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

We describe an expanded, second iteration of a study to explore how human-AI cooperation may be impacted by the belief that data used to train an AI system is racialized, that is, was trained on data from a specific racialized group of people. Understanding how the racialization of AI systems may impact the way people interact with them during tasks is important, given both the ubiquity of those systems and the sustained impact of systems of racism on societies. Despite the continued anthropomorphization of AI systems, the potential impact of racialization is understudied. During this study, participants completed a human-AI cooperation task and completed a survey questionnaire afterward. Similar to the first iteration of the study, statistical analysis of the task behavior revealed a statistically significant effect of the self-identified race of the participant, as well as the interaction between this self-identified race and the treatment condition (i.e., the way in which the agent was racialized). Additionally, we've constructed an initial cognitive model of the task, that will allow us to have a cognitive-level, process-based explanation of the results found from the study.

Keywords: Human-AI interaction, Race, Cooperation, ACT-R.

1 Introduction

Understanding how the racialization of AI systems may impact the way people interact with them during tasks is important, given both the ubiquity of those systems and the sustained impact of systems of racism on societies. Despite the continued anthropomorphization of AI systems, the potential impact of racialization is understudied. An important aspect of this racialization is the way sociocultural structures may be cognitively represented by people to result in the development and persistence of racism, sometimes viewed through the lens of implicit biases. These biases can affect perceptions and satisfaction and impact cognitive processes and behavior when interacting with AI artifacts. Several such cases have been demonstrated in studies highlighting shooter bias [e.g., see [5] and [7], where the White participants were quick to shoot the Black targets. Another study from [6] has shown that racial stereotypes impact how consumers interact with anthropomorphic AI agents, with different racialized AI bots being perceived and treated based on existing stereotypes. In a study conducted to see if people trust partners from different races, [8,11] also found that participants placed higher trust in partners who shared the same self-identified race. Given the ubiquity of AI systems and that an AI system may be explicitly racialized or implicitly associated with Whiteness (e.g., as argued by [3]), it is important to continue to study how systems of racism interact with an increased prevalence of human-AI interaction.

Research from the previous iteration of this study showed how the racialization of an AI agent could affect behavior during a human-AI cooperation task [1]. In that study, Atkins et al. [1] found that participants' self-identified race interacted with how the AI agent was racialized (i.e., the AI agent was said to have learned by observing data from a particular racialized group of people) to ultimately impact their performance while cooperating with an AI agent during a modified version of the Stag Hunt task [4] called Pig Chase; this task was inspired from *Pig Chase by the Microsoft Malmo Collaborative AI Challenge* [14]. Here, we describe the results from an expanded version of that study that included participants from a wider range of self-identified racial groups (i.e., beyond White and Black), and treatment conditions that included pictures of racialized individuals (which presumably added a more explicit visual phenotypical racial association to the AI agent).

2 Methods

2.1 Participants

The participant sample included over 950 participants, all recruited through Prolific.co. Participants were paid $10 for their participation. Prolific's automatic participant filters were used to specify participants' demographics for their self-reported race and being from and located within the United States to achieve a balanced set of people who identified as "Black/African American", "White/ Caucasian", or ["Asian", "Mixed", or "Other"].

2.2 Design

The participants were asked to play the pig chase game with an AI agent for fifteen trials. The participants could either trust their companion, an AI agent, and decide to catch the pig for a higher reward or could exit for a lower reward.

Each participant was placed into one of seven possible treatment groups where they were told that the AI learned by observing behavior of people who identify as a certain racial category or a control condition that didn't mention race:

1. Black or African American, with one of two possible pictures displayed for reference (these pictures were obtained from the Chicago Face database [2] and different one was used for each treatment condition).
2. Black or African American, but with no picture shown.
3. White or Caucasian, and with one of two possible pictures displayed for reference (these pictures were obtained from the Chicago Face database [2] and different one was used for each treatment condition).
4. White or Caucasian, but with no picture was shown.

