UKGM
UNIVERSITÄTSKLINIKUM
GIESSEN UND MARBURG

# Trustworthy AI - Beyond Standardization and Certification

**Nadine Schlicker**

**Institute for AI in Medicine, University Hospital of Gießen and Marburg, Philipps-University of Marburg, Germany**

Schlicker, N., Uhde, A., Baum, K., Hirsch, M. C., & Langer, M. (2022). *Calibrated Trust as a Result of Accurate Trustworthiness Assessment – Introducing the Trustworthiness Assessment Model*. PsyArXiv. 10.31234/osf.io/qhwvx

**AI Quality Summit, November 2nd, 2022,** Frankfurt, Germany

# Human-System-Interaction: Aiming for Calibrated Trust

- **Over-trust** = users rely always on the system disregarding system limitations
- **Under-trust** = users do reject the system and diminish the potentials of automated systems
- ✓ **Calibrated trust** = user's adequately trust and distrust a system (advice)

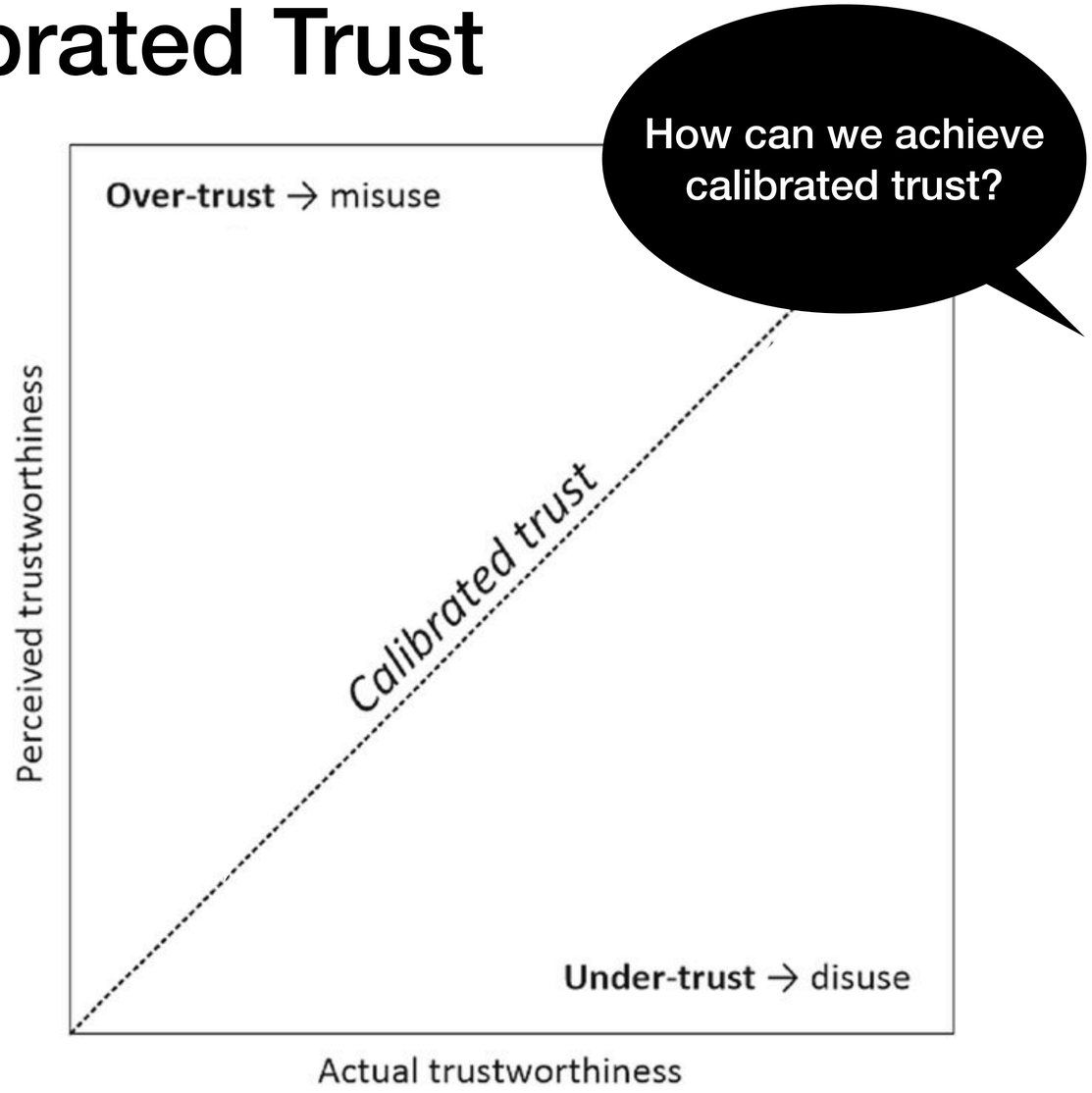
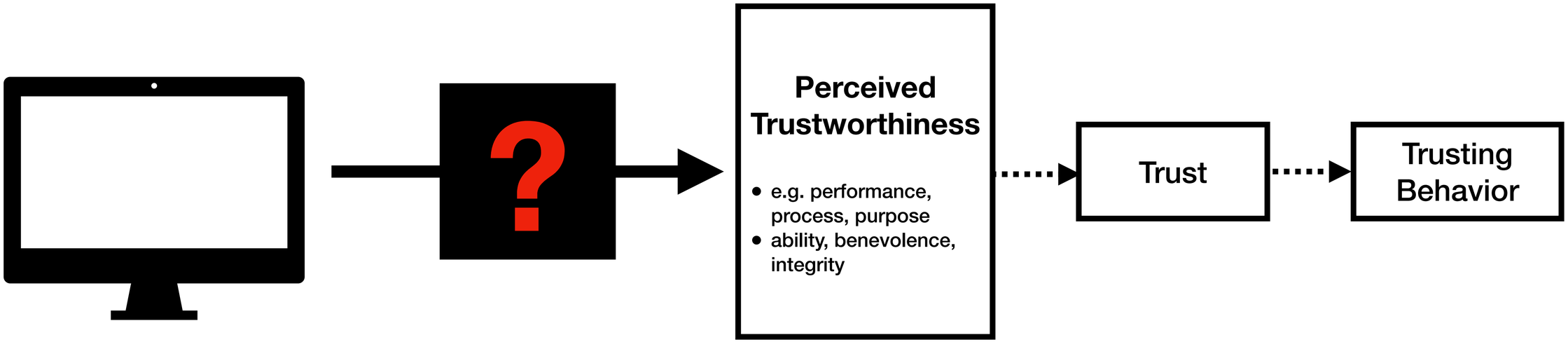
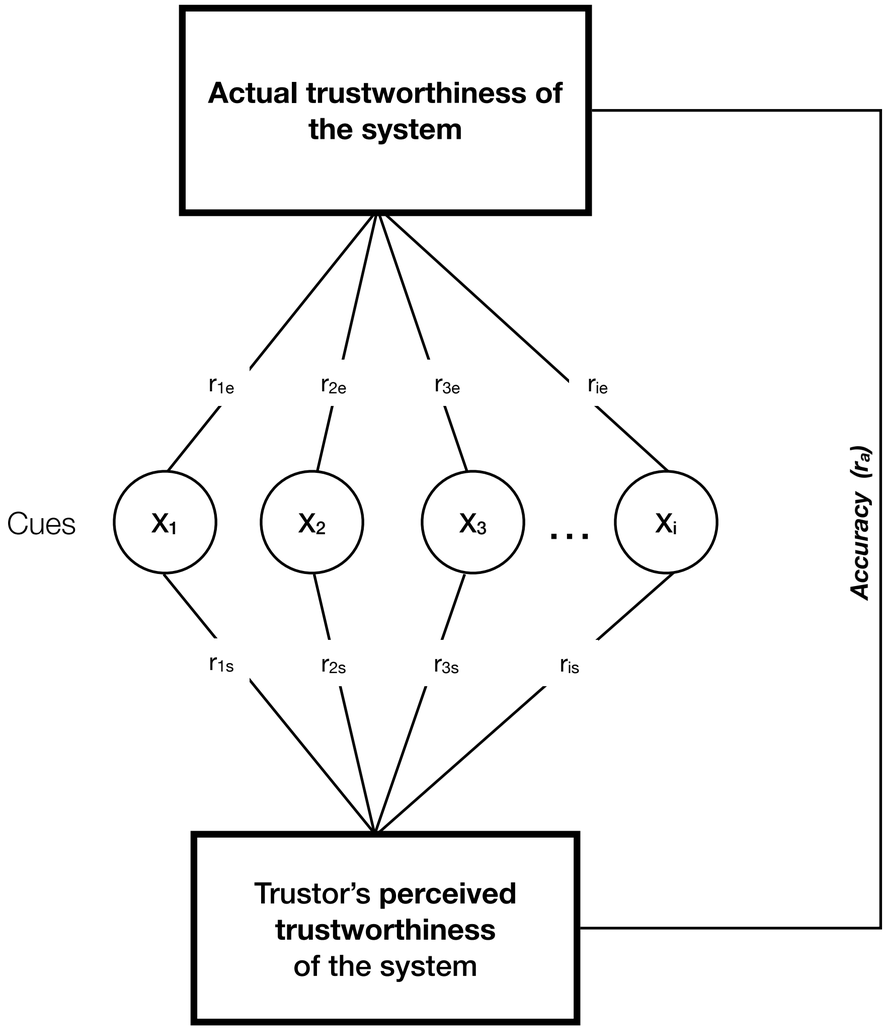



