Cite as:
Atkins, A. A., Brown, M. S., & Dancy, C. L. (2021). 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, 279-288.

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-80387-2_27

Examining the Effects of Race on Human-AI Cooperation

Akil A. Atkins, Matthew S. Brown, Christopher L. Dancy

Bucknell University, Lewisburg, PA 17837, USA
{aaa014,msb027,christopher.dancy}@bucknell.edu

Abstract. Recent literature has shown that racism and implicit racial biases can affect one's actions in major ways, from the time it takes police to decide whether they shoot an armed suspect, to a decision on whether to trust a stranger. Given that race is a social/power construct, artifacts can also be racialized, and these racialized agents have also been found to be treated differently based on their perceived race. We explored whether people's decision to cooperate with an AI agent during a task (a modified version of the Stag hunt task) is affected by the knowledge that the AI agent was trained on a population of a particular race (Black, White, or a non-racialized control condition). These data show that White participants performed the best when the agent was *racialized* as White and not racialized at all, while Black participants achieved the highest score when the agent was racialized as Black. Qualitative data indicated that White participants were less likely to report that they believed that the AI agent was attempting to cooperate during the task and were more likely to report that they doubted the intelligence of the AI agent. This work suggests that racialization of AI agents, even if superficial and not explicitly related to the behavior of that agent, may result in different cooperation behavior with that agent, showing potentially insidious and pervasive effects of racism on the way people interact with AI agents.

Keywords: Human-AI, Race, Stag Hunt, Cooperation.

1 Introduction

There has been a variety of discussions related to the ways in which implicit racial biases affect interactions between people (e.g., [8], [11]) and interactions between people and artifacts (e.g., [1], [3], [12]). Correll et al. [8] analyzed a specific situation where these biases could play a critical role: study participants were tasked with "shooting" armed targets and to "not shoot" those who were unarmed. White participants were quicker to shoot armed targets if they were Black but were quicker to make the decision to not shoot an unarmed target if they were White [8].

Stanley et al. [11] gave participants $10 and asked them to give whatever amount of money they wished to a partner; the partner would receive quadruple the amount and had already decided whether they would split their earnings evenly or keep it to themselves. During the experiment, participants were shown an image of either a Black man or White man to represent their partner [11]. Before completing the tasks, participants were also given implicit bias tests to evaluate whether they were "pro-black", "pro-

white", or neither [11]. The researchers found that those who were identified as pro-white, which was about eighty percent of the participants, were more likely to trust their partner and therefore give them more money if they were white and vice versa [11]. These findings demonstrated that people were more likely to trust and work with those that align with their own racial identity.

As our technology continues to become more ubiquitous in society, there will be different ways in which racism can continue to foundationally affect behavior. Focusing on human interaction with robots, Strait et al. [12], conducted a study and found that robots racialized as Black and Asian were subject to almost double the dehumanizing comments that the robot racialized as White received [7]. This example shows that the social construct of race plays a critical role in the ways in which not only people are treated but other agents as well. In another study that used the shooter bias incident model, but with robots as well as humans, Bartneck et al. [3] showed that even though those potentially being shot in the study were robots, the results were almost identical to the human-based study described in Correll et al. [8]. Participants were still quicker to shoot Black agents that were armed, whether they were a human or a robot [3].

The present study investigates how participants will interact with an AI agent that has been explicitly racialized, or not racialized at all. To measure their level of cooperation with the AI agent, participants were tasked to play a game called *catch the pig* (a modified version of the Stag Hunt task [15]), where the goal was to score as many points as possible. While completing the task, participants' in-game movements and scores were tracked to measure level of cooperation and after completing the trials participants were asked 3 qualitative questions regarding any strategies they had and their perceived level of intelligence that the AI agent pertained.

2 Methods

2.1 Participants

The participant sample included 186 participants, all recruited through Prolific.co. Participants were told they would be paid $6.50 per hour for their participation in the study and they could be eligible for up to $7.50 per hour based on their performance in the study (in reality all participants were given $7.50 regardless of performance). Prolific's automatic participant filters were used to specify participants' demographics (their reported race as well as being from and located within the United States) to guarantee a balanced sample of people who identified as either "Black/African American" or "White/Caucasian".