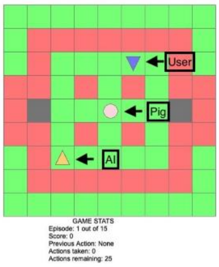
5. Didn’t mention race and was told that AI was trained by observing people’s behaviors without any race-specific details or pictures.

Like the previous experiment run by Atkins et al. [1] the AI with which participants interacted was not trained on human behaviors and instead used an A* algorithm to select actions (where to go) on the map. Individual task-related behaviors, such as keys pressed, reaction times, and score were collected for each round during the task.

2.3 Procedure

Participants began the experiment through Prolific and were assigned a treatment condition using Prolific.co. They then were directed to a Qualtrics page that included specific instructions for the game and details of their randomly assigned treatment condition. They were informed that the first three trials in the study were for practice to help them understand the game. Additionally, they were told to exit through the rightmost gray square on the eighth trial to ensure they were paying attention.

After reading the instructions, participants were redirected to play the game hosted through Pavlovia.org, where they controlled a blue, triangular game piece and collaborated with an "AI-controlled" yellow, triangular game piece to catch a pink, circular game piece representing a pig (as shown in Fig. 1). All pieces were on a 9x9 grid but could only move within a 5x5 area, which was blocked by red square tiles. Each move on the green squares deducted 1 point from the score.



GAME STATS
Episode: 1 out of 15
Score: 0
Previous Action: None
Actions taken: 0
Actions remaining: 25

Fig. 1. A screenshot displaying the starting position of the game for each trial.

Participants could score points in two ways: by working with the AI to catch the pig or by exiting through the gray squares on either side of the board. Each trial began with the human-controlled piece in the upper right corner, the AI in the lower left, and the pig in the center, as shown in Fig. 1. If participants worked with the AI to catch the pig, they earned 25 points. This required surrounding the pig, so it had no valid moves. Alternatively, they could choose to exit through the gray blocks for 5 points.

After participants completed all 15 trials, they were redirected from Pavlovia to Qualtrics to answer five qualitative questions about their actions in the experiment. The questions included, “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?" and were asked to "Rate the level of intelligence the AI exhibited during the experiment:" using a slider with the leftmost value representing "no intelligence" and the rightmost value representing "very intelligent", followed by "How did the way the AI agent (yellow triangle) was trained affect your behavior during the trials?" and lastly, "How/When did you decide to use the exit block instead of trying to catch the pig on any given trial?"

3 ACT-R model

We have developed an initial cognitive model of behavior during the task to begin to understand the cognitive processes and knowledge mediating behavior during the Pig Chase task. We used the ACT-R cognitive architecture to develop this model due to its strong theoretical foundation in human memory and the potential importance of memory to the task treatment (i.e., associations related to racialization.)

3.1 ACT-R architecture

The ACT-R cognitive architecture [9], contains functional modules for visual, aural, vocal, perception, motor functions, and declarative memory, all of which are linked to a procedural memory system. The procedural memory system uses procedural memory represented as productions that encapsulate knowledge on how to perform specific tasks. Buffers within ACT-R serve to as a form of communication between any given module and the procedural memory system, with the contents of these buffers at any given time reflecting ACT-R's current state. The procedural memory system is driven by a pattern matcher that identifies the production that most closely aligns with the current state of the buffers (or the environment).

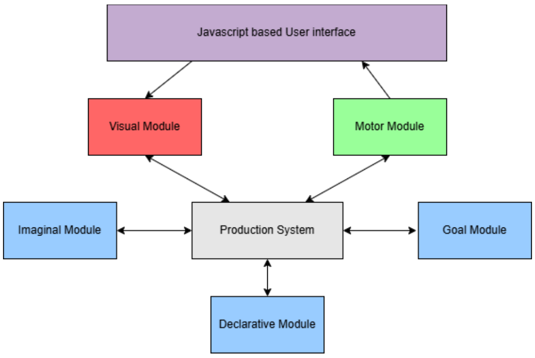


Fig. 2. Depicting the ACT-R modules that were used in our model.

