Towards a socio-cultural perspective on Human-AI cooperation with PigChase Task

Swapnika Dulam[1] and Christopher L. Dancy[1,2]
szd5775@psu.edu, cdancy@psu.edu

PennState College of Engineering

1. Department of Computer Science and Engineering,
2. Department of Industrial & Manufacturing Engineering
Pennsylvania State University, University Park PA, USA

INTRODUCTION

- The Computers are Social Actors (CASA) paradigm (Nass et al., 1994) describes how humans implicitly apply the same social heuristics of human-human interactions to human-computer interactions.
- AI systems are ubiquitous, and racial biases in human-AI interactions reflect the lasting impact of racism on society.
- Expanded study by Atkins et al. (2021), who identified that racial biases influence cooperation with AI based on AI's perceived race by adding more subtly nuanced treatment conditions and building a cognitive model.
- Participants completed a human-AI cooperation task based on a pig chase game (Johnson et al., 2016), as shown in Fig. 1, followed by a survey.
- Developed an initial cognitive model to provide a cognitive-level, process-based explanation of the results.

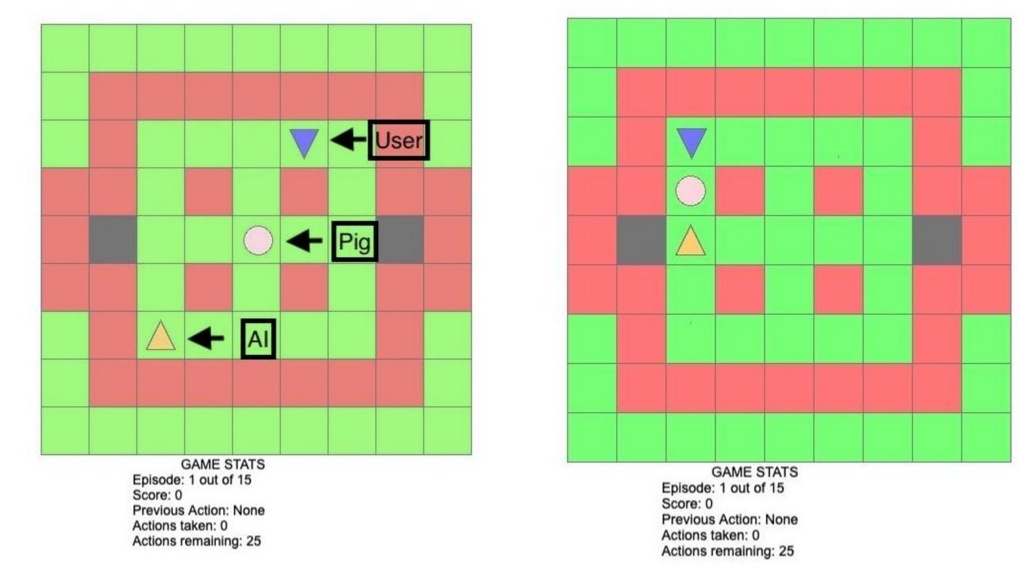

Fig. 1. Pig chase game with all game pieces in starting position on left and after capturing pig on the right

RESEARCH QUESTIONS

- What strategies did participants use when they interacted with AI?
- What impact might the race of AI have on their strategy?
- Are these strategies transferable from the game to other real-world agents?

METHODS

- Collected data from over 950 participants.
- **Participant demographics included:**
  - Black/African American,
  - White/Caucasian
  - Non-White
- **Pig Chase game:**
  - Trust the AI and collaborate with it to catch the pig to get 25 points.
  - Can exit to get 5 points.
  - Get -1 for every step taken.
  - Played for 15 trials, where the first three trials were for practice
  - 8th trial: Attention trial (mandatory exit)

Treatment Conditions:

