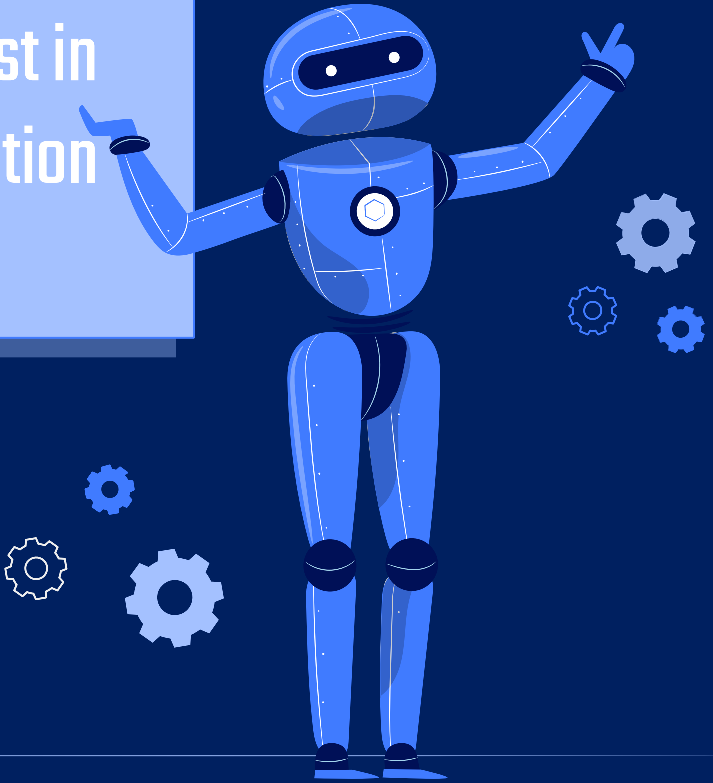
Understanding appropriate trust in Human-AI interaction



Siddharth Mehrotra
PhD Candidate (2nd year)
TU Delft, The Netherlands

INTRODUCTION

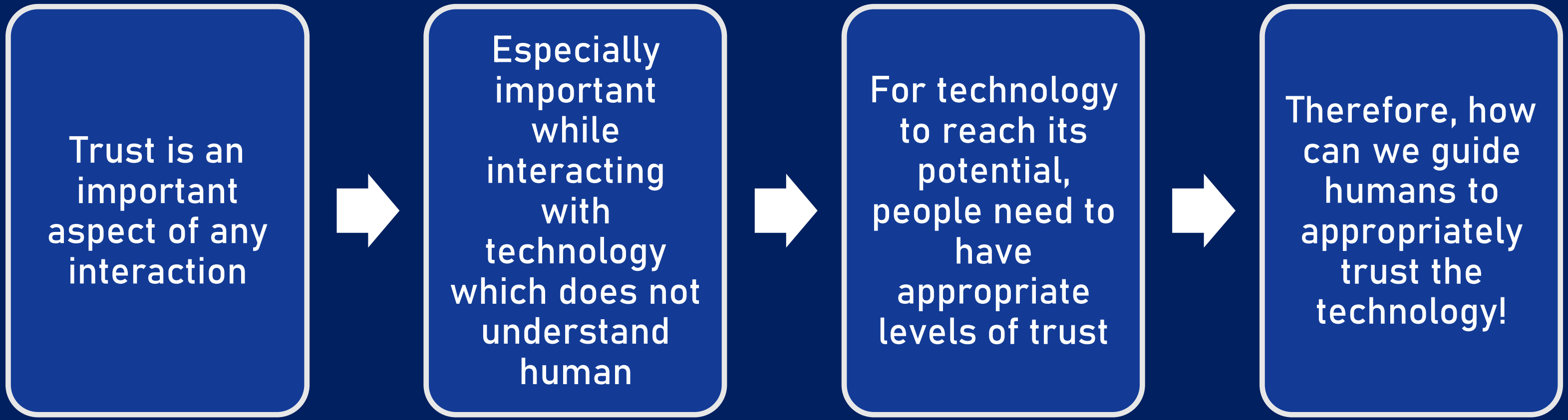
01.