3.2 Architecture of the model

The pig chase game for the study was coded using JavaScript and used socket communication to interact with ACT-R server on a designated port [13]. The game positions were sent to the visual module in ACT-R, which used production system and other modules like goal module, imaginal module and declarative module to come up with a move that can be taken, which is communicated back to the interface with the help of motor module. Fig. 2 shows an architectural overview of the communication process.

3.3 Model Implementation

For this model we used a basic strategy that was heavily reliant on the imaginal module to keep track of visual locations and choose rules accordingly. There could be several other strategies that rely on different modules. The current model was run 150 times to collect the scores. Within the model, there are over a hundred production rules that use the other modules for a decision-making strategy. The hidden utility module helps form better associations between which rule needs to be chosen under a given condition.

The execution of the model begins with an empty goal buffer, which proceeds by trying to locate the pig, and both the agents on UI. After locating these game pieces, the model chooses two possible directions based on their specific location. Once these directions are chosen, the next step is to assess if the position ahead is blocked based on the current orientation of the ACT-R agent (the blue triangle). The flow diagram for these rules is displayed in Fig [3].

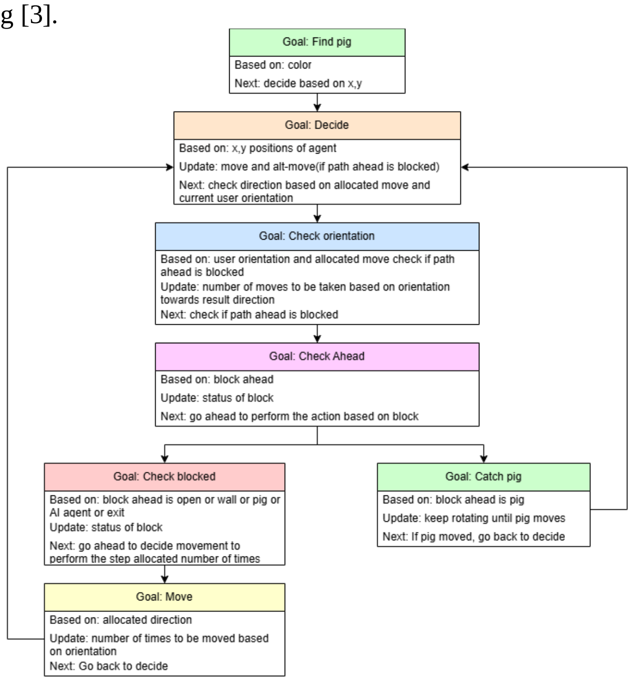

Goal: Find pig
Based on: color
Next: decide based on x,y
Goal: Decide
Based on: x,y positions of agent
Update: move and alt-move(if path ahead is blocked)
Next: check direction based on allocated move and current user orientation
Goal: Check orientation
Based on: user orientation and allocated move check if path ahead is blocked
Update: number of moves to be taken based on orientation towards result direction
Next: check if path ahead is blocked
Goal: Check Ahead
Based on: block ahead
Update: status of block
Next: go ahead to perform the action based on block
Goal: Check blocked
Based on: block ahead is open or wall or pig or AI agent or exit
Update: status of block
Next: go ahead to decide movement to perform the step allocated number of times
Goal: Catch pig
Based on: block ahead is pig
Update: keep rotating until pig moves
Next: If pig moved, go back to decide
Goal: Move
Based on: allocated direction
Update: number of times to be moved based on orientation
Next: Go back to decide

Fig. 3. Showing the flowchart of actions used by the model

The possible moves based on the orientation are:

a) If the path ahead is blocked by a path blocker (red square) or AI, the agent chooses an alternative direction.
b) If the path ahead is blocked by a pig, it means the pig is found.
c) If the path ahead is an exit, the agent exits.
d) If the path ahead is not blocked by anything, it moves in that direction and continues its pursuit of finding the pig.

