Assessing perceived Humanness of Artificial Intelligence in Chess

A Turing Test experiment using Think Aloud and Eye Tracking methods

MSc. Thesis

Robin Koerts

TU Delft

Assessing perceived Humanness of Artificial Intelligence in Chess

A Turing Test experiment using Think Aloud and Eye Tracking methods

by

Robin Koerts

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday April 2 2024, at 10:45

Student number:	4580990
Project duration:	February 2023 - April 2024
Institution:	Delft University of Technology
	Faculty of Mechanical Engineering
Thesis committee:	Yke Bauke Eisma (Chair)
	Joost de Winter
	Dimitra Dodou

TUDelft

Assessing perceived Humanness of Artificial Intelligence in Chess: a Turing Test experiment using Think Aloud and Eye Tracking methods

Robin Koerts | **Joost de Winter** | **Yke Bauke Eisma**

ABSTRACT

With the advancement of Artificial Intelligence leading to increasingly human-like outputs, assessing a machine's ability to exhibit human-like intelligence has become more essential than ever. This study aims to investigate how human-like chess players perceive four conditions: one human opponent and three different types of algorithms. One of these algorithms, Maia, has been trained on human data and aims to play the most human-like move. In a custom-designed experiment similar to a Turing test, chess players faced off against Maia, Stockfish and a human without knowing their opponent's nature. After each game, the chess player assessed how human-like the moves of the opponent were and estimated whether they played against an engine or a human opponent. During the game, participants were asked to think aloud about their next move and react towards the moves of the opponent. Additionally, the gaze of the player was captured with the SR EyeLink Portable Duo at 1000Hz, with the goal of finding differences within the player's gaze while participants tried to discover the nature of their opponent. Results from the experiment revealed that, based on responses to a subjective questionnaire, the perceived humanness of Maia is statistically similar to a human and different from the other two chess engines. From the analysis of the voice recordings, categories of sentences were identified that could suggest recognition of the opponent, specifically: "expected", "unexpected", "human-like" and "engine-like". From the eye-tracking results, the average fixation duration and pupil diameter changes following the opponent's move were compared for each condition, but showed no statistical differences between conditions. In summary, Maia was perceived more human-like compared with other chess engines. However, differences in underlying cognitive processes on how the human perceived this difference in a Turing Test experiment were not identified.

Introduction

Artificial Intelligence (AI) is getting increasingly intelligent, including models which mimic human output.[1] Generative AI specifically has produced more sophisticated human-like output, including video generation models such as Sora[2,3], face generation with GANs[4], art models[5], music models[6] and large language models (LLMs) like ChatGPT-4.[7]

The GPT-4 model by OpenAI was trained using an unprecedented scale of data and exhibits remarkable capabilities across a variety of domains and tasks.[8] As a byproduct of the improving language skills of LLMs, researchers have explored the intriguing possibility that Theory of Mind may have emerged, previously considered exclusive to humans[9]. Distinguishing text written by humans from that generated by algorithms has become so difficult that the creators of GPT-4 have failed to create an AI classifier obtaining high accuracy[10], showing that the capability of machines to output human-like intelligence and recognising human output within the era of LLMs is more relevant than ever.[11]

Assessing a machine's capability to demonstrate human-like intelligence goes back more than 70 years.[12] Alan Turing's "imitation game", better known as the Turing Test, can be seen as a practical way of measuring a machine's ability to exhibit intelligent behavior. One algorithm that shows progress in recreating human-like decision-making is the chess engine Maia. Its goal is to facilitate collaboration between humans and algorithms by imitating human behaviour on the chessboard at different levels of expertise in a tunable way.[1] Maia uses the AlphaZero deep neural network framework to predict human moves at a higher accuracy than existing algorithms such as Stockfish and Leela.[13] The model bases its choice on its training data, consisting of 12 million chess games per model, provided through Lichess, an open-source chess website.

Historically, chess has been a key area of study in the development of artificial intelligence algorithms, with researchers using it as a model system to explore the potential of AI.[14] Chess was seen as the ultimate test for how well an algorithm could compute a problem, with Alan Turing writing one of the first chess programs exploring heuristics.[15,16] Debates and discussion about the capabilities of Artificial Intelligence increased with Deep Blue beating the at the time world champion Garry Kasparov in a six-game match in 1997[17,18] Some researchers argued that Deep Blue was the first program to pass the Turing Test, emphasizing its ability to exhibit intelligence in chess.[19]

The game of chess can also be seen as an accessible, familiar and relatively simple experimental framework which could be used to produce knowledge about other, more complex systems.[14] The game of chess has been used in an experiment setting to explore cognitive processes such as perception, problem solving and memory[20], including experiments in which the gaze of the player and verbal analysis were used to gain a comprehensive understanding of how chess experts think, evaluate positions, and make decisions during games.[21,22]

The simplicity and predefined nature of chess rules, coupled with the abundance of data makes chess a research area where the development of algorithms can excel.[23] Due to these simple and predefined rules, producing a human-like output on the chessboard could be seen as easier compared to other domains. If Maia turns out to be an algorithm which is capable of mimicking human behaviour on the chessboard indistinguishable from a human player, it could potentially be foreshadowing an increase in human-like output from algorithms in several other areas. While the research team of Maia tested the algorithm on big datasets, no within-subject experiment has been done with the maia algorithm yet.

Using chess as an experimental framework, the aim of this study is to create a Turing-test alike experiment in which the participating chess players are asked to recognise the nature of their opponent and comment upon how human-like the moves of the opponent were perceived. In this study, the perceived humanness of algorithm Maia is tested and compared with a human opponent and engine of similar skill and a state-of-the-art chess engine such as Stockfish 16[24] through the use of a questionnaire.

Furthermore, this study aims to gain an understanding of the underlying cognitive processes while chess players are trying to distinguish chess engines from human opponents through the use of modern day measurement techniques. Notably, the Turing test has not been conducted with the use of these modern measurement techniques. If measurements representing underlying cognitive processes could give a consistent indication whether a human believes they are interacting with a human or machine, it might increase relevancy of the Turing test combined with measuring cognitive processes in future robotics research.

Comprehending the cognitive processes underlying decision-making can be challenging due to the incommensurability of cognitive processes with measurement techniques. Although it is impossible to observe human cognitive processes directly, eye tracking data is believed to be a good proxy for reflecting them.[25,26,27] Eye-tracking data could provide insights into cognitive processes, as it is proposed in the mind-eye assumption that the duration of eye fixations corresponds to the time spent on mental processing.[28] Potential valuable measurements can be derived from the pupil diameter, as research shows that dilation of the pupil can indicate emotions such as surprise.[29,30] Pupil dilation scales with the level of surprise[31] and with the difficulty of the problem presented[32], suggesting that pupil-linked brain systems track the content of perceptual events, reaching a peak after about 1.5 seconds.[29,33] Measuring the pupil dilation of the chess players and comparing them per condition might potentially derive insights into differences in cognitive processes while facing off against either an engine or human opponent.

Besides assessing the perceived humanness via questionnaire data & cognitive processes while the participant is trying to distinguish between human and machine, the participant could be asked to think aloud during the experiment. The voice recordings while thinking aloud represent a part of the current thoughts of the participants, potentially giving insights similar to previously conducted think aloud experiments by De Groot.[21]

In summary, this thesis aims to not only assess the humanness of chess algorithms, but also to investigate potential differences within measurements which could represent underlying cognitive processes when participants try to distinguish between computer and human.

The goal of this paper is to answer the following research question:

How human-like do chess players perceive algorithms such as Maia compared to traditional engines and human opponents, and can differences in cognitive processes be measured through eye tracking or think-aloud recordings when distinguishing between computer or human?

This paper contains the methods used to create this experiment, after which an analysis of the obtained results will be presented. The research question will be examined while discussing the results, ending with the conclusion and potential future prospects of this research.

Methods

Participants

A total of 24 participants took part in the experiment, of which 23 male and one female. The participants ranged in age from 14 to 59 years old (*M = 26.3*, *SD = 9.3*). Out of these 24 participants, most played chess regularly(*M = 5.5*, *SD = 1.2*, where 1 indicates few times a year and 7 daily). Participants typically played more chess against human opponents than against chess engines, usually playing against engines only a few times a year (*M = 1.5*, *SD = 0.7*, where 1 indicates a few times a year and 7 daily). 23 out of 24 participants were rated online, which after converting ratings of the participants towards a LiChess Blitz rating[34] resulted in a mean rating of 1645 and a standard deviation of 348. Six participants had a FIDE rating (*M = 1910*, *SD = 279*). For a more detailed overview of the participants's characteristics , see Appendix 2.

The experiment received approval from the Delft Human Research Ethics Committee, with each participant providing written informed consent before participating in the experiment.

Apparatus

The experiment used a 17.3-inch monitor of the laptop model ROG Zephyrus S17 GX701 GX701LXS-XS78 with a total display area of 383x215 mm and a screen resolution of 1920×1080 pixels.[35] The root of the screen of the laptop was placed 71 cm from the table.

Speech was planned to be recorded using the Philips DVT2810 Voice Tracer audio recorder.[36] Halfway during the fifth participant, this device broke down. As a last minute replacement, the iPhone 14 was used for participant 5, 6 and 7.[37] For the last 17 participants, the Olympus VP-20 was used to record speech.[38]

Eye movements were recorded using the SR Research EyeLink Portable Duo.[39] With the Portable Duo, no head support is needed to capture eye-tracking movements. This is vital within the design of this experiment, because participants have to be able to speak during the experiment. The Portable Duo uses specialized algorithms for head free-to-move and head-stabilized tracking modes.[40] The Portable Duo was placed 65 cm from the table and captured eye movements up to 1000 Hz in head free-to-move mode, which was used within this experiment.

To optimize the freedom of movement of the participants and the range of measuring the gaze of the chess player, the angle of the Portable Duo was changed to face directly towards the face of the participant. Due to low errors in the accuracy obtained in head free-to-move mode (Mean = 0.26°, Median = 0.22° and Standard Deviation = 0.16°)[40], changing the angle of the device before calibrating will result in minimal changes on the eye tracking data.

Experiment setup

The experiment was conducted via WebLink, a screen recording software solution which records eye movements while participants view and interact with static or dynamic media such as a website.[41] The software captures the screen, eye movements like fixations, saccades, blinks and pupil diameter, but also browser navigation and history, key presses, mouse clicks and mouse positions.[40]

During the experiment, the participant is guided through several instructions via Weblink before calibrating the Portable Duo. When calibrated, the participant is sent to the LiChess website. When the game is finished, the experimenter would close the LiChess.org page and the participant would be guided to questionnaire page on google forms to answer questions about the just played game of chess.

LiChess.org was chosen due to its ability to start relatively easy from a pre-made position, an advantage for this experiment, explained in greater detail within section "Positions Played From". LiChess also has a ZEN mode, removing all excessive elements on the webpage.[42] This results in less possibilities for distraction & less objects to analyse for the participants, possibly advantageous for eye-tracking experiments.

