Errare humanum est? A Pilot Study to Evaluate the Human-Likeness of a AI Othello Playing Agent

Enrico Lauletta
enricolauletta@gmail.com
Sapienza University of Rome
Rome, Italy

Beatrice Biancardi
bbiancardi@cesi.fr
LINEACT CESI
Nanterre, France

Antonio Norelli
norelli@di.uniroma1.it
Sapienza University of Rome
Rome, Italy

Maurizio Mancini
m.mancini@di.uniroma1.it
Sapienza University of Rome
Rome, Italy

Alessandro Panconesi
ale@di.uniroma1.it
Sapienza University of Rome
Rome, Italy

ABSTRACT

Olivaw is an AI Othello playing agent which autonomously learns how to improve its gameplay by playing against itself. Some top-notch players (including former World Champions) reported that they had the impression that Olivaw's gameplay was human-like. To better investigate the processes related to these impressions, we conducted a pilot study using the Othello Game Evaluation App, a computer application we developed to evaluate pre-recorded Othello games in a controlled setting while assuring an adequate user experience. An exploratory analysis of the results shows that the participants mostly evaluated Olivaw as a human. When asked for a motivation for their choice, some of them reported that they evaluate poor game moves (and, consequently, losing the game) as an indication of the human-likeness of the player.

CCS CONCEPTS

• **Applied computing → Computer games**; • **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

AI agent, board game, Othello, human-likeness

ACM Reference Format:
Enrico Lauletta, Beatrice Biancardi, Antonio Norelli, Maurizio Mancini, and Alessandro Panconesi. 2022. Errare humanum est? A Pilot Study to Evaluate the Human-Likeness of a AI Othello Playing Agent. In *ACM International Conference on Intelligent Virtual Agents (IVA '22), September 6–9, 2022, Faro, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3514197.3549699

1 INTRODUCTION

Othello, also known as Reversi, is a popular 2-players board game. Olivaw [2] is an AI playing agent exploiting the AlphaGo Zero's paradigm [3] in the Othello game. Compared to AlphaGo Zero, Olivaw was developed and trained, reaching the level of the best human players, using limited resources, such as commodity hardware and free cloud services. Olivaw was also tested against other strong AI Othello players, such as Edax [1], beating them several times. That result is remarkable, as Olivaw's game tree consists of about a couple of thousands positions, which is much lower than the Edax's one, consisting of more than ten millions.

Some interesting feedback was collected in the study reported in [2]. In particular, the feedback of experienced Othello players watching some games between Olivaw and top-notch human players, including World and Italian Champions, seems to indicate some sort of human-likeness in the way Olivaw plays. Such comments include, for example "[Olivaw] did play more like a human, btw", "Those were more human style moves", "Also super human style", "[Olivaw] struggled with the edge/corner region play", followed by "Which is also exactly how it is for humans...", "[Olivaw] seems to do a really nice job of being good, but human like, and therefore a bit more "fun" to play against than most bots".

These comments are in line with the fact that human-likeness in virtual agents is not only elicited by verbal and non-verbal behavior, but also by playing strategies, especially in computer games context [4]. The fact that Olivaw is perceived as more human-like than other AI agents may be related to the way it learned. Indeed, reinforcement learning is similar to how humans learn, i.e., it involves trial and error, instead of relying on handcrafted rules based on human experts. Encouraged by this feedback, we aim at better understanding the factors contributing to the human-likeness of Olivaw. As a first step, we need to develop a more controlled setting and eliciting the best user experience to investigate these impressions. We designed and implemented the Othello Game Evaluation App, whose final version is depicted in Figure 1. The app design process followed an iterative approach. We first built a simple prototype, based on the main interface of the app WZebra[1]. Then, we conducted several brainstorming sessions and 2 interviews (one per iteration) with a former finalist of the Italian Othello Championship, asking him to try out and evaluate the current app prototype, following a cooperative evaluation UI design approach. We present here a pilot study conducted on the app providing first insights about observers' perception of Olivaw's human-likeness.

IVA '22, September 6–9, 2022, Faro, Portugal

ACM ISBN 978-1-4503-9248-8/22/09.
https://doi.org/10.1145/3514197.3549699

[1] http://www.radagast.se/othello/download.html

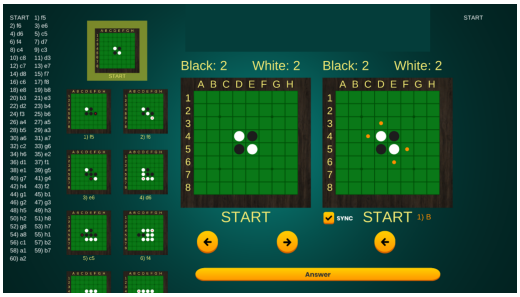

Figure 1: a screenshot of the Game Evaluation App. From the left: list of the pre-recorded game moves; corresponding board configurations; current board configuration; editable board, to try different moves from the current configuration.

