FZI Research Center for
Information Technology

Distinguishing Between Truth and Fake – Using Explainable AI to Understand and Combat Online Disinformation

Isabel Bezzaoui, Jonas Fegert and Christof Weinhardt

Contact: bezzaoui@fzi.de

About the Presenter

- **Name**: Isabel Bezzaoui
- **Course**: Ph.D. Candidate
- **Background**: Sociology and Information Systems
- **Affiliation**: FZI Research Center for Information Technology and KIT Karlsruhe
- **Research Interests**:
  - Disinformation detection
  - Critical media literacy
  - Trust in artificial intelligence

Agenda

- Introduction
- Combating Disinformation Using Machine Learning-Based Systems
- Combating Disinformation – Critical Media Literacy
- Fostering Trust in Artificial Intelligence
- Method and First Activities in Design Science Research
- Conclusion and Future Work

Introduction

- Disinformation is defined as false information spread with the intention to deceive

- **DeFaktS** intends to empower actual users **across various platforms** to critically question news and social media posts

- The project team will develop an **(X)AI** for a participation platform that aims to **combat online disinformation campaigns** and **foster critical media literacy** among users by **informing** them about the occurence of fake news in a **transparent** and **trustworthy** way

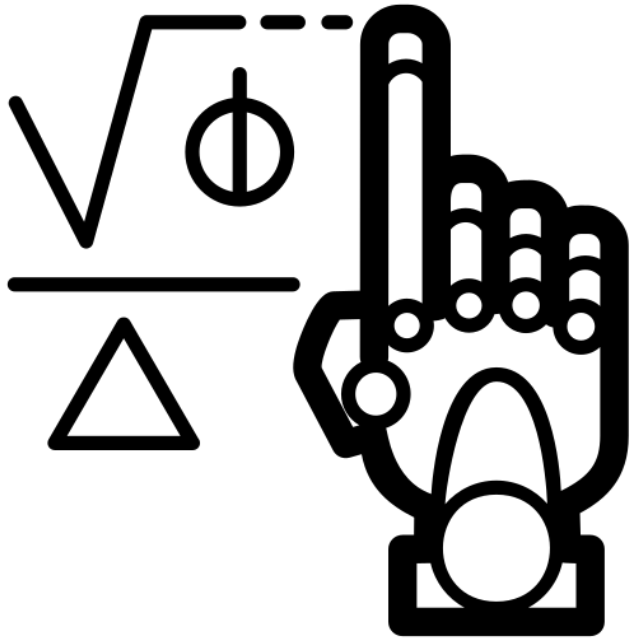
SPONSORED BY THE
Federal Ministry of Education and Research

Combating Disinformation Using Machine Learning-Based Systems

- MLS fake news detection is a rapidly expanding field of research

  → Focus is mostly on extracting multiple features, putting them into classification models and selecting the best classifier

- Research gap: empirical evaluations of when classifiers are put into practice with **real users** and of what **benefits and impact** the developed tools may have