Due to cheating avoidance, LiChess bans all kind of engine usage. This is why, for this experiment, the account "TuringTest001" was converted to an account with "BOT" status.[43] This function exists for people playing with the help of engines, also called advanced chess, cyborg chess or centaur chess.

Figure 1. (a) The seating place of the participant, with on the table the Olympic VP-20 and the Portable Duo placed upon the laptop model ROG Zephyrus S17 GX701 GX701LXS-XS78. (b) The view of the participant during the experiment. (c) The seating place of the researcher. (d) The view of the researcher, containing from left to right 1) a monitor displaying the chess GUI, here for condition Stockfish3500 2) a laptop where the researcher plays chess against the participant 3) a laptop provided by SR Research, where the researcher is able to check whether the Portable Duo is still able to collect gaze data.

Experiment task

While playing chess, the main task for the participant of this Turing-test alike experiment is to identify the nature of the opponent: Human or Engine. The participant did not know how many of their opponents would be an engine, as participants were told that the proportion of human and engine opponents might vary.

Each chess game was limited in time, 5 minutes a side with no seconds added to the block for each move that is played. The participant could spent their time however they would like. Their opponent was forced to move every 10 seconds.

The participants were not encouraged or discouraged to win. Participant were allowed to play moves which were considered non optimal, but were not encouraged or discouraged to do so.

Experimental procedures

When the participant arrived, they were welcomed by the researcher and were asked to sign an consent form including some informative instructions and fill in the questionnaires in appendix 2.

Within this questionnaire, the participant was asked whether they think engines played differently than humans, where 23 out of 24 participants answered "Yes".

With the intent to let people think about the experiment before participating, all chess players were asked how they would try to recognise the difference between an engine and a human opponent in the game of chess. All answers are displayed in Appendix 2 but in general, participants highlighted focusing on unusual moves, tactical superiority and deviations from expected patterns in the opening or mid-game.

After filling in the pre-game questionnaire, the participants were given vocal instructions about the experiment, starting about the game of chess and the task of the participant, which was not to win, but to recognise the nature of the opponent while playing chess. Additionally, the participant was instructed to think aloud in English while the voice recorder is enabled and giving some examples to talk about, for example the move the participant is considering & reflecting on the move of the opponent, specifically if the participant expected the move or not, was logical or not or gave away the nature of the opponent. Furthermore, the participant was instructed to sit as still as possible after the eye tracker has been calibrated. Within the weblink environment, instructions similar to the vocal instructions were repeated on screen at the start of the experiment. For more information about the specific instructions given and the order in which they were given, please consult Appendix 3.

Before the calibration of the Portable Duo, the participant was asked to sit comfortably. During the experiment, the participant had to sit in a certain range for the Portable Duo to register their eye movements on the screen. During the experiment, the researcher could check whether the eye gaze was being registered. If this was not the case, the researcher would ask if the participant could move in a specific way in a soft voice to avoid disturbing the voice recording.

After calibration of the Portable Duo, the weblink application sent the participant to the LiChess website and received an invitation from the researcher. From the moment the participant accepted the position, the voice recorder was activated by the researcher. The participant would first get used to talking aloud by analysing the position. When finished analysing, the participant played their first move and the game continued until either a checkmate occurred, time on either side ran out or the participant resigned.

High level engines occasionally play optimal moves in less than a second, potentially giving away their nature in complex situations where humans would have to think thoroughly. To increase difficulty in recognising the opponent for the participant, the researcher would play a chess move every ± 10 seconds.

To ensure no mistakes are made by the researcher, the three different engines were controlled from three different interfaces: Stockfish3500 from the Nibbler chess GUI[44], Stockfish1500 from the Scid vs PC chess GUI[45] and Maia1 from the LiChess website interface.[46]

The order in which the conditions would be played by the researcher were determined before the experiment, randomly assigning all possible combinations with one repetition. Please consult appendix 1 for the full overview of the order of the conditions.

When the game was finished, the researcher would deactivate the voice recorder and make sure the participant was guided towards a google forms, where the participant could give their opinion about the game just played.

In total, eight games were played per participant, so every participant encountered all conditions twice. After the instructions given and four games were played, a break was held. Every experiment, including instructions, calibration and breaks took about 2 hours to complete.

Positions Played From

Within this experiment, the chess game was not started from the well known starting chess beginning position. This was chosen due to a number of reasons:

1. If a skilled chess player plays a theoretical main opening line, usually chess players have played this line before and perhaps thought out several variations. This would potentially limit their need to scan the chessboard,

since the player already knows the next possible moves. One of the goals of not starting from the normal chess position is to force chess players to scan the board to enforce good/better eyetracking data.

2. High-level engines consistently play theoretical opening lines, easily recognisable by high-level chess players which also know these ’optimal’ sequences of moves.

3. Low-level engines can play "weird" opening lines, no decent skilled human would ever attempt.

4. Mistakes known as "blunders" can be prevented within the first few moves when choosing a positional position and not a tactical position, with intended consequence that the participant will not achieve a chess position which is lost within a few moves.

5. From a tactical situation, engines could potentially be recognised relatively easily by their strong tactical play. This is why extreme tactical positions such as a forced capture or a situation where only a limited number of first moves leads to an equal game are avoided as a starting position.

All positions played from were games played on LiChess from the highest Rapid rated player at the time of preparing the experiment, named GM Drvitman. A PNG was downloaded containing Rapid games.[47] All games were evaluated by Stockfish 16 from white’s perspective, where each point is worth a full pawn. 0.01 point can be seen as a centipawn, a hundredth of a pawn. The evaluation measures how distant the evaluation of the move a player made in a game is from the suggested computer’s best move.[48]

The downloaded games were evaluated after 9 moves and filtered on the following requirements to be used within this experiment:

- The position should be rated equal by Stockfish 16. Any evaluation between -1.0 and 1.0 will be seen as equal enough for this experiment.

- At least 2 pawns have to be moved from both sides.

- At least 2 out of the 4 minor pieces (knight and bishops) should be developed from both sides.

- The position must be positional, not tactical. This means all positions which contain one forced move are eliminated, giving the participant several options as a resonable first move.

Figure 2. An example of a typical starting position (Study 1 Chapter 61). Stockfish analysis shows that the the five best moves to be played for white (a4, f4, Bf4, Qd3 and Be3) result from a +0.2 to -0.1 chess position with plenty of moves resulting into in an equal start.

All selected position were recreated in LiChess studies. In total, 240 positions were selected to supply games for a maximum of 30 participants playing 8 games. Since there is a maximum of 64 positions within a study, four studies have been created. For the 24 participants that participated, the first three studies were used in total. Links to all created positions can be found within Appendix 1.

Independent variables

The independent variable, or the variable that is manipulated by the researcher, is the opponents the participant played against for this study. To draw a fair comparison, the skill level of the conditions should be about equal.

There are four types of opponents, henceforth called conditions:

- **Human**: To compare perceived humanness of algorithms with an actual human, the first condition is a human chess player. To standardize this condition, the same chess player is used in all experiments. To assess the skill level of the human player, his chess.com account rating was consulted: 944 Rapid chess.com. This converts roughly to a 1100 LiChess Blitz rating.[34]

- **Maia**: Maia is a human-like neural network chess engine.[49] Its goal is to play the human move of a certain skill level - not to play the "best" move. Nine skill-varying models have been trained, each on 12 million LiChess games.[1] Within this experiment, the algorithm Maia1 was chosen to replicate a similar skill level as the human, as Maia1 has been trained upon LiChess games of players rated around 1100.

- **Stockfish3500**: Stockfish is a free and open-source chess engine.[24] Stockfish has ranked first in 13 of the last 15 Top Chess Engine Championship.[50] As of January 2024, it is considered the strongest CPU chess engine with an estimated Elo rating of 3634.[51] It is able to search promising variations to a greater depth due to the use of heuristic tree search algorithms.[52] Within this experiment, this condition

- **Stockfish1500**: The very same open-source chess engine has an option to lower its skill level, which this condition represents. Within this experiment, the skill level of Stockfish was set to 4/20 to replicate a skill level relatively similar to the human condition, chosen after play testing.**??**

Dependent variables

The dependent variables or the measured variables are categorised into four categories: Chess results, Subjective questionnaire metrics, voice recording metrics and eye tracking metrics.

Chess results

The results of the chess games might provide insights into the occurred events during the games of chess, potentially deriving into insights to better understand the rest of the results. The following data can be obtained from the chess games and are shown per condition:

- **Percentage of games won**: Number of points obtained against each condition, from the participant's perspective.

- **The number of moves** played against each condition.

- **Time thought participant**: The number of seconds a participant took before making a move.

- **Time thought researcher**: The number of seconds the researcher took before making a move. This variable is mainly calculated as a general check to ensure consistency by the researcher.

Subjective questionnaire metrics

After each chess game, the participant was asked to fill in a questionnaire to measure the following metrics:

- **Perceived humanness of the moves of the opponent**: To measure the perceived humanness of the moves, the participant gave an indication after each chess game. A 7-point Likert-scale is used[53], with 1 representing computer and 7 representing computer.

- **Expected opponent**: As a second indication of the perceived humanness of the opponent, a binary question was used to indicate the nature of the opponent: Human or Engine.

- **Confidence in recognition of your opponent**: The participant gave an indication of their confidence of successful recognition of the opponent. A 7-point Likert-scale is used[53], with 1 representing not confident at all and 7 representing extremely confident.

- **Strength of the opponent**: The participant gave an indication of the estimated strength of their opponent. A 7-point Likert-scale is used[53], with 1 representing weak and 7 representing strong.

The first two questions intent to measure perceived humanness. Although the two questions are alike, a subtle difference does exist. Since Maia's nature is an engine which plays the most human-like move it can find, two different questions are asked: one about the moves and one about the nature of the opponent.

Although the two questions are alike, a subtle difference exist within the questions. The first question tries to measure how human-like the moves of the opponent were perceived on a scale on 1 till 7. Here, the participant can indicate how human-like the opponent was perceived due to the scale. The second question is a binary one, forcing the participant to choose between only two options: Human or Engine.

The difference in these two questions is interesting on another level: Can participants, even though they think the moves of Maia are human-like, still discover that they are playing against an engine? Difference these two questions could indicate other factors playing a role than just how human-like the moves were perceived.

The third question was intended as a general test for the entire experiment. If all participants would be unsure after every game, it might have been possible to conclude that the differences found within the data were coincidental.

The fourth question was intended as an after experiment strength comparison of the used algorithms. Within the design of this experiment, the goal was to increase difficulty in recognising the nature of the opponent by using algorithms with similar strength as the human playing.

The result section shows the average over the two questionnaires when the participant played against the same condition.

Voice recording metrics

During this think-aloud experiment, the voice of the participant was recorded. The thoughts of participants are represented by their vocal analysis done during the experiment, which can be compared per condition.

