# From Input to Interaction

## The Role of Human Involvement in AI Development for Trustworthiness & Dependence

Lina Kluy*, Barbara Deml* & Eileen Roesler**
* Institute of Human and Industrial Engineering, Karlsruhe Institute of Technology, Germany
** Human-Agent Collaboration Lab, George Mason University, USA

lina.kluy@kit.edu

Does informing that humans interactively trained an AI-based support system affect trustworthiness perception and behavior?

Which effect does experiencing the system have?


## Background

**Perceived trustworthiness** through information & transparency [1] [2] [3]

**Descriptions of AI** elicit different mental models & beliefs [4]

**Information on AI training & development** is valued & human trained systems are perceived as more trustworthy [5]

**Social transparency** by infusing human elements and providing context [6]

**BUT:**

**Actual system experience** might have a more profound influence [6] [8]

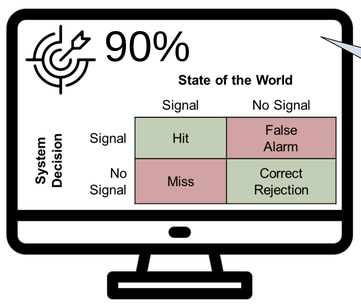
### Human involvement:

higher initial perceived trustworthiness,
more dependence behavior,
but the subjective effect vanishes after experience

## Methods

### 2x2-between-within-design + control group

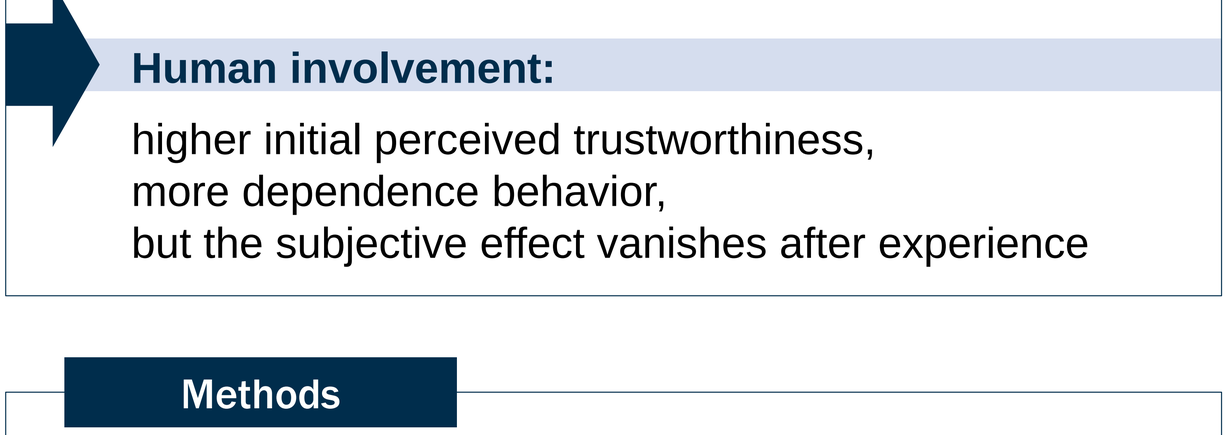
Human involvement (low, high) x
Time of measurement (post description, post experience)




"Find the **odd image** - if it exists."

Duration:
*M* = 30 min
*SD* = 14 min

40x




*N* = 122 AI lay people
61♀ 58♂ 2
77% (self-)employed

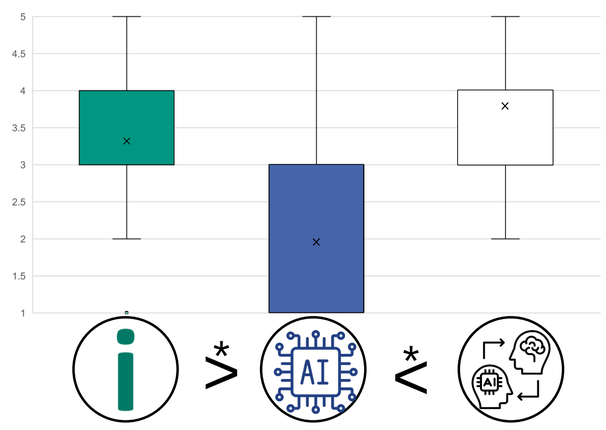
$M_{age}$ = 39 *SD* = 11.37
$M_{AI}$ = 3.30 *SD* = 1.39 (Scale: 1-7)
$M_{nfc}$ = 4.07 *SD* = .90 (Scale: 1-7)