Figure adapted from: de Visser et al., 2020

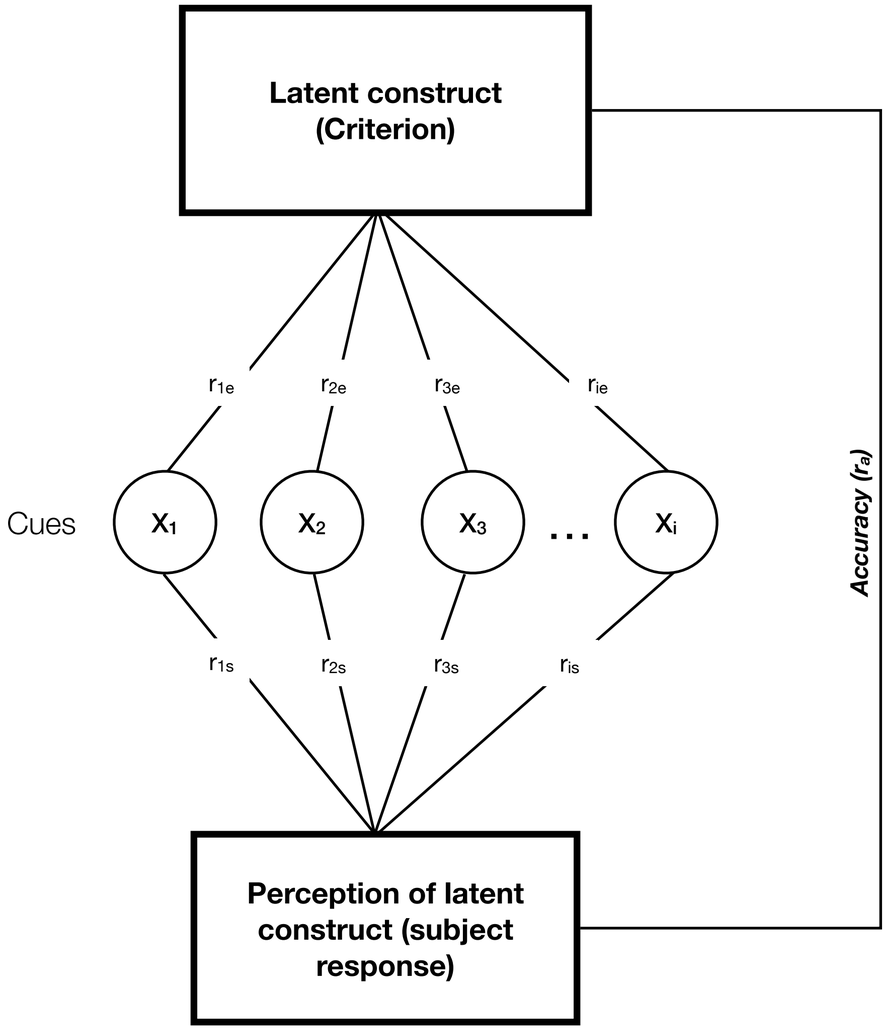
# Related Work



# Basic Idea

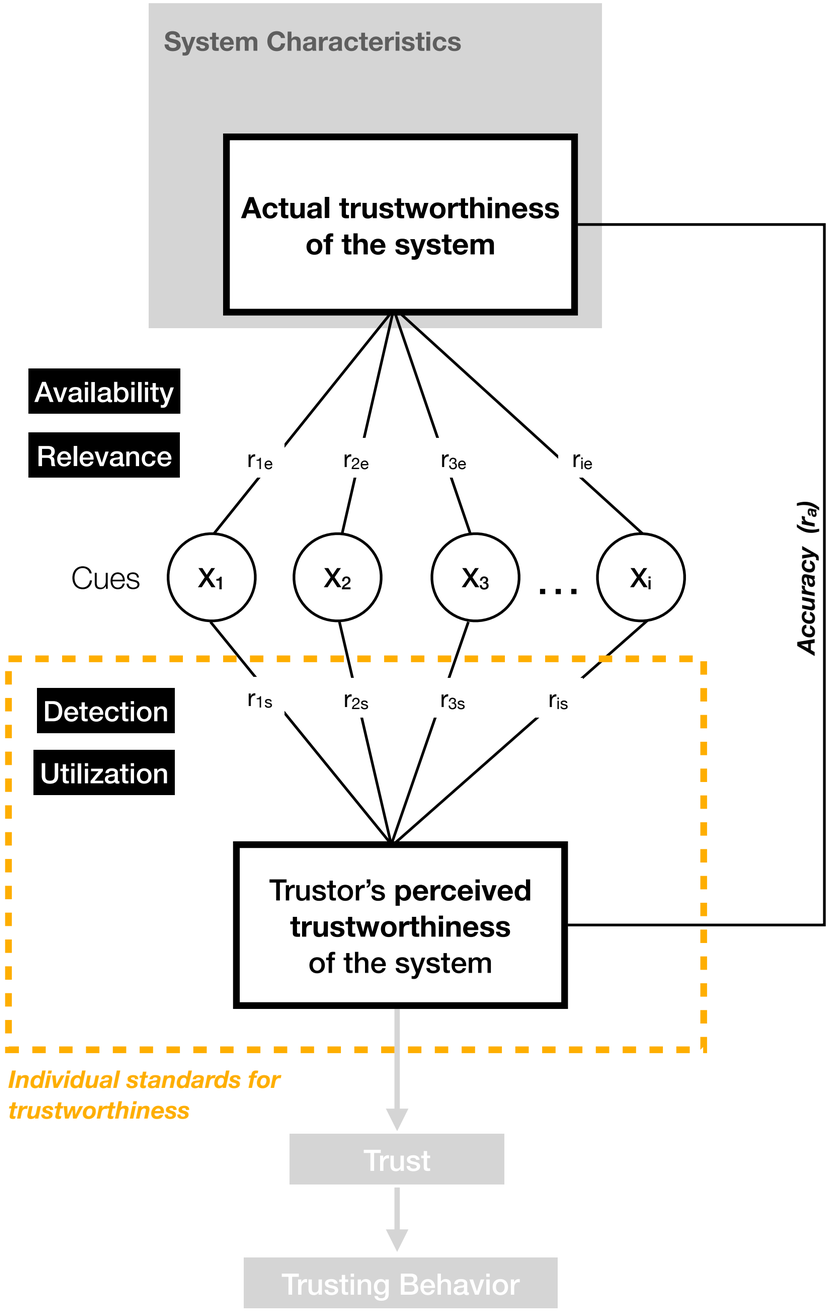

Actual trustworthiness of the system
$r_{1e}$
$r_{2e}$
$r_{3e}$
$r_{ie}$
Cues
$X_1$
$X_2$
$X_3$
...
$X_i$
$r_{1s}$
$r_{2s}$
$r_{3s}$
$r_{is}$
Trustor's perceived trustworthiness of the system
Accuracy ($r_a$)