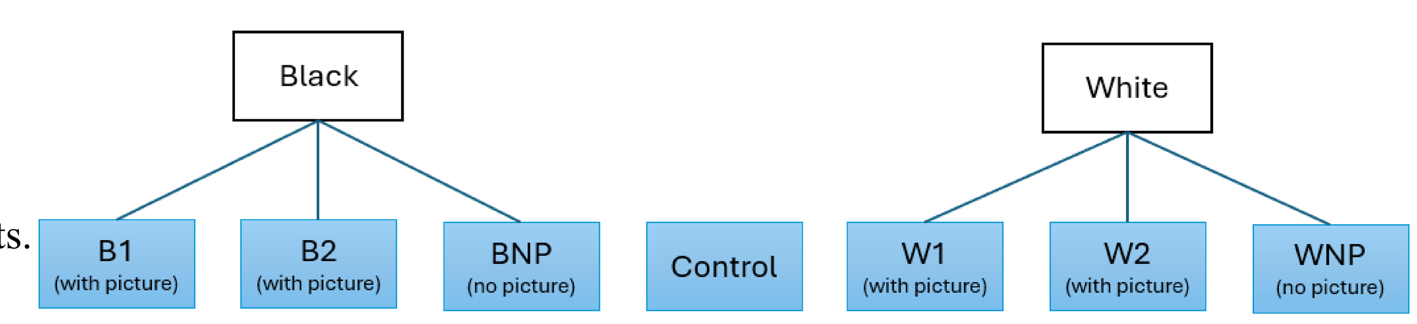

Table 1. Participant demographics for each treatment condition.

Treatment	Group	Black	White	Non-White
B1	*Black*	47	48	48
B2	*Black*	46	47	44
BNP	*Black*	46	44	43
Control	*Control*	42	48	43
W1	*White*	44	48	44
W2	*White*	47	45	32
WNP	*White*	42	47	40

- The AI agent used an A* pathfinding algorithm and was not trained on human behaviors, like Atkins et al. (2021).
- After completing 15 trials, participants answered the post-game survey with five questions to understand participant strategies, their perception of AI, the impact of race in the interaction, and their intelligence estimate of the AI agent.

RESULTS

- After cleaning, 935 records were analyzed using a two-way ANOVA.
- Significant effects:
  - Participant Demographic: $F(2,935) = 6.85$, $p < .005$
  - Treatment X Demographic: $F(2,935) = 2.22$, $p < .01$
- No significant effect for treatment alone $F(2,935) = 1.66$, $p = .12$ indicating an impact of demographics on the scores obtained.