## Results

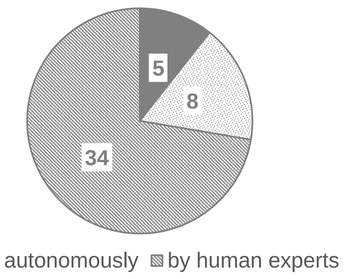
preliminary

### Did the manipulation work?

**Perceived human involvement**
(Scale:1-5, not at all-completely relying on human)

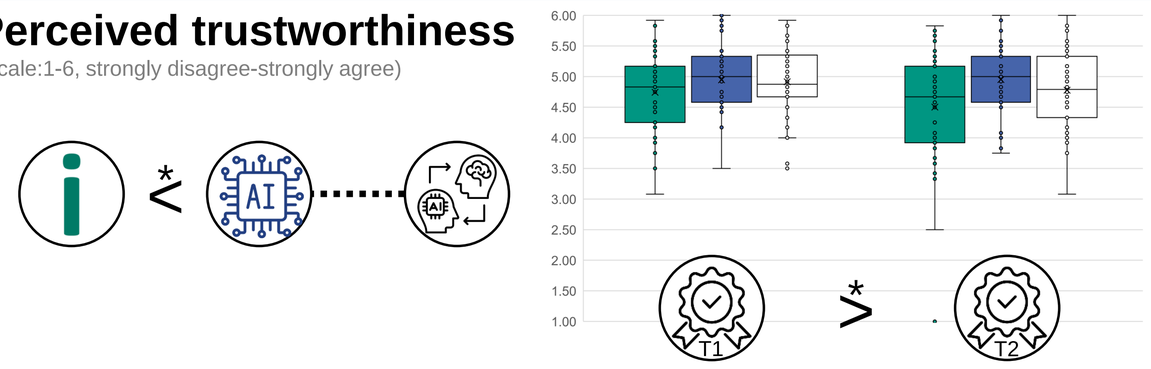



**Mental model about training**



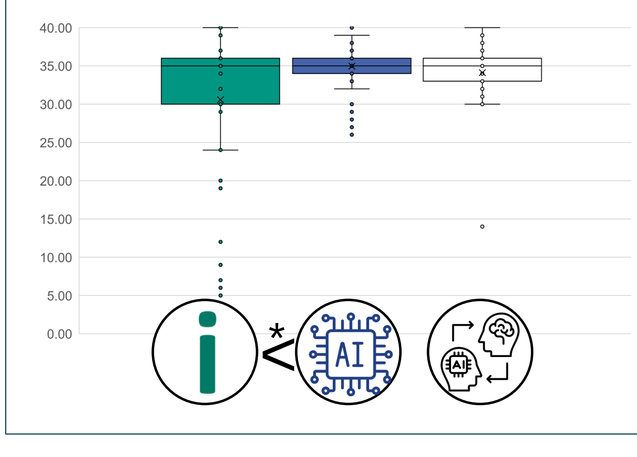

### Did human involvement affect evaluation?

**Perceived trustworthiness**
(Scale:1-6, strongly disagree-strongly agree)




### How did people behave?

**Dependence**




"I essentially just followed what the AI told me."

**Human+AI accuracy**
= (H+CR)/(H+CR+FA+M)
*M* = 88.55%
*SD* = 16%

"I studied the images myself carefully."

**Overruling**
= min. 1 of 2 M / FA
M 78.7%
FA 70.5%

## Conclusion

**Informing people of human involvement** worked, but actual content might be irrelevant for perceived trustworthiness.

**Perceived trustworthiness** was quite high, but lower after experience.

**Decision Behavior** was accurate, but diverse, e.g. complete dependence vs. verification.

**Sources:** [1] Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., & Langer, M. (2022). *A Micro and Macro Perspective on Trustworthiness: Theoretical Underpinnings of the Trustworthiness Assessment Model (TrAM).* https://doi.org/10.31234/osf.io/qhwvx , [2] Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent Transparency: A Review of Current Theory and Evidence. *IEEE Transactions on Human-Machine Systems*, *50*(3), 215–224. https://doi.org/10.1109/THMS.2020.2965529 , [3] Rajabiyazdi, F. (2023). *The Effects of Automation Transparency on Human Performance* [Dissertation]. , [4] Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022). "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?, *5*, 1–28. https://doi.org/10.1145/3491102.3517527 , [5] Ashoori, M., & Weisz, J. D. (2019, December 5). *In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes*. http://arxiv.org/pdf/1912.02675v1 , [6] Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding Explainability: Towards Social Transparency in AI systems. In Y. Kitamura et al. (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). ACM. https://doi.org/10.1145/3411764.3445188 , [7] Liu, B. (2021). In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384–402. https://doi.org/10.1093/jcmc/zmab013 , [8] Mariadassou, S., Klesse, A.-K., & Boegershausen, J. (2024). Averse to what: Consumer aversion to algorithmic labels, but not their outputs? *Current Opinion in Psychology, 58.*. https://doi.org/10.1016/j.copsyc.2024.101839



Supp. Material:
https://osf.io/xwm48