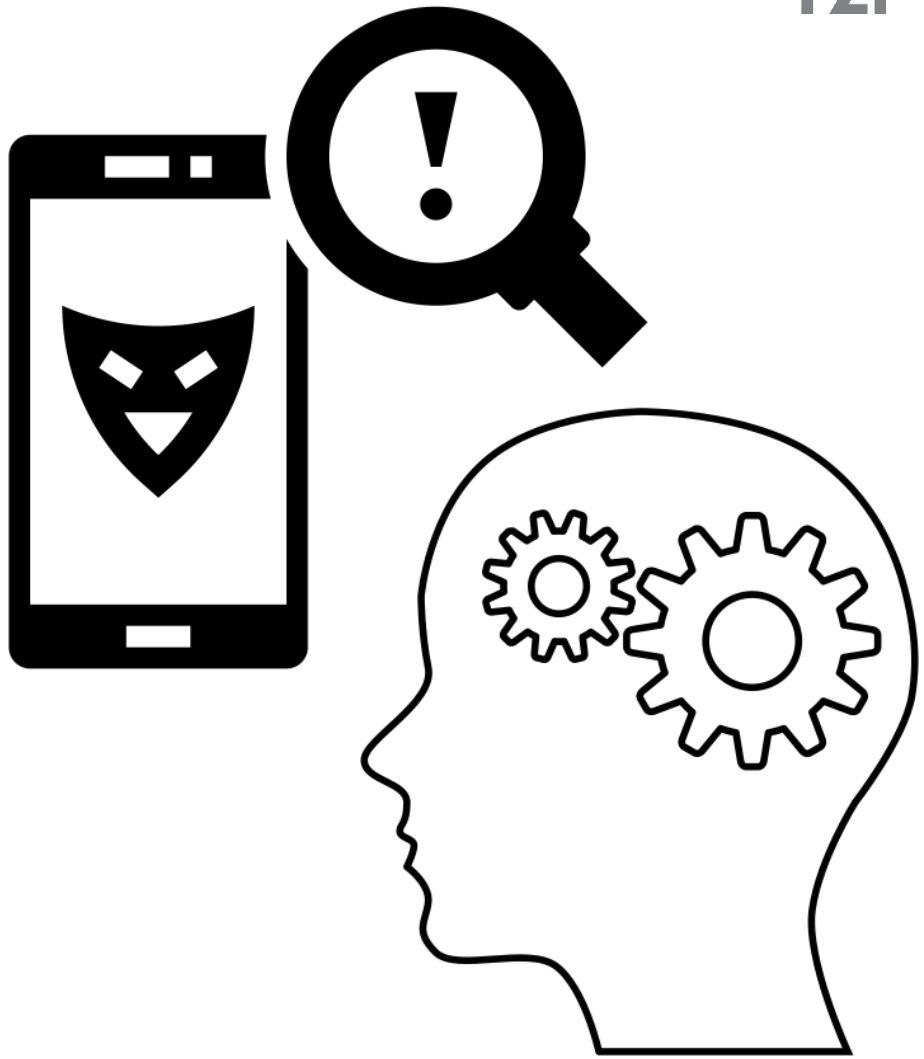


RQ: How to design an artifact for the detection of online disinformation that helps to foster an informed and critically thinking citizenry?

Combating Disinformation – Critical Media Literacy

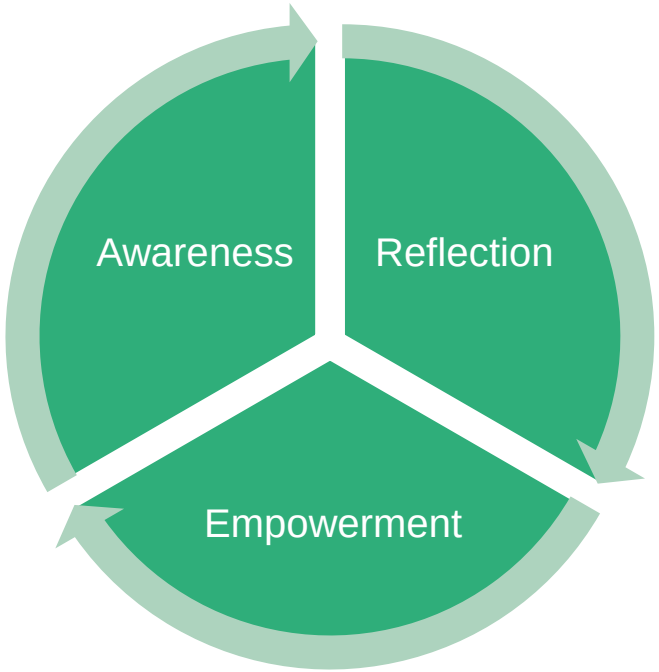
- Promoting critical media literacy (CML) can help people **judge the accuracy** of online content more accurately
- **Susceptibility** to fake news is driven mostly by poor critical thinking
- CML **assists individuals** to use media responsibly, to discern and assess media content, to critically examine media forms, to explore media effects, and to deconstruct alternative media

→ It seems crucial to investigate the **potential of MLS detection tools** for promoting CML among social media users