4 Results

The results are based on a two-fold analysis: a quantitative analysis based on participants' scores and a qualitative analysis based on answers to the survey questions at the end of the experiment.

4.1 Task Performance

The participants scored points in trials 67% of the time and exited through the designated block only 12% of the time, exhausted their steps 21% of the time, and timed out only 1% of the time, indicating the game was taken seriously with an objective to score points. This remained consistent with the treatment-wise statistics where the participants in all treatment groups captured the pig 60-70% of the time and chose to exit 9-12% of the time and did not score in the remaining trials. Although we observe consistency with respect to catching the pig, the number of steps taken to catch the pig affects the scores secured by the players. A greater number of steps taken will lead to a lesser score. While the ACT-R model captured pigs 85% of the time, exited 3% of the time, and exhausted 13% of the time.

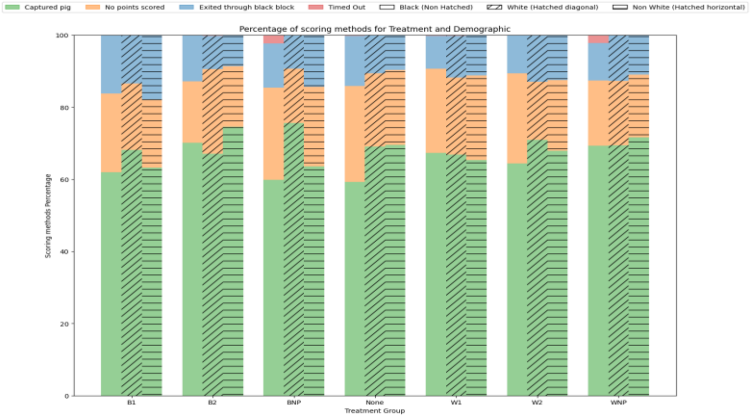


Fig. 4. Method of scoring for all treatment groups for all demographics in all Treatments.

After cleaning the data, we removed records with duplicate IDs, multiple attempts leaving us data from 939 participants. We then removed the outliers for every treatment by calculating the z-score and removing records based on the score field where the absolute value of the z-score was greater than 3. This removed a total of 4 records from the 939 records. For the 935 records as displayed in Table 1, a two-way ANOVA was used to test the effect of the factors, treatment group, and demographics of the participants on the cumulative score. The assumptions to support the validity of ANOVA were conducted to test for Normality using the Shapiro-Wilk test. The homogeneity of variances was verified using Levene's test. Followed by a two-way ANOVA on the treatment group and participant demographic.

Table 1. Distribution of participants in treatment groups.

Treatment	Black/African American	White/Caucasian	Non-White
B1	47	48	48
B2	46	47	44
BNP	46	44	43
Control	42	48	43
W1	44	48	44
W2	47	45	32
WNP	42	47	40

The results for treatment X demographic showed a statistically significant effect of participant demographic $F(2,935) = 6.85$, $p<0.005$, and treatment X demographic $F(2,935) =2.22$, $p < 0.01$ *but* not treatment $F(2,935) =1.66$, $p=.12$ indicating an impact of demographic on the scores obtained.

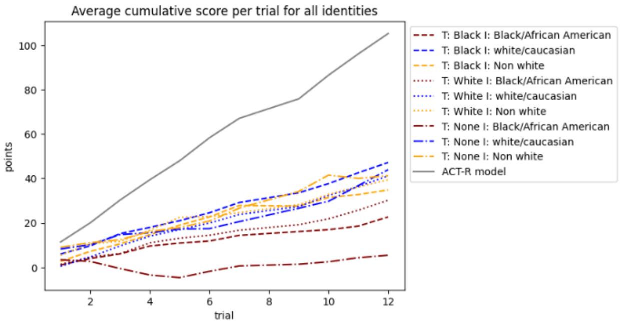


Fig. 5. Showing average cumulative scores for the experiment after excluding initial and attention trials.

4.2 Qualitative data

We analyzed the responses to the survey questions to gain insights behind the decision-making strategies of users. We categorized the data into seven labels, out of which trends observed in three labels are discussed here.