My research interest lies in understanding appropriate trust in Human-AI interaction.

02.

I use techniques from Human Computer/Robot Interaction, AI & Psychology domains to guide humans to appropriately trust their AI systems .

Research Question Formulation



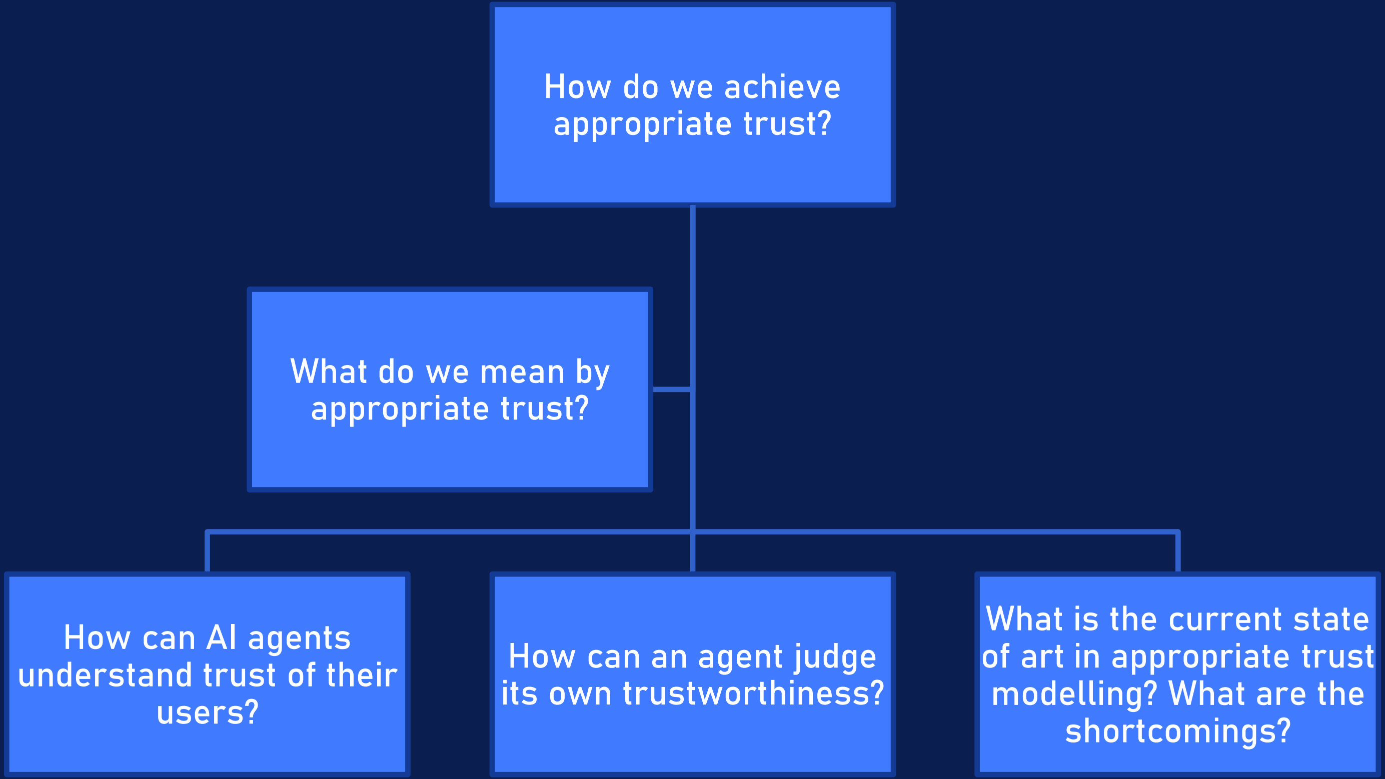
Trust is an important aspect of any interaction

Especially important while interacting with technology which does not understand human

For technology to reach its potential, people need to have appropriate levels of trust

Therefore, how can we guide humans to appropriately trust the technology!

Research Plan



How do we achieve appropriate trust?

What do we mean by appropriate trust?

How can AI agents understand trust of their users?

Role of Human Values

How can an agent judge its own trustworthiness?

Looking beyond agent ability, role of inverse trust

What is the current state of art in appropriate trust modelling? What are the shortcomings?

