How does an artificial agent's behavior affect human trustworthiness?

Paul Lindhorst <p.lindhorst@student.tudelft.nl>

Supervisor: Carolina Jorge
Responsible Teacher: Dr. Myrthe Tielman

Hypothesis: A conflicting agent decreases human trustworthiness

1. Background

Human/AI collaboration

Human & Agent work together to achieve common goal → Coactive Design

Trustworthiness

Inherent property of a person; how much one is motivated to do good to another party; cooperate; help; how good someone is at achieving a task. [1]

Trust

Perceived trustworthiness; directional; subjective. [2]

Measuring trustworthiness

ABI model [2] :
- **Ability**: skill/competence to achieve a task
- **Benevolence**: caring/communicative/willingness to cooperate
- **Integrity**: honorable/keeping promises

2. Experiment

Environment Simulation

- USAR: Urban Search & Rescue
- MATRX package in Python
- Goal: fetch and drop injured people to a drop-off, 10 min time limit

Trustworthiness metrics

Objective measures:
- Ability: time, total of moves, game completion
- Benevolence: number of messages, human helps the agent, agrees to agent suggestions
- Integrity: amount of lies

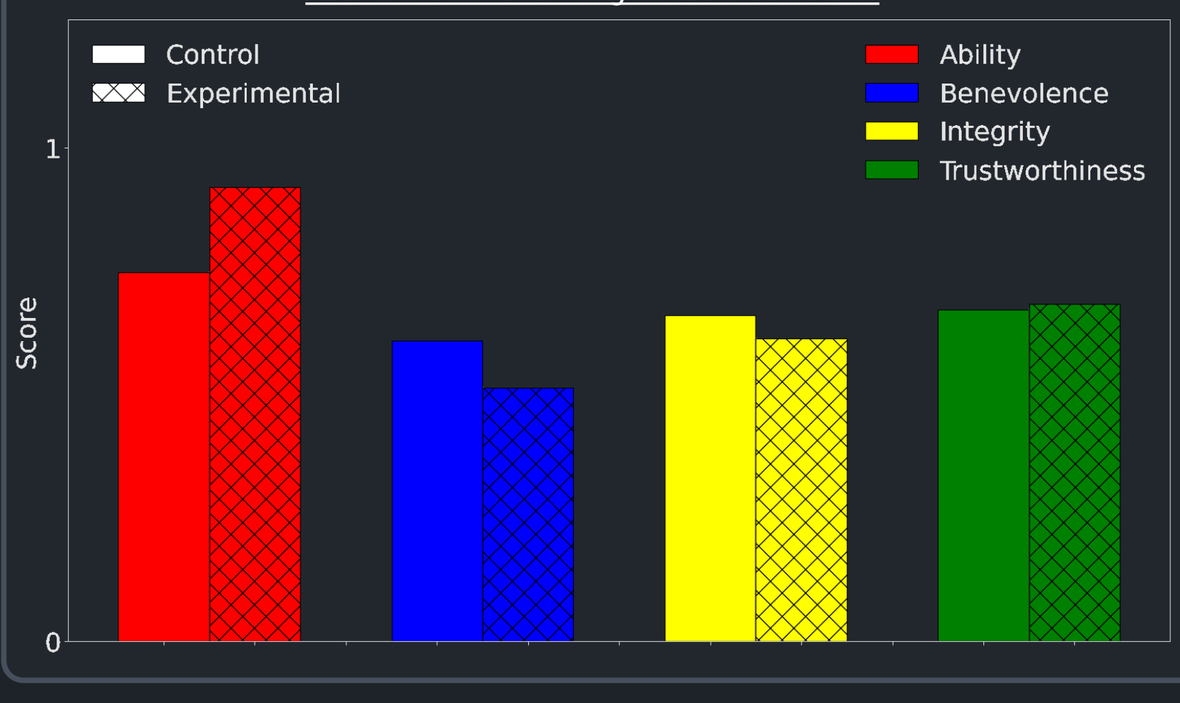
Subjective measures: Questionnaire [3]
- Ability: "I was qualified to do my job"
- Benevolence: "I communicated often"
- Integrity: "I kept my promises"