# Basic Idea




Kuncel, 2017

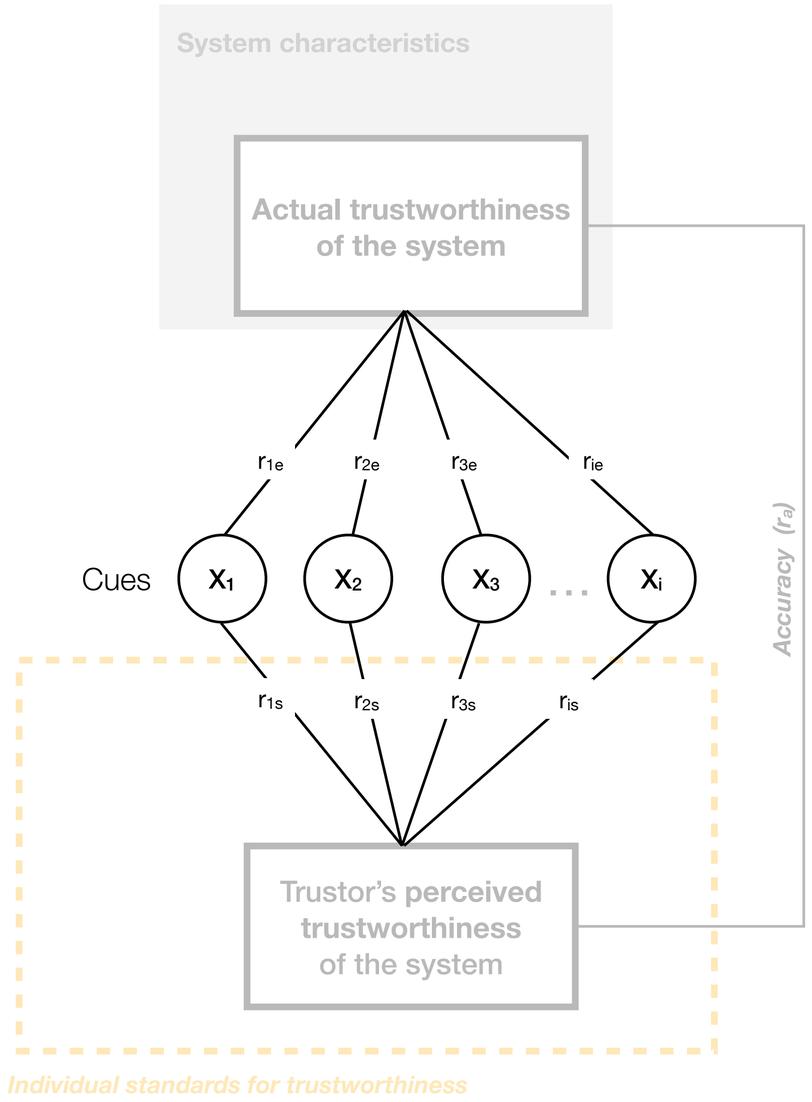
# Our Trustworthiness Assessment Model




Funder, 1995
Kuncel, 2017
Schlicker et al., 2022

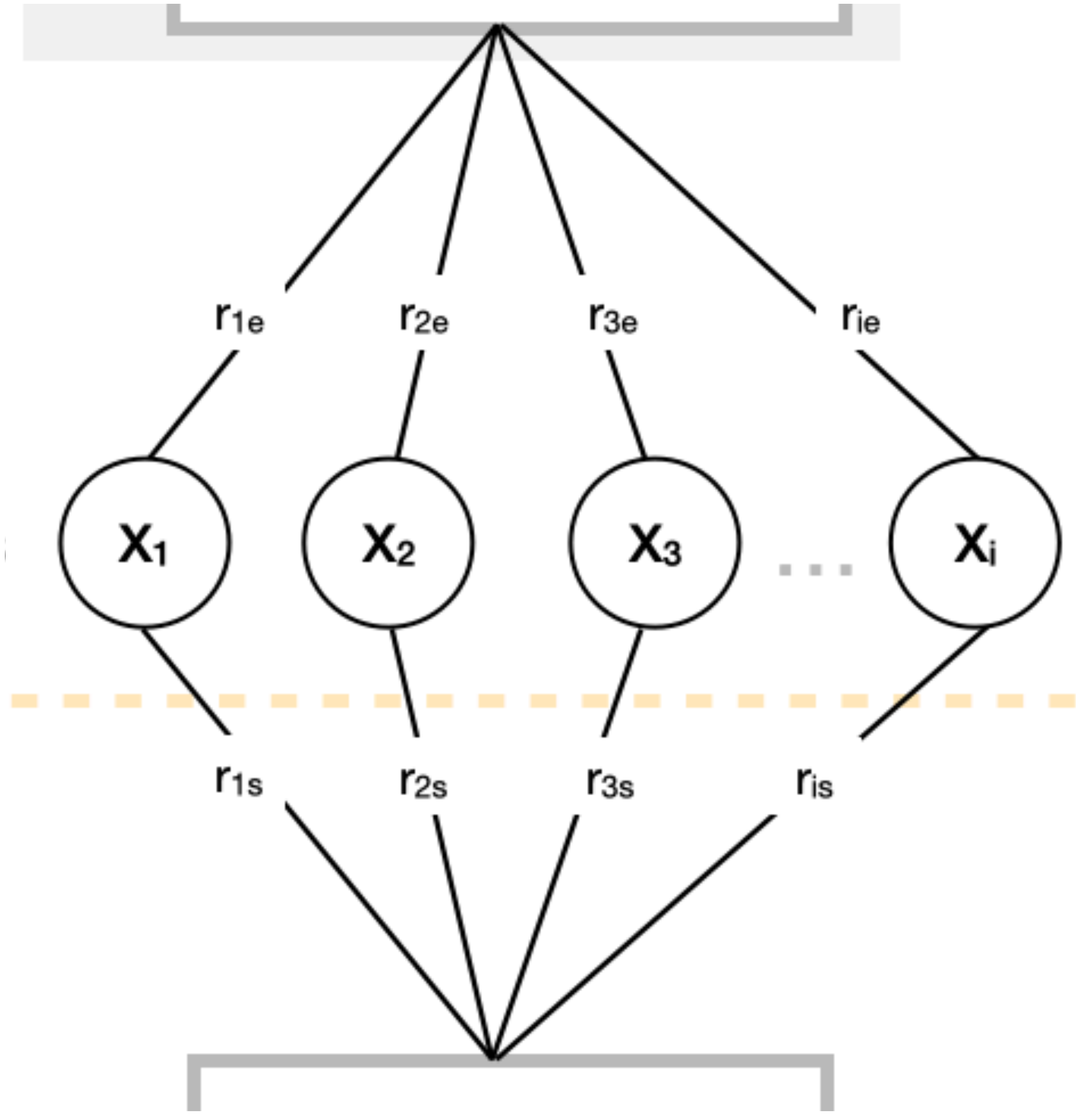
# Cues

- **Pieces of information that presumably provide insights regarding the *Actual Trustworthiness (AT)* of a system**



# Cues

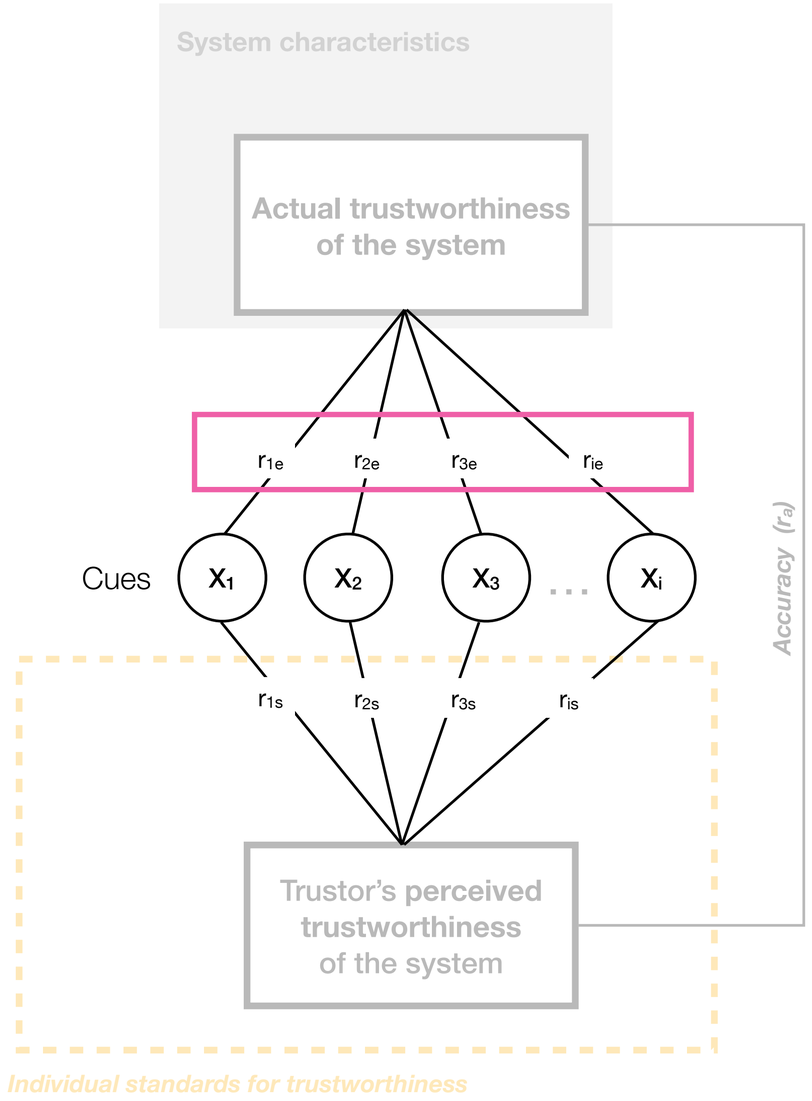
- **Examples:**
  - Logo of a company
  - Marketing
  - Information in the system’s user manual
  - Single outputs that a system produces
  - Testimonies of colleagues
  - Certification seal for "trustworthy AI”
  - …



# Cues

- Pieces of information that presumably provide insights regarding the *Actual Trustworthiness (AT)* of a system