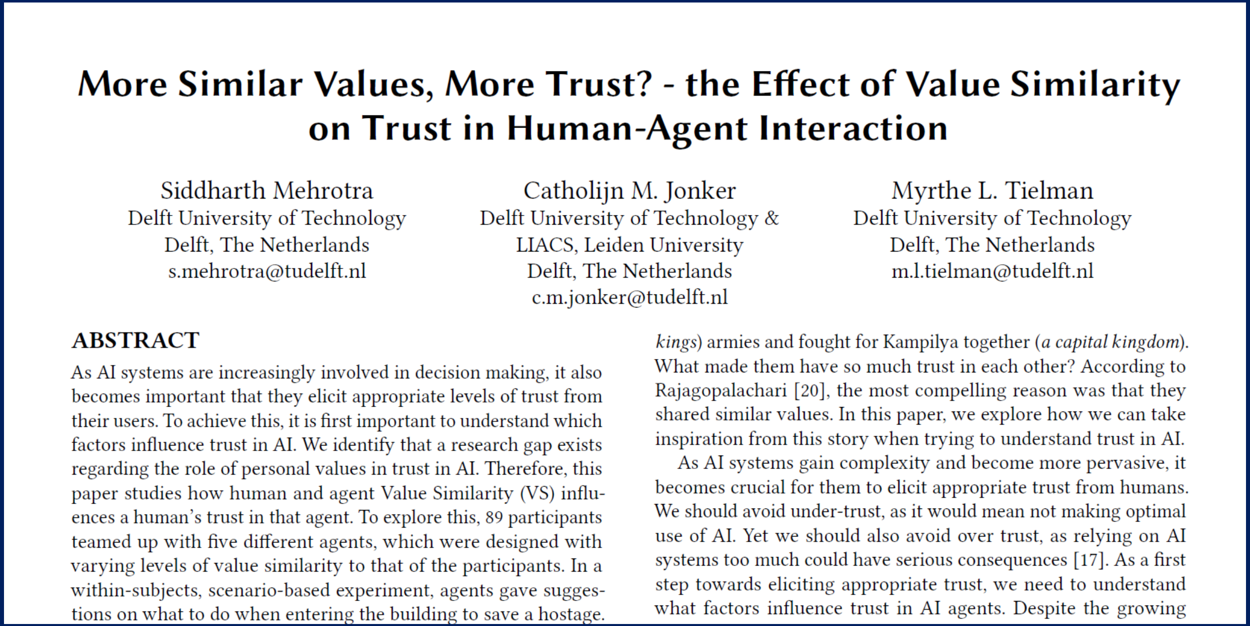
A Systematic Review

How can AI agents understand trust of their users?

Role of Human Values (when we think of values, we think of what is important to us)

Current Progress

Mehrotra S., Tielman M. L. and Jonker C.M., More similar values, more trust? - Effect of Value Similarity on Human-AI trust in proceedings of *4th AAAI/ACM Conference on AI, Ethics, and Society (AIES)* - to appear – full paper, 2021.



More Similar Values, More Trust? - the Effect of Value Similarity on Trust in Human-Agent Interaction

Siddharth Mehrotra
Delft University of Technology
Delft, The Netherlands
s.mehrotra@tudelft.nl

Catholijn M. Jonker
Delft University of Technology &
LIACS, Leiden University
Delft, The Netherlands
c.m.jonker@tudelft.nl

Myrthe L. Tielman
Delft University of Technology
Delft, The Netherlands
m.l.tielman@tudelft.nl

ABSTRACT

As AI systems are increasingly involved in decision making, it also becomes important that they elicit appropriate levels of trust from their users. To achieve this, it is first important to understand which factors influence trust in AI. We identify that a research gap exists regarding the role of personal values in trust in AI. Therefore, this paper studies how human and agent Value Similarity (VS) influences a human's trust in that agent. To explore this, 89 participants teamed up with five different agents, which were designed with varying levels of value similarity to that of the participants. In a within-subjects, scenario-based experiment, agents gave suggestions on what to do when entering the building to save a hostage.

kings) armies and fought for Kampilya together (*a capital kingdom*). What made them have so much trust in each other? According to Rajagopalachari [20], the most compelling reason was that they shared similar values. In this paper, we explore how we can take inspiration from this story when trying to understand trust in AI.

