An Intentional AI for Hanabi: Reproduction

MJ Johns Adam Khan

Introduction

In this reproduction, we consider the paper "*An intentional AI for hanabi*" (Eger, Martens, and Cordoba 2017) and propose the following research questions which reproduce and extend the findings of the prior paper:

- RQ1: Does an intentional AI agent perform better in a collaborative game with a human than a baseline AI agent? (as measured by score)
- RQ2: Is the player's perception of the AI agent (e.g. perceived skill and/or enjoyment of playing together) correlated with higher score, regardless of implementation?
- RQ3: Can we use Machine Learning to predict human responses to improve game-playing AI for more compatible cooperation?

All code for this project is publicly available on GitHub: https://github.com/mjohns2355/stat204_final

Background and Motivation

Game-playing AI has proven to be a useful test-bed for exploring decision-making AI implementations, and has attracted a wide range of researchers and research interests (Yin et al. 2023; Hu et al. 2024). Similarly, AI within games offer a valuable opportunity to understand human perception of interacting, and collaborating, with AI (Laird 2002; Yannakakis 2005). As we rapidly approach a state where humans and AI interact and cooperate on a daily basis, it is becoming increasingly important to (1) develop AIs that meaningfully cooperate with humans, and (2) understand what aspects of human-AI interaction are enjoyable, challenging, or unpleasant for humans. With this paper, we aim to contribute to further understanding AI implementations and human-AI interaction through an analysis of an intentional AI for the game of hanabi. We used the following R packages: ggplot2, dplyr, patchwork, and vtable.

Dataset Description

We are using two datasets for our analysis. First is the results of 240 games along with survey information from participants, which is publicly available on GitHub: https://github.com/yawgmoth/HanabiData/ The second dataset is the turn-by-turn game logs from all games scoring over 20 points. While this data is not publicly available, we were granted access by one of the original authors. The game logs use one-hot-encoding to store the current game state (e.g. cards in each players' hand, cards that have been played or discarded, and what the current player knows from hints) and for the player's action on that turn (e.g. play card 1, discard card 4, etc.). The data is a matrix with dimensions 192731 x 393. The first 373 features represent Game State, the last 20 represent Player Action.

Study Objectives

Our study has three objectives. First, to reproduce the results from the original paper using the extended dataset. Second, to extend the research to consider player perception of the AI (e.g. skill, likability). Lastly, to train a machine learning model on game logs to predict the player's next move from the current game state. This model could then be a proposed addition to the current AI implementation, offering a better collaborator with the human.

Exploratory Data Analysis

We begin our reproduction by first exploring the data. The dataset consists of 240 rows/games and 14 columns, 10 of which being survey responses such as the player's perception of ai skill (1-5 scale) and enjoyment of playing with the ai (1-5 scale), while the others consist of game data, such as game score or the type of ai played with. We first checked for missing values and counted a total of 27 rows containing at least one missing value and thus decided to discard these rows for our subsequent analysis. Following this, we examined the distribution of scores, which can be seen in Figure 1. Inspecting this, we see that the distribution appears somewhat bimodal, with the most frequent score(s) being 17 and close to 17, and with distinct peaks at 5 and 9 as well. Table 1 shows a summary of the data. Plotting the scores across different AI implementation in Figure 2 further motivates the investigation that the paper proposes. Additionally, figure 3 and 4 motivate our own investigation into the relationship between people's perception of their AI partner's skill, and the final score. We notice that game score tends to increase with perceived skill. Additionally, figures 5 and 6 lead us to examine whether enjoyment varies with game score, as we notice that game score tends to increase with enjoyment.

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Likability	213	2.7	1.2	1	2	4	5
Perceived.Skill	213	2.4	0.95	1	2	3	5
Perceived.Intention	213	3.1	0.85	1	3	4	5
Final.Score	213	12	5.4	0	8	16	21