First, 185 voice recordings have been transcribed from audio to text using Whisper.[54] Due to some unintentional Dutch speech by the participants during some of the experiments, the "Medium_EN" model has been used, currently the most accurate Whisper model for English language detection.[55]

Through the use of whisper, two type of files have been generated for every available voice recording:

- **PXCX Condition.rtf** - containing all sentences spoken during the experiment.

- **PXCX Condition_Segmented.txt** - containing all sentences spoken during the experiment & Timing of beginning and end of the sentence, rounded to 0.5 seconds.

To obtain quantifiable differences within how the conditions were perceived, the .txt files containing all spoken sentences have been analysed individually by the OpenAI model 'gpt-4-0125-preview'[56] whether each sentences spoken could be put into either of these categories: 1) an unexpected, unlogical, surprising, weird or strange event happens in the game 2) an expected or logical event occurs in the game 3) The participants suspects the opponent is an engine, computer, unhuman, pc or machine 4) the participants suspects the opponent is human or a person.

For the first two categories, sentences were excluded when they contained information about how to potentially discover the opponent is either human or engine. The the last two categories, sentences were excluded when potential future moves would be either expected or unexpected.

When the sentence has been matched with a certain category, this sentence is outputted into a specific excel file as a json. This allows us to compare conditions by counting the number of sentences spoken within each category.

Due to inconsistencies within the LLM, each category has been analysed 10 times. From this, the element wise average has been calculated, which have been used in further calculations.

From this element wise average per participant, the average number of sentences spoken within the two recorded trials against a certain condition has been counted per category such that conditions can be compared. The averages and standard deviations of these counts are shown per condition within the results.

Due to different number of sentences being spoken when playing against a certain condition per participant, a second number has been reported in the result section. The average number of sentences spoken within the two recorded trials has been didived by the total number of sentences each participant spoke when playing against each condition. Then, the averages and standard deviations of these counts have also been reported in the results section to compare conditions in categories.

To provide the most accurate prompt, all sentences of 10 .rtf files were marked by hand whether they could be put into either category. With these 10 files serving as a benchmark, the system instructions and prompts could be tested on accuracy as the hand checked files could provide accuracy of these system instructions and promps by comparing the sentences from the hand checked files by the sentences outputted by the LLM. An example of these hand checked files are visible in Appendix 8. The exact system instructions and prompts used to put sentences in a specific category can be found in Appendix 7.1.

Additionally, the LLM 'gpt-4-0125-preview'[56] was used within python to discover which categories spoken sentences can potentially be placed in, including listing these categories from most frequently occurring to least occurring. All sentences spoken for a specific condition were included into the prompt given to the LLM, potentially revealing differences in categories within the four conditions. The specific system instruction and prompt given to the LLM to find differences in categories of spoken sentences per condition given can be found in Appendix 7.2.

Eye tracking metrics

After extracting the 1000Hz eye tracking data for all participants, the eye tracking data has first been filtered on timestamps where a chessboard is displayed on screen, discarding eye tracking data obtained when weblink instructions, LiChess home page or the questionnaire were visible on screen. Secondly, all timestamps including missing gaze data, such as blinks and the participant not looking at the screen were excluded from the analysed data, including an 100ms gap before and after each missing timestamp.

When filtering on fixations within the eye tracking data, fixations can be found within the data between two saccades. Within the extraction of the data, weblink has the option to output for every data point 1 for saccade and 0 for no saccade. These can be found in column 26, 27 and 28 of the eye tracking data for the average, left eye and right eye respectively. The thresholds used for finding saccades within the eye tracking data are as follows: 30 °/s velocity & 8000 $°/s^2$ acceleration.

To obtain data from after the opponent makes a move, first the timestamps of the moves played within the eye tracking data need to be extracted. This can be done by using the MoviePy library in python, extracting sounds from the screen captured videos when it exceeds a certain threshold. Due to the length of the sound being constant for moves and captures, the timing of the moves made can be extracted with a higher accuracy compared to PNG data of the chess games.

Unfortunately, participants 8, 19 and 22 pressed the backwards and forwards key several times to analyse previously played moves within this experiment, creating the same sounds as a played move when analysing previous positions. This is why these participants were excluded from the after-move analysis.

Fixation Duration

With the intent to represent an indication about the time spent on mental processing, the average fixation duration has been extracted per experiment by the following method:

$$\textbf{Average fixation duration per trial} = \frac{\text{Total ms in fixation per trial}}{\text{Number of fixations per trial}}$$

To give an indication whether the fixation duration changes after the opponent made a move, the fixation duration has been extracted for every move made by the opponent by the following method:

$$\textbf{Fixation duration after opponent made a move} = \frac{\text{Total ms in fixation 5000ms after the opponent made a move}}{\text{Number of fixations 5000ms after the opponent made a move}}$$

5000 ms was chosen to give the best indication of think time, considering the high probable variation within the time in second elapsed before a participant will make a move.

Pupil diameter

Per move of the opponent, the absolute values of the pupil diameter in the next 2000ms have been captured per participant per experiment. Extracting the values of the pupil diameter of all moves played by a certain condition, an element-wise average can be obtained of all the absolute pupil diameter values per participant per condition.

With these averages, it is possible to plot the absolute and relative pupil diameter change over a time span of 2000ms. For both the absolute and relative plots, transparent grey lines are plotted for each individual average per participant per condition and the average of these 20 lines is plotted in red. To prevent inconsistencies, temporal data of the pupil diameter where the participant blinked in the 2000ms following a move from the opponent have been excluded from these plots and further pupil diameter analysis.

To analyse the change in pupil diameter across all participants, the peak of the average per participant has been found. To reduce the impact of outliers on the results, the index of the peak has been found and the average of the pupil diameter values 100ms before and after this peak have been obtained, over which the average has been taken. This average is compared with the starting value of the pupil diameter of the average per participant to obtain relative change of the pupil diameter after the opponent has made a move:

$$\Delta\text{Pupil Diameter}_{\text{participant}} = \frac{\overline{\text{Pupil Diameter}_{200}} - \text{Pupil Diameter}_{\text{move}}}{\text{Pupil Diameter}_{\text{move}}}$$

where

- $\Delta\text{Pupil Diameter}_{\text{participant}}$ represents the relative change per participant in pupil diameter,
- $\overline{\text{Pupil Diameter}_{200}}$ is the average of 200 absolute pupil diameter values around the peak found within the average per participant,
- $\text{Pupil Diameter}_{\text{move}}$ is the absolute pupil diameter value of the average per participant at the timestamp of the opponent's move. (t = 0 ms)

These values have been reported in the result section by means of tabular data and a box plot.

Statistical analysis

Statistical analysis has been done for all of the results, to display significance within the differences within the obtained results.

It should be noted that not all data has been recorded, specifically voice recordings are missing for Participant 5 Conditions 1 - 4, Participant 8 Condition 8, Participant 9 Condition 1 and Participant 24 Condition 3. Missing 7 voice recordings is far from ideal, however, it was estimated that the remaining 185 voice recordings represent the participant group well enough. Due to the 7 missing recordings, the results of the, voice recordings are displayed per recorded condition to normalize the displayed results.

Unfortunately, one of the 192 eyetracking recordings has been corrupted as well, specifically Participant 5 Condition 5.

Participants 8, 19 and 22 have been excluded from the after move analysis due to the fact that during the experiment, these participants looked at previous moves. Within the current method used, these were labeled as moves. To maintain certainty that results show actual after moves data, these participants are excluded from statistical analysis looking at eye tracking data after a move from the opponent.

Due to interruption of the experiment for participant 13, two eye tracking files, 13B and 13C, have been combined for participant 13 into 13B. However, since the gaze measurements were very noisy, participant 13 was also excluded from the after move analysis.

To indicate significance between two conditions in the subjective questionnaire data, the paired t-test was used since the experiment done within this study is an within-subjects experiment. For a comparison between two conditions, a t-stat and p-value are displayed between all conditions. Generally, a threshold of P = 0.05 is used to determine whether effect is statistically significant for a paired t-test.[57,58,59]

To indicate significance within the objective measurements, the voice recordings and eye tracking data, the paired t-test was also used to indicate statistically significant difference by showing a t-stat and p-value between all conditions for all categories.[57,58,59]

Results

Chess games

The number of wins out of 48 games played against each condition is displayed in Table 1 and the mean and standard deviation of the number of moves played against each condition, the seconds thought by the participant against each condition and the seconds thought by the researcher playing as a condition are displayed in Tables 2, 3 and 4 respectively.

Condition	Number of wins
Human	36
Maia	32.5
Stockfish1500	13.5
Stockfish3500	1

Table 1. Number of wins obtained from white's perspective, out of 48 games played against each condition. A draw counts for half a point.

Condition	Mean	SD
Human	23.29	5.21
Maia	21.88	6.19
Stockfish1500	22.19	5.93
Stockfish3500	17.07	4.77

Table 2. The number of moves played against each condition, equal to the number of plies from the participant.

Table 3. Time in seconds that elapsed before a participant made a move, with mean and standard deviation calculated for each condition.

Condition	mean	sd
Human	10.38	3.42
Maia	10.85	3.31
Stockfish1500	12.94	5.22
Stockfish3500	13.43	3.76

Table 4. Time in seconds that elapsed before the researcher made a move, with mean and standard deviation calculated for each condition.

Condition	mean	sd
Human	9.96	0.42
Maia	10.49	0.44
Stockfish1500	9.98	0.62
Stockfish3500	10.06	0.27

On average, the researcher managed to play move around ±10 seconds for each condition and the participant needed less time thinking when playing against Human & Maia compared to the Stockfish opponents.

Questionnaire

Figure 3 shows a box plot where is displayed per condition how "human-like" each participant estimated the moves of their opponent. Table 5 displays the mean, median and standard deviation of these results and Table 6 displays a paired t-test, where the different opponents are compared to each other to note statistical difference found within the questionnaire data.

Table 5. Perceived humanness of the moves for different opponents.

Opponent	Mean	Median	SD
Human	5.229	5.25	1.207
Maia	4.979	5.0	1.036
Stockfish1500	3.583	3.5	1.336
Stockfish3500	3.229	3.25	1.479

Table 6. Paired T-Test Results comparing the questionnaire data on the perceived humanness of the moves for different opponents.

