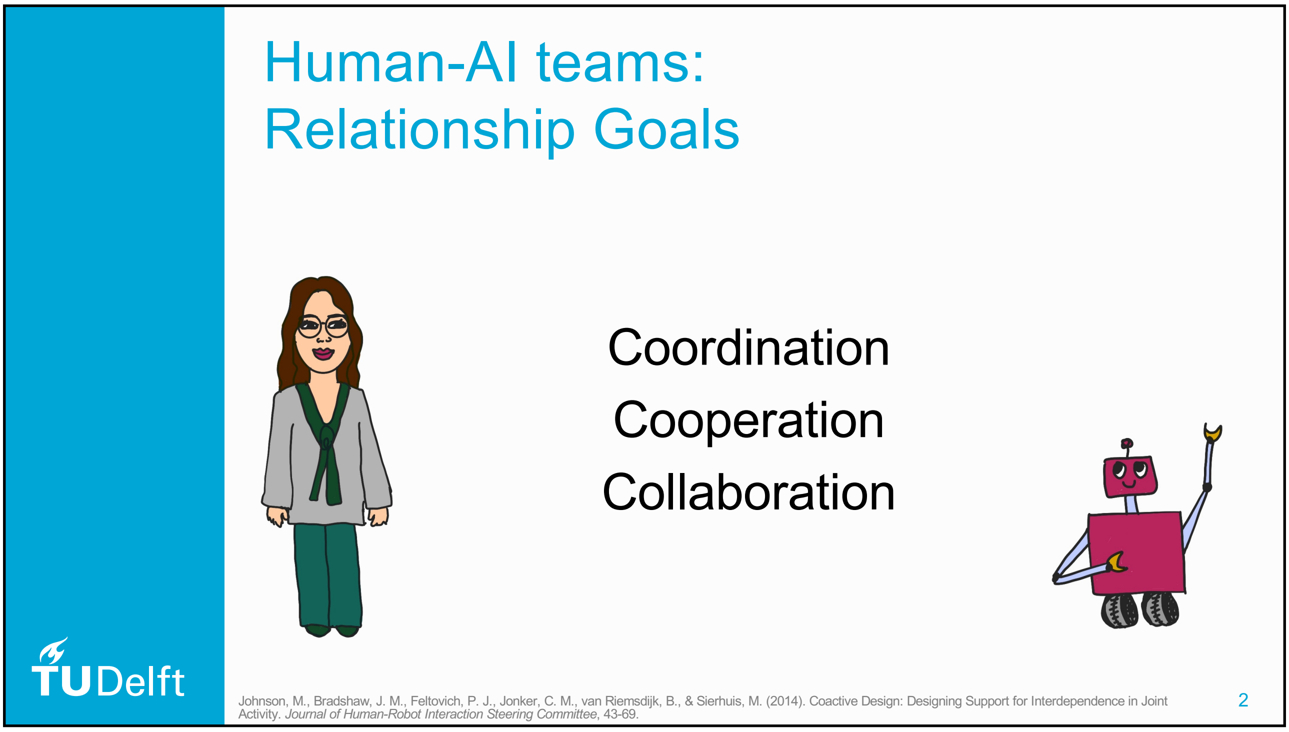
Towards Modelling Appropriate Mutual Trust in Human-Agent Teams

Carolina Centeio Jorge
C.Jorge@tudelft.nl
Siddharth Mehrotra
Catholijn M. Jonker
Myrthe L. Tielman

1



Human-AI teams: Relationship Goals

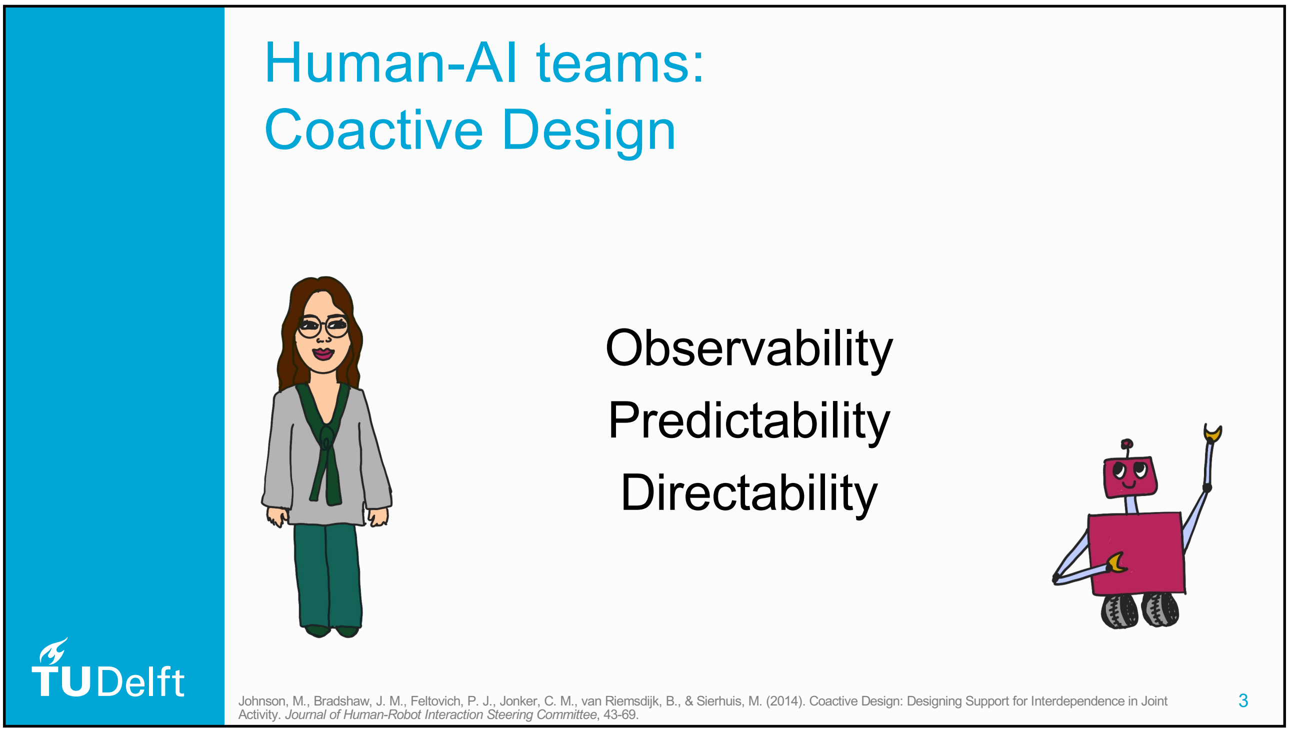
Coordination

Cooperation

Collaboration

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, B., & Sierhuis, M. (2014). Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of Human-Robot Interaction Steering Committee*, 43-69.

