Improving Trust Strategies in Simulation Trough Forgiveness: Off-Policy Evaluation for Human Choice Prediction

Diana Morgan
diana.morgan@campus.technion.ac.il

Veronika Sorochenkova
veronika@campus.technion.ac.il

Abstract

Recently Large Language Models (LLM) have experienced several significant breakthroughs enabling them to simulate human-written text with unseen before proficiency as well as demonstrating an impressive ability to respond to life human inquiries. These developments allow for a wide variety of research into using LLM-based agents for interactive problems with both human-machine and human-human communication. Our focus will be on language-based persuasion games utilizing artificial agents, where artificial agents attempt to persuade others using text-based communication.

Our paper is built upon the work of Shapira et al. (2023) in which they attempted to predict human behavior in language-based persuasion game between an *expert* and a *decision-maker* (DM). We attempt to improve strategic behavior of a simulated DM by introducing two additional strategies, inspired by human forgiving behavior. Our goal is to enhance the training set in order to improve the model's predictive ability.

1 Introduction

This work is largely an extension of the paper presented by Shapira et al. (2023) and so we encourage the reader to familiarize themselves with the paper. However, we will introduce the main ideas relevant to our work in this section.

Most of the recent research has focused on the development of agents for zero-sum games where the focus is to maximize utility. This focus has produced noticeable results; however, it contradicts with the nature of most real-world economic interactions which tend to be non-zero sum. These real-world economic interactions often involve non-cooperative, non-zero-sum elements where traditional approaches of maximizing solutions are not performing well enough. As a result, there is a growing interest in developing strategies for non-cooperative games, such as language-based persuasion games.

Shapira et al. (2023) based their research on a game based around hotel recommendation system using strategically chosen hotel reviews. Each iteration involves the expert (travel agent) selecting a hotel review from a limited selection available to him that is intended to convince the decision-maker (DM) to go to the hotel in question. This setup intends to predict and influence DM's choices, simulating realistic persuasion scenario. The research aims not just to predict optimal strategies but to understand and predict human decision-making when faced with various persuasive tactics by artificial agents.

The research extends into off-policy evaluation (OPE), where the focus is on predicting how humans respond to unknown agents in non-cooperative settings. The data was collected with a mobile application simulating a persuasion game, where human participants interacted with various artificial agents. The resulting data is intended for the development of predictive models that offer insights into human behavior in unfamiliar scenarios.

2 Data

We have enhanced the results of Shapira et al. (2023) using their dataset, fully described in their paper under "The Human-Bot Interaction Dataset." Here you can find a summary.

2.1 Reviews

Hotel reviews were sourced from Booking.com, covering 1,068 hotels with seven scored reviews

each. Half of these hotels were considered good, with a median score of 8.01.

2.2 Simulated Agent

To simulate agent behavior, simple bots with decision trees (up to depth 2) were created, resulting in 1,179 strategies. Two dissimilar sets (A and B, with six bots each) were chosen for human interaction. Data was collected via a mobile app for nine months from May 2022, with 87,204 decisions recorded (71,579 for set A, 15,625 for set B) by 245 players who completed the game.

In the game, each human played 10 rounds against six agents (from either set A or B). Agents received a positive payoff if the buyer chose the hotel, regardless of quality, while buyers got a positive payoff only for correct decisions (choosing good hotels and avoiding bad ones), with a maximum score of 10. Players needed 8-10 points to advance to the next agent, based on the difficulty level subjectively determined by the authors. The goal was to "win" against all six agents by achieving the target payoff for each.

3 Model

Shapira et al. (2023) combined actual interaction data with simulated data to enhance their predictive model. They proposed a simulation scheme generating users based on defined, but random, rules and recorded interactions with all 1,179 bots, not just predefined groups A and B. This section reviews their original simulation scheme and explains our modifications.

3.1 Game definition

In Shapira et al. (2023), each bot-DM interaction simulation involves randomly selecting six expert strategies from a strategy space. For each simulated DM-bot interaction, 10 hotels are randomly sampled, one for each of the 10 rounds. During each round, the expert uses its strategy to select a review from the hotel's review set. The DM plays the 10-round game against the same expert until achieving a payoff of 9 points, then moves on to the next expert. The DM uses the review's textual content and estimated numerical score to make decisions.

3.2 Simulated DM Strategies

The simulation relies on two probability vectors: the nature vector (p1, p2, p3), which sets initial probabilities for three basic strategies (Trustful, Language-based, and Random), and the temperament vector (pt0, pt1, pt2, pt3), updated each round. This vector consists of Oracle strategy and nature vector. The nature vector strategies include Trustful (based on past behavior and outcomes), Language-based (learning from the current hotel's review), and Random (inherent human behavior randomness). The DM strategies involve estimating review scores with added noise for realism and using an LLM to predict scores. For more details, please refer to Shapira et al. (2023).

