Towards an understanding of the effects of racializing AI on human-AI cooperation

Swapnika Dulam and Christopher L. Dancy

The Pennsylvania State University, University Park PA 16802, USA

{szd5775, cdancy}@psu.edu

INTRODUCTION

- AI systems are ubiquitous, and racial biases in AI interactions reflect the lasting impact of racism on society.
- Expanded study by Atkins et al. [1], who identified that racial biases influence cooperation with AI based on AI's perceived race.
- Added:
  - More racial groups
  - More treatment conditions
  - Pictures.
- Participants completed a human-AI cooperation task based on a pig chase game [2,3] followed by a survey.
- Developed an initial cognitive model to provide a cognitive-level, process-based explanation of the results.

METHODS

- Over 950 participants.
- Participant demographics included:
  - Black/African American,
  - White/Caucasian
  - Non-White
- Rules:
  - Trust the AI to catch pig for 25 points or exit for 5 points.
  - -1 for every step taken.
  - Played for 15 trials.
  - First 3 trials: Practice
  - 8th trial: Attention trial (mandatory exit)

Fig. 1. Treatment groups for the experiment

- The AI used A* algorithm, not trained on human behaviors, same as Atkins et al. [1].
- After completing 15 trials, participants answered five questions:
  - "Did you think the AI agent was using a certain strategy to play the game? If so, could you explain it?"
  - "Generally, how did you choose your own behavior during the trials?"
  - "Rate the level of intelligence the AI exhibited during the experiment:"

  - " How did the way the AI agent (yellow triangle) was trained affect your behavior during the trials? "
  - " How/When did you decide to use the exit block instead of trying to catch the pig on any given trial? "

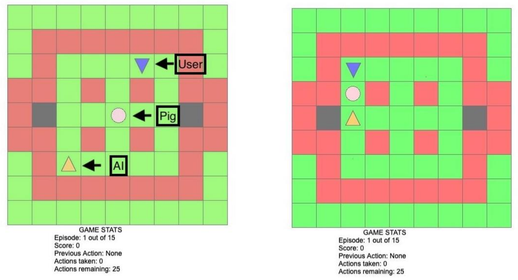

Fig. 2. Pig chase game with all game pieces in starting position on left and after capturing pig on the right.

ACT-R MODEL

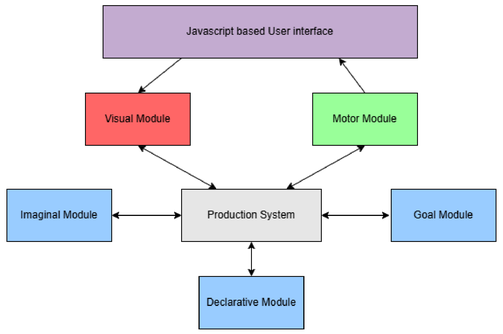

Fig. 3. The ACT-R modules that were used in our initial model.

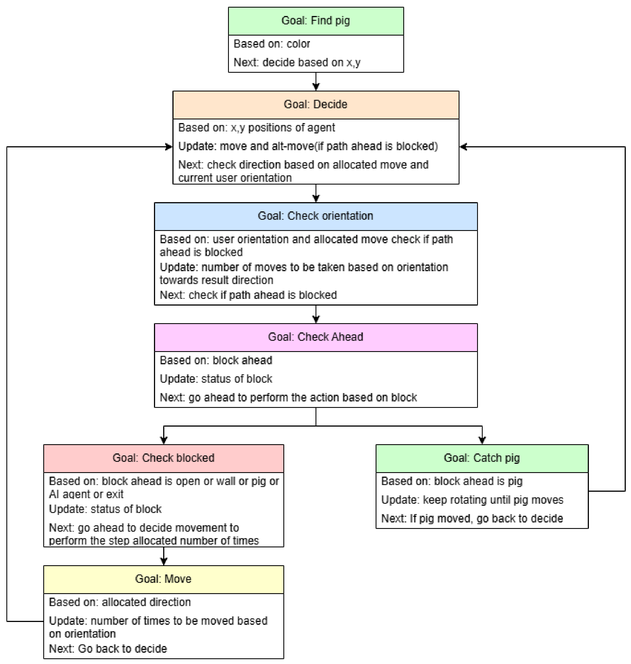

Fig. 4. Flowchart describing action sequence for the ACT-R model

RESULTS & DISCUSSION

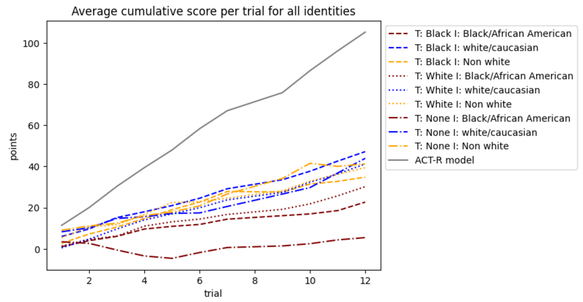

Fig. 5. Showing average cumulative scores for the experiment after excluding initial and attention trials.