2.2 Design

Each participant was placed into one of three groups (all containing the same total amount of participants): (1) A *Black* treatment condition where participants were told that the AI agent, they would be cooperating with was trained on a predominantly Black population; (2) A *White* treatment condition where participants were told that the AI

Cite as:
Atkins, A. A., Brown, M. S., & Dancy, C. L. (2021). 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, 279-288.

The final authenticated version is available online at https://doi.org/10.1007/978-3-030-80387-2_27

agent they would be cooperating with was trained on a predominantly White population; (3) A *control (no race)* condition where the race of behavioral data was not mentioned. Despite what the participants were led to believe, the AI had not been trained on any human behaviors and used an A* algorithm to complete the task. During the task, individual task related behavior was collected during each round. Task-related behavior included score during each trial, the keys pressed for each trial, as well reaction times.

2.3 Procedure

Once participants started the experiment through Prolific, they were directed to a Qualtrics survey. The Qualtrics survey contained the condition-specific text for their randomly assigned condition and instructions on navigating the game. Participants were instructed that the first three trials in the study were for practice to ensure they understood the game, and they were also told to immediately exit through the rightmost gray square on the eighth trial to ensure that they were paying attention.

After participants read through the instructions, they began the experiment. Participants were tasked with controlling a blue, triangular game piece and to work with an "AI controlled" orange, triangular game piece to catch a pink, circular game piece that represented a pig. All game pieces were placed in a 9x9 grid in which they could only move within a 5x5 square within the grid. Inside the 5x5 grid, participants could move on any of the green squares, with each movement subtracting 1 point from the score. There were two methods for participants to score points which were to either work with the AI agent to catch the pig or exit through one of the gray squares on either side of the board. For each trial, the human controlled piece was placed in the upper right corner, the "AI" was placed in the lower left corner, and the pig was placed in the middle of the game board, as seen in Fig. 1.

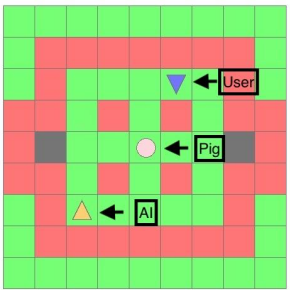

Fig. 1. The starting positions for the players (human and AI) and the pig did not change.

Once in the game, the participant could choose to work with the AI agent to attempt to catch the pig, in which case if they were successful, they would be awarded 25 points. Successfully catching the pig is only possible while cooperating with the AI agent as the Human and AI agent must surround the Pig on a block where the Pig has no valid moves. However, participants could also choose to not work with the AI agent by electing to exit through the grey blocks on either side of the board, in which case they would only be awarded 5 points.

After participants completed all 15 trials within Pavlovia, they were instructed to return to Qualtrics to answer three qualitative questions about their actions in the experiment. Participants were asked "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 lastly, they 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".

3 Results

To understand how the experiment participants performed in cooperation with the AI agent given the treatment condition, we analyzed recorded task performance across trials. We also collected answers to open-ended survey questions to understand why they may have exhibited that behavior.

3.1 Task performance

A 2x3 ANOVA (participant demographic, treatment group) of participants' final scores showed a significant effect of participant demographic ($F=5.81$, $p < .05$), but not treatment ($F=.45$, $p=.64$) or demographic x treatment ($F=1.81$, $p=.17$). Fig. 2 shows all of the average running total scores for each demographic-by-treatment interaction and a discernable difference between Black participants' and White participants' average running total scores across trials.