AI cooperated with Human: Compared to other demographics, Black participants believed AI cooperated with them in all treatment conditions. White participants reported that AI cooperated better in Black treatments than in White treatments. Non-White participants believed AI cooperated better in B2, Control, and all White treatments.

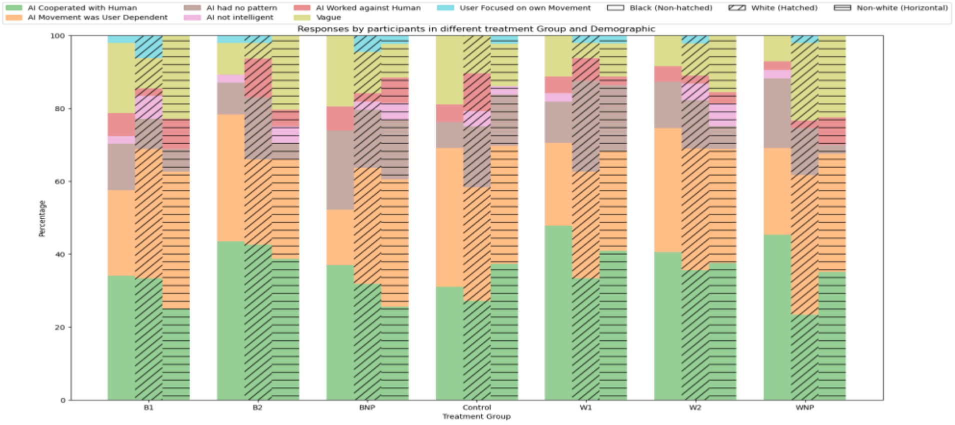


Fig. 6. Percentage-wise categorized responses in all treatment groups for all demographics.

AI not intelligent: A small fraction of Black participants in treatment conditions B1, B2, W1, and WNP reported that AI was not intelligent. White participants, in a somewhat higher proportion, believed AI to be lacking in intelligence in B1 and, to a lesser extent, in W2, Control, and BNP. Non-White participants reported a relatively higher fraction in W2 and smaller in B2, BNP, and Control.

AI Worked against Human: Except in the B2 treatment condition, Black participants believed AI worked against them more in other black treatment groups and the control group as compared to those participants in the White treatment groups. For White participants, the order from highest to lowest fraction of AI worked against Human can be seen in B2, followed by Control, W1, and less in B1, BNP, W2, WNP. Non-White participants on an average believed AI worked against them higher in Black treatment groups and less in White treatment groups and did not work against them in control treatment group.

4.3 Discussion

Trends observed from Black participants: From Fig. 4, we can see that Black participants tried to capture the pig higher in White treatments than in Black treatments which explains why the scores are higher in White treatment groups as shown in Fig. 7. Their average perceived intelligence for all treatment groups is in the range of 55-60 indicating, a similar view of the "intelligence" of the AI agent despite differences seen in performance between treatment groups as shown in Fig. 7. From qualitative analysis in Fig. 6, we can also observe that they tend to believe more that AI cooperated with them, which conforms to their perceived intelligence of the AI agent.

Trends observed from White participants: The perceived intelligence of the AI agent by White participants was relatively high in treatments where they played with a perceived Black AI agent when no picture was displayed (BNP). This was in direct correlation with higher mean scores achieved by players. However, from Fig. 6, we can see that participants believed AI co-operated better in W2 than in BNP indicating a potential bias against Black treatment conditions.

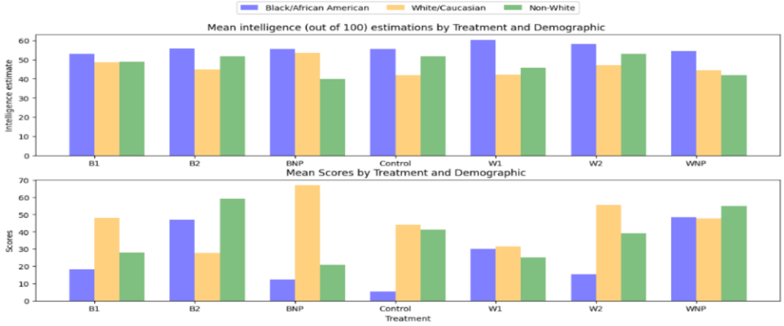


Fig. 7. Depicting the estimated intelligence of AI agents and average scores of the participants.