2



Human-AI teams: Coactive Design

Observability

Predictability

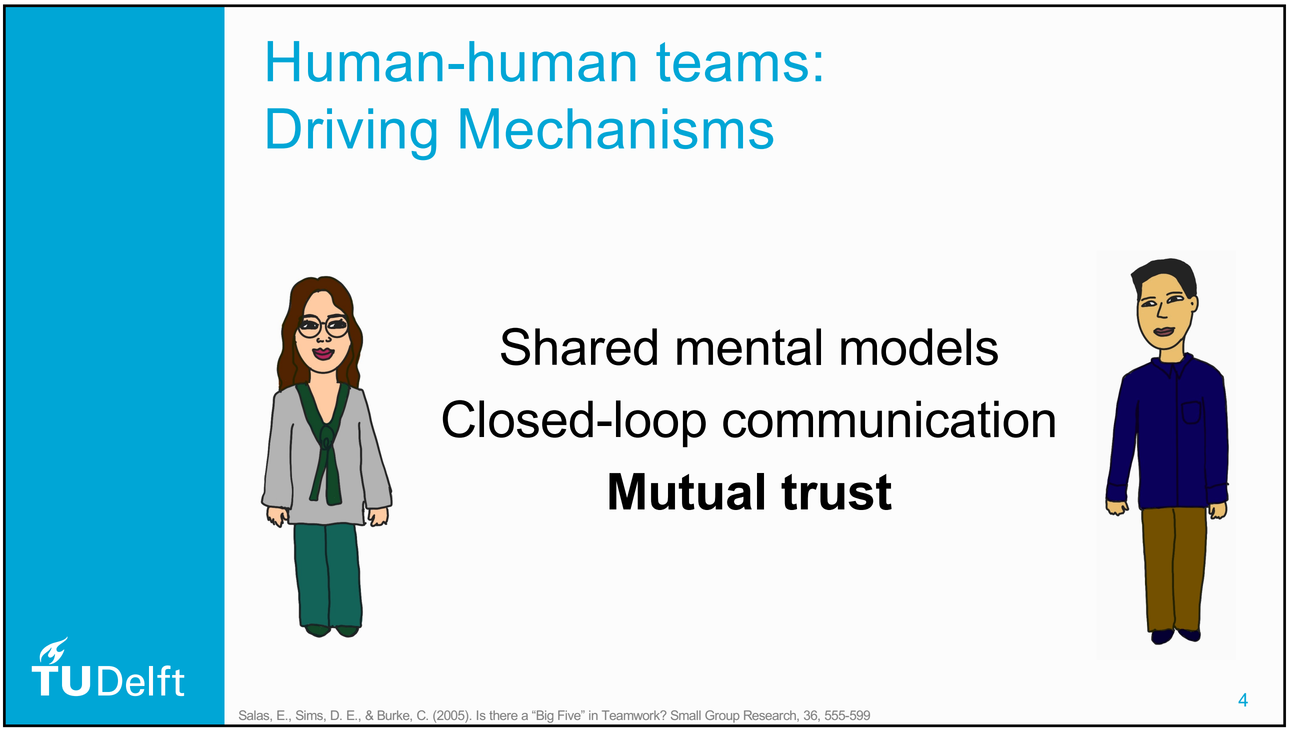
Directability

TU Delft

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, B., & Sierhuis, M. (2014). Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of Human-Robot Interaction Steering Committee*, 43-69.

3

3



Human-human teams: Driving Mechanisms

Shared mental models

Closed-loop communication

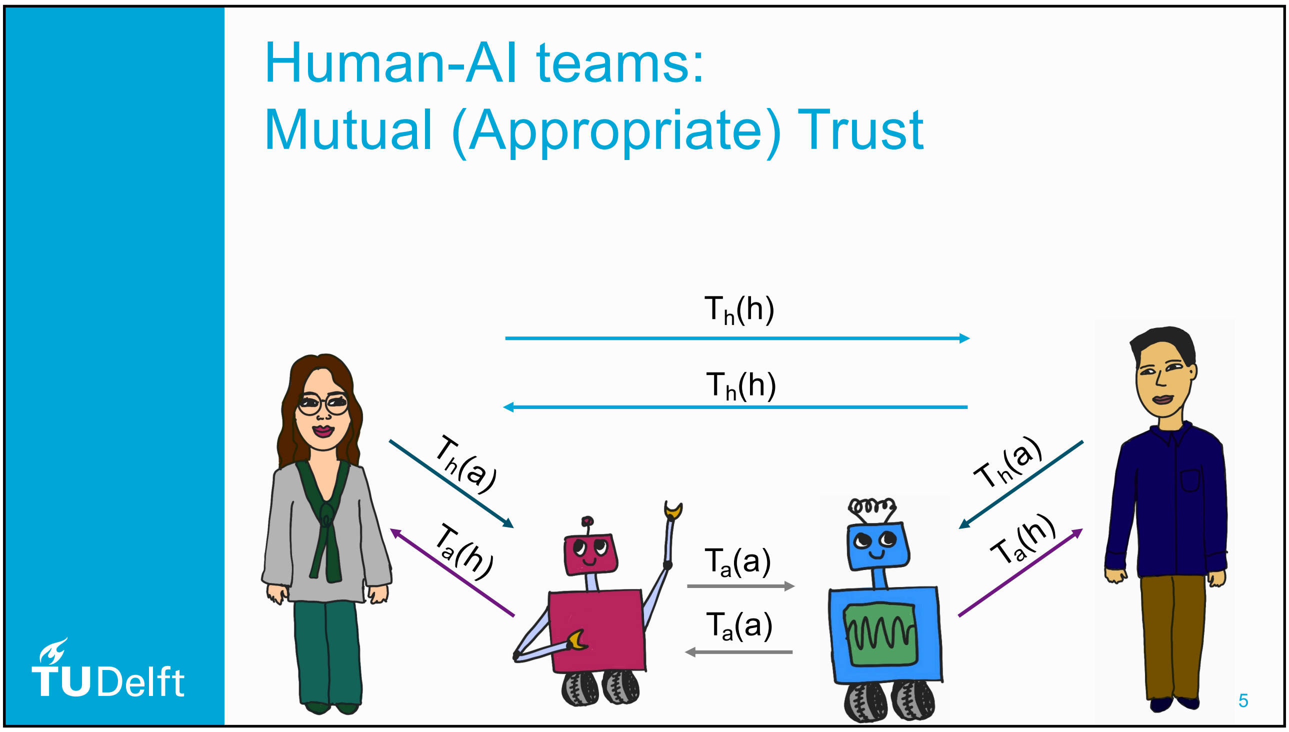
Mutual trust

TU Delft

Salas, E., Sims, D. E., & Burke, C. (2005). Is there a "Big Five" in Teamwork? Small Group Research, 36, 555-599

4

4



Human-AI teams: Mutual (Appropriate) Trust

TU Delft

5

5

Human-AI teams: Why should an agent appropriately trust?