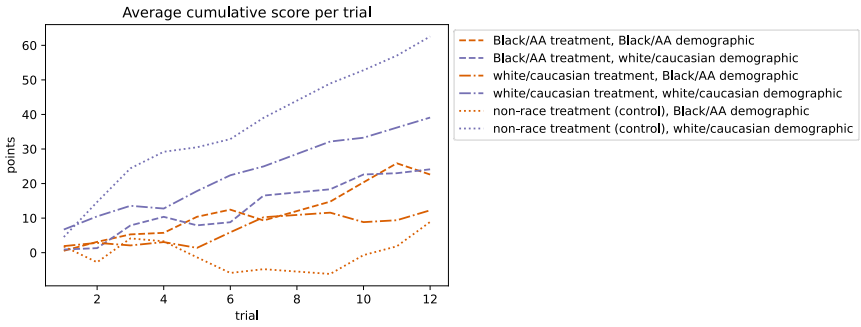

Fig. 2. Average of running total score for each treatment group x demographic interaction.

Looking at the ordering of the average running total scores within each participant demographic (Fig. 2), those data show a more consistent ordering across trials between treatment groups for the White participants (with the control condition showing the highest score and the Black condition showing the lowest score), though Black participants did show a less consistent opposite ordering (with the Black condition showing the highest score and the control condition showing the lowest score).

Though we did not find significant results with the treatment or demographic x treatment in the ANOVA, a more direct comparison between participants from different demographics may prove more apt given the trends shown in Fig. 2. There was little difference between the average running total score across blocks for self-identified Black participants and self-identified White participants who were in the Black/Africana American treatment group (Fig. 3). Black participants and White participants in this treatment group showed an average final score of 22.6 and 24.1 (respectively).

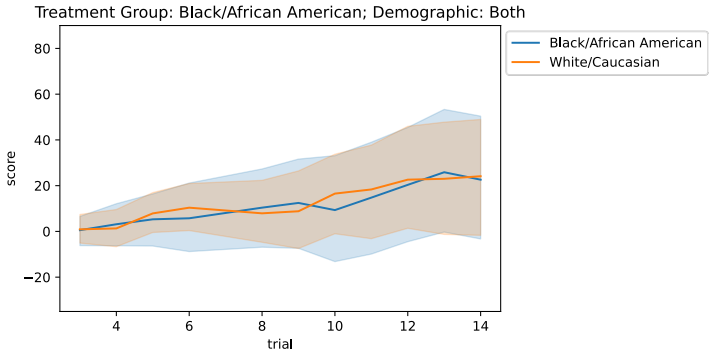

Fig. 3. Average of running total score for Black treatment group with 95% confidence interval bands.

When looking at Black and White participants in the White treatment group, the running total score begins to diverge as the task progresses. Despite Black and White participants showing a greater distance between average final scores (12.3 and 39.1, respectively), those data still showed an overlap in confidence intervals (albeit lesser when compared to the Black treatment).

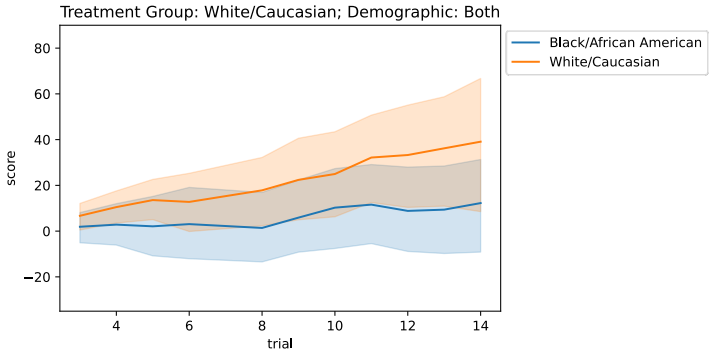

Fig. 4. Average of running total score for White treatment group with 95% confidence interval bands.

When comparing the average running score of Black participants and White participants in the control (non-race) condition, these data show a greater divergence in performance than the other two conditions. Indeed, by the end of the final trial, average Black and White participant score was 8.9 and 62.6 (respectively).