As AI systems gain complexity and become more pervasive, it becomes crucial for them to elicit appropriate trust from humans. We should avoid under-trust, as it would mean not making optimal use of AI. Yet we should also avoid over trust, as relying on AI systems too much could have serious consequences [17]. As a first step towards eliciting appropriate trust, we need to understand what factors influence trust in AI agents. Despite the growing

More Similar Values, More Trust? - The Effect of Value Similarity on Trust in Human-Agent Interaction – AAAI/ACM AIES'21 (to appear)

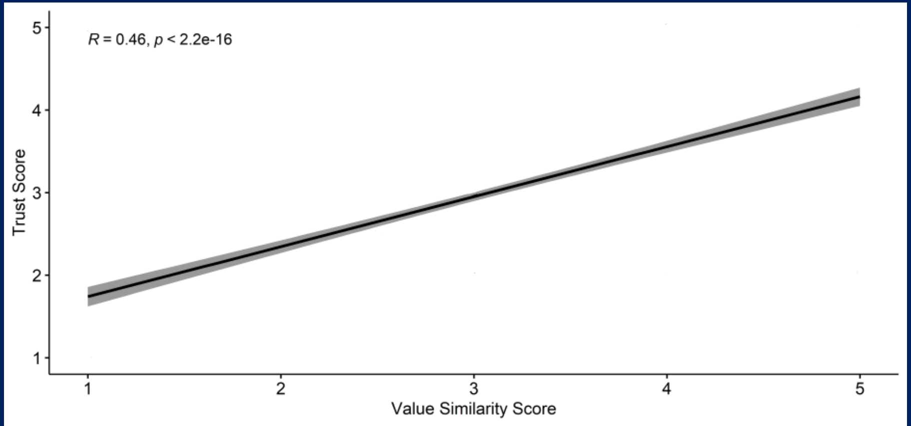
- **Hypothesis:** " Value similarity between the user and the agent positively affects the trust a user has in that agent. "

- **Scenario**: Rescue **a hostage from a building with the help of an agent**.

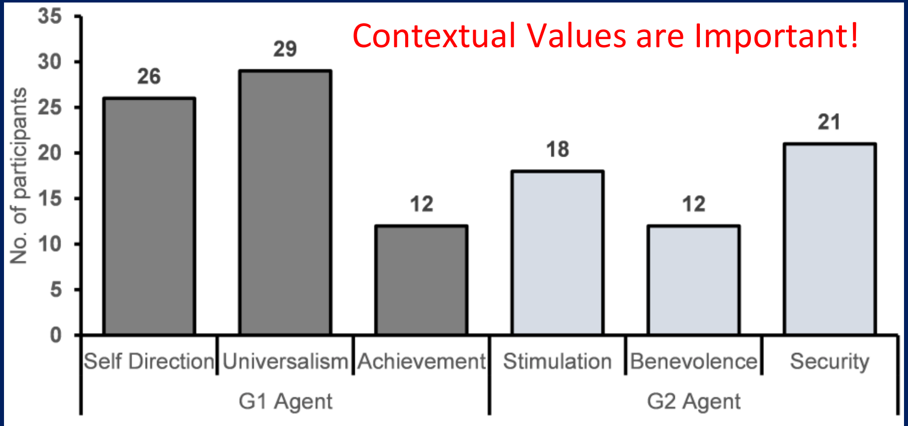
- **Interaction**: Agent explanations are based on values which they possess and are of varying similarity with that of user.

- **Measures**: Value Similarity Questionnaire (Siegrist, '05) and Human Computer Trust Scale (Gulati, '19).

More Similar Values, More Trust? - The Effect of Value Similarity on Trust in Human-Agent Interaction – AAAI/ACM AIES’21 (to appear)



Value similarity and trust are significantly moderately correlated.



(values ranked 1 and 2 of participant)

(values ranked 3 and 4 of participant)

The numbers on the top of the histogram represent how often those values occurred

This study shows that value similarity between an agent and a human is positively related to how much that human trusts the agent.

