VolkswagenStiftung

technische universität dortmund

UNIVERSITÄT DES SAARLANDES

On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness

Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, Sarah Sterz

Explainable Intelligent Systems

www.eis.science

THE EXPLAINABILITY-TRUST-HYPOTHESIS

[. . .] in many, if not most, cases, the explanation is beneficial [. . .] to foster better trust [. . .]." [Richardson & Rosenfeld, 2018]

[. . .] explainability will also enhance trust at the user side [. . .]." *[Nalepa et al., 2018]*

"*Artificial agents need to explain their decision to the user in order to gain trust [. . .].*" *[Pieters, 2011]*

"*In order for humans to trust black-box methods, we need explainability [. . .].*" [Gilpin et al., 2018]

Explainability ⟶ ***Explainability-Trust-Hypothesis:*** Explainability is a suitable means for facilitating trust in a stakeholder. ⟶ Trust

EMPIRICAL EVIDENCE

Langer et al., 2021:	Providing information on analysis for personnel decisions has both positive and negative effects on perceived fairness
Schlicker et al, 2021:	No effect of explanations on perceived justice of automated decisions
Kizilcec, 2016:	Too much information can erode trust
Papenmeier et al, 2019:	Explanations have either no or a negative effect on trust

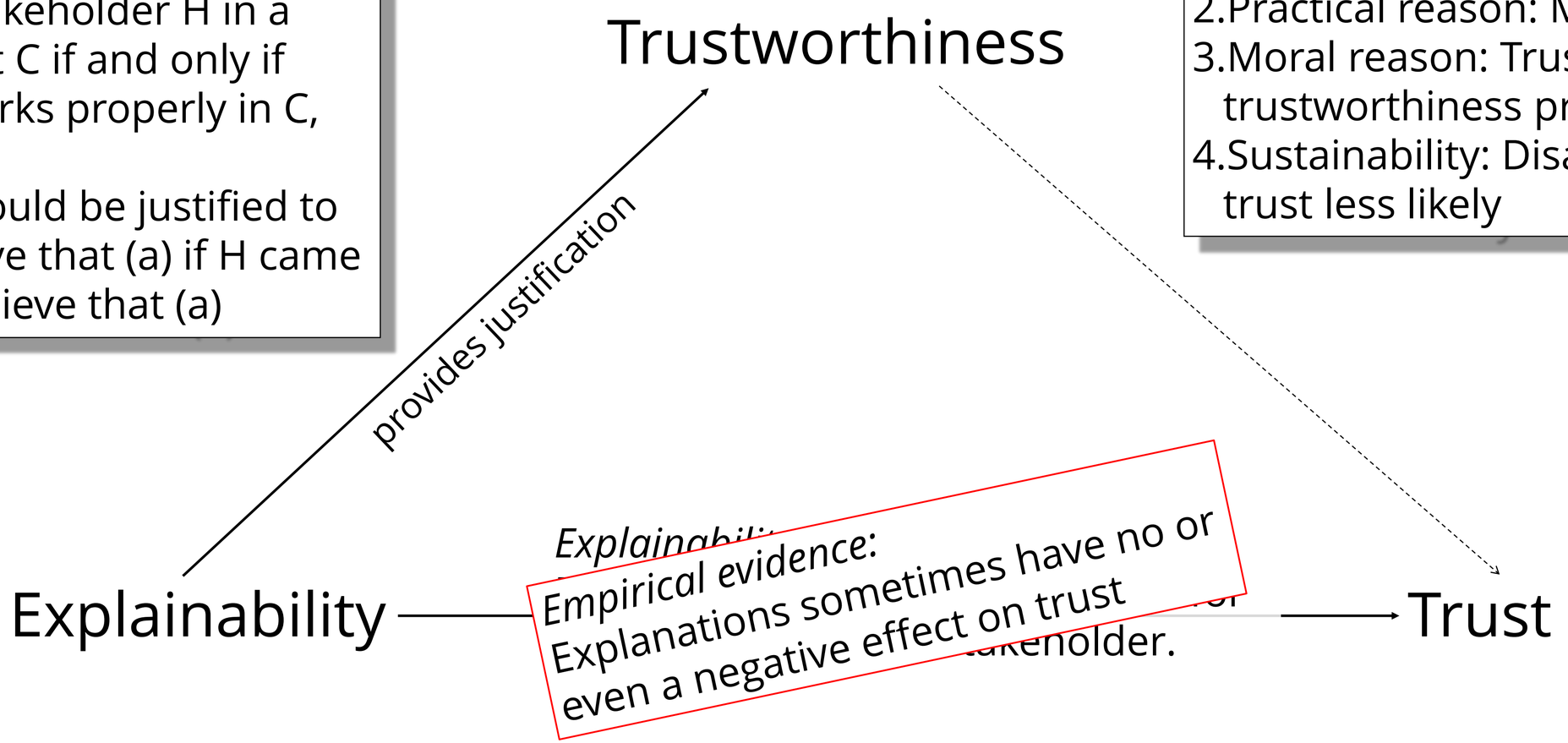
Possible reasons:

1) High initial trust
2) Explanation reveals problem
3) Explanations incomprehensible

FROM TRUST TO TRUSTWORTHINESS

Definition Trustworthiness:
A system S is trustworthy to a stakeholder H in a context C if and only if
a) S works properly in C, and
b) H would be justified to believe that (a) if H came to believe that (a)

Trustworthiness > Trust:
1. Goal: Warranted trust
2. Practical reason: More control
3. Moral reason: Trust without trustworthiness problematic
4. Sustainability: Disappointed trust less likely



FUTURE RESEARCH DIRECTIONS

1. Finding the best operationalization of „trustworthiness“
2. Specifying the relationship between explainability and trustworthiness
3. Investigating the impact of explanations on trust
4. Examining the link between trustworthiness and trust

THANK YOU!

This presentation was created within the project Explainable Intelligent Systems (EIS) funded by the Volkswagen Foundation.

QUESTIONS?