Comparison	T-Stat	P-Value
Human vs Maia	0.835	0.412
Human vs Stockfish1500	4.317	2.556×10^{-4}
Human vs Stockfish3500	5.112	3.537×10^{-5}
Maia vs Stockfish1500	3.714	1.143×10^{-3}
Maia vs Stockfish3500	4.132	4.049×10^{-4}
Stockfish1500 vs Stockfish3500	0.947	0.354

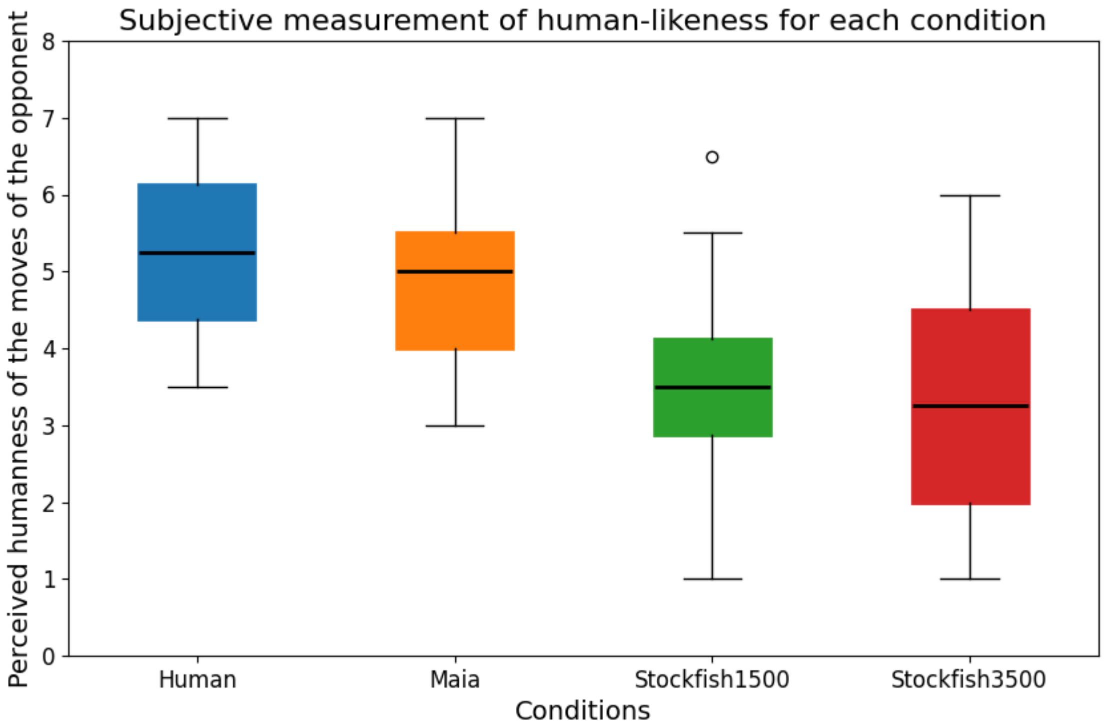


Figure 3. Boxplot displaying the subjective measurement of the perceived humanness from each participant of the moves of the opponent for each condition, ranging from 1 (computer-like) till 7 (human-like), answered by each participant after each game.

The averages of the answers to the next three questions from the questionnaire are displayed in Table 7, deriving from the questionnaire visible in Appendix 4. The first given value for each condition displays the number of times Human was chosen divided by the total responses per condition, which was 48. For the second and third questions, representing Confidence Recognition Opponent and Strength Opponent respectively, the average of the 48 responses per condition to the questionnaire are displayed.

Table 7. Number of perceived human opponent, average confidence in recognition of opponent and average observed strength displayed per type of opponent. Confidence ranges from 1 (not confident at all) till 7 (extremely confident) & Strength ranges from 1 (weak) till 7 (Strong).

Condition	# Human	Confidence Recognition Opponent	Strength Opponent
Human	34/48	5.1250	3.7292
Maia	33/48	4.9792	3.7917
Stockfish1500	18/48	4.9375	4.5208
Stockfish3500	12/48	5.3542	6.25

For the number of humans identified, a second t test is done, visible in Table 8.

When combining results from Table 6 and 8, it is visible that Maia is most similarly affiliated with the human opponent, with paired t-tests ($t(23) = 0.835$, $p = 0.412$) & ($t(23) = 0.225$, $p = 0.824$) indicating no statistical significance difference between how human-like Maia and the human were estimated. However, when comparing both Stockfish conditions with Maia or the human opponent, statistical significance difference can be found when

Table 8. Paired T-Test Results for the number of perceived human opponents.

Comparison	T-Statistic	P-Value
Human vs Maia	0.225	0.824
Human vs Stockfish1500	3.391	0.0025
Human vs Stockfish3500	4.836	7.009×10^{-5}
Maia vs Stockfish1500	3.315	0.0030
Maia vs Stockfish3500	4.764	8.393×10^{-5}
Stockfish1500 vs Stockfish3500	1.661	0.110

comparing the conditions. When comparing the two Stockfish conditions, paired t-tests ($t(23) = 0.835$, $p = 0.412$) & ($t(23) = 0.225$, $p = 0.824$) indicates similarities within the perceived humanness.

For 58% of the answers given, a certain move or moment in the game was noted where the participant felt that he/she recognised the nature of the opponent. Due to the large variety in these answers, these 124 responses were not used within this study, but could potentially be useful for cross-data analysis.

For all questionnaire questions, see Appendix 4. An .csv file is available within the published dataset, where all answers can be found.

Voice recording

Due to GPT-4 not producing the exact same output, even with temperature setting set to zero, the element wise average of these 10 runs has been used for further calculations. From these element wise averages, the number of sentences which a participant has spoken on average from the two trials when playing against each condition are shown per category per condition in column 2 and 3 of Tables 9, 10, 11 and 12.

Furthermore, since the number of sentences spoken in total differ per condition, displayed in Table 17, the number of sentences spoken per participant when playing against a certain condition have been used to show a fairer comparison. Per condition, these total number of sentences spoken give a percentage of the number of sentences which a participant has spoken when playing against a certain condition, shown per category per condition in column 4 and 5 of Tables 9, 10, 11 and 12.

Table 9. The number of sentences mentioning Expected or Logical events per condition. The first set of mean and standard deviation indicates the average sentences spoken per participant against each condition. The second set shows the percentage of sentences spoken per participant against each condition. The element-wise average of 10 runs is computed for each participant.

Condition	Mean	SD	% Mean	% SD
Human	3.98	4.72	2.52%	2.40%
Maia	3.03	3.83	1.95%	1.70%
Stockfish1500	4.84	4.69	3.07%	3.04%
Stockfish3500	3.91	3.95	2.57%	2.36%

Table 10. The number of sentences mentioning Unexpected or Illogical events per condition. The first set of mean and standard deviation indicates the average sentences spoken per participant against each condition. The second set shows the percentage of sentences spoken per participant against each condition. The element-wise average of 10 runs is computed for each participant.

Condition	Mean	SD	% Mean	% SD
Human	3.72	2.91	2.82%	2.74%
Maia	3.92	2.34	3.22%	1.92%
Stockfish1500	6.43	3.86	4.44%	2.16%
Stockfish3500	3.98	3.35	2.92%	1.94%

Table 11. The number of sentences mentioning a suspected Engine opponent per condition. The first set of mean and standard deviation indicates the average sentences spoken per participant against each condition. The second set shows the percentage of sentences spoken per participant against each condition. The element-wise average of 10 runs is computed for each participant.

Condition	Mean	SD	% Mean	SD
Human	1.94	2.31	1.45%	1.83%
Maia	1.23	1.62	0.82%	1.23%
Stockfish1500	2.61	3.55	1.86%	2.58%
Stockfish3500	1.79	2.49	1.47%	2.12%

Table 12. The number of sentences mentioning a suspected Human opponent per condition. The first set of mean and standard deviation indicates the average sentences spoken per participant against each condition. The second set shows the percentage of sentences spoken per participant against each condition. The element-wise average of 10 runs is computed for each participant.

Condition	Mean	SD	% Mean	SD
Human	1.63	2.14	1.00%	1.23%
Maia	1.99	2.52	1.66%	2.11%
Stockfish1500	1.93	2.42	1.52%	2.21%
Stockfish3500	0.69	1.01	0.58%	0.94%

Visual representation of the data given in Tables 9, 10, 11 and 12 are computed in box plots, found in the appendix 7.3 in figures 22, 23, 24 and 25. To indicate statistical significance difference between each of the conditions, Tables 13, 14, 15 and 16 show paired t-tests between each condition for each category done on the number of sentences being mentioned in a specific category.

Comparison	T-statistic	p-value
Human vs. Maia	1.08	0.2935
Human vs. Stockfish1500	-0.88	0.3891
Human vs. Stockfish3500	0.09	0.9327
Maia vs. Stockfish1500	-2.17	0.0407
Maia vs. Stockfish3500	-0.88	0.3895
Stockfish1500 vs. Stockfish3500	1.13	0.2705

Table 13. Paired t-test comparing all conditions on the number of sentences containing Expected or Logical events.

Comparison	T-statistic	p-value
Human vs. Maia	-0.29	0.7754
Human vs. Stockfish1500	-3.67	0.0013
Human vs. Stockfish3500	-0.30	0.7680
Maia vs. Stockfish1500	-3.49	0.0020
Maia vs. Stockfish3500	-0.11	0.9106
Stockfish1500 vs. Stockfish3500	3.48	0.0020

Table 14. Paired t-test comparing all conditions on the number of sentences containing Unxxpected or Illogical events.

Comparison	T-statistic	p-value
Human vs. Maia	1.32	0.2014
Human vs. Stockfish1500	-0.84	0.4119
Human vs. Stockfish3500	0.25	0.8050
Maia vs. Stockfish1500	-2.25	0.0340
Maia vs. Stockfish3500	-1.09	0.2859
Stockfish1500 vs. Stockfish3500	1.17	0.2559

Table 15. Paired t-test comparing all conditions on the number of sentences containing suspected Engine opponent.

Comparison	T-statistic	p-value
Human vs. Maia	-0.70	0.4906
Human vs. Stockfish1500	-0.63	0.5319
Human vs. Stockfish3500	1.79	0.0858
Maia vs. Stockfish1500	0.09	0.9269
Maia vs. Stockfish3500	2.30	0.0311
Stockfish1500 vs. Stockfish3500	1.98	0.0603

Table 16. Paired t-test comparing all conditions on the number of sentences containing suspected Human opponent.

Tables 13, 14, 15 and 16 shows statistical difference between Maia and Stockfish1500 in the category Expected and Logical events ($t(23)$ = -2.17, p = 0.0407), between Stockfish1500 and the other conditions in the category Unexpected and Illogical events ($t(23)$ = -3.67, p = 0.0013) & ($t(23)$ = -3.49, p = 0.0020) & ($t(23)$ = 3.48, p = 0.0020), between Maia and Stockfish1500 in the category suspected Engine opponent ($t(23)$ = -2.25, p = 0.0340) and between Maia and Stockfish3500 in the category suspected Human opponent ($t(23)$ = 2.30, p = 0.0311)

Total Spoken Sentences	
Human	3688
Maia	3104
Stockfish1500	3582
Stockfish3500	3335

Table 17. Total Spoken Sentences when by Different Entities

Appendix 7.2 shows the system instruction and prompt for finding potential new categories. It also shows the found categories for each condition, which are the exact same conditions in the same order of frequency.

Eye tracking

Average Fixation Duration

In figure 4, the average fixation duration during this experiment is visually displayed, where one data point is equal to the average of the average fixation duration of two full trials of the experiment. The mean and standard deviation of the average fixation duration are displayed in table 18.

Figure 5 displays the average fixation duration in the 5000ms after the opponent plays a move, with Table 19 displaying the mean and standard deviation.