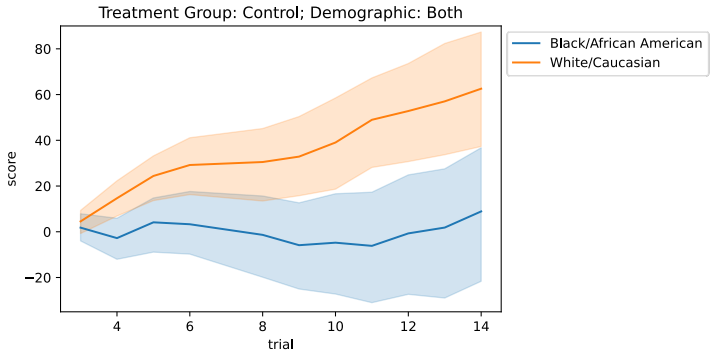

Fig. 5. Average of running total score for Control treatment group with 95% confidence interval bands.

3.2 Qualitative data

The qualitative survey dataset provides more insight into how the treatment group might have affected the participants' sentiments towards the AI agent. Similar to the task-related performance data in the previous section, Black participants appeared to have more consistency across some key categories of responses. Fig. 6 shows that of the 19 participants in the Black treatment group whose qualitative responses indicated that they believed that the AI was cooperating with the participant, a large proportion were Black (15 total as opposed to the 4 White participants who indicated that they believed the AI was cooperating with them in the task.) Within this same treatment group, of the 15 participants who indicated that the AI was not intelligent or that the AI had no pattern of behavior, the large majority were White participants (12 total) as compared to the (3) Black participants. Additionally, of the 4 total participants who indicated that they believed the AI was working against them, 3 were Black and 1 was White.

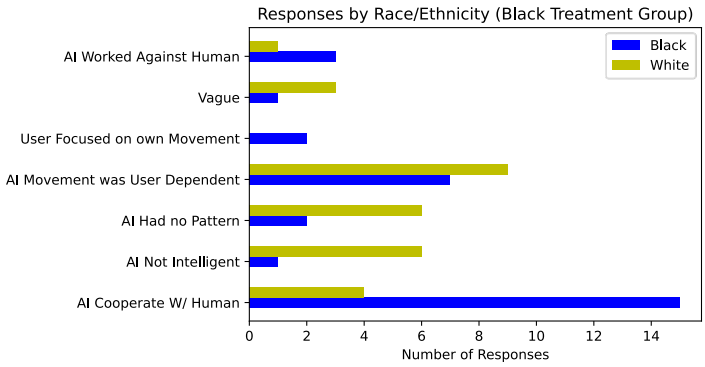

Fig. 6. Counts of categorized written responses from participants in the Black treatment group.

Black and White participants in the White treatment group show a more equal proportion of responses within the previously mentioned categories (Fig. 7). Within the White treatment group, as compared to the Black treatment group, more White participants

indicated that they believed that the AI agent was cooperating with them during the task (11 of the total 21 within this condition) and subsequently there were fewer White participants who indicated that they believed the AI agent was not intelligent or had no pattern of behavior (4 participants of the 8 total). Thus, White participants' responses aligned more with responses from Black participants. Interestingly, while none of the Black participants indicated that they believed the AI was working against them in this condition, 3 White participants indicated as such in their survey response.

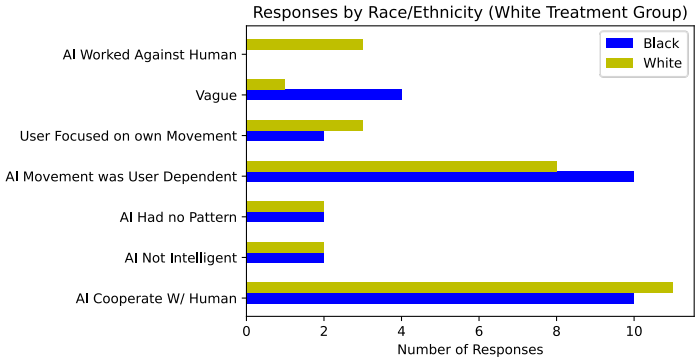

Fig. 7. Counts of categorized written responses from participants in the White treatment group.

