Directability's influence on human-agent trust

Author: Yuxin Jiang Y.Jiang-5@student.tudelft.nl
Supervisors: Ruben Verhagen, Carolina Jorge **Responsible professor**: Myrthe Tielman

Background

- **Proper trust** makes agents more reliable
- Previous studies didn't look into **directability's difference representations'** influence on trust

Procedure

Does directbility of an agent improve trust in that agent?

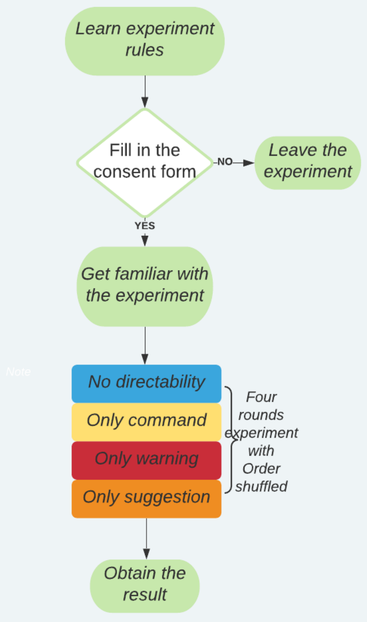
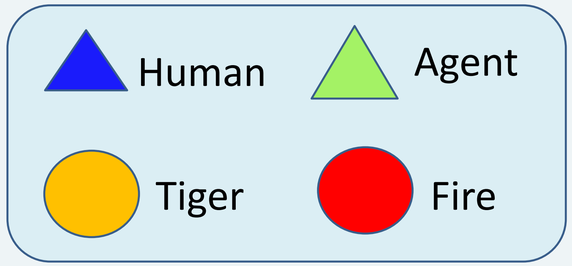


Figure 1: Experiment procedure

Experiment setup (BW4T)



- **Directability:** Human can send messages to agent, including **Commands, Warnings, Suggestions**
- **Rules:** Tiger can eat human, fire can burn both human and agent
- **Risk-taking behavior:** the human relies on the agent to pick up target blocks

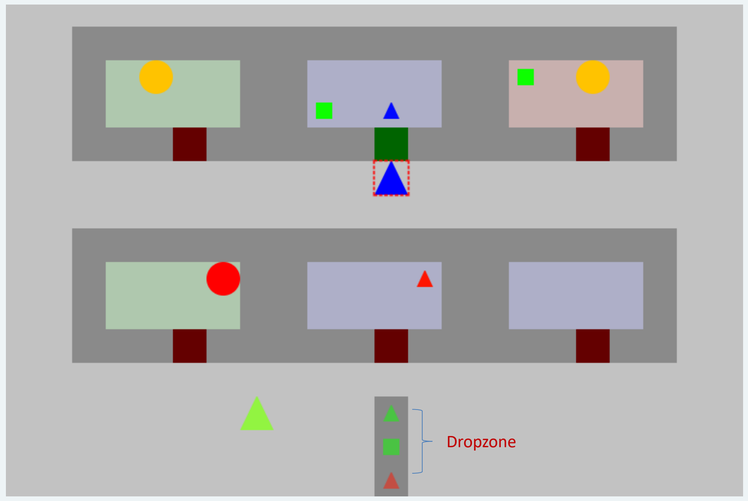


Figure 2: the BW4T world

Results

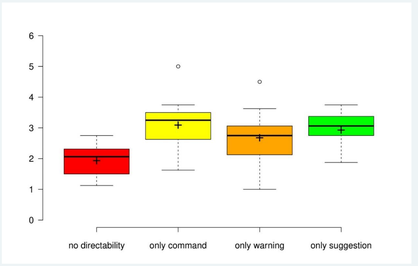


Figure 3: Result from questionnaire

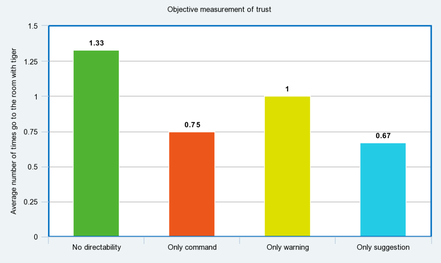


Figure 4: risk-taking behaviours

- Two measurements: **number of times human goes to the tiger room** and **questionnaires** after the game does not completely match
- According to **ANOVA analysis**, The mean for no directability is way lower than the other three representations

Conclusion

- **Limitations** include number of participants, confounding factors such as game familiarity
- Directability improve trust from human to agent
- It is suspected that **commands** and **suggestions** are better at boosting trust than **warnings**