Ensure team’s goal

Mitigate possible risks

Decision making

TU Delft

6

6

What is appropriate trust?

TUDelft

7

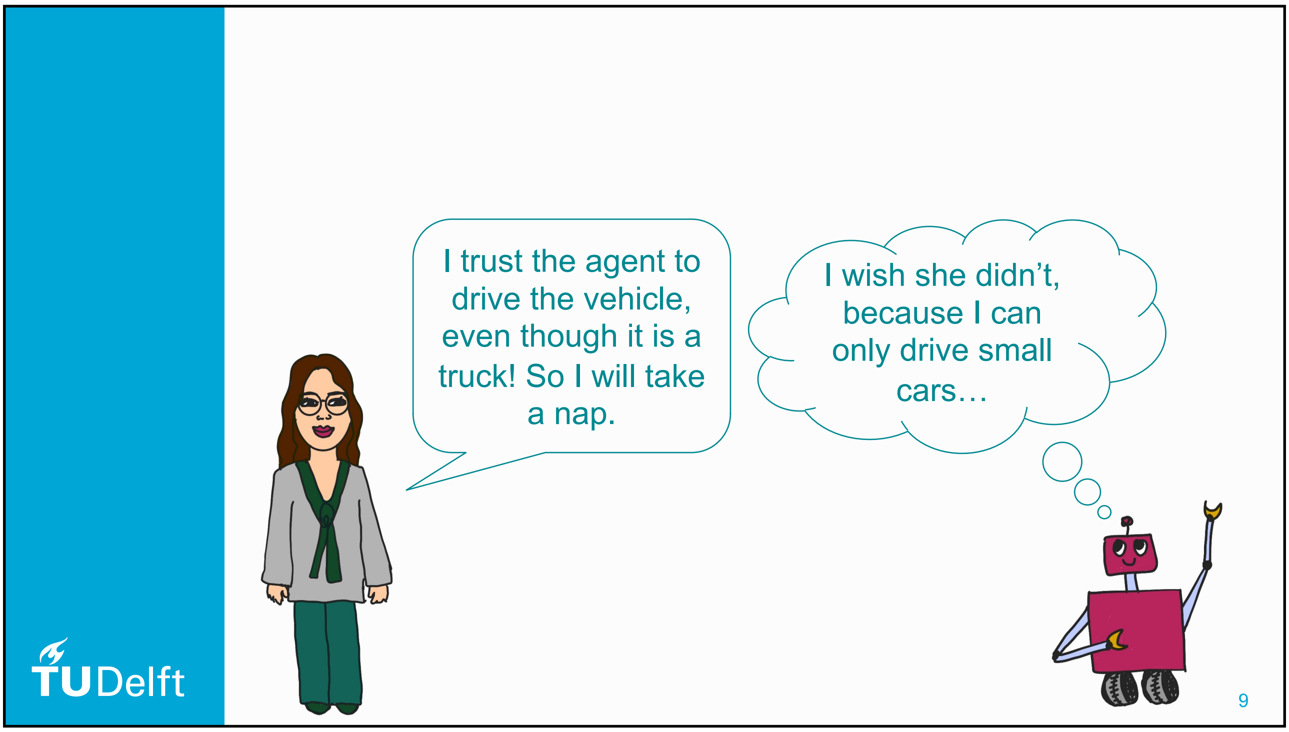
7

What is **in**appropriate trust?

