Building Appropriate Trust in AI: The Significance of Integrity-Centered Explanations

Definition: ***Appropriate trust*** in teams happens when one teammate's trust towards another teammate corresponds to the latter's actual trustworthiness.

I have found many papers looking at 'ability' as a core factor for trust in AI systems.

So, if I make a system with enormous computing power, people will trust it!

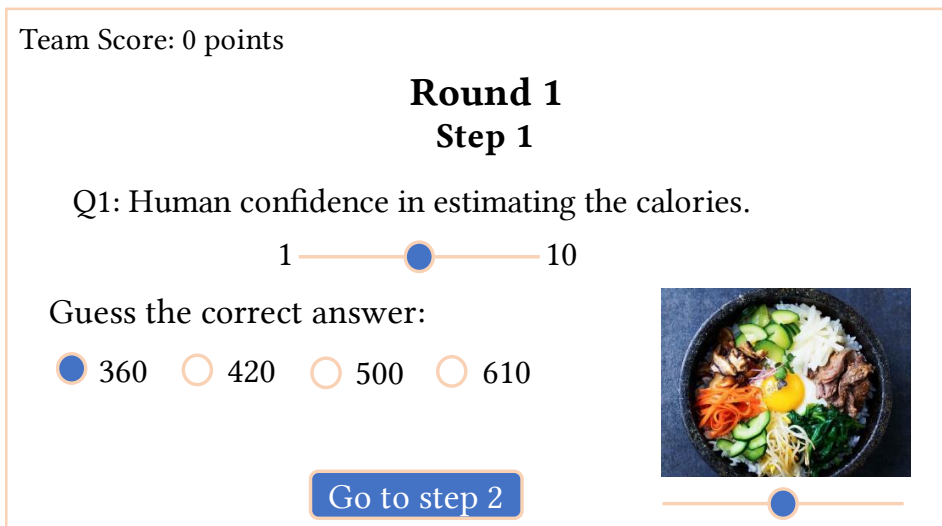
Elements of Human Trust

- Ability
- **Integrity**
  - *Honesty*
  - *Transparency*
  - *Fairness*
- Benevolence

Well! That's not true! Many researchers from psychology and sociology links integrity and benevolence as more important elements of trust than the 'ability' of the AI system.

Research Question: How does the expression of different principles of integrity through explanation affect the appropriateness of the trust of a human in that agent?

Our approach: We consider three principles related to integrity to express through explanations:
(1) **Honesty** about the system's capabilities and confidence.
(2) **Transparency** about the process of decision making.
(3) **Fairness** in terms of sharing what risks such as biases exist.



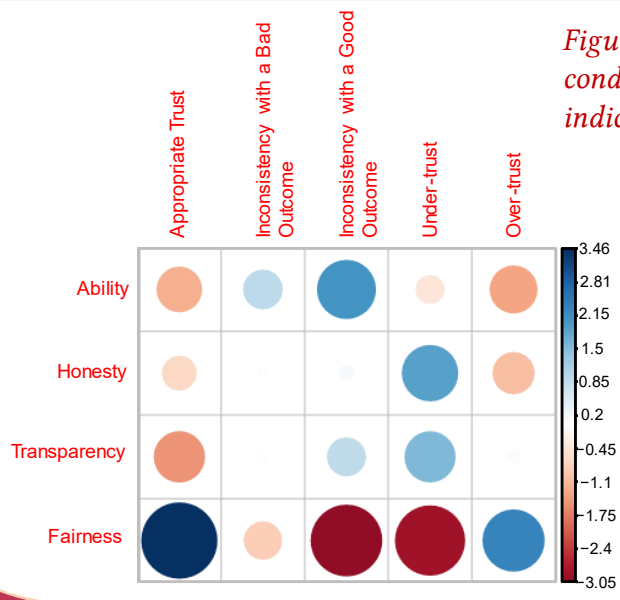


Figure 1 (on bottom left) shows a correlation plot between trust categories and integrity conditions. The relative contribution of each cell to the total Chi-square score provides an indication of the nature of the dependency between trust categories and conditions.

Our results show that

(a) An agent who displays its integrity by *being explicit about potential biases* achieved **appropriate trust** more often compared to being honest about capability or transparent about the decision-making process

(b) **subjective trust** builds up and recovers better with *honesty-like integrity explanations.*

Siddharth Mehrotra, Carolina C. Jorge, Catholijn M. Jonker and Myrthe L. Tielman

TU Delft

Hybrid Intelligence