How can an agent judge its own trustworthiness?

Looking beyond agent ability,
role of inverse trust

Current Progress

Jorge C., **Mehrotra S.**, Jonker C.M. and Tielman M. L., Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust (Position Paper) in Human-Agent Teams in proceedings of 22nd edition *of TRUST workshop* at 20th International Conference on Autonomous Agents and Multiagent Systems (*AAMAS*), 2021.

Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams

Carolina Centeio Jorge[1]
C.Jorge@tudelft.nl

Siddharth Mehrotra[1]
S.Mehrotra@tudelft.nl

Catholijn M. Jonker[1,2]
C.M.Jonker@tudelft.nl

Myrthe L. Tielman[1]
M.L.Tielman@tudelft.nl

[1] Department of Intelligent Systems, TU Delft
[2] Leiden Institute of Advanced Computer Science, Leiden University

Abstract

In human-agent teams, how one teammate trusts another teammate should correspond to the latter's actual trustworthiness, creating what

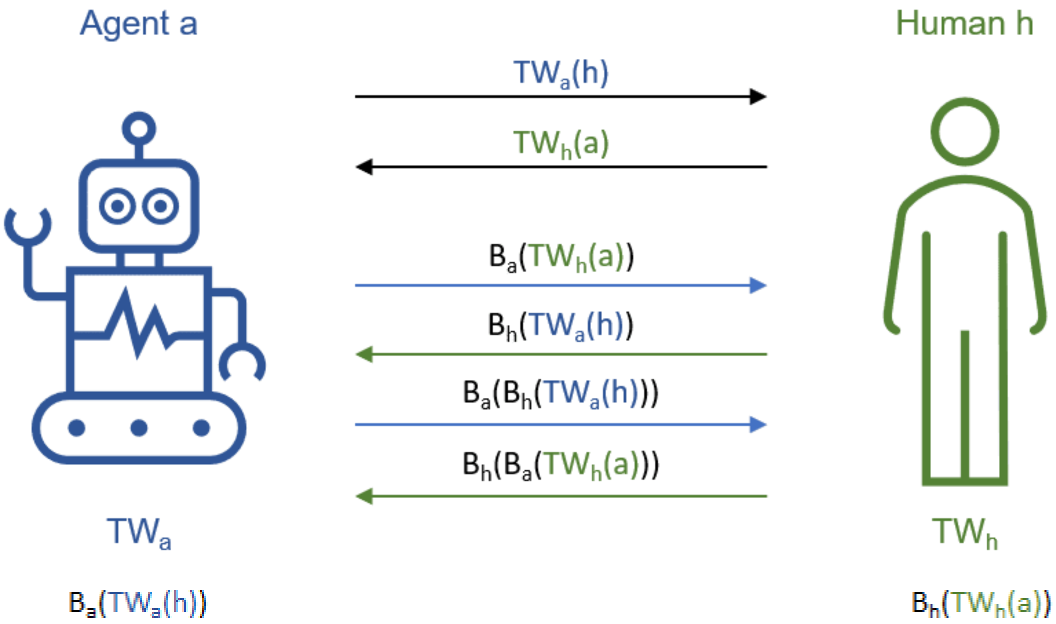
Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams – Trust Workshop @ AAMAS ‘21

- Appropriate Trust: When the belief regarding an entity's trustworthiness (to do a task) corresponds to their actual trustworthiness.

- Agents can use beliefs about trustworthiness to represent how they trust their human teammates, as well as to reason about how their human teammates trust them.

- Agents can define their own trustworthiness, using the concepts of ability, benevolence and integrity.

Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams – Trust Workshop @ AAMAS '21



$$\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_a(h))) \equiv \mathcal{B}_h(\mathcal{TW}_a(h))$$

Synopsis

- As a first step towards eliciting appropriate trust, we need to understand what factors influence trust in AI agents.

- Despite the growing attention in research on trust in AI agents, a lot is still unknown about people ' s perceptions of trust in AI agents.

- Our notion of understanding appropriate trust is divided in 3 different interconnected ways

1. Agents understanding trust of their users.
2. Agent ' s reflection of its own trustworthiness
3. Reviewing already established appropriate trust models