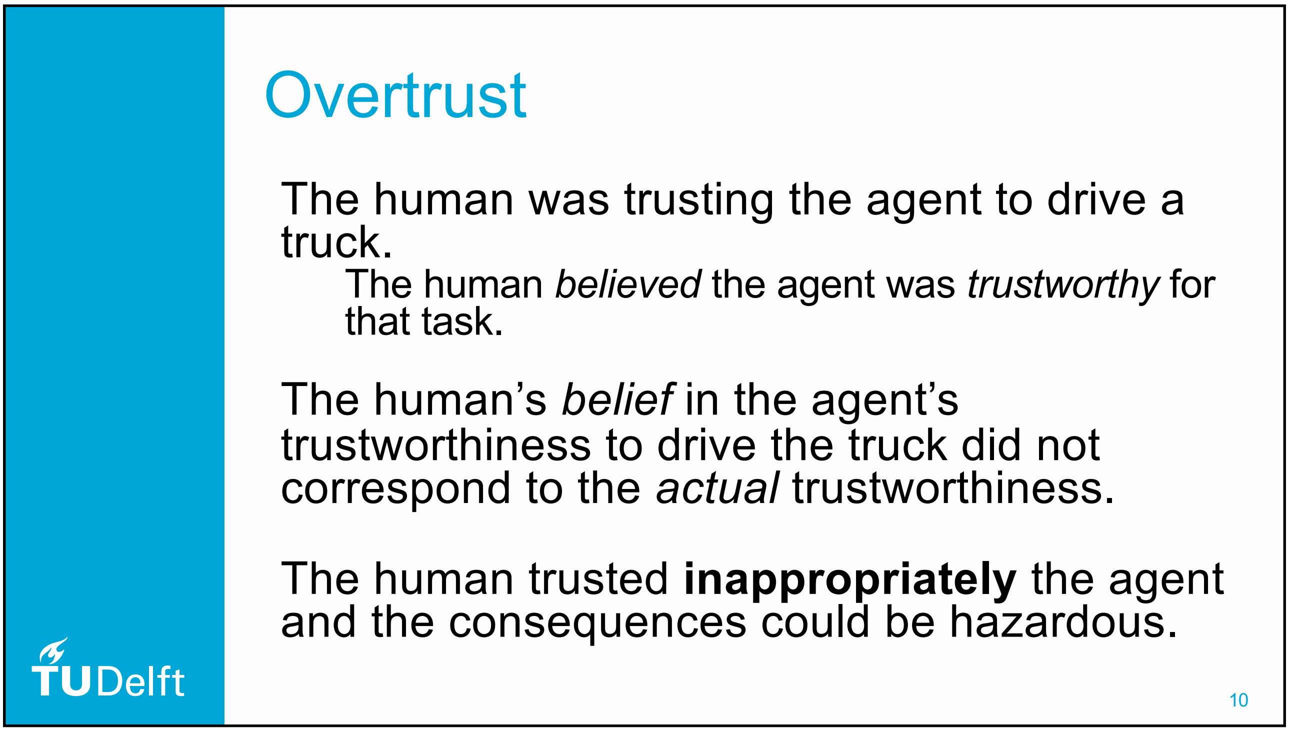
TUDelft

8

8


I trust the agent to drive the vehicle, even though it is a truck! So I will take a nap.
I wish she didn't, because I can only drive small cars…

9



Overtrust

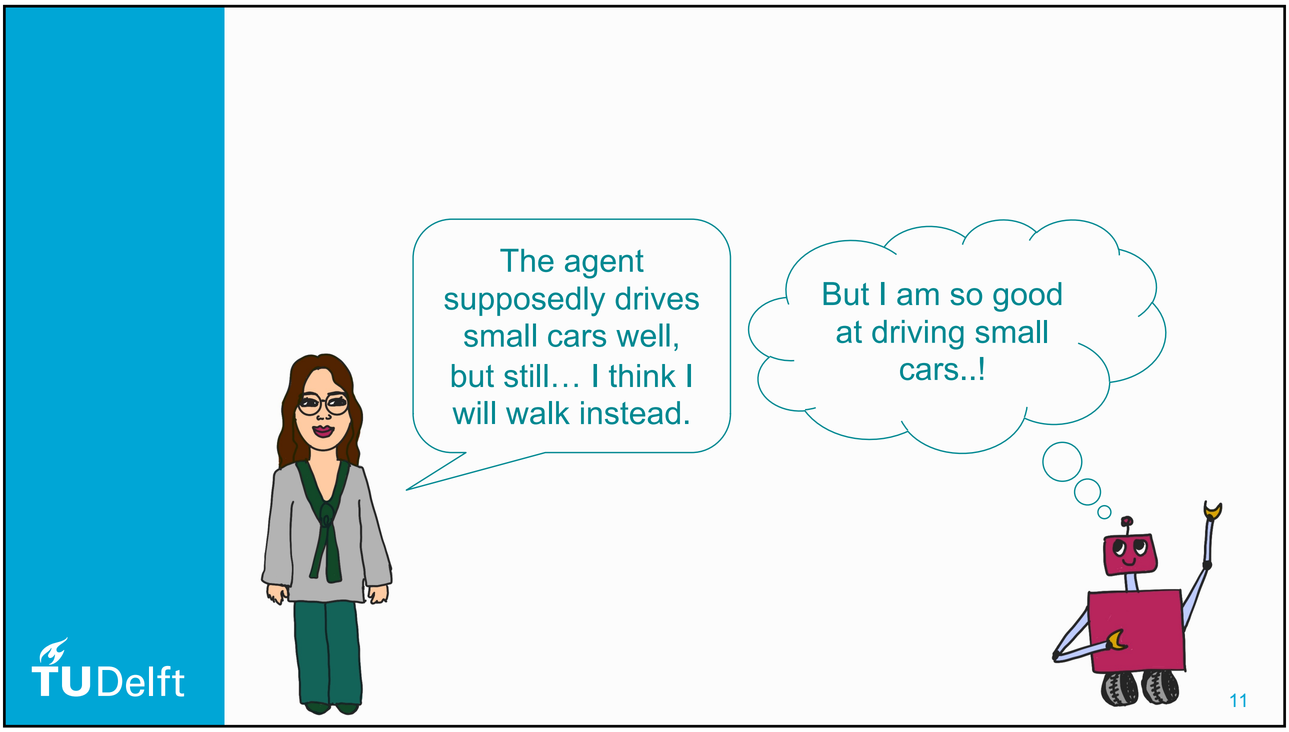
The human was trusting the agent to drive a truck.

The human *believed* the agent was *trustworthy* for that task.

The human's *belief* in the agent's trustworthiness to drive the truck did not correspond to the *actual* trustworthiness.

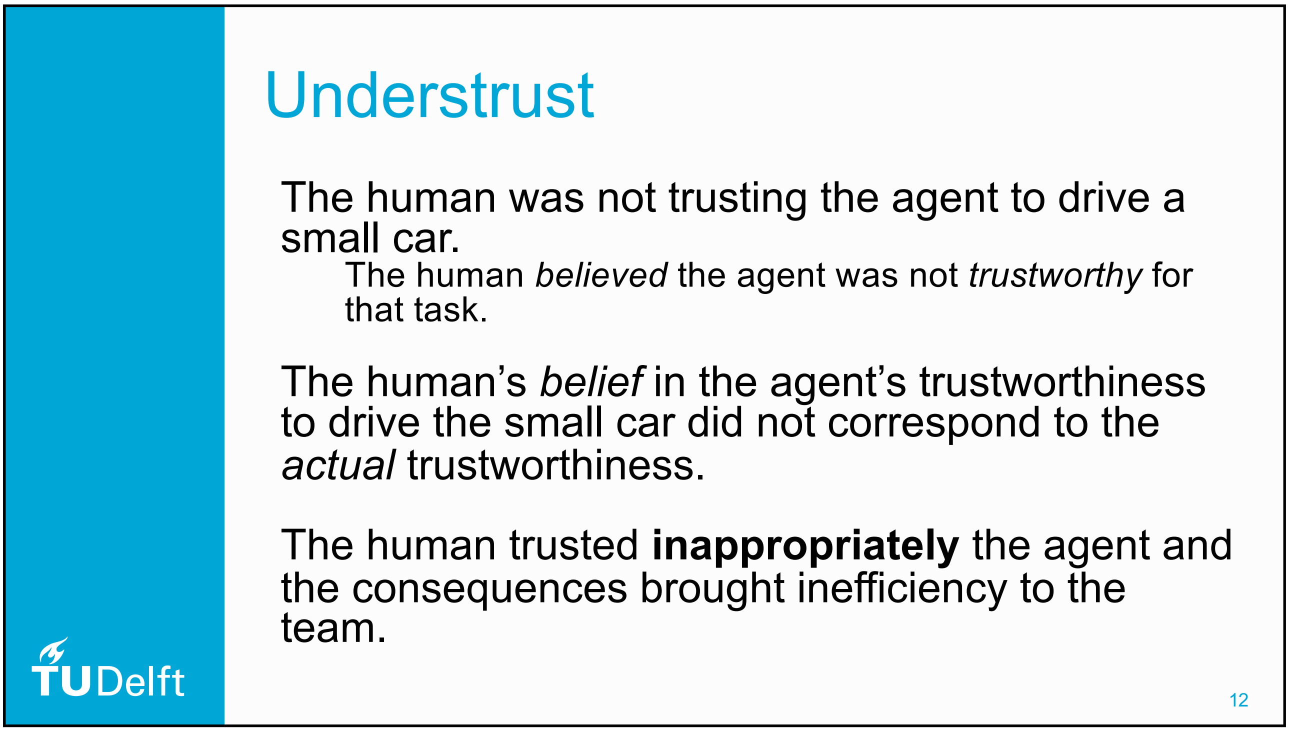
The human trusted **inappropriately** the agent and the consequences could be hazardous.