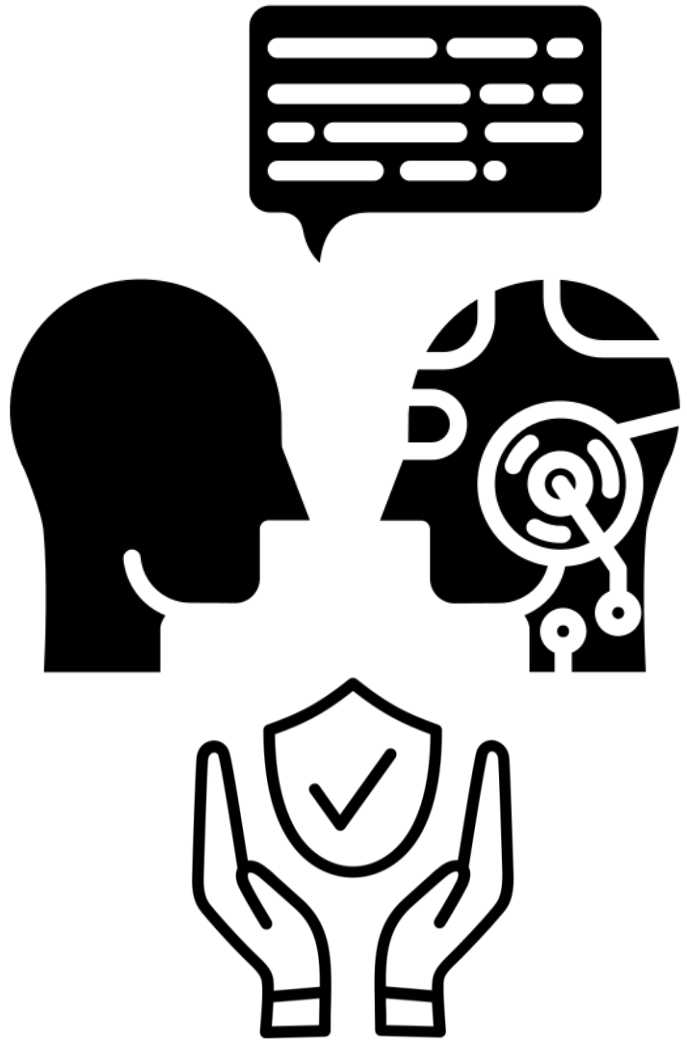
Combating Disinformation – Critical Media Literacy

RQ1.1: (How) Does the tool promote critical media literacy by helping users identify disinformation more accurately?



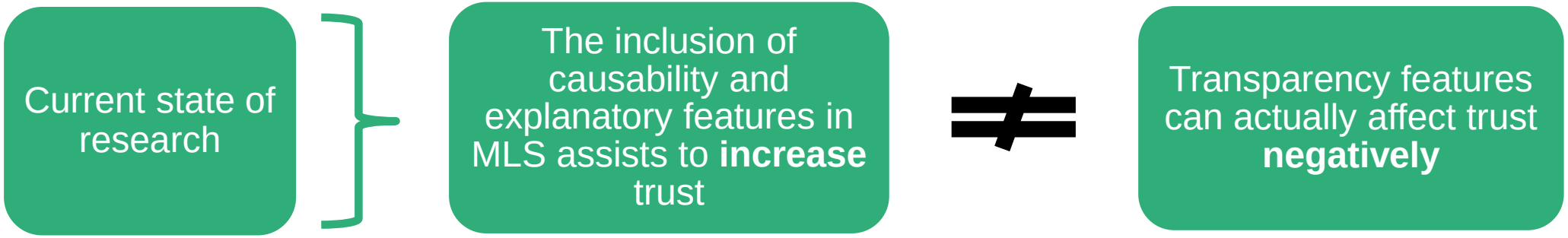
Fostering Trust in Artificial Intelligence

- Previous research has demonstrated the importance of **trust** for the **acceptance** and **perceived usefulness** of ICT tools, and MLS in particular
  - → **Transparency** is an important aspect when it comes to dealing with disinformation
- The implementation of an **XAI-approach** into the development process seeks to make
  - the system‘s internal dynamics more **transparent**
  - the analysis‘ conclusions more **understandable** and hence **trustworthy** to the user
- Need to examine the **effect of XAI elements** on user trust and thus acceptance and perceived usefulness of the final tool