2 PILOT STUDY

We introduce the pilot study, whose goal is to provide us with some first insights about the outcome of our larger future experiment (i.e., whether or not Olivaw is recognized as a human player). This is a first step to answer the question: "*How is* Olivaw*'s behavior perceived compared to a human player in terms of human-likeness?*" In addition, it allowed us to test the app with real users and check if it could be used to collect data in our future experiment.

2.1 Procedure

Nine participants (all male, members of the Italian Othello association, age 30-50yo) took part in the pilot study. Each participant judged 4 pre-recorded games (in a counter-balanced order) by browsing through the moves using the Othello Game Evaluation App. For each game, one player, i.e., the *human*, was always indicated to be a human. The participants were asked to rate the human-likeness of the other player, i.e., the *opponent*.

2.2 Experimental Design

In order to reduce the complexity of the experimental design, we had to fix some variables: the competence level of the players (Master level), and the output of the game (the *human* player always wins). The independent variables manipulated in this pilot study are:

- **color** of the *opponent* player: black (**B**), white (**W**) (it also determines who starts the game, i.e., the black player);
- the **identity** of the *opponent* player: human (**H**), Olivaw (**O**).

The resulting 4 conditions, each represented by a game taken from the Othello Quest platform, are:

- **BH**: the *opponent* player is black (i.e., playing first) and human;
- **WH**: the *opponent* is white (i.e., playing second) and human;
- **BO**: the *opponent* player is black and Olivaw (high generation 20, search depth = 400) [2];
- **WO**: the *opponent* player is white and Olivaw.

The dependent variable measured in the pilot study is the participants' evaluation of the *opponent*'s human-likeness. After exploring each game, the following question was asked:

"*Knowing that the [**color**] player is human, how would you define the identity of the [**other color**] player?*"

The answer was given by selecting a percentage pair through a sliding bar on a fine-grained scale ranging from "*100% human - 0% computer*", to "*0% human - 100% computer*". This provides additional information about the level of uncertainty of the participant about their rating. For example, the answer "*70% human - 30% computer*" suggests that the participant is quite confident about their evaluation, while the answer "*50% human - 50% computer*" suggests that the participant is not sure about the identity of the *opponent*.

2.3 Discussion

The sample size being quite small (n=9), we did not apply statistical methods but we rather focused on qualitative observations from quantitative and qualitative answers given by the participants, that will be useful for the future experiment.

2.3.1 Quantitative Answers. We look at the data under two perspectives: by considering the frequency of correct answers about the *opponent* identity, and the mean percentage about the *opponent*'s human-likeness. From the contingency table (Table 1 left) we can see that Olivaw is often mistaken for a human player, even more frequently than the human players themselves (Table 1, right).

Table 1: [left] Contingency table showing the relationship between the real identity of the *opponent* (on the rows) and the identity selected by the participants (on the columns). The *opponent* is considered as "human" if the relative percentage assigned was greater than 50%. [right] Human-likeness percentages for each condition.

	Human	Computer
Human	13	5
Olivaw	15	3

Condition	Human-likeness
BH	74.4%
WH	67.8%
BO	75.6%
WO	75.6%

2.3.2 Qualitative Answers. From the participants' free comments, it has emerged that in general they tend to assume that a computer is "strong", so when a player loses a game, especially by a large margin, or has made trivial mistakes, we tend to identify them as a human. An important criterion on which the Othello players rely on is to see if a player makes moves that are apparently inexplicable at a certain moment of the game, but then turn out to be useful many moves ahead. This kind of moves can be the result of a deep analysis of the possible future configurations, which can be done only by a computer, or they can be done by high-level players who have learned from experience, which is generally rarer.

3 CONCLUSION AND FUTURE WORK

These exploratory results encourage us to conduct further experiments using the Othello Game Evaluation App. In our future experiment, we will evaluate games when the *opponent* player wins, to check if Olivaw is perceived as "human" regardless of the output of the game. We are also interested in investigating how much Olivaw is judged "not human" compared to a traditional software, e.g., Edax [1]. This will help understand the factors contributing to the human-likeness of Olivaw.

REFERENCES

[1] Richard Delorme. 1998. *Edax*. https://github.com/abulmo/edax-reversi
[2] Antonio Norelli and Alessandro Panconesi. 2022. OLIVAW: Mastering Othello without Human Knowledge, nor a Penny. *IEEE Transactions on Games* (2022).
[3] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.
[4] Iskander Umarov and Maxim Mozgovoy. 2014. Creating believable and effective AI agents for games and simulations: Reviews and case study. In *Contemporary Advancements in Information Technology Development in Dynamic Environments*. IGI Global, 33–57.