Table 18. Average fixation duration per participant in seconds for each condition for the whole trial

Condition	Mean	SD
Human	0.277	0.043
Maia	0.275	0.046
Stockfish1500	0.272	0.040
Stockfish3500	0.272	0.040

Table 19. Average fixation duration per participant in seconds for each condition after a move of the opponent

Condition	Mean	SD
Human	0.248	0.045
Maia	0.243	0.039
Stockfish1500	0.245	0.045
Stockfish3500	0.242	0.032

The results of the paired t-test for the fixation duration are displayed in Table 20 for the whole trial per condition & in Table 21 for the fixation duration 5000ms after the opponent makes a move. Results indicate no statistical differences found between the conditions.

Comparison	T-statistic	p-value
Human vs. Maia	0.67	0.5096
Human vs. Stockfish1500	1.73	0.0967
Human vs. Stockfish3500	1.66	0.1113
Maia vs. Stockfish1500	1.27	0.2164
Maia vs. Stockfish3500	1.17	0.2528
Stockfish1500 vs. Stockfish3500	-0.19	0.8524

Table 20. Paired t-test comparing all conditions on fixation duration for the whole trial.

Comparison	T-statistic	p-value
Human vs. Maia	0.69	0.5005
Human vs. Stockfish1500	0.20	0.8469
Human vs. Stockfish3500	0.76	0.4592
Maia vs. Stockfish1500	-0.25	0.8040
Maia vs. Stockfish3500	0.10	0.9185
Stockfish1500 vs. Stockfish3500	0.29	0.7714

Table 21. Paired t-test comparing all conditions on fixation duration after opponent's move.

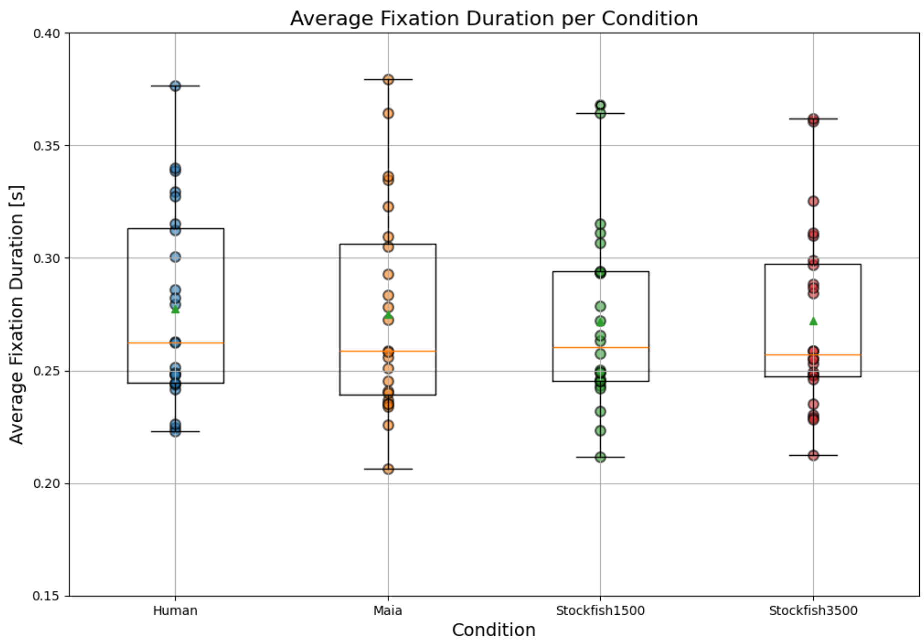


Figure 4. Average fixation duration shown for every condition. Every point within this graph translates to the average fixation duration of one participant during one full chess game.

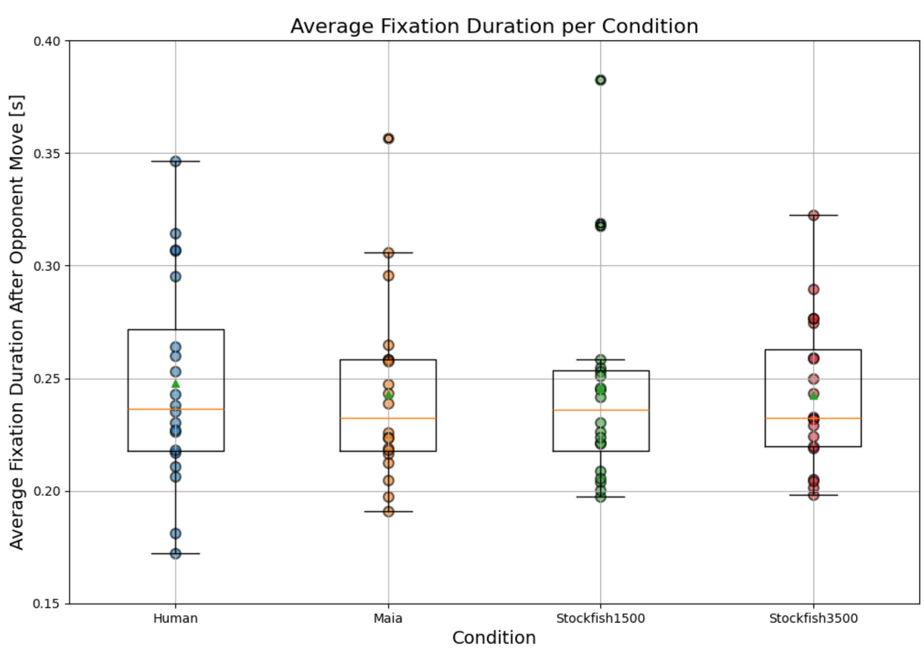


Figure 5. Average fixation duration shown 5000ms after the opponent made a move shown for every condition. Every point within this graph translates to the average fixation duration of one participant.

The average fixation duration was also compared per participant per condition as an extra analysis in Appendix 6 in figures 15, 16, 17 and 18, showing the 4 highest FIDE rated players in red, indicating that for a full game of chess, no pattern in fixation duration can be found between skill of the players.

Pupil Diameter change after move opponent

In figure 6, the average relative change of the pupil diameter is shown for every condition, where 100 pupil diameter values around the peak found within the averaged absolute pupil diameters are compared with the starting absolute pupil diameter per participant. The mean and standard deviation are displayed in Table 22 & paired t-test results in Table 23.

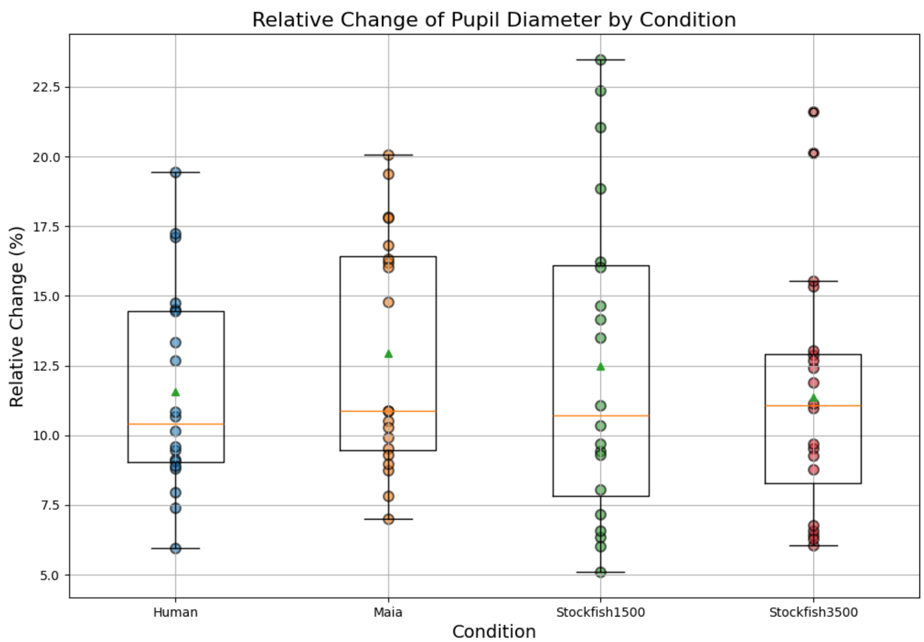


Figure 6. Average relative change of the pupil diameter shown for every condition. Every point within this graph translates to one participant, where 100 absolute pupil diameter values before and after the maximum value are averaged and compared with the starting value of the pupil diameter.

Condition	Mean	SD
Human	11.58%	3.59%
Maia	12.96%	4.12%
Stockfish1500	12.47%	5.56%
Stockfish3500	11.36%	4.26%

Table 22. Relative pupil diameter increase values for each condition.

Comparison	T-statistic	p-value
Human vs. Maia	-2.01	0.0585
Human vs. Stockfish1500	-1.05	0.3086
Human vs. Stockfish3500	0.22	0.8287
Maia vs. Stockfish1500	0.48	0.6333
Maia vs. Stockfish3500	1.63	0.1190
Stockfish1500 vs. Stockfish3500	1.00	0.3295

Table 23. Paired t-test results for relative pupil diameter increase values, comparing each condition.

Results of the paired t-test do not indicate significant difference when comparing conditions. The results do indicate that, on average, the pupil dilated most when playing against Maia, followed by Stockfish1500, then

Stockfish3500 and lastly the Human condition. It is also notable that the average of each participant shows dilation of the pupil diameter, with none showing miosis.

In order to show the full picture of the pupil diameter change, all pupil diameter values 2000 ms after the opponent made a move are captured and and are plotted continuously. For each participant, the absolute values of the averages of the pupil diameter have been plotted in figures 11a, 12a, 13a and 14a for the conditions Human, Maia, Stockfish1500 and Stockfish3500 respectively, shown in Appendix 6.

To show the relative change of the pupil diameter, each value has been divided by the starting value of the average to show relative change over time and are shown in figures 11b, 12b, 13b and 14b for the conditions Human, Maia, Stockfish1500 and Stockfish3500 respectively, shown in Appendix 6.

Per condition, the absolute averages of all pupil diameters are shown in Figure 7 and the relative averages of all pupil diameters value are shown in Figure 8.

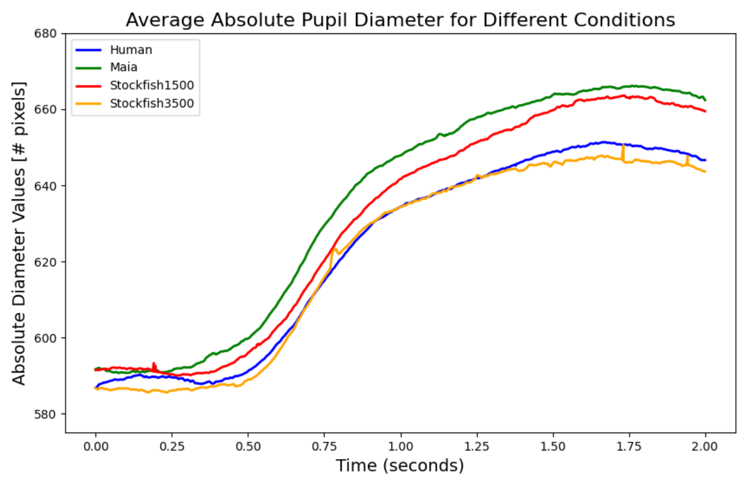


Figure 7. Continuous absolute change of the pupil diameter shown for every condition. Each plotted line displays the average absolute pupil diameter change.