Fig. 8 shows that while there were a higher number of responses in the control group that indicated a belief that the AI was cooperating with the participant as compared to the other two treatment groups (25 total), Black participants showed a higher number of responses in this category (15) than White participants (10). Black participants in the control group had a number of responses that indicated they did not believe the AI agent was intelligent or that it did not have a discernable pattern (2 for both) that was similar to the number of responses by Black participants in the Black and White treatment groups.

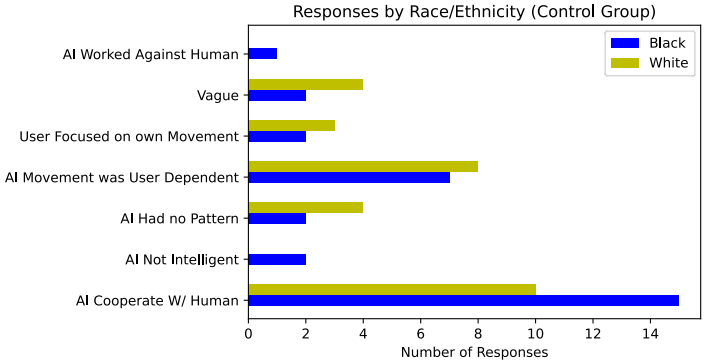

Fig. 8. Counts of categorized written responses from participants in the control treatment group.

3.3 Discussion

The task performance results showed an interesting difference between Black and White participants: though White participants (on average) performed better on this task, this appeared to be somewhat mediated by the treatment. Though Black and White participants showed nearly identical average running total scores within the Black/African American treatment group, the treatments show opposite ordering for the White and control conditions. Thus, while Black participants in the Black treatment group performed the best among the Black participants, White participants in the Black treatment group performed the *worst* among the White participants. Given that a participant achieves a higher score when they work with the AI agent to "capture the pig" (as opposed to just exiting the screen), the lowest average score by the White participants in the Black treatment group likely indicates a decreased willingness to cooperate with the AI agent (as compared to other treatment groups), and conversely the highest average score by Black participants in the Black treatment group indicates an increased willingness to cooperate with the AI agent (as compared to other treatment groups).

Interestingly, the qualitative data indicate that White participants were much less likely to judge the AI agent as cooperative than both the White participants in other treatment groups and the Black participants in their same treatment group and much more likely to judge the AI agent as not intelligent or as not showing a discernable pattern (with external judgement of discernable patterns being another way one might judge *intelligence*, Haugeland [10] , pp 5). While the Black participants showed a fairly consistent (and low) proportion of responses that fell in the "cooperate" and "not intelligent" or "no pattern" categories, the White participants proportions of responses that consider the AI agent not intelligent appear to be more similar between the White and control treatment conditions, with the Black treatment condition standing out as high among all the demographic, treatment interactions.

Cave [6] describes intelligence as a value-laden term and argues that throughout history it has been used to preserve the power of the "white, male elite". The implementation of IQ tests and SAT provide context for these statements as these two forms of testing were used to justify the oppression of Black people, as well as other people from historically marginalized communities [6]. These methods of measuring intelligence have become dominant in our society; however, they represent a very narrow view of what intelligence actually can be. Given the historical use of intelligence tests to justify oppression through the painting of certain groups of people as less intelligent, some participants (particularly the White participants given the performance and qualitative data) may have assumed the Black AI would be less intelligent and would not allow for participants to score the maximum number of points within the study.

Stanley et al. [11] found that people were more willing to work with and trust those they relate to more (including by race), which is another way in which the results can be explained. Another explanation may be tied to the "dominant mode" of *being human* (that is, what it means to be human) being ascribed to White Western male, a "genre of the human" that excludes Black people [14]. The historical, sociocultural knowledge that defines *the human* in terms of whiteness perhaps led to participants in the study to

relate the *whiteness* of the AI agents to the *humanness* and the *intelligence* of that agent [14]. Thus, White participants may have trusted the AI agent racialized as White more than the AI agent racialized as Black. This also could explain why White participants were more likely to work with the AI agent in the control condition, than in the Black condition. Though participants weren't told that the AI agent was trained on a "specific demographic" in the control condition, AI agents have been popularized as *White*, perhaps increasing the White participants likelihood to see the AI agent as trustworthy and cooperative during the task, and consequently making the Black participants less likely to cooperate with the agent during the task (even if those participants weren't more likely to report that the AI agent was *not intelligent* or *without a pattern*) [7].