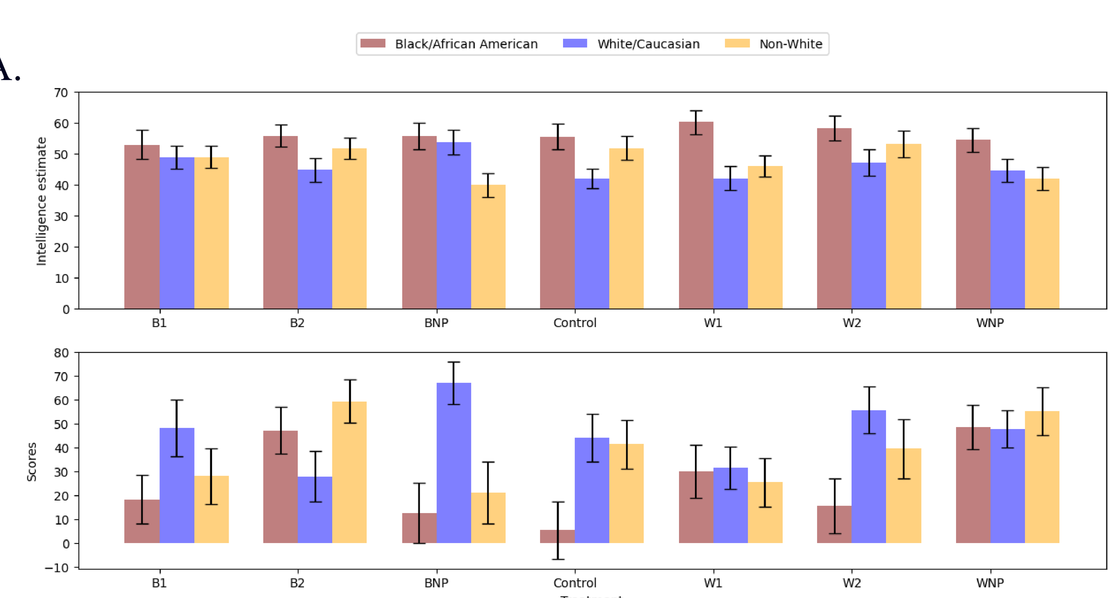

Fig. 2. Estimated intelligence of the AI agent and the average scores

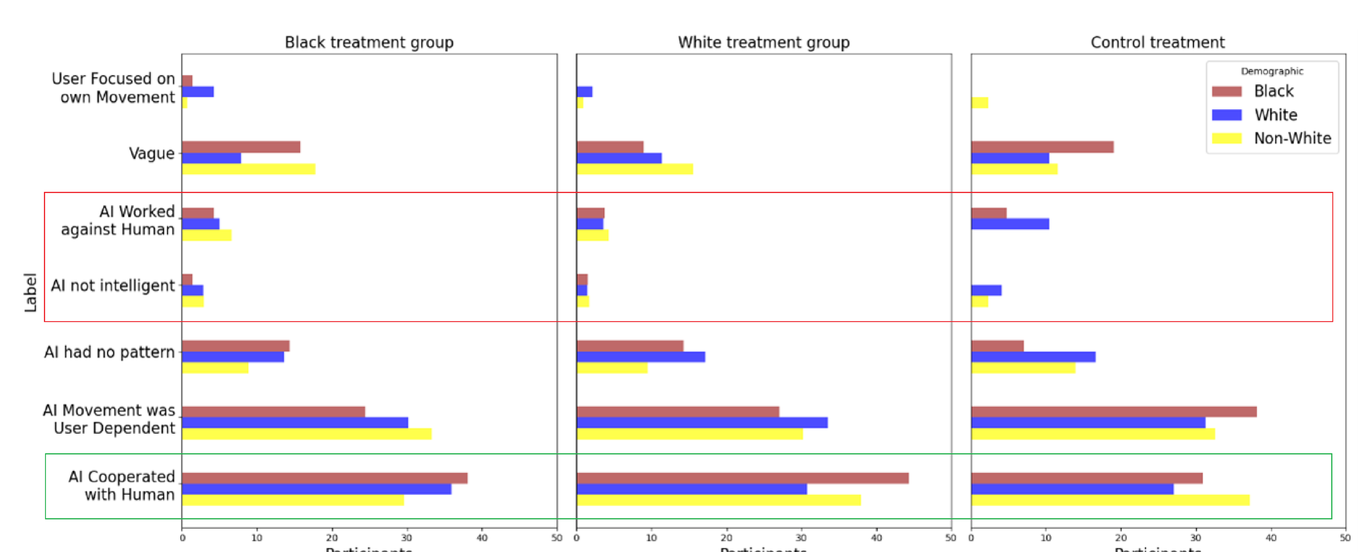

Fig. 3. Categorized responses in all treatment groups for all participants

ACT-R Model

- Computational cognitive modeling through ACT-R cognitive architecture (Anderson, 2007) was used to better understand the cognitive processes involved in these decision-making strategies.
- The model interacted with a JavaScript-based environment and was run over 150 times for different parameter settings with a reward scheme approximating game rules
- The current model achieved a closer fit to participants, as shown in Table 2.