10



The agent supposedly drives small cars well, but still… I think I will walk instead.

But I am so good at driving small cars..!



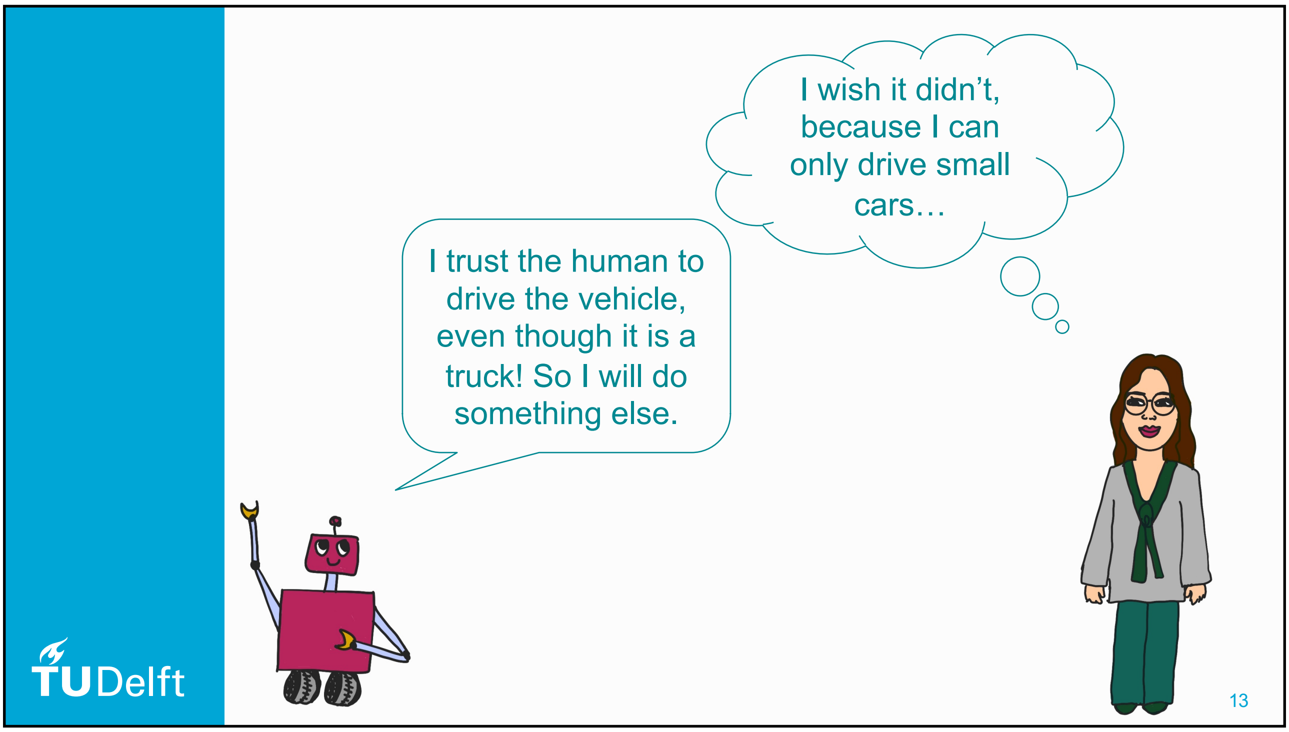
Understrust

The human was not trusting the agent to drive a small car.

The human *believed* the agent was not *trustworthy* for that task.

The human's *belief* in the agent's trustworthiness to drive the small car did not correspond to the *actual* trustworthiness.

The human trusted **inappropriately** the agent and the consequences brought inefficiency to the team.