- **Cues are more or less closely related to the AT**



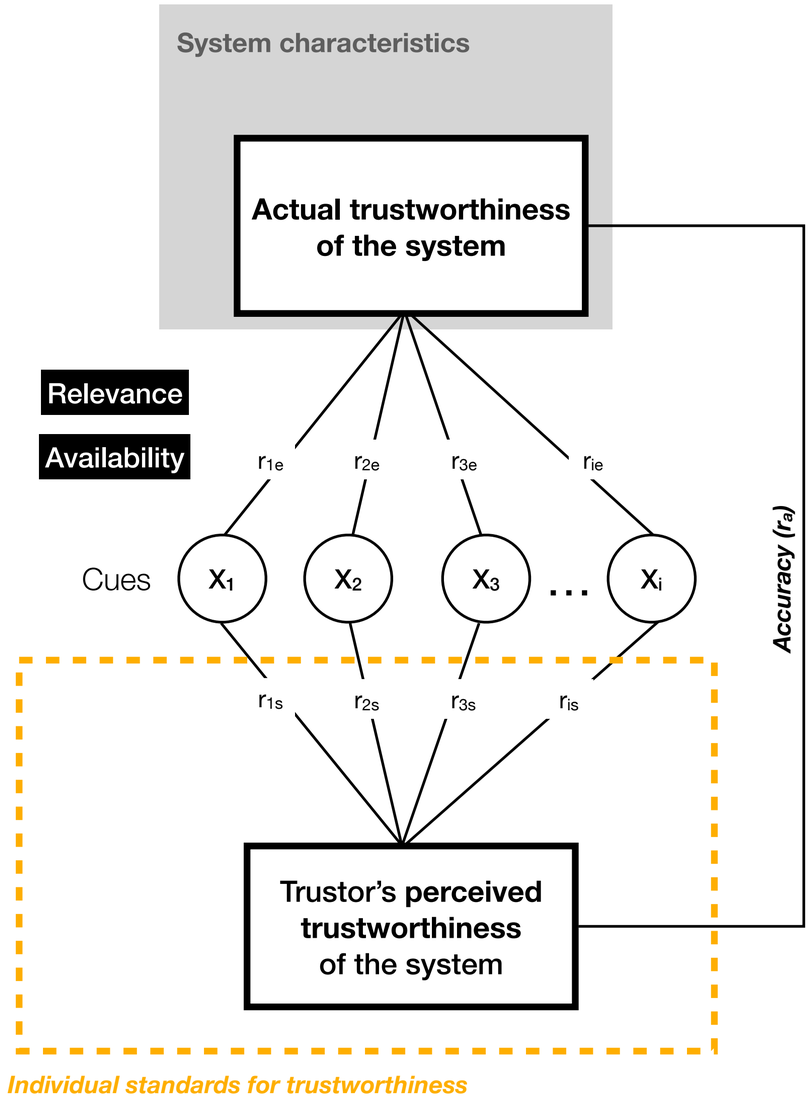
# Relations between the Model Components

**Cue relevance**

- Determines how indicative a cue is for the system's AT
- Relevant cues correlate strongly with the AT of a system (e.g. information regarding a system's performance in a task)
- Less relevant cues correlate low with the AT (e.g. popularity of a brand, ubiquitous presence of advertisement)

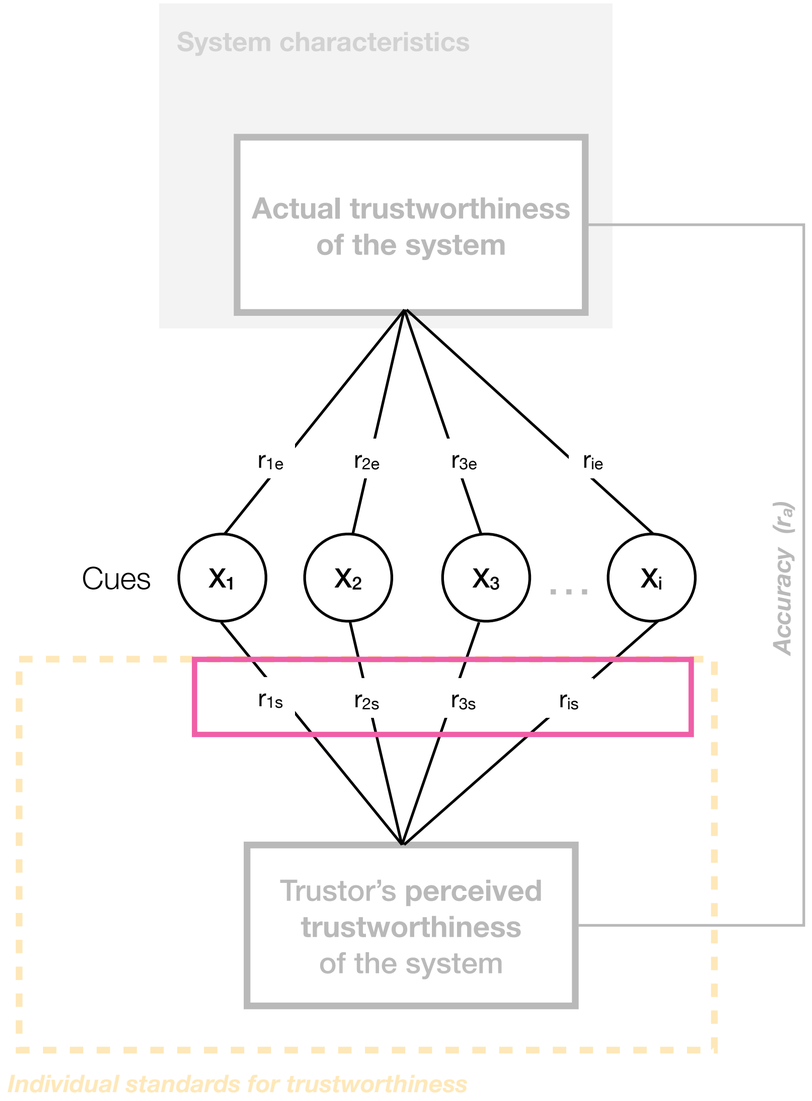
**Cue availability**

- Refers to the fact that cues can only be detected when they are accessible to the trustor (e.g. restricted access rights for training data)



# Cues

- Pieces of information that presumably provide insights regarding the *Actual Trustworthiness (AT)* of a system
- Cues are more or less closely related to the AT
- **Selection and weighting of the individual cues determines trustor's *Perceived Trustworthiness (PT)***