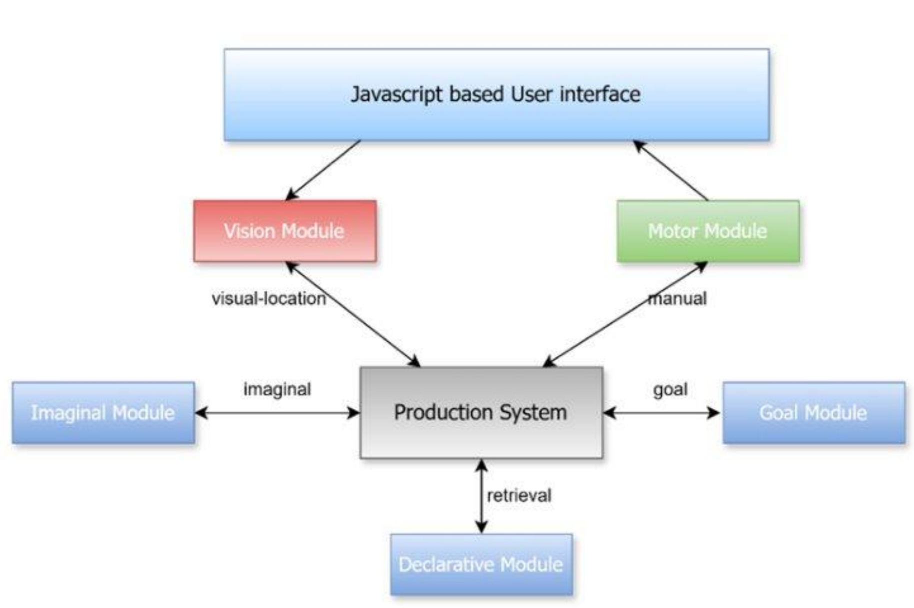

Fig. 4. The ACT-R architecture diagram in our model

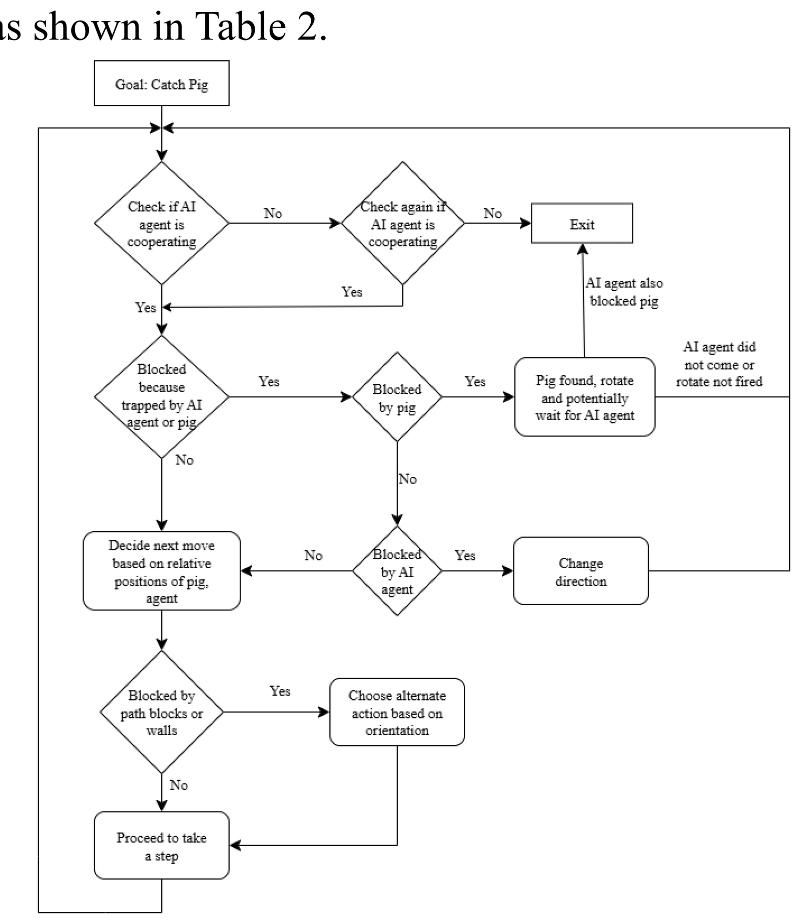

Fig. 5. Flowchart for action sequence in the ACT-R model

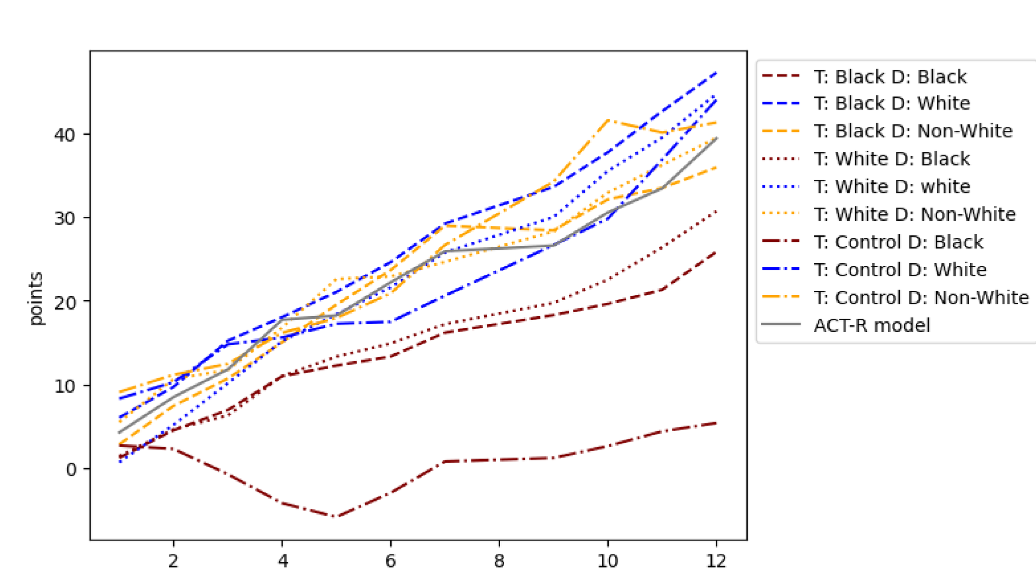

Fig. 6. Average cumulative scores for all treatments and ACT-R model.

Table 2. R^2 values for ACT-R model fit with grouped treatments.

Treatment	Black	White	Non-White
Black	-0.45	0.84	**0.97**
White	0.42	**0.93**	**0.96**
Control	-45.5	0.90	0.83

CONCLUSION AND FUTURE WORK

- The initial ACT-R model closely fit the average score of the participants of Non-White demographics.
- Our findings, like those in Atkins et al. (2021), reveal distinct behavioral patterns showing racialization
  influences behavior, particularly based on participants' race.
  - Black participants had a positive opinion towards AI and indicated that AI seemed intelligent, but believed AI worked against them in the control treatment.
  - White participants did not perceive AI as intelligent compared to other demographics.
  - Non-White participants explicitly favored AI agents in White treatment conditions over Black treatment conditions.
- Adjusting some parameters to reflect the difference in treatment conditions can help us obtain a better fit for Black demographics with all treatment conditions.
- Integrating a holographic declarative memory (Kelly et.al., 2020) can help us better understand implicit biases within participant responses.

REFERENCES

- Nass, C., Steuer, J., & Tauber, E. R. (1994). *Computers are social actors* (pp. 72–78). In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94). https://doi.org/10.1145/191666.191703
- Atkins, A. A., Brown, M. S., & Dancy, C. L. (2021). Examining the Effects of race on Human-AI cooperation. In *Lecture notes in computer science* (pp. 279–288). https://doi.org/10.1007/978-3-030-80387-2_27
- Johnson, M., Hofmann, K., Hutton, T., & Bignell, D. (2016). The Malmo Platform for Artificial Intelligence Experimentation. In *Ijcai* (Vol. 16, pp. 4246-4247).
- Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* https://doi.org/10.1093/acprof:oso/9780195324259.001.0001
- Kelly, M. A., Arora, N., West, R. L., & Reitter, D. (2020). Holographic Declarative Memory: Distributional semantics as the architecture of memory. *Cognitive Science*, *44*(11). https://doi.org/10.1111/cogs.12904