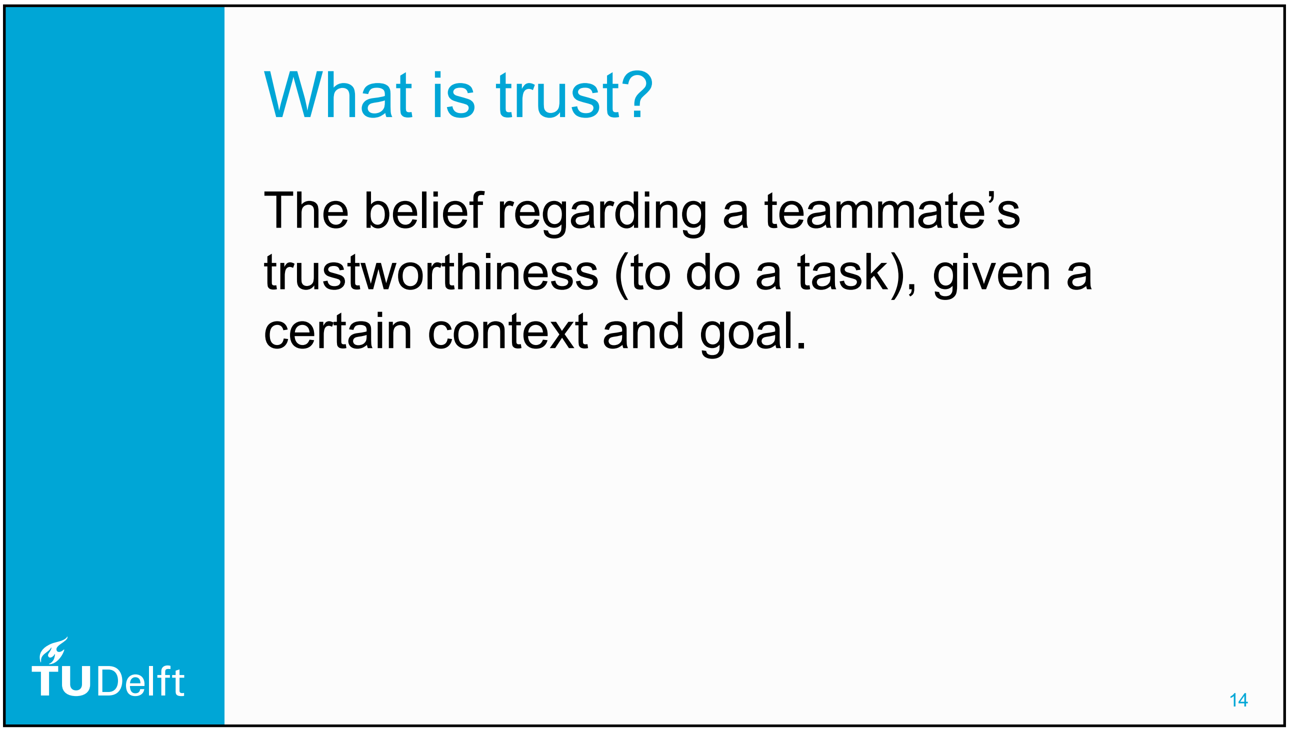


I wish it didn’t, because I can only drive small cars…

I trust the human to drive the vehicle, even though it is a truck! So I will do something else.

13

13

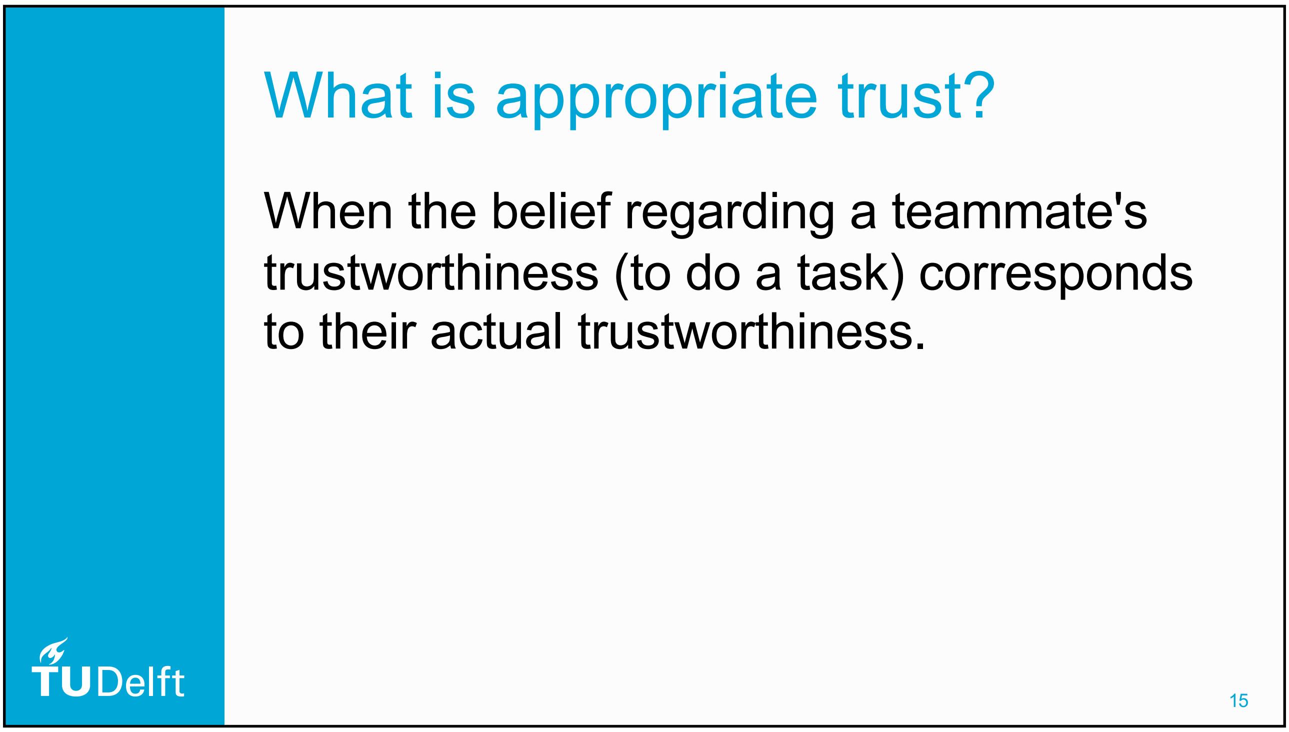


What is trust?

The belief regarding a teammate’s trustworthiness (to do a task), given a certain context and goal.