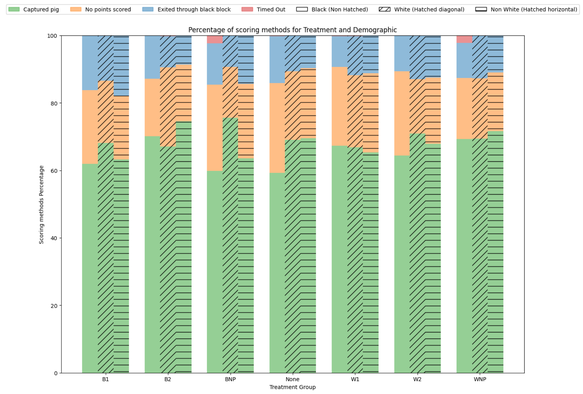

Fig. 6. Method of scoring for all treatment groups for all demographics in all Treatments.

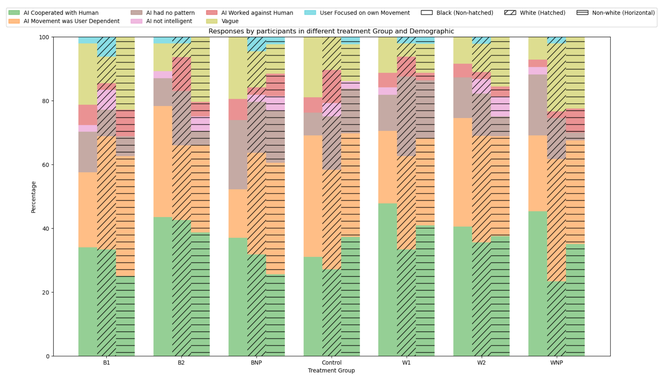

Fig. 7. Percentage-wise categorized responses in all treatment groups for all demographics.

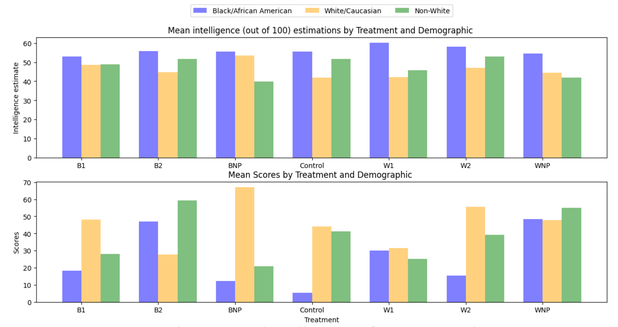

Fig. 8. Depicting the estimated intelligence of AI agents and average scores of the participants

- After cleaning, 935 records were analyzed using a two-way ANOVA.
- Significant effects:
  - Participant Demographic $F(2,935) = 6.85$, $p<0.005$
  - Treatment X Demographic $F(2,935) = 2.22$, $p < 0.01$
- No significant effect for treatment alone $F(2,935) = 1.66$, $p=.12$ indicating an impact of demographic on the scores obtained.

CONCLUSION & FUTURE WORK

- The initial ACT-R model [4], based on observing the pig's movement, outperformed the average participant.
- Further development is needed to explore participant strategies.
- Our findings, like those in Atkins et al. [1], reveal distinct behavioral patterns showing racialization influences behavior, particularly based on participant's race.
  - Black participants performed well in White treatments, indicating that racialized treatment of AI agents did not affect their behavior.
  - White participants did not perceive AI as intelligent compared to other demographics.
  - Non-White participants mostly believed AI did not co-operate well in Black treatments.
- Implicit Association Task (IAT) could help explore individual implicit biases.

REFERENCES

1. Atkins, A. A., Brown, M. S., & Dancy, C. L. Examining the Effects of Race on Human-AI Cooperation. In proceedings of the 14th International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation, Virtual, pp. 279-288 (2021).
2. Yoshida, W., Dolan, R. J., & Friston, K. J. Game theory of mind. PLOS Computational Biology, 4(12), e1000254 (2008).
3. Johnson, M., Hofmann, K., Hutton, T., Bignell, D.: The Malmo platform for artificial intelligence experimentation. In: IJCAI International Joint Conference on Artificial Intelligence 2016, pp. 4246–4247 (2016).
4. Anderson, J. R. How can the human mind occur in the physical universe? OUP, New York, NY (2007).