Figure 1

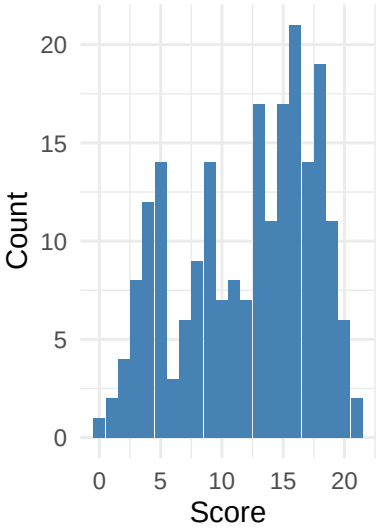


Figure 2

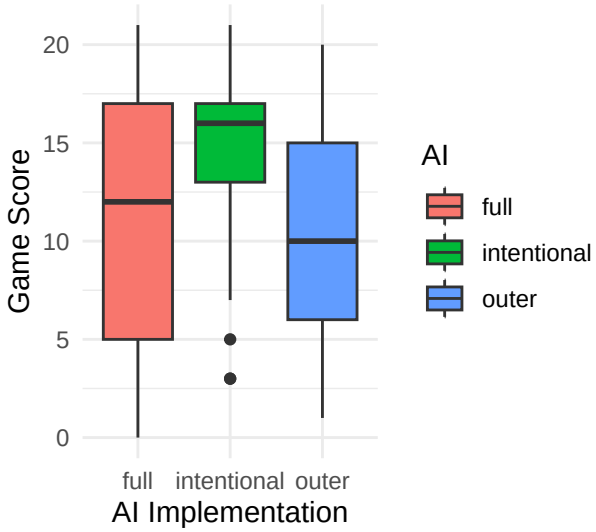


Figure 3

Perceived Skill vs Game Score

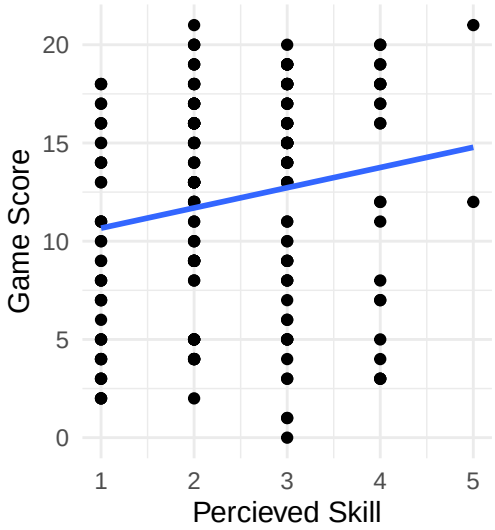


Figure 4

Perceived Skill vs Mean Score

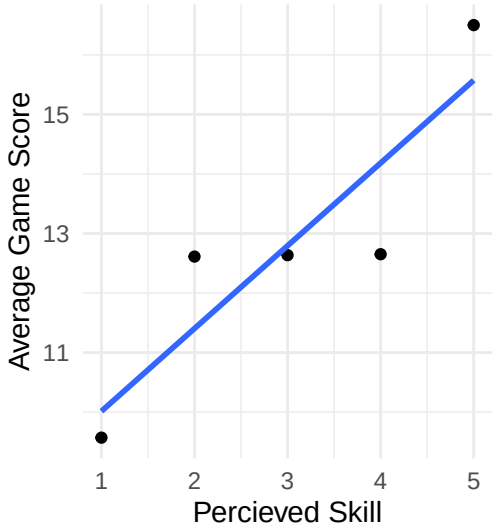


Figure 5

Enjoyment vs Game Score

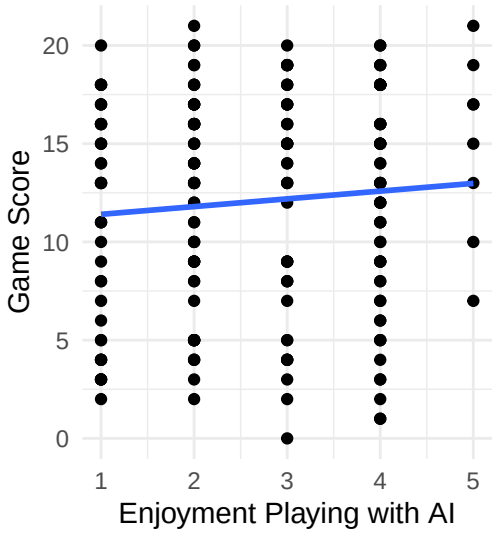
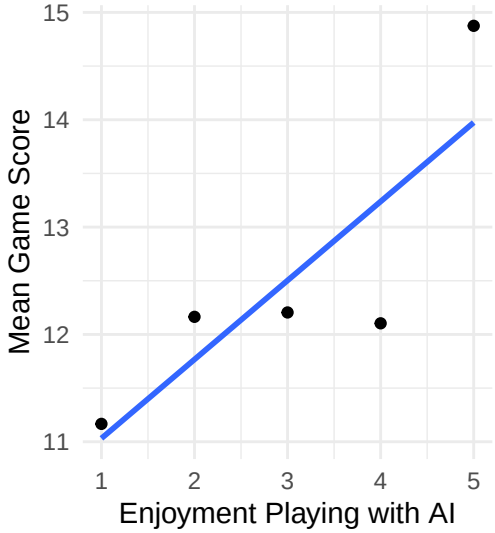