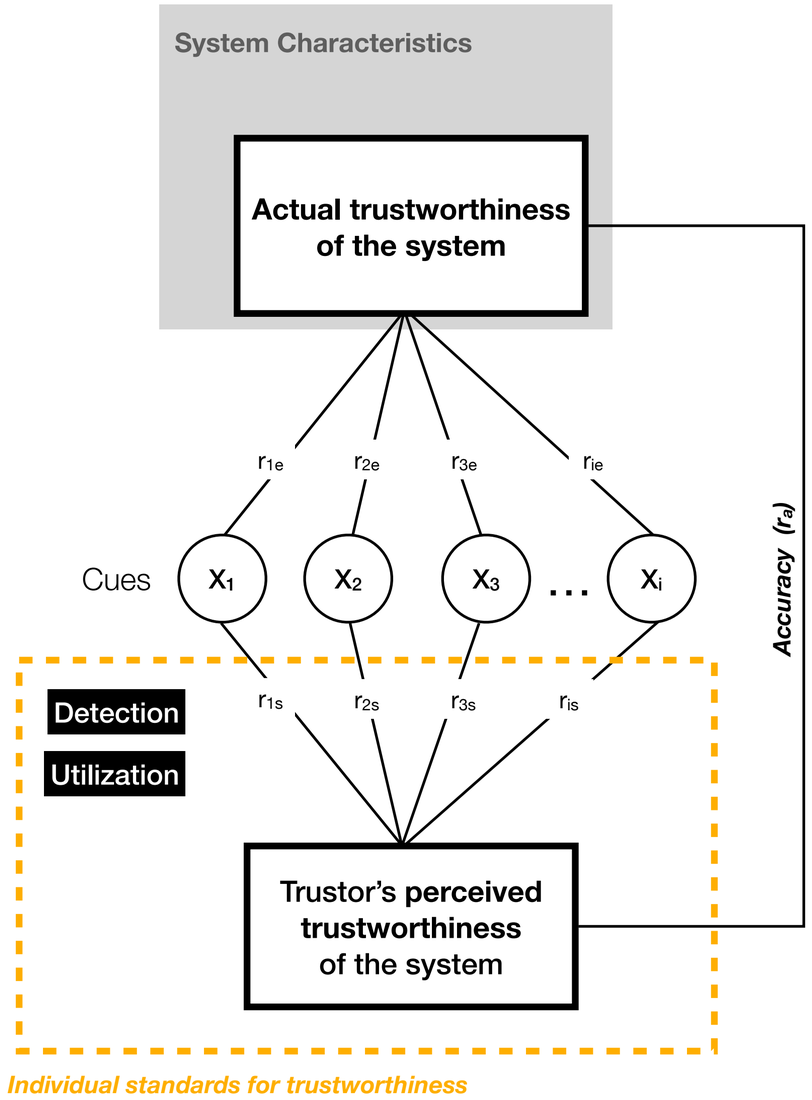
# Relations between the Model Components

**Cue detection**

- Relevant and available cues must be detected by the trustor
- Potential influencing factors:
  - Trustor’s attention capacities (Hawkins,1990), their situation awareness (Endsley,1995), time pressure, or their experience with a system
  - UI properties

**Cue utilization**

- Relevant, available, and detected cue must be correctly interpreted by the trustor
- Potential influencing factors
  - Domain knowledge
  - User's experience with system(s)