3.3 Trustful Strategy

In the Trustful strategy, the decision maker (DM) chooses to visit the hotel only if the estimated review score aligns with the hotel's quality feedback over the past K rounds. Here, K is a randomly determined parameter unique to each DM. Unlike typical human-bot interactions, this strategy relies on the DM using the numerical review score. To replicate the challenge humans, encounter in precisely estimating scores from text, the estimated score is calculated as $\hat{s} + x$, where $\hat{s}$ represents the actual review score and $x \sim \text{Normal}(0, \epsilon)$ is a noise variable that follows a normal distribution.

The feedback on hotel quality is based on the average of the hotel's review scores.

3.4 Language-Based Strategy

In the Language-based strategy, the DM uses a large language model (LLM) to predict the review score. If the LLM predicts a score of 8 or higher, the DM decides to stay at the hotel. These predictions were generated using Text-Bison Anil et al. (2023), which evaluated each review on a scale from 1 to 100, considering a hotel to be good if it received a score of 80 or above. These scores were then adjusted to fit the 1-10 range.

3.5 Random Strategy

In the Random strategy, the DM's decision is made randomly, representing the natural variability and unpredictability in human behavior.

3.6 Oracle strategy

The Oracle strategy is designed to enable decision makers (DMs) to make the correct choice with an increasingly higher probability over time. It simulates human ability to learn and improve without explicitly defining the way in which

learning occurs. This strategy is a valuable step towards simulating human behavior.

3.7 Our additions

In our attempt to improve on the current performance of Shapira et al. (2023) model we implemented two strategies. Both of our attempts are based around forgiving human tendency we justify in related work section of this paper. The strategies attempted to simulate DM's proclivity to "give another chance" in two different ways.

3.8 Trustful up to a number of lies

This strategy (Strategy 1) modifies "trustful" strategy of the original manuscript with a "Forgiveness Threshold" – a random variable determining maximum number of lies a DM is willing to permit the expert to tell within the history window. As long as the threshold is not reached – a "lie" (mismatch between the hotel quality and the review ranking received from the travel agent) is disregarded and the decision is made as if the expert was truthful so far.

3.9 Trustful up to the lie severity

Our second strategy (Strategy 2) incorporates both the original "truthful" strategy and the "Language Based" approach. We attempted to judge the severity of the lie told by the expert by analyzing language signal from each of the previous reviews received within the history window. If the discrepancy between true hotel value and DM's LLM estimation does not exceed a "Forgiveness Threshold" – maximal permitted distance between truth and estimation – we disregard the deception and proceed as if the expert has remained truthful.

4 Related Work

This paper is largely based on the work of Shapira et al. (2023).

Our primary enhancement of the work by Shapira et al. was inspired by the seminal research of Axelrod (1980) and Axelrod and Hamilton (1981). Specifically, we drew upon sections discussing the Prisoner's Dilemma, TIT-FOR-TAT strategy, TIT-FOR-TWO-TATS strategy and the concept of forgiveness, integrating these principles to refine and expand upon the original framework.

Prisoner's Dilemma is a fundamental problem in game theory that demonstrates why two rational individuals might not cooperate even if it is in their best interest to do so.

Axelrod's research involved a series of computer tournaments where different strategies competed in repeated rounds of the Prisoner's Dilemma. One of the key outcomes of this study was the identification and analysis of various strategies that participants used.

In their 1981 paper, Robert Axelrod and William D. Hamilton extended the analysis of the Prisoner's Dilemma by exploring the evolution of cooperation. They introduced and evaluated the TIT-FOR-TAT strategy, which emerged as a highly effective approach in the iterated version of the Prisoner's Dilemma.

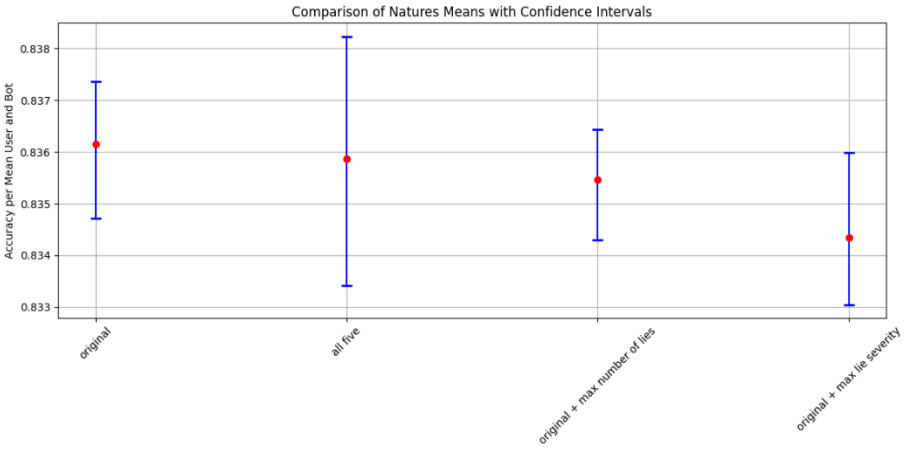
A key aspect of the TIT-FOR-TAT strategy, and its variation – TIT-FOR-TWO-TATs, is their concept of forgiveness. This trait means that while the strategy responds to defection with defection, it returns to cooperation if the opponent cooperates again. This characteristic prevents cycles of retaliation and encourages sustained cooperation. Forgiveness is crucial because it allows for the restoration of cooperation after conflicts, making long-term collaboration more viable.