4 Conclusion

Future work should expand upon these findings to dig deeper into these race/ethnicity and treatment effects. Other traditional categories of race within the existing (US) system of racism should be included to understand how *proximity to* whiteness (e.g., see Bonilla-Silva [5]) may affect interaction with knowledge of how the agent was trained (i.e., the *Whiteness* of the AI agent). More work could also be done to understand whether the effect of the treatment may be modulated by the presentation of the knowledge of the demographics of the people that had their behavior recorded to train the agent. In this study, the treatment was a statement on how the agent was trained that may not have proved particularly salient (or may not have had as large of an effect as possible on the experiment). Given the likely decay of the *subsymbolic* value of memory associated with the treatment (e.g., Anderson [2]), increasing the effects of the treatment on memory may be possible by repeating how the agent was trained during the task.

The current study makes use of a collaborative task, to examine whether participants would be more likely to trust and cooperate with an AI racialized as White and an AI that isn't explicitly racialized rather than one that is racialized as Black. Moreover, instead of selecting a phenotypical representation for the racialization (e.g., changing the color of the agent), we sought to see how *knowledge of* racialization may affect cooperation behavior. We found that the demographics of the participant and the racialization of the agent affected not only task performance, but also how the participants *perceived* AI agent behavior (particularly for white participants in the latter case). This work suggests that racialization of AI agents, even if superficial and not explicitly related to the behavior of that agent, may result in different cooperation behavior with that agent, showing potentially insidious and pervasive effects of racism on the way people interact with AI agents.

Acknowledgments. This work was supported by the National Science Foundation under grant No. 1849869.

References

1. Addison, A., Bartneck, C., & Yogeeswaran, K. (2019). Robots Can Be More Than Black And White: Examining Racial Bias Towards Robots. *In proceedings of the the AAAI/ACM Conference On Artifical Intelligence, Ethics, and Society*, Honolulu, HI, 493-498.
2. Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: OUP.
3. Bartneck, C., Yogeeswaran, K., Ser, Q. M., Woodward, G., Sparrow, R., Wang, S., & Eyssel, F. (2018). Robots And Racism. *In proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*, New York, NY.
4. Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new jim code*. Medford, MA: Polity Press.
5. Bonilla-Silva, E. (2015). The Structure of Racism in Color-Blind, "Post-Racial" America. *American Behavioral Scientist, 59*(11), 1358-1376.
6. Cave, S. (2020). The Problem with Intelligence: Its Value-Laden History and the Future of AI. *In proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, 29–35.
7. Cave, S., & Dihal, K. (2020). The Whiteness of AI. *Philosophy & Technology*.
8. Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*(6), 1314-1329.
9. Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*: MIT Press.
10. Haugeland, J. (1997). *Mind design II: philosophy, psychology, artificial intelligence*. Cambridge, MA: MIT Press.
11. Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences, 108*(19), 7710.
12. Strait, M., Ramos, A. S., Contreras, V., & Garcia, N. (2018, 27-31 Aug. 2018). Robots Racialized in the Likeness of Marginalized Social Identities are Subject to Greater Dehumanization than those racialized as White. *In proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 452-457.
13. Beasley, M. (2017). There Is a Supply of Diverse Workers in Tech, So Why Is Silicon Valley So Lacking in Diversity? Retrieved from Center for American Progress: https://www.americanprogress.org/issues/race/reports/ 2017/03/29/429424/supply-diverse-workers-tech-silicon-valley-lacking-diversity/
14. Wynter, S. (2003). Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation - An Argument. *CR: The New Centennial Review*, 3(3), 257-337.
15. Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLOS Computational Biology*, 4(12), e1000254.