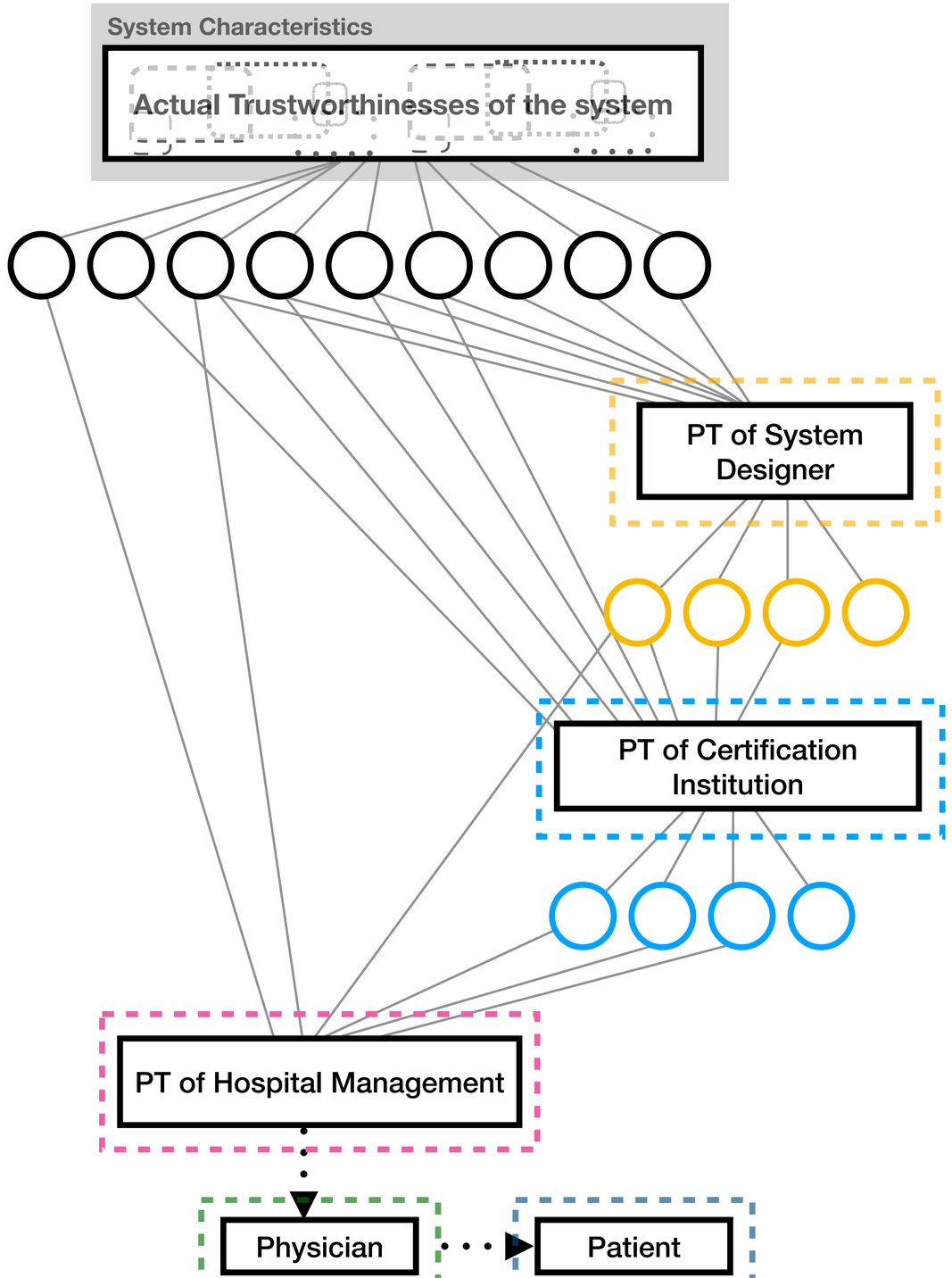


# Zoom Out

A macro level perspective

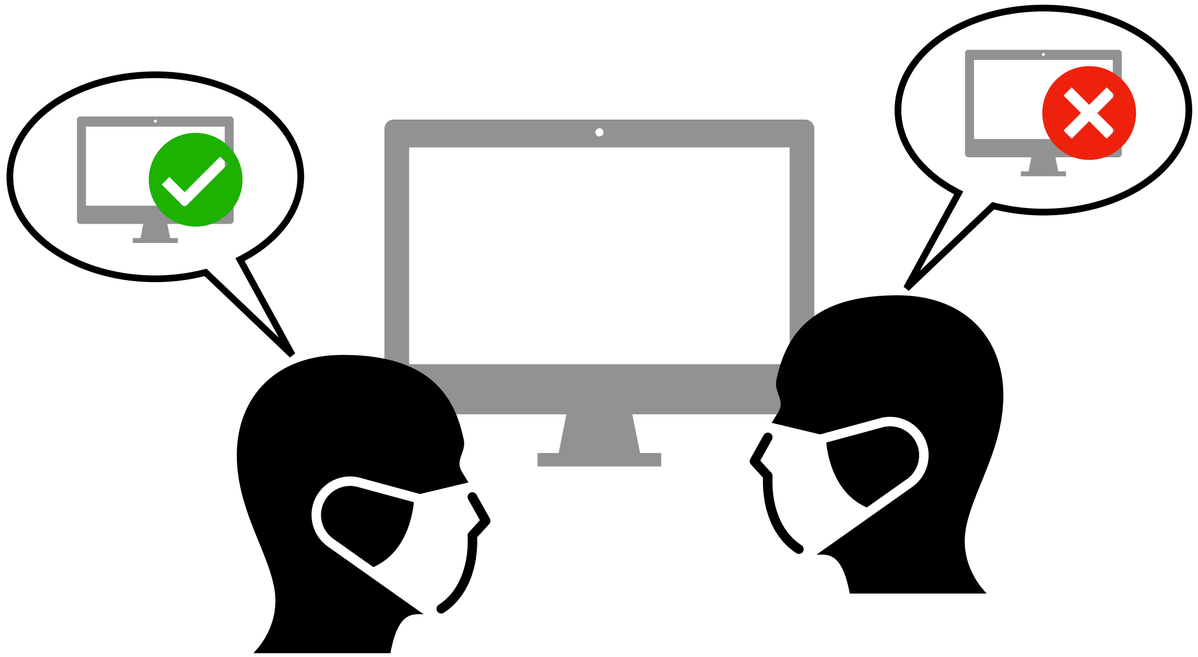
# Macro Level

- Trustworthiness assessment **proliferates**
- The described process on a micro level proceeds for different trustors
- The perceived trustworthiness of a trustor produces new cues for other trustors