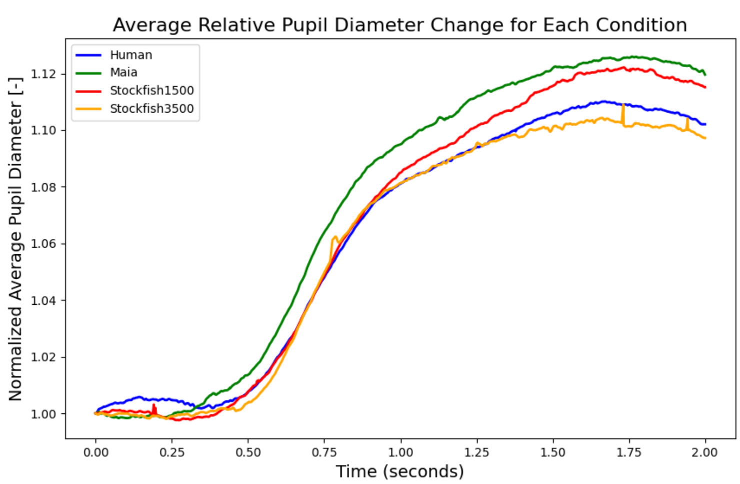


Figure 8. Continuous relative change of the pupil diameter shown for every condition. Each plotted line displays the average relative pupil diameter change.

Discussion

This study aimed to explore the perceived humanness of different chess algorithms and comparing them with a human chess player in a Turing test experiment setting. In a within-subject experiment with 24 chess players of different skill levels, 192 chess games were played and participants assessed how human-like chess algorithms were perceived. Through a questionnaire, indications of how human-like the algorithms played were obtained and whether participants could accurately estimate the nature of their opponent. The results indicate that chess engine Maia was perceived similar as the human opponent, with questionnaire data showing similarities in perceived human-likeness in both the moves played and estimated opponent. Additionally, Maia was perceived as more human-like compared to both versions of the Stockfish engine, while the perceptions of human-likeness of the Stockfish conditions were statistically similar according to the conducted paired t-test.

A potential explanation for the differences in perceived humanness can be found when analysing the outcomes of the chess games. Table 1 shows that participants achieved more victories against Maia and the human condition compared with the Stockfish conditions. Additionally, Table 3 shows the amount of time in seconds elapsed before making a move was higher compared to playing against the Stockfish conditions, suggesting the need for participants to think longer. This longer time was potentially used to either determine whether the move made by the opponent seemed human-like or not, or possibly by needing more time to figure out how to respond on the chessboard. The latter would indicate stronger play from the Stockfish conditions. The perceived strength of the algorithms displayed in 7 indicates that the Stockfish conditions are perceived stronger compared to the Human and Maia condition, especially Stockfish3500, potentially leading the participants to discover the nature of the opponent by estimating its strength.

When reflecting on the existing literature, Young et al.[1] tested nine Maia algorithms and several other chess engines on nine training dataset containing 120.000 games played by human players at a certain skill level to note that Maia chooses the human-made move more often compared to other engines. When Maia has been trained upon a certain chess rating, its peak performance of move-matching caps out at around 51%-53% at a specific skill level, displayed in figure 9. When comparing this with the maximum of Stockfish and Leela, 41% and 43% respectively, this is an improvement in human-like behaviour on the chessboard, but does not indicate how closely the play of Maia represents that of a human of similar skill. It can also be noted that Maia peaks at a certain skill level, similar of that how a human chess player of a certain skill level would, indicating a more human-like output compared to Stockfish and Leela.

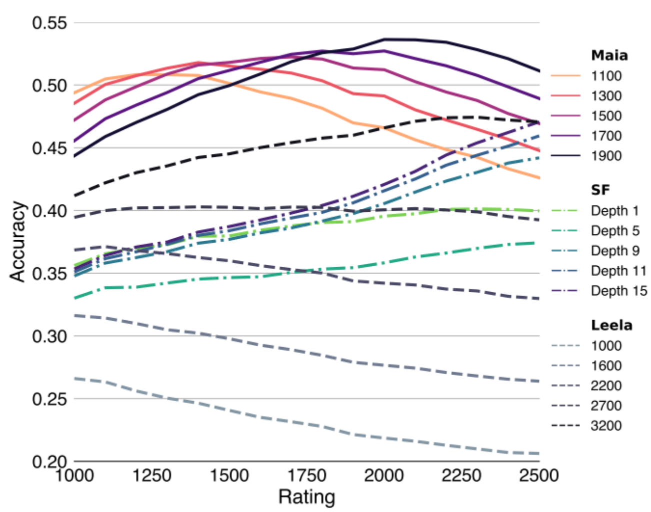


Figure 9. Comparison of move-matching performance for Maia, Stockfish, and Leela models[1]

When reflecting the results of this study and comparing them with Young et al.[1], this study demonstrates that Maia is potentially perceived more human-like than previously considered. While the increase in move-matching shown in figure 9 is significant, the within-subject experiment used in this study to statistically compare Maia with a human chess player and other chess engines on perceived humanness obtains results which show similarities between a human chess player and the Maia algorithm.

This study also aimed to explore whether measurements indicative for cognitive processes show difference when distinguishing between interacting with a human or computer in the game of chess. If measurements indicating cognitive processes give a significant difference when participants are distinguishing between interacting with a human or computer, it could potentially signify relevance for future Turing test experiments within the LLM era of robotics. However, results indicate that there are no significant differences observed in fixation duration and changes in pupil diameter when comparing the conditions against which participants played. This implies that fixation duration or pupil diameter change measurements does not lead to a measurable difference of cognitive processes when trying to figure out whether the participant is interacting with a human or computer.

When observing the relative pupil diameter change shown in Figures 7 and 8 and their relative increased values shown in Table 22, the average relative pupil diameter has been modulated by over 10% around its peak, indicating that the pupil diameter does dilate on average. One potential explanation would be that, when the opponent plays a move, new information is visible compared to the usually static chessboard. As the opponent plays a move, a new visual situation and potential problem appears, potentially leading to pupil dilation.[60,61] Another possible explanation could be that since the absolute values of the pupil diameter is extracted in weblink as the number of on screen pixels captured by the Portable Duo. If on average, the chess player moves closer towards the eye tracker after the opponent makes a move, it might result in a bigger relative pupil diameter measurement after the opponent makes a move. Since the luminance was constant throughout each individual experiment, this could not explain the sudden change in pupil diameter.

According to literature, pupil diameter does increases with the difficulty of the problem.[32] However, Table 23 containing statistical paired t-test results does not indicate a significant difference between conditions, suggesting that the change in task difficulty faced by participants is not significantly detectable.

Comparing this with the average peak of the pupil diameter changes within the condition, it can be visible that the conditions Maia and Stockfish3500 obtain on average a slightly bigger pupil dilation. One possible explanation for this slight dilation difference could be due to the fact that these algorithms blunder less compared to the Stockfish1500 or human condition, potentially leading to a game that requires more consistent, higher-level thinking and strategising from the participant.

Another measurement indicative for cognitive processes were the sound recordings while participants spoke their thoughts aloud. Upon examining the voice recordings, statistical significant differences can be found when comparing the Stockfish1500 condition within the category unexpected, illogical, strange or weird events displayed in Tables 10 and 14. This could suggest that the nature of the condition Stockfish1500 is estimated computer-like due to the number of inhuman moves the Stockfish algorithm makes.

It is beyond the scope of this research to conduct research considering the timing of the sentences spoken. The timestamps of when sentences are spoken have been included in the dataset, as are specific sentences within categories found. As temporal aspects can be considered significant within eye tracking data, cross-data analysis could potentially signify patterns in the eye tracking data where the participants notes a certain category in the voice recordings.

Also beyond the scope of this research has been to conduct research considering the gaze of the player, to potentially discover whether patterns can be discovered from the specific coordinates the chess player has looked at or the order of these specific coordinates. Analysing where the chess player have looked can be challenging due to the dynamic environment of the chessboard, containing an ever changing structure. To discover patterns from where the chess players have looked on the chessboard, cross-data analysis raises new challenges as the dynamics of the chessboard should be captured autonomously.

However, for the purpose of answering the research question, the findings of the human-likeness for Maia are nonetheless valid due to statistical significance demonstrated on the human-likeness of the algorithm conducted through subjective questionnaire data.

Conclusion

In this paper, the following research question was examined:

How human-like do chess players perceive algorithms such as Maia compared to traditional engines and human opponents, and can differences in cognitive processes be measured through eye tracking or think-aloud recordings when distinguishing between computer or human?

Changing the opponent in a chess match and letting participants fill in subjective questionnaire to assess whether the moves of the opponent were human-like & estimating if the opponent was a human or computer gave the insight that, in the results of the within-subjects experiment of this study, Maia was perceived similar as the human opponent. Both Stockfish algorithms were also perceived as less human-like compared to Maia. The experiment done in this study could show to be a significant addition for the Maia team, since this study tests their algorithm in a within-subject experiment. A within-subject experiment design typically has a higher statistical power than a between-subject experiment design.[58] A within-subject experiment reduces the variability caused by differences between subjects, as the same participants interacted with all conditions.

However, the change in opponent did not significantly influence measurements of which the goal was to measure a difference in underlying cognitive processes. While this study did analyse the eye tracking data and voice recordings, no statistical notable difference was found between the four conditions considering fixation duration and pupil diameter change.

Future work

Looking ahead, the extensive dataset produced by this study could perhaps yield several more insights, since an open source dataset containing eye tracking data while playing chess is not available.

To better understand the implications of these results, future studies could address the combination of the voice recording and eye tracking data. Cross-data analysis could provide insight in a specific time in the chess game, where mentioned in the voice recording, where a specific is expressed. Overlapping the voice recording with the gaze of the player and pupil diameter could potentially yield more insights by separating the sentences in which a specific emotion such as surprise has been mentioned and looking at the differences in eye tracking compared to non-verbal situations.

Also, due to the abundance of data available and previous studies done on the experimental framework of chess, this substantial dataset containing modern-day measurements could potentially provide insights into studies done on the decision making of chess players.

Acknowledgements

I would like to extend my gratitude for the excellent council provided by my supervisor, Yke Bauke Eisma & to Joost de Winter, who always liked to discuss the topic of this thesis when required, becoming an unofficial second supervisor.

I express my sincere gratitude to the TU Delft, specifically the Human-Robot Interaction department within Cognitive Robotics for their welcoming attitude and providing the essential measurement equipment for this study.