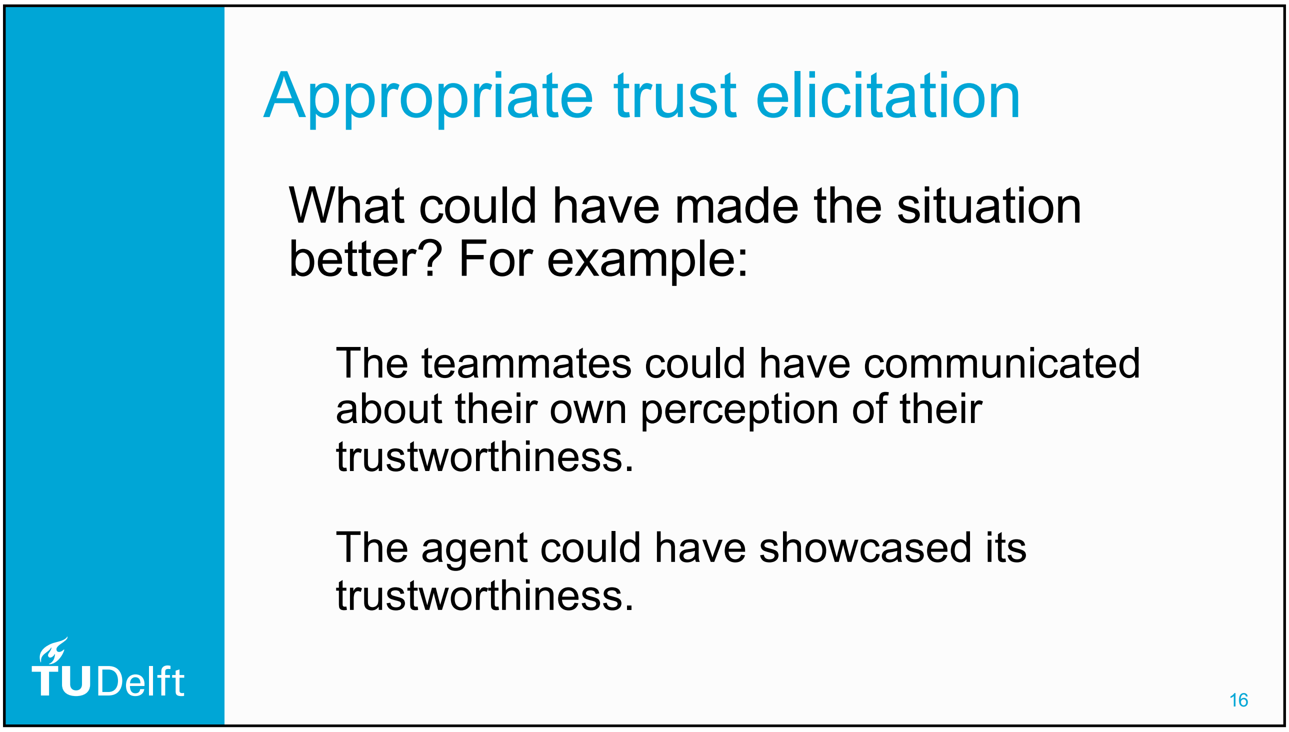
14

14



What is appropriate trust?

When the belief regarding a teammate's trustworthiness (to do a task) corresponds to their actual trustworthiness.

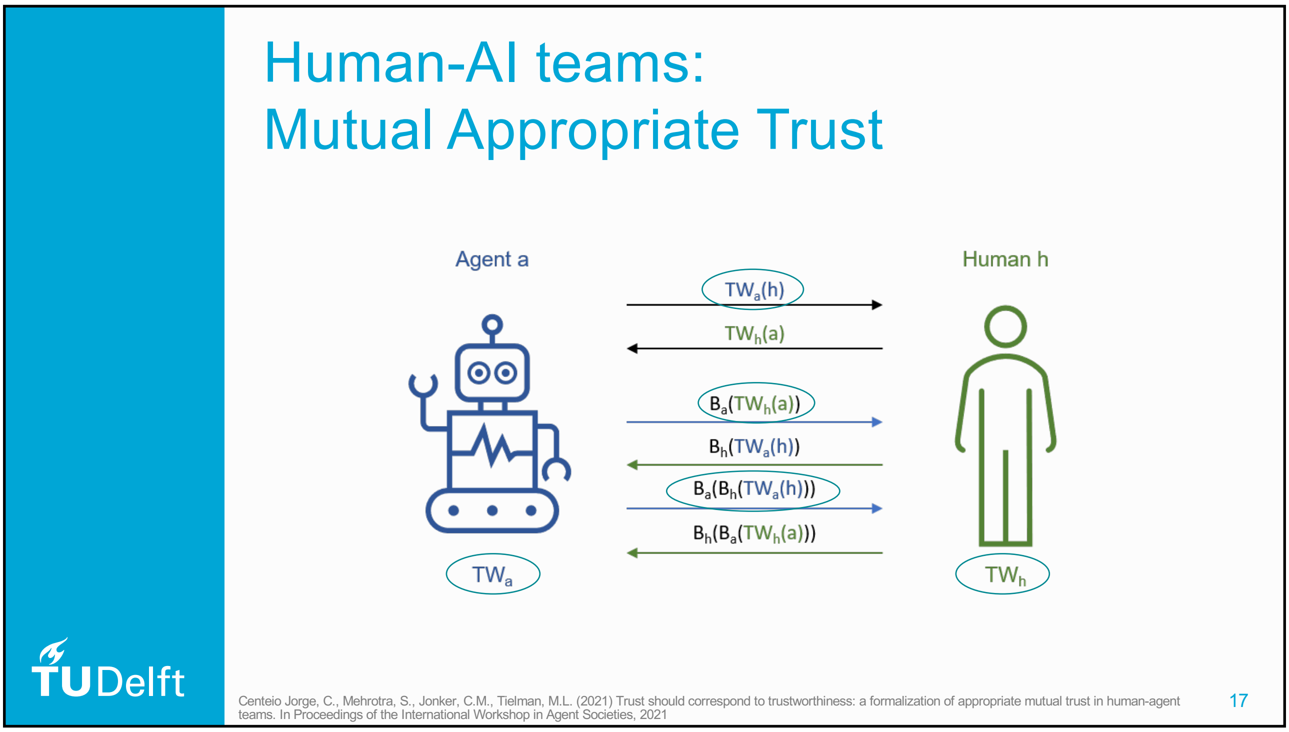


Appropriate trust elicitation

What could have made the situation better? For example:

The teammates could have communicated about their own perception of their trustworthiness.

The agent could have showcased its trustworthiness.