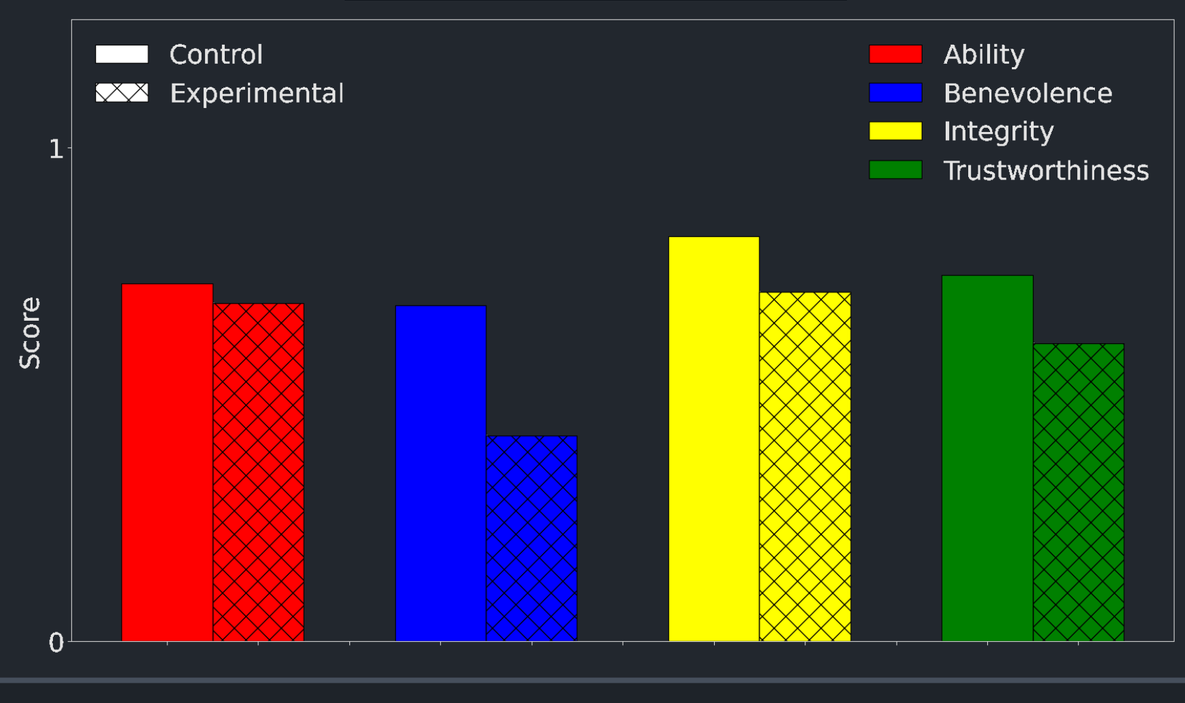
Conflicting agent

- Randomly drops victims
- Lies about (not-)finding people
- Gives bad suggestions

3. Results

Participants:
- Control group: 20 (normal agent)
- Experiment: 20 (conflicting agent)





4. Analysis

Use *statistical inference* to test hypothesis: $\bar{x}_{exp} < \bar{x}_{ctrl}$ (→ compare mean scores of experiment/control group)
- T-Test (parametric) / Mann-Whitney U Test (non-parametric)
- Shapiro-Wilk Test checks for normality

	Objective metrics				Questionnaire			
	Significant	p-value	t-statistic	Test	Significant	p-value	t-statistic	Test
Ability	No	1.0	26.0	Mann-Whitney U	No	0.156	1.026	T-Test
Benevolence	No	0.056	1.63	T-Test	Yes	0.001	3.433	T-Test
Integrity	No	0.217	0.792	T-Test	Yes	0.017	278.5	Mann-Whitney U
Trustworthiness	**No**	0.6	-0.254	T-Test	**Yes**	0.003	2.962	T-Test

- Objective metrics: Hypothesis does not hold → $\bar{x}_{exp} \not< \bar{x}_{ctrl}$
- Subjective metrics: Hypothesis holds → $\bar{x}_{exp} < \bar{x}_{ctrl}$ with 95% confidence

5. Conclusion

Human *subjective* trustworthiness ***decreases*** when paired with a conflicting AI. However, *objective* trustworthiness has not been negatively affected.

→ Human has low self trustworthiness when paired by the conflicting agent, but this does not affect the search & rescue task.

Limitations

- False positives: lies/no communication/no game completion ⇔ no experience/slow learning
- Number of participants → scale game online to recruit more people

References

1. Hardin, R. (2002), *Trust and Trustworthiness*.
2. Mayer et al (1995), *An Integrative Model of Organizational Trust*.
3. Mayer et al (1999), *The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-Experiment*.