References

1. McIlroy-Young, R., Sen, S., Kleinberg, J. & Anderson, A. Aligning superhuman AI with human behavior. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*, DOI: 10.1145/3394486.3403219 (ACM, 2020).
2. Liu, Y. *et al.* Sora: A review on background, technology, limitations, and opportunities of large vision models (2024). 2402.17177.
3. OpenAI. Sora: Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators.

4. Goodfellow, I. J. *et al.* Generative adversarial networks (2014). 1406.2661.
5. Gatys, L. A., Ecker, A. S. & Bethge, M. A neural algorithm of artistic style (2015). 1508.06576.
6. Melechovsky, J. *et al.* Mustango: Toward controllable text-to-music generation (2023). 2311.08355.
7. OpenAI *et al.* Gpt-4 technical report (2024). 2303.08774.
8. Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with gpt-4 (2023). 2303.12712.
9. Kosinski, M. Evaluating large language models in theory of mind tasks (2024). 2302.02083.
10. OpenAI. New ai classifier for indicating ai-written text.
11. Li, H. Turing test in the era of LLM. In *Submitted to Peking University Course: Cognitive Reasoning* (2024). Under review.
12. Turing, A. M. Computing machinery and intelligence. *Mind* **59**, 433–60, DOI: 10.1093/mind/lix.236.433 (1950).
13. Silver, D. *et al.* A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* **362**, 1140–1144, DOI: 10.1126/science.aar6404 (2018). https://www.science.org/doi/pdf/10.1126/science.aar6404.
14. Ensmenger, N. Is chess the drosophila of artificial intelligence? a social history of an algorithm. *Soc. Stud. Sci.* **42**, 5–30, DOI: 10.1177/0306312711424596 (2012). PMID: 22530382, https://doi.org/10.1177/0306312711424596.
15. Rojas, R. *The First Code for Computer Chess*, 191–201 (Springer Nature Switzerland, Cham, 2023).
16. Turing, A. Digital computers applied to games. In *Faster than thought* (1953).
17. Fast, E. & Horvitz, E. Long-term trends in the public perception of artificial intelligence. *Proc. AAAI Conf. on Artif. Intell.* **31**, DOI: 10.1609/aaai.v31i1.10635 (2017).
18. Campbell, M., Hoane, A. & hsiung Hsu, F. Deep blue. *Artif. Intell.* **134**, 57–83, DOI: https://doi.org/10.1016/S0004-3702(01)00129-1 (2002).
19. Krol, M. Have we witnessed a real-life turing test? *Computer* **32**, 27–30, DOI: 10.1109/2.751325 (1999).
20. Charness, N. The impact of chess research on cognitive science. DOI: https://doi.org/10.1007/BF01359217 (1992).
21. de Groot, A. D. *Thought and Choice in Chess* (Amsterdam University Press).
22. Chase, W. G. & Simon, H. A. Perception in chess. *Cogn. Psychol.* **4**, 55–81, DOI: https://doi.org/10.1016/0010-0285(73)90004-2 (1973).
23. Frey, P. W. An introduction to computer chess (1983).
24. Stockfish. Stockfish 16: strong open source chess engine. https://stockfishchess.org.
25. Ikhwantri, F., Putra, J. W. G., Yamada, H. & Tokunaga, T. Looking deep in the eyes: Investigating interpretation methods for neural models on reading tasks using human eye-movement behaviour. *Inf. Process. Manag.* **60**, 103195, DOI: https://doi.org/10.1016/j.ipm.2022.103195 (2023).
26. Richardson, D. C., Dale, R. & Spivey, M. J. Eye movements in language and cognition. *Hum. Cogn. Process.* 323–344, DOI: 10.1075/hcp.18.21ric (2007).
27. König, P. *et al.* Eye movements as a window to cognitive processes. *J. Eye Mov. Res.* **9**, DOI: 10.16910/jemr.9.5.3 (2016).
28. Just, M. A. & Carpenter, P. A. A theory of reading: From eye fixations to comprehension. *Psychol. Rev.* **87**, 329–354, DOI: 10.1037/0033-295x.87.4.329 (1980).
29. Maria Theobald, E. G.-K. & Brod, G. Predicting vs. guessing: the role of confidence for pupillometric markers of curiosity and surprise. *Cogn. Emot.* **36**, 731–740, DOI: 10.1080/02699931.2022.2029733 (2022). PMID: 35077310, https://doi.org/10.1080/02699931.2022.2029733.

30. Oliva, M. & Anikin, A. Pupil dilation reflects the time course of emotion recognition in human vocalizations. *Sci. Reports* **8**, DOI: 10.1038/s41598-018-23265-x (2018).

31. Kloosterman, N. *et al.* Pupil size tracks perceptual content and surprise. *Eur. J. Neurosci.* **41**, DOI: 10.1111/ejn.12859 (2015).

32. Hess, E. H. & Polt, J. M. Pupil size in relation to mental activity during simple problem-solving. *Science* **143**, 1190–1192, DOI: 10.1126/science.143.3611.1190 (1964). https://www.science.org/doi/pdf/10.1126/science.143.3611.1190.

33. Preuschoff, K., t Hart, B. & Einhäuser, W. Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Front. neuroscience* **5**, 115, DOI: 10.3389/fnins.2011.00115 (2011).

34. Chessgoals. https://chessgoals.com/rating-comparison (2023).

35. Asus. Rog zephyrus s17: Laptops: Rog global. https://rog.asus.com/laptops/rog-zephyrus/rog-zephyrus-s17-series.

36. Philips. Audio video recorders. https://www.dictation.philips.com/nl/producten/audio-video-recorders/voicetracer-audiorecorder-dvt2810.

37. Apple. Iphone 14 and iphone 14 plus - technical specifications. https://www.apple.com/iphone-14/specs/.

38. Olympus. New olympus audio recorder vp-20 captures voice and sound in superb quality. https://www.olympus.nl/company/nl/nieuws/persberichten/2020-02-12t10-00-00/new-olympus-audio-recorder-vp-20-captures-voice-and-sound-in-superb-quality-25094-nl.html.

39. Ltd., S. R. Eyelink portable duo. https://www.sr-research.com/eyelink-portable-duo/ (2023).

40. Ltd., S. R. Eyelink portable duo brochure. https://www.sr-research.com/wp-content/uploads/2021/06/eyelink-portable-duo-brochure.pdf.

41. SR-research. Weblink. https://www.sr-research.com/weblink (2022).

42. LiChess. Lichess zen mode. https://lichess.org/page/zen.

43. LiChess. Lichess bots. https://lichess.org/blog/WvDNticAAMu_mHKP/welcome-lichess-bots.

44. rooklift. Chess gui nibbler. https://github.com/rooklift/nibbler.

45. Hudson, S. Chess gui scidvspc. https://scidvspc.sourceforge.net/.

46. Bot maia1 lichess. https://lichess.org/@/maia1.

47. Duplessis, T. Lichess game download. https://github.com/smartchessguiapp/ligamedl.

48. Leite, R. V. & de Oliveira, A. V. C. Expected human performance behavior in chess using centipawn loss analysis. In Fang, X. (ed.) *HCI in Games*, 243–252, DOI: 10.1007/978-3-031-35979-8_19 (Springer Nature Switzerland, Cham, 2023).

49. McIlroy-Young, R. Maia: A human-like neural network chess engine. https://maiachess.com/.

50. Wikipedia. Top chess engine championship. https://en.wikipedia.org/wiki/Top_Chess_Engine_Championship.

51. CCRL. Computer chess rating lists 40/15 download & statistics. https://computerchess.org.uk/ccrl/4040/index.html.

52. Maharaj, S., Polson, N. & Turk, A. Chess ai: Competing paradigms for machine intelligence. *Entropy* **24**, 550, DOI: 10.3390/e24040550 (2022).

53. Sauro, J. & Dumas, J. S. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, 1599–1608, DOI: 10.1145/1518701.1518946 (Association for Computing Machinery, New York, NY, USA, 2009).

54. Radford, A. *et al.* Robust speech recognition via large-scale weak supervision (2022). 2212.04356.

55. OpenAI. Whisper. https://github.com/openai/whisper.

56. OpenAI. Overview openai api models. https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo.

57. McCrum-Gardner, E. Which is the correct statistical test to use? *Br. J. Oral Maxillofac. Surg.* **46**, 38–41, DOI: https://doi.org/10.1016/j.bjoms.2007.09.002 (2008).

58. de Winter, J. & Dodou, D. *Human subject research for engineers: A practical guide* (2017).

59. scipy.stats.ttestrel. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html.

60. Privitera, C. M., Renninger, L. W., Carney, T., Klein, S. & Aguilar, M. The pupil dilation response to visual detection. In Rogowitz, B. E. & Pappas, T. N. (eds.) *Human Vision and Electronic Imaging XIII*, vol. 6806, 68060T, DOI: 10.1117/12.772844. International Society for Optics and Photonics (SPIE, 2008).

61. Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C. & Krejtz, I. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE* **13**, 1–23, DOI: 10.1371/journal.pone.0203629 (2018).

1 Appendix A: Experiment setup

Order of conditions played

	C1	C2	C3	C4	C5	C6	C7	C8
P1	Maia	SF3500	Human	SF1500	Maia	Human	SF3500	SF1500
P2	SF1500	Human	Maia	SF3500	SF1500	Human	SF3500	Maia
P3	Maia	Human	SF1500	SF3500	SF3500	SF1500	Maia	Human
P4	Human	SF1500	Maia	SF3500	Human	Maia	SF3500	SF1500
P5	SF1500	Maia	SF3500	Human	Maia	SF1500	Human	SF3500
P6	SF3500	Human	Maia	SF1500	SF1500	SF3500	Maia	Human
P7	Human	SF3500	SF1500	Maia	SF3500	Maia	SF1500	Human
P8	SF3500	Maia	Human	SF1500	Maia	SF3500	SF1500	Human
P9	SF1500	SF3500	Human	Maia	Human	SF1500	SF3500	Maia
P10	Human	Maia	SF1500	SF3500	SF1500	Maia	Human	SF3500
P11	SF3500	SF1500	Human	Maia	Maia	SF1500	SF3500	Human
P12	SF3500	Human	SF1500	Maia	Human	SF3500	Maia	SF1500
P13	Maia	Human	SF3500	SF1500	Maia	SF3500	Human	SF1500
P14	SF1500	Human	SF3500	Maia	SF1500	Human	Maia	SF3500
P15	SF3500	SF1500	Maia	Human	Maia	Human	SF1500	SF3500
P16	Human	Maia	SF3500	SF1500	Human	SF1500	Maia	SF3500
P17	Maia	SF1500	Human	SF3500	SF1500	Maia	SF3500	Human
P18	SF1500	SF3500	Maia	Human	SF3500	Human	Maia	SF1500
P19	SF3500	Maia	SF1500	Human	Human	SF3500	SF1500	Maia
P20	Maia	SF3500	SF1500	Human	SF3500	Maia	Human	SF1500
P21	Human	SF1500	SF3500	Maia	SF1500	SF3500	Human	Maia
P22	SF1500	Maia	Human	SF3500	Human	Maia	SF1500	SF3500
P23	Maia	SF1500	SF3500	Human	SF3500	SF1500	Human	Maia
P24	Human	SF3500	Maia	SF1500	SF3500	Human	SF1500	Maia

Table 24. Randomized conditions for all 24 participants. Each row indicates a participant and each column indicates which part of the experiment corresponds with with condition the participant played against. SF is an abbreviation for Stockfish.