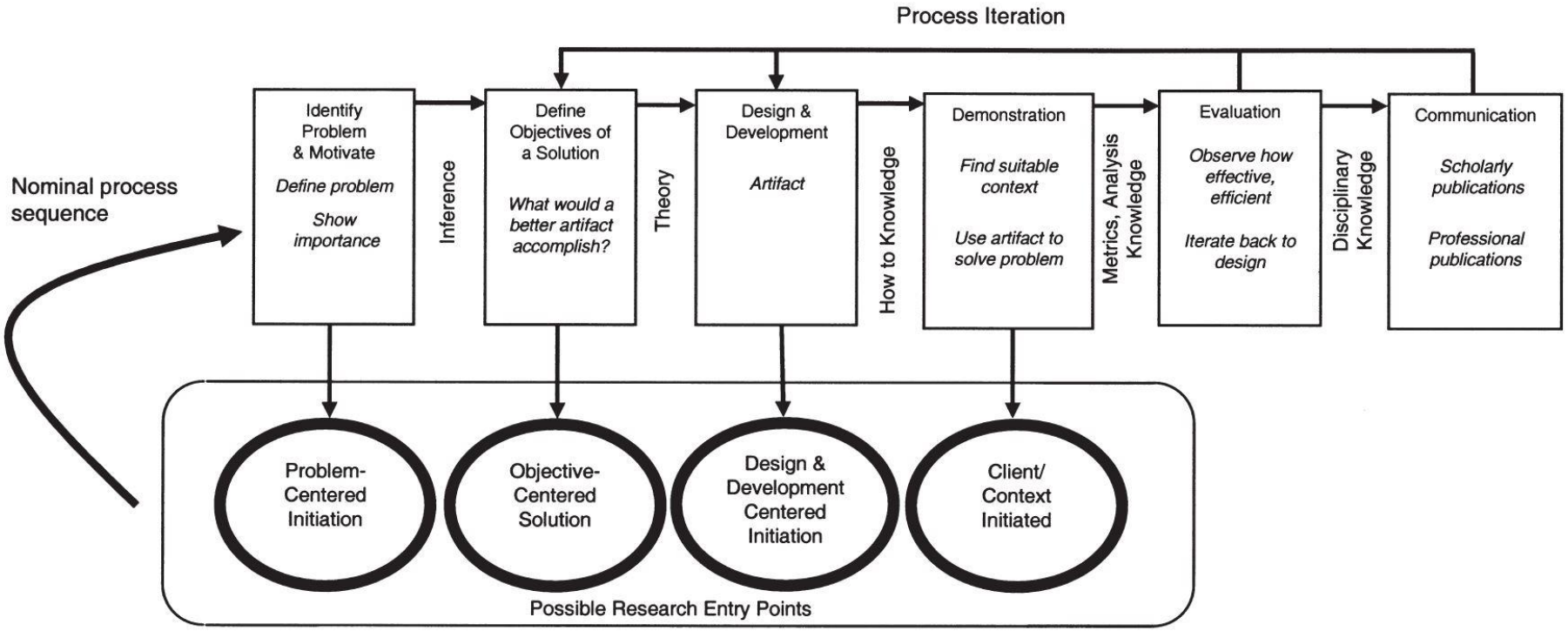
Fostering Trust in Artificial Intelligence

RQ 1.2: (How) Does the tool‘s XAI-component help users trust the algorithm‘s assessment?



- In the DeFaktS project, this controversy will be addressed through the evaluation of **whether, and if so which,** XAI elements increase user trust in the application

Method and First Activities in Design Science Research



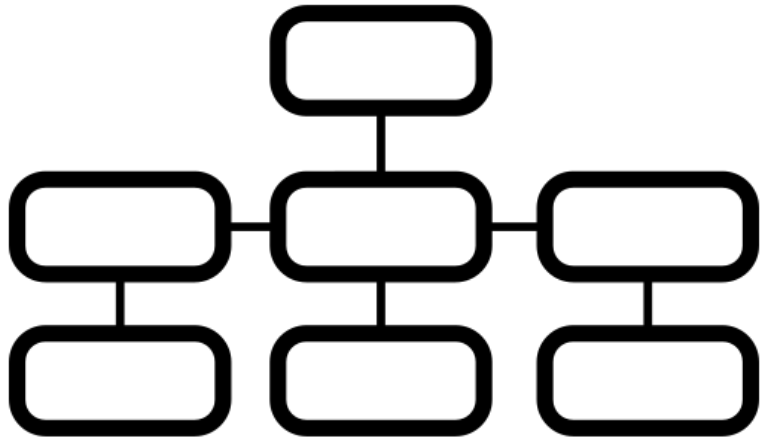
Method and First Activities in Design Science Research

