014). Some observations on mental models. In *Mental models* (pp. 15-22). Psychology Press.
- Rabold, J., Deininger, H., Siebers, M., & Schmid, U. (2019). Enriching visual with verbal explanations for relational concepts—combining LIME with Aleph. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 180-192). Springer, Cham.
- Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical biosciences*, 23(3-4), 351-379.
- Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975). Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and biomedical research*, 8(4), 303-320.
- Vortrag Sameer Singh: <https://de.slideshare.net/0xdata/explaining-blackbox-machine-learning-predictions>
- Web of Science: https://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch&SID=C4y2i4DRIGVuNBj5Ulm&preferencesSaved=
- Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (pp. 7-9).
- Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2020). "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, 1-12.
- Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

Zum Ausprobieren und Anwenden

- **Awesome-explainable-AI GitHub Repro mit einem Überblick über verschiedene XAI Methoden**
<https://github.com/wangyongjie-ntu/Awesome-explainable-AI>
- **Local Interpretable Model-Agnostic Explanations (LIME) GitHub Repro der Entwickler:innen**
<https://github.com/marcotcr/lime>
- **iNNvestigate neural networks GitHub Toolbox mit den gängigsten XAI Methoden**
<https://github.com/albermax/innvestigate>
- **LRP Demo zum Ausprobieren im Browser** <https://lrpserver.hhi.fraunhofer.de/>
- **Konzepte (TCAV) als XAI Methode** <https://github.com/tensorflow/tcav>
- **SHAP (Shapley Additive exPlanations) Repro:** <https://shap.readthedocs.io/en/latest/>



Universität Augsburg
Fakultät für Angewandte
Informatik



Menschzentrierte
Künstliche Intelligenz
Institut für Informatik

Katharina Weitz
Lehrstuhl Menschzentrierte Künstliche Intelligenz
Universität Augsburg

katharina.weitz@informatik.uni-augsburg.de
www.hcm-lab.de