Figure 6

Enjoyment vs Mean Score



Methods

In the original study, participants were told they would be playing with an AI, but not which AI implementation they would be matched with. The original analysis involved an ANOVA comparing score based on AI implementation (Intentional, Outer, or Full). The original study's ANOVA resulted in ($p<0.0001$) showing statistical significance for a relation between AI implementation and final score. To compare the specific implementations, the original study used a Tukey test to determine that there was a statistically significant difference between the intentional AI and the outer state AI (baseline), as well as between the outer state AI and the full AI, however they did not find statistical significance when comparing the intentional to the full AI.

Beyond reproducing the original ANOVA and Tukey tests using the larger dataset, we extend this research by considering the potential correlation of player perception of AI on the resulting score regardless of AI implementation. To measure this correlation, we use Spearman's Rank considering the 5-point Likert responses to perceived AI Skill and and how much the player liked their AI companion as potential variables. This method is used to test how strongly two sets of ranks are correlated, in this case we hypothesize that both perceived AI skill and likability will be correlated with higher game score.

Lastly, we investigated whether machine learning models could be used to predict player action based on current game state. Since the game log data was structured in a tabular format and labeled, we trained an XGBoost and Neural Network multiclass classifiers (80/20 train-test split) on the game states and evaluated their performance.

Results

To reproduce the original results, we conducted an ANOVA and Tukey Test. We verified the assumptions of the ANOVA with diagnostic plots. The Q-Q Residuals plot shows potential concern for skewness, however we have a large sample size and ANOVA is robust to mild normality violations.

Table 2: Omnibus results from the One-Way ANOVA

	Degrees of Freedom	Sum of Squares	Mean Square	F Statistic	p-value
AI	2	589	294.66	11.19	$< .001$
Reiduals	210	5531	26.34		

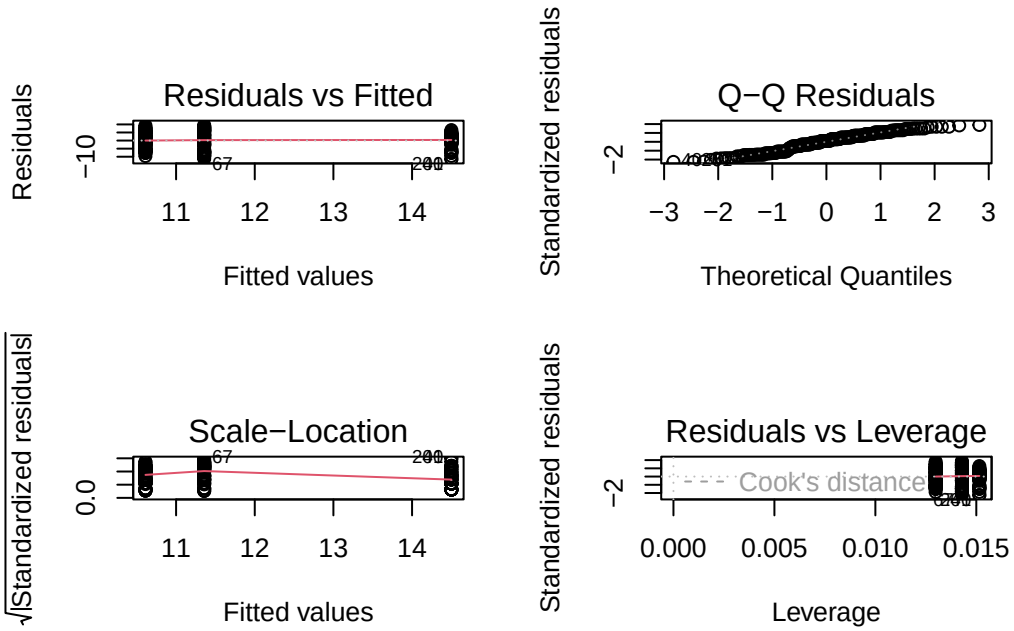


Table 3: Pairwise Comparison Results from the Tukey Test

AI	Difference	p-value
Intentional-Full	3.14	.001
Outer-Full	-0.75	.65
Outer–Intentional	-3.89	<.001

Our results align with the findings of the original paper: There is a statistically significant difference favoring Intentional AI implementation scores compared ot the Outer or Full implementation scores to a 0.05 significance level. On the other hand, there is no statistically significant difference between the scores of the Outer and Full implementations as the original authors noted.