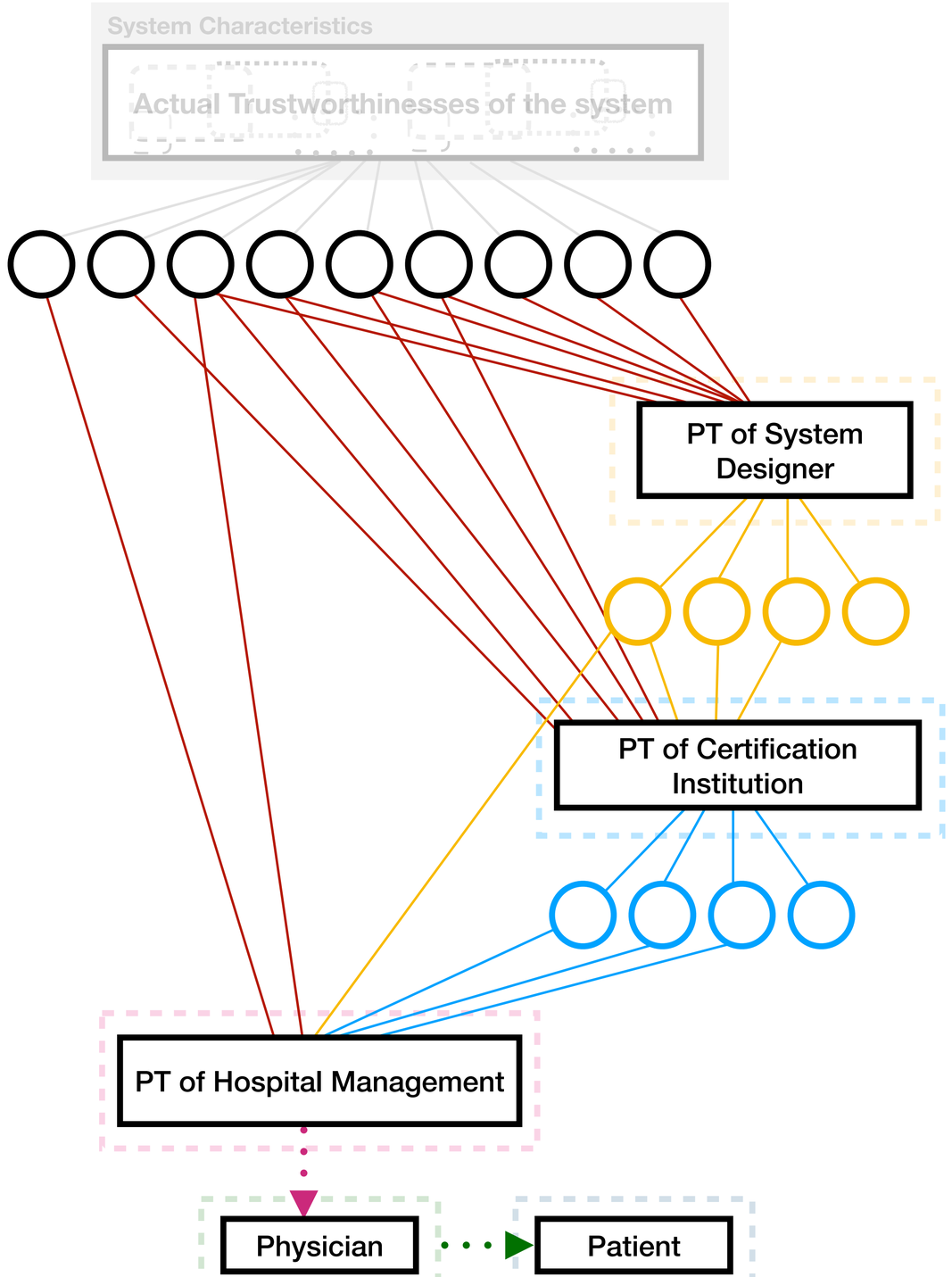
# Different trustors and different perceptions of trustworthiness

- Different trustors can come to different perceptions of system trustworthiness, although the cues and the system characteristics stay the same
    - Due to **different weighting of cues** and **different individual standards**