Trends observed from Non-White participants: From Fig. 6, we can see that they believed AI cooperated better in White and control treatments, than in Black treatments. They also perceived that AI was not intelligent higher in Black treatments compared to White treatments and believed AI worked against them mostly in Black treatments.

5 Conclusion

The initial ACT-R model, based on the strategy of observing the pig's movement, scored higher points compared to the average participant's performance. This suggests that the strategy used by the ACT-R model is different from that of the players. Therefore, the model needs further development to understand the potential strategies used by the participants during the game. Additionally, this model did not consider treatment effects, which also should be accounted for in the future models.

Our findings, similar to those in Atkins [1], reveal distinct behavioral patterns that suggest the influence of racialization, particularly in relation to the self-identified race of the participant. The data from our experiment also suggests that White participants did not perceive AI as intelligent compared to other demographics. Interestingly, Black participants performed well in White treatments, indicating that racialized treatment of AI agents did not affect their behavior. However, this was not the case for Non-White and White participants.Despite receiving a higher score on the Black No Picture treatment, White participants did not think AI cooperated better than the other White Picture 2 treatment, which also recorded a higher score. This indicates how knowledge of racialization may affect how participants perceive AI, something that is not clearly visible. Additionally, Non-White participants showed a higher negative correspondence to Black treatments. They believed AI did not cooperate as much and was not as intelligent and worked against them more in Black treatments than in Control or White treatments.

Existing tasks such as the Implicit Association Task (IAT) [12] may prove useful for obtaining additional data on implicit biases particular participants show and how those biases are related to behavior at an individual level [10]. A fine-grained analysis to survey questions might reveal more insights into participants' decision-making.

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. Ma, D.S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. Behavior research methods 47, pp. 1122–1135 (2015).
3. Cave, S., & Dihal, K. The Whiteness of AI. Philosophy & Technology, 33, pp. 685-703 (2020).
4. Yoshida, W., Dolan, R. J., & Friston, K. J. Game theory of mind. PLOS Computational Biology, 4(12), e1000254 (2008).
5. Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. Journal of Personality and Social Psychology, 83(6), pp. 1314-1329 (2002).
6. Davis, N., Olsen, N., Perry, V. G., Stewart, M. M., & White, T. B. I'm only human? The role of racial stereotypes, humanness, and satisfaction in transactions with anthropomorphic sales bots. Journal of the Association for Consumer Research, 8(1), pp. 47-58 (2023).
7. Bartneck, C., Yogeeswaran, K., Ser, Q. M., Woodward, G., Sparrow, R., Wang, S., & Eyssel, F. Robots And Racism. In proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18), New York, NY (2018).
8. Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. Implicit race attitudes predict trustworthiness judgments and economic trust decisions. Proceedings of the National Academy of Sciences, 108(19), 7710 (2011).
9. Anderson, J. R. How can the human mind occur in the physical universe? OUP, New York, NY (2007).
10. Jennifer T. Kubota; Uncovering Implicit Racial Bias in the Brain: The Past, Present & Future. Daedalus 153 (1): pp. 84–105 (2024). doi: https://doi.org/10.1162/daed_a_02050
11. Eyssel, F., & Loughnan, S. "It don't matter if you're black or white"? Effects of robot appearance and user prejudice on evaluations of a newly developed robot companion. In Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings 5 (pp. 422-431). Springer International Publishing (2013).
12. Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology, 74(6), pp. 1464–1480 (1998). https://doi.org/10.1037/0022-3514.74.6.1464
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019).
14. 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).