Human-AI teams: Mutual Appropriate Trust

Agent a

Human h

$TW_a(h)$

$TW_h(a)$

$B_a(TW_h(a))$

$B_h(TW_a(h))$

$B_a(B_h(TW_a(h)))$

$B_h(B_a(TW_h(a)))$

TW_a

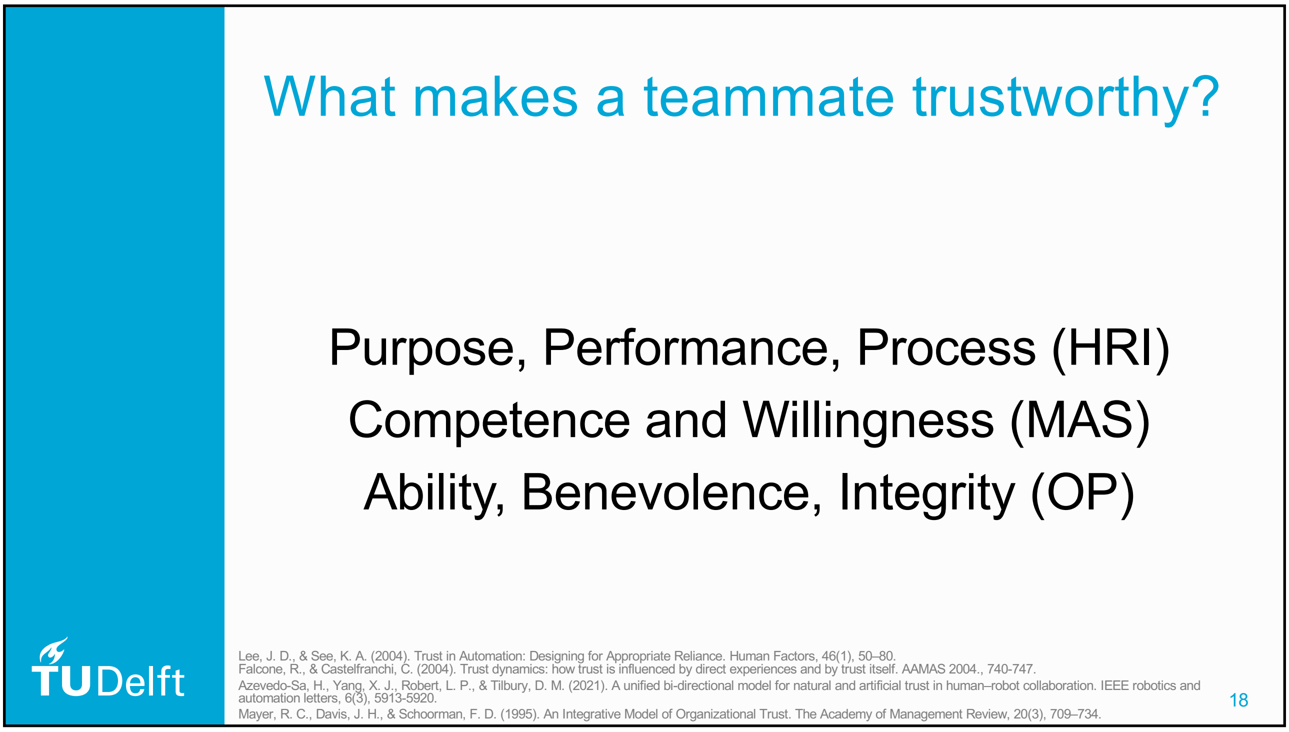
TW_h

TU Delft

Centeio Jorge, C., Mehrotra, S., Jonker, C.M., Tielman, M.L. (2021) Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams. In Proceedings of the International Workshop in Agent Societies, 2021

17

17



What makes a teammate trustworthy?

Purpose, Performance, Process (HRI)

Competence and Willingness (MAS)

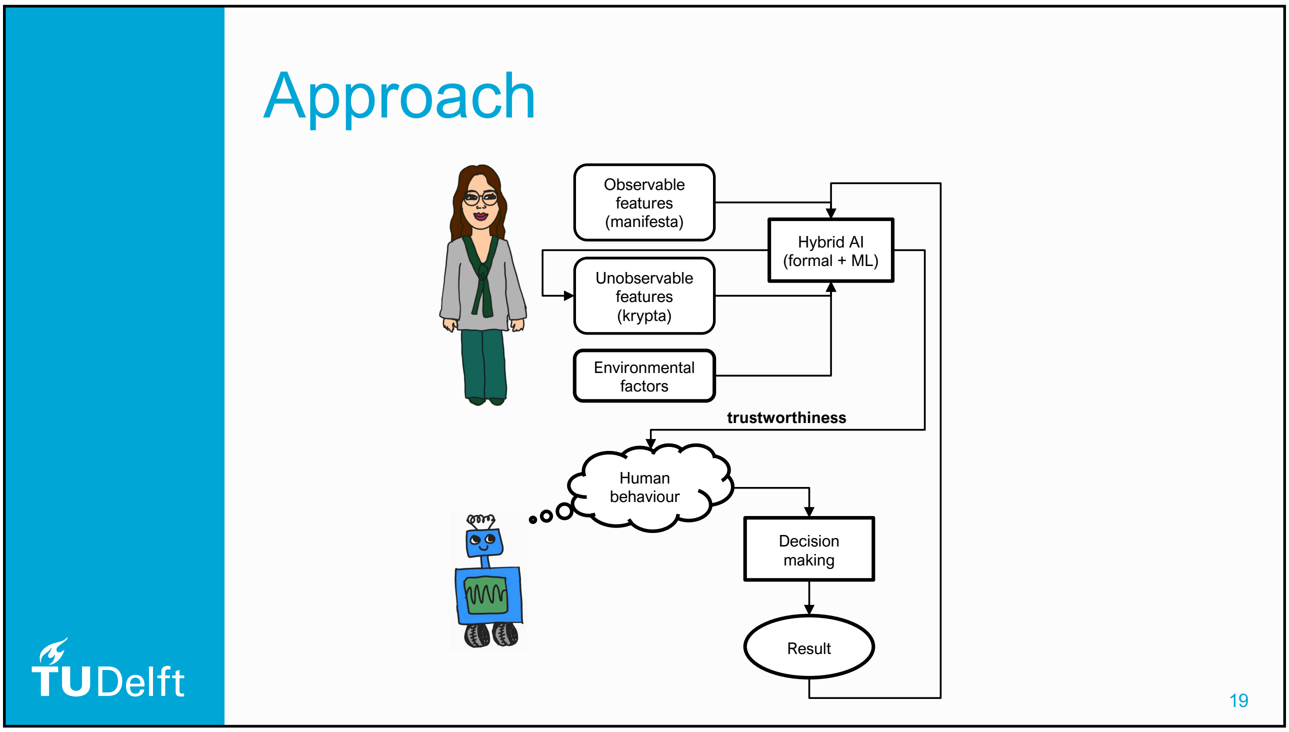
Ability, Benevolence, Integrity (OP)

TU Delft

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors, 46(1), 50–80.
Falcone, R., & Castelfranchi, C. (2004). Trust dynamics: how trust is influenced by direct experiences and by trust itself. AAMAS 2004., 740-747.
Azevedo-Sa, H., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2021). A unified bi-directional model for natural and artificial trust in human–robot collaboration. IEEE robotics and automation letters, 6(3), 5913-5920.
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.

18

18



Approach

Observable features (manifesta)

Unobservable features (krypta)

Environmental factors

Hybrid AI (formal + ML)

trustworthiness

Human behaviour

Decision making

Result

19

Thank you!

20