Development of a ‚Fake News Taxonomy‘

- that entails linguistic features and dimensions of disinformation
- to facilitate and ensure the quality of the data labeling process

Goal: **Create a taxonomy of fake news that**

- encompasses broad and event-independent dimensions and characteristics of disinformation
- is still specific enough to precisely identify and label deceiving content



Conclusion and Future Work

- In this research-in-progress, we contribute to the knowledge base of fake news detection using MLS by **developing an XAI-artifact and evaluating its performance** in the context of fostering critical media literacy and trust among social media users

- **Innovative aspects**:
  - Non-expert users shall be enabled to **understand, trust, and utilize** the tool‘s interpretation and explanation of detection results
  - The DeFaktS-tool shall increase **overall critical thinking and awareness** of online disinformation, cultivating an informed citizenry and fostering political participation

- **Aim of the paper**: show initial approaches to researching and combating online disinformation using MLS

Thank you for your attention!

References (1/2)

H. Alsaidi, and W. Etaiwi, “Empirical evaluation of machine learning classification algorithms for detecting COVID-19 fake news”, *Int. J. Advance Soft Compu. Appl, 14(1),* pp. 49-59, 2022.

W. H. Bangyal et al., “Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches”, *Computational and Mathematical Methods in Medicine,* pp. 1-13, 2021.

L. Bozarth, and C. Budak, “Toward a better performance evaluation framework for fake news classification”, *Proceedings of the international AAAI conference on web and social media,* 14, pp. 60–71, 2020.

P. Dahlgren, “Media and political engagement: Citizens, communication, and democracy”,

J. Groshek, and K. Koc-Michalska, “Helping populism win? Social media use, filter bubbles, and support for populist presidential candidates in the 2016 US election campaign”, *Information Communication and Society, (20:9)*, pp. 1389-1407, 2017.

A. M. Guess et al., “A digital media literacy intervention increases discernment between mainsitream and false news in the United States and India”, *PNAS*, *117*(27), pp. 15536–15545, 2022.

C. Lai et al., “Fake news classification based on content level features”, *Applied Sciences, 12(3)*, p. 1116, 2022.

M. Mahyoob, J. Al-Garaady, M. Alrahaili, “Linguistic-based detection of fake news in social media.” Forthcoming, *International Journal of English Linguistics, 11(1),* pp. 99-109*,* 2020.

G. Pennycook, and D. G. Rand, “Lazy, not biased: Suceptibility to partisan news is better explained by lack of reasinong than by motivated reasoning”, *Cognition*, pp. 1–12, 2018.

References (2/2)

D. Ribes Lemay et al., “Trust indicators and explainable AI: A study on user perceptions”, *IFIP Conference on Human-Computer Interaction - INTERACT 2021*, pp. 662–671, 2021.

T. Schmidt, F. Biessmann, T. Teubner, „Transparency and trust in artificial intelligence systems”, *Journal of Decision Systems*, *29*(4), pp. 260–278, 2020.

J. B. Schmitt, D. Rieger, J. Ernst, H. J. Roth, „Critical media literacy and Islamist online propaganda: The feasibility, applicability and impact of three learning arrangements”, *International Journal of Conflict and Violence*, *12*, pp. 1–19, 2018.

D. Shin, “The effects of explainability and causability on perception, trust and acceptance: Implications for explainable AI”, *International Journal of Human-Computer Studies*, *146,* pp. 1-11, 2021.

K. Shu, A. Bhattacharjee, F. Alatawi, T. H. Nazer, K. Ding, M. Karami, H. Liu, “Combating disinformation in a social media age”, *WIREs Data Mining and Knowledge Discovery*, *10*, pp. 1–23, 2020.

K. Siau, and W. Wang, “Building trust in artificial intelligence, machine learning, and robotics”, *Cutter Business Journal*, *31*(2), pp. 47-53, 2018.

P. K. Verma, P. Agrawal, I. Amorim, R. Prodan, “WELFake: Word embedding over linguistic features for fake news detection”, *IEEE Transactions on Computational Social Systems, 8(4),* pp. 881–893, 2021.

S. Yu, and D. Lo, “Disinformation detection using passive aggressive algorithms”, *ACM Southeast Conference, Session 4*, p. 324f, 2020.