TIT FOR TWO TATS was noted for being more forgiving than TIT-FOR-TAT, making it potentially more resilient in environments with noise or misunderstandings, where occasional defections might not necessarily indicate an uncooperative opponent.

Forgiveness in general is a key aspect of human interactive behavior we aimed to simulate with our strategic approach. According to Pollack and Bosse (2014) forgiveness is vital in investor – entrepreneur relationship, where continued relationship with the defected entrepreneur is considered valuable.

5 Experiments and Results

We concentrated our experiments on comparison between different basic natures performances, including and excluding strategies to compare their impact. The other parameters of the problem were kept at best performance determined by Shapira et al. (2023) in their original parameter tuning.



1. original – as described in original manuscript
2. all five – original and both our additions
3. original + max number of lies – original and new strategy 1
4. original + max lie severity – original and new strategy 2

Figure 1: Comparison between four DM basic natures. Does addition of one (or two) of our strategies improve the performance of the model

We concentrated on two main metrics –

1. Does addition of one (or two) of our strategies improve the performance of the model (results in Figure 1).

2. Does replacing the original manuscripts trustful strategy by our modification improve the performance of the model (results in Figure 2).

We observed marginally better performance of basic nature performing similarly to the original manuscripts nature but replacing the original "truthful" strategy with our strategy 1 (strategy restricting possible number of lies). However, the statistical t-test failed to showcase significant improvement in models' performance (p-value=0.297, T-statistic=0.548).

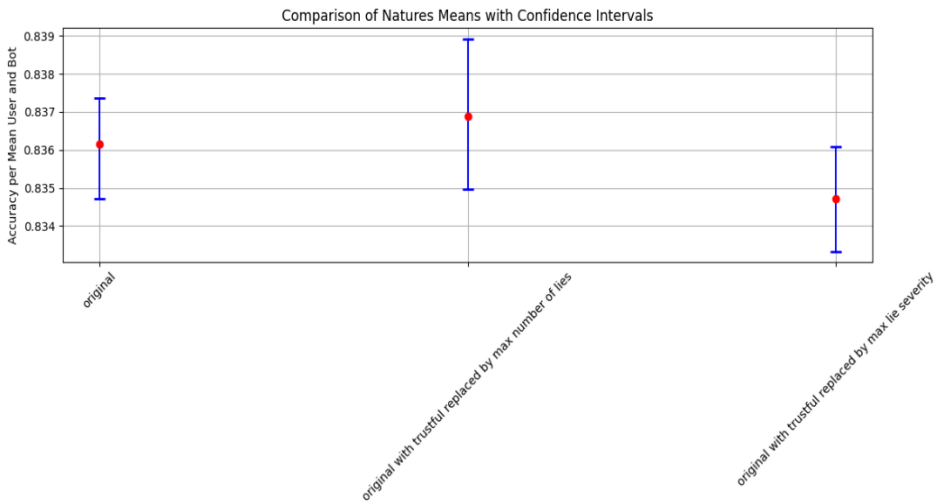
6 Discussion

While our analysis was inconclusive, we believe that adding trust-simulating parameters to the problem is a fruitful research direction and we encourage further attempts to improve this model via trust-based strategic development.

7 References

Shapira, E., Madmon, O., Reichart, R., & Tennenholtz, M. (2024). Can Large Language Models Replace Economic Choice Prediction Labs?. arXiv preprint arXiv:2401.17435.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent



1. original – as described in original manuscript
2. original with trustful replaced by max number of lies – original where trustful was replaced by strategy 1
3. original with trustful replaced by max lie severity – original and new strategy 2

Figure 2: Comparison between three DM basic natures. Does replacing the original manuscripts trustful strategy by our modification improves the performance of the model.

El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report. ArXiv:2305.10403 [cs].

Pollack, J. M., & Bosse, D. A. (2014). When do investors forgive entrepreneurs for lying?. Journal of Business Venturing, 29(6), 741-754.

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. Journal of conflict resolution, 24(1), 3-25.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. science, 211(4489), 1390-1396.