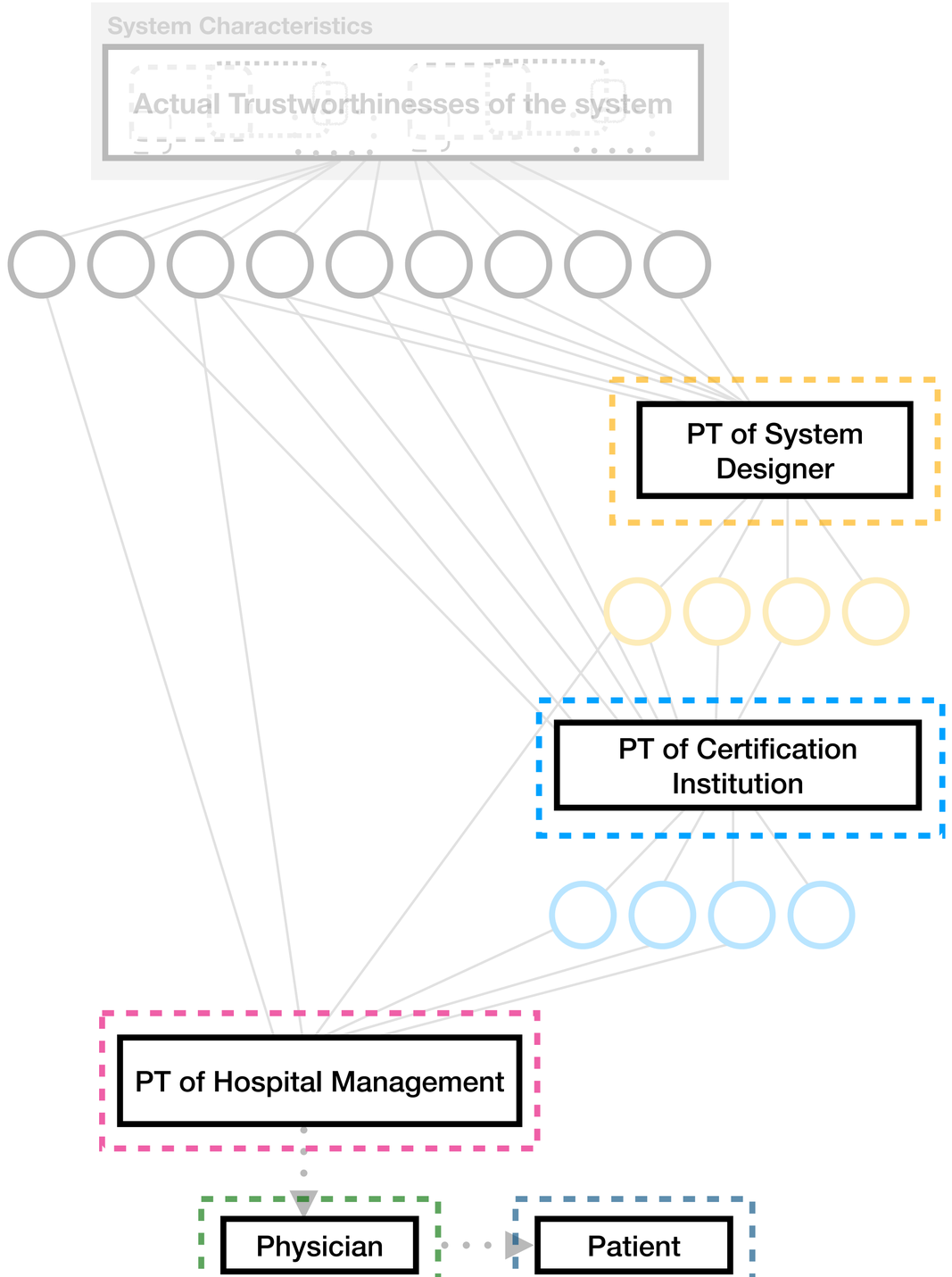
# Weighting of Cues

- **Relevance / Availability / Detection / Utilization**
  - Trustors at the beginning of the trustworthiness proliferation use proportionally **more primary cues**
    - Primary cues are available and they have the expertise to detect and utilize them correctly
  - Downstream trustors use more **secondary cues** (cues that have been processed, interpreted, and filtered by another trustor)
    - Relevance of secondary cues (e.g. marketing)
    - Availability (e.g. data access)
    - Detection & Utilization (lack of expertise)



# Individual standards

- Individual standards are explicitly or implicitly answering the question: **“What is a trustworthy system for me?”**
- Individual standards depend on e.g.
  - Goals and interests (System Designer vs. Hospital management vs. Physicians vs. Patients)
  - Cultural background (e.g. collectivistic vs. individualistic cultures)
  - Normative / Regulatory frame in which trustors operate (e.g., under the influence of the GDPR)



UKGM
UNIVERSITÄTSKLINIKUM
GIESSEN UND MARBURG

Thank you!

# Nadine Schlicker


nadine.schlicker@uni-marburg.de

https://www.linkedin.com/in/nadine-schlicker-87a132201/

https://www.researchgate.net/profile/Nadine-Schlicker

@SchlickerNad


**Get in touch!**

# References


- de Visser, E., Peeters, M. M. M., Jung, M., Kohn, S., Shaw, T., Pak, R., & Neerincx, M. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. International Journal of Social Robotics, 12. https://doi.org/10.1007/s12369-019-00596-x

- Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned From Human–Automation Research. Human Factors: The Journal of the Human Factors and Ergonomics Society, 59(1), 5–27. https://doi.org/10.1177/0018720816681350

- Funder, D. C. (1995). On the Accuracy of Personality Judgment:A Realistic Approach. 19.

- Hawkins, H., Hillyard, S., Luck, S., Mouloua, M., Downing, C., & Woodward, D. (1990). Visual Attention Modulates Signal Detectability. Journal of Experimental Psychology. Human Perception and Performance, 16, 802–811. https://doi.org/10.1037/0096-1523.16.4.802

- High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI.

- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

- Kuncel, N. R. (2018). Judgment and Decision Making in Staffing Research and Practice. In D. Ones, N. Anderson, C. Viswesvaran, & H. Sinangil, The SAGE Handbook of Industrial, Work and Organizational Psychology: Personnel Psychology and Employee Performance (pp. 474–487). SAGE Publications Ltd. https://doi.org/10.4135/9781473914940.n17

- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. The Academy of Management Review, 20(3), 709–734. https://doi.org/10.2307/258792

- Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. Mensch Und Computer 2021, 325–329. https://doi.org/10.1145/3473856.3474018

- Schlicker, N., Uhde, A., Baum, K., Hirsch, M. C., & Langer, M. (2022). *Calibrated Trust as a Result of Accurate Trustworthiness Assessment–Introducing the Trustworthiness Assessment Model.* PsyArXiv. 10.31234/osf.io/qhwvx