Table 4: Spearman's Rank for Perceived Skill vs Final Score

S	Rho	p-value
1297231	0.194	.004

Table 5: Spearman's Rank for Perceived Likability vs Final Score

S	Rho	p-value
1470958	0.087	.208

As seen in Table 4, the Spearman's Rank Correlation shows that the correlation (rho = 0.1945) between Perceived Skill and Final Score is statistically significant to an alpha = 0.05 significance level even with a rather small correlation coefficient. As a diagnostic, we notice that the data is seemingly monotonically increasing (per figure 3), confirming the validity of the test. The test found no statistically significant correlation between enjoyment/like playing with the AI and Final Score to an alpha = 0.05 significance level.

Machine Learning: We used XGBoost and a Neural Network to train two potential models for predicting the next move based on current game state. We used 373 features, with One-Hot-Encoding to describe the current game state, at 20 output variables with One-Hot-Encoding representing the next move. The games states in the dataset were only obtained from games that scored greater than 20 points, so our model was trained on relatively well played games of Hanabi. In our first pass proof-of-concept models, the Nerual Network had a 41% accuracy at correctly predicting the next move, while XGBoost had a 52% accuracy and 81% top-3 accuracy. We do want to note that in a game with randomness such as Hanabi, top-3 accuracy is a very fitting metric to evaluate a turn action model as there are typically multiple good moves that can drive the game towards a higher score. We also understand that a best, or even good move in Hanabi may not always be known during each turn. This is due to the randomness of drawing from a deck in Hanabi, while a game such as chess (with no random elements) always has an associated best or good moves. The code for both is available in the Appendix. One limitation of our approach is that we did not account for repeat players in the game logs, so one player who played many times may have had undue influence on the training data.

Discussion & Conclusion

Based on our ANOVA and Tukey tests we are able to reproduce the results of the original study, and using the larger dataset provides evidence that the results are generalizable. This offers continued evidence to support the value of intentionality in AI design.

Based on our Spearman's Rank test, liking/enjoying the AI companion was not statistically significantly correlated with a higher score, however perceived AI skill was correlated with a higher score ($p < .005$). This suggests that performing well in the game is not heavily reliant on the player *liking* their AI companion, and also that *liking* their companion is not heavily reliant on how well the AI is playing, however players who perceive the AI as more skillful were more likely to have high scoring games.

Lastly, our implementation of a Machine Learning model offers a proof-of-concept for future work that can integrate prediction into the intentional AI, allowing the AI to consider the likelihood of future moves based on the current game state.

References

Eger, Markus, Chris Martens, and Marcela Alfaro Cordoba. 2017. "An Intentional AI for Hanabi." In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, 68–75. https://doi.org/10.1109/CIG.2017.8080417.

Hu, Chengpeng, Yunlong Zhao, Ziqi Wang, Haocheng Du, and Jialin Liu. 2024. "Games for Artificial Intelligence Research: A Review and Perspectives." *IEEE Transactions on Artificial Intelligence* 5 (12): 5949–68. https://doi.org/10.1109/TAI.2024.3410935.

Laird, John E. 2002. "Research in Human-Level AI Using Computer Games." *Commun. ACM* 45 (1): 32–35. https://doi.org/10.1145/502269.502290.

Yannakakis, Georgios N. 2005. "AI in Computer Games: Generating Interesting Interactive Opponents by the Use of Evolutionary Computation," December. https://era.ed.ac.uk/handle/1842/879.

Yin, Qi-Yue, Jun Yang, Kai-Qi Huang, Mei-Jing Zhao, Wan-Cheng Ni, Bin Liang, Yan Huang, Shu Wu, and Liang Wang. 2023. "AI in Human-Computer Gaming: Techniques, Challenges and Opportunities." *Machine Intelligence Research* 20 (3): 299–317. https://doi.org/10.1007/s11633-022-1384-6.