Positions played from

All of the played positions can be found in the first three studies:

`TuringTest001's Study - https://lichess.org/study/4NIrHjdr`
`TuringTest001's Study_2 - https://lichess.org/study/6DXv6dP6`
`TuringTest001's Study_3- https://lichess.org/study/AawyoalP`
`TuringTest001's Study_4- https://lichess.org/study/eSSgWmvg/zQdMibnu`

A fourth one was created to potentially add 6 more participants.

Figure 10. An example of a typical starting position (Study 1 Chapter 61). Stockfish analysis shows that the the five best moves to be played for white (a4, f4, Bf4, Qd3 and Be3) result from a +0.2 to -0.1 chess position, resulting in an equal start.

2 Appendix B: Pre-Experiment Questionnaire Data

Questionnaire before experiment

The 24 participants which participated in the experiment were asked to fill in a pre-experiment questionnaire. The questionnaire contains the following questions:

- **Participant number**
- **Full name**
- **Age**: *Mean = 26.3, SD = 9.3*
- **Gender**: *23 Male, 1 Female*
- **How often do you play chess (1 Few times a year - 7 Daily)**: *Mean = 5.5, SD = 1.2*
- **How often do you play chess against a human opponent? (1 Few times a year - 7 Daily)** *Mean = 5.5, SD = 1.4*
- **How often do you play chess against an engine? (1 Few times a year - 7 Daily)** *Mean = 1.5, SD = 0.7*
- **Do you have a FIDE or online chess rating?**
- **If yes, what is your FIDE rating?** *Mean = 1910, SD = 279*
- **If yes, what is your online rating? Please include on which website & time format.** LiChess Blitz rating: *Mean = 1645, SD = 348*
- **"Do you think humans and engines play different within the game of chess?"** *23 Yes / 1 No*
- **"How would you try to recognise the difference between an engine and a human opponent?"**

All individual answers are displayed in table 25.

To make the participant think about how to approach the task before conducting the experiment, one more question was asked: "*How would you try to recognise the difference between an engine and a human opponent?*" This question was excluded from table 25 due to fitting issues.

The answers to this question per participant were as follows:

1. Looking for unusual mistakes an engine would not make
2. giving free pieces which taking them is not necessarily the shortest path to victory
3. Silent moves, nothing to do with the position an engine will wait for human mistake
4. If I see a move very different from my expectations I usually think is a engine or someone much better than me.
5. Certain ideas and strategies are very tough to find for humans but are findable through hard calculation of computers. These ideas include (multiple) backwards moves as well as (multiple) sacrifices for long term compensation.
6. Sharpness and continuity
7. They have better tactical skills. They can see tactical tricks very fast compared to humans. Also better theory since they usually have an opening database. I would check for its tactical tricks.
8. Engines might tend to make weird moves that doesn't make sense as a human being
9. When I play chess I also think for my oponent, this means that I try to predict his or her moves. When moves deviate from these predictions this would increase the probability of me saying it is a computer.
10. Computers will use their time differently from humans. Computers (especially the good ones) can very easily enter into complex tactical positions, whereas humans will often avoid such positions or make mistakes. Computers break general principles easier than most humans.
11. looking for 'defending' moves that don't seem logical but an analysis will show they're slightly better in some weird way
12. Within openings, humans play a standard way of chess. Good engines replicate those openings and bad engines deviate. In mid game humans sometimes miss a good move but a good engine might not miss a good move. At the same time a bad engine might also miss a good move so it is hard to differentiate.
13. Human opponents (at my level) generally play more according well known patterns.
14. Making of blunders, playing with a clear strategy (human)
15. Understanding if the thought process of the player I'm playing against resembles mine.
16. Engines often do not make 'dumb' mistakes and the game often lasts longer.
17. Time usage, odd tactical behaviour
18. a smart engine wins slowly and improves its position slowly. analyzing such a game would give a cruve where the human player slowly loses. Also an my experience with playing against an engine is that dubious unknown strategies do not work.
19. engines will play more theoretical moves and humans can possibly fall for traps or discovered attacks.
20. engines can react instantly to the moves I make, while humans need to think about that for a certain period of time.
21. no idea
22. Distribution of thinking time between moves, whether the moves are "natural", playing strength
23. Playing style: Creation of weaknesses and strategy
24. Humans play more intuitive and coherent moves, engines just play the best.

The original questionnaire data is also available within the .csv file.

	Age	Male	How often do you play chess? 1 - 7	How often do you play chess against a human opponent? 1-7	How often do you play chess against an engine? 1-7	Fide rating	Online rating	Do you think humans & engines play different within the game of chess?
P1	25	Male	5	5	1	-	1250 Rapid LiChess	Yes
P2	26	Male	5	5	2	-	1000 Rapid Chess.com 750 Blitz Chess.com	Yes
P3	24	Male	5	5	2	-	1400 Bullet Chess.com	Yes
P4	24	Male	7	7	1	-	1400 Blitz Chess.com	Yes
P5	25	Male	6	6	2	1600 KNSB	1200 Rapid Chess.com	Yes
P6	25	Male	3	3	1	-	1600 Blitz Lichess	Yes
P7	32	Male	6	6	1	-	1850 Rapid Chess.com 1400 Blitz Chess.com	Yes
P8	24	Male	7	7	2	-	1250 Rapid Chess.com	Yes
P9	27	Male	5	5	2	-	1300 Rapid LiChess	Yes
P10	25	Male	6	6	1	-	1750 Rapid LiChess	Yes
P11	23	Male	4	4	3	-	1000 Bullet Chess.com 1500 Bullet LiChess	Yes
P12	26	Male	4	5	2	-	1000 Rapid Chess.com	Yes
P13	25	Male	7	6	1	1850 FIDE	1900 Blitz Chess.com	Yes
P14	22	Female	3	6	1	1650 FIDE	1500 Blitz Chess.com 1600 Rapid Chess.com	No
P15	24	Male	5	5	2	-	1450 Blitz Chess.com	Yes
P16	26	Male	5	5	1	-	1050 Blitz Chess.com	Yes
P17	50	Male	7	7	1	1800 FIDE	2100 Blitz Lichess 2200 Rapid Lichess	Yes
P18	24	Male	7	7	3	-	1700 Blitz Chess.com	Yes
P19	21	Male	7	7	3	-	950 Bullet Chess.com 1100 Blitz Chess.com 1350 Rapid Chess.com	Yes
P20	22	Male	5	7	1	-	900 Rapid Chess.com	Yes
P21	59	Male	5	1	1	-	-	Yes
P22	24	Male	6	6	1	-	1800 Rapid LiChess	Yes
P23	14	Male	5	5	1	2150 FIDE	2150 LiChess Blitz	Yes
P24	15	Male	7	7	1	2400 FIDE	2800 Chess.com Blitz	Yes

Table 25. Questionnaire before the event. All names are excluded & all ratings are rounded to the closest 50 or 100 for privacy reasons.

The original questionnaire data is also available within the .csv file.

3 Appendix C: Instructions given to the participants

3.1 Weblink Instructions

1. **Welcome** Today, you will be playing chess from 8 pre-selected positions. Please click once to continue to the next instructions.
2. **About the chess game** You will always play with white. It's move 10 in the game and the position is rated equal. You and your opponent will both get 5 minutes on the clock.
3. **About the chess game** To exit the chess game, press "q". Please do not press "q" while playing.
4. **Recognise: Human or Engine** Your opponent will be either a human or an engine. Your main job will be to recognize if your opponent is human or an engine.
5. **Recognise: Human or Engine** To make recognition more difficult, your opponent will move every ±10 seconds. The order of human & engine opponents will be random. You should not expect same proportion of human & engine opponents, as they might vary.
6. **Speech & Eyes** During your games, we will monitor both your speech and eyes. It will be expected that you talk about your game.
7. **Speech** Please only speak in English. The experimenter will enable the voice recorder.
8. **Speech** Some talking points:
   - How are you trying to figure out whether the opponent is an engine or human?
   - Which moves are expected from the opponent?
   - Which moves are you considering?
9. **Eyes** Please sit still while playing calibrating and playing chess, such that we can capture your eye's movement. To make sure we capture your eye-tracking data correctly, we will now calibrate the eye-tracker
10. **Reminder, please talk about:**
    - How are you trying to figure out whether the opponent is an engine or human?
    - Which moves are expected from the opponent?
    - Which moves are you considering?
11. If everything is clear, please continue. If not, please ask the experimenter.

    After your next click, you will be send through to lichess, where you will be invited to your first game.

    First, analyse the position in about one minute. When you have finished, please start playing.
12. After your next click, you will be send through to lichess.
13. **After position 1** Thank you for playing the first position. By clicking, you will enter a google forms. Please fill in the form.
14. **After form 1** With the next click, you will go to lichess for the second position
15. **After position 2** Thank you for playing the second position. By clicking, you will enter a google forms. Please fill in the form.
16. **After form 2** With the next click, you will go to lichess for the third position
17. **After position 3** Thank you for playing the third position. By clicking, you will enter a google forms. Please fill in the form.
18. **After form 3** With the next click, you will go to lichess for the fourth position
19. **After position 4** Thank you for playing the forth position. By clicking, you will enter a google forms. Please fill in the form.
20. **After form 4** With the next click, the experiment has concluded.

The participant is linked through to LiChess.org after instruction 12, 14, 16 and 18, where an invitation is given. The participant is linked through to a google forms questionnaire after instruction 13, 15, 17 and 19.

3.2 Vocal Instructions

Vocal instructions given to the participants:

1. The participants plays with the white chess pieces
2. The chess game does not begin at the starting position, starts at an roughly equal position after 9 moves from both white and black.
3. First, get used to the position. Please think aloud about your next move(s) and perhaps also a plan of approach on how to discover the nature of your opponent.
4. Your job will not be to try to win the game of chess, but to recognise if the opponent is human or engine. After each game, you can fill in a questionnaire to estimate the nature of your opponent.
5. During the game, it would be appreciated if you could think aloud to the best of your capabilities.
6. Main examples of what to talk about are for example the move you are considering and why & reflecting on the move of the opponent, specifically if the move you expected was played or not or gives you a clue about whether you are playing against an engine or human opponent.
7. Please only speak aloud in English while the voice recorder is enabled.
8. The researcher will enable the voice recorder, to make sure the timing is correct.
9. Please sit as still as possible during the experiment.
10. The researcher is able to monitor if your eyes are still tracked. If you move much, the researcher will whisper you to move. Please only react to this with body language, better not to say sorry as the voice recorder will pick up your voice.

4 Appendix D: Questionnaire after each game

Questions asked after each game:

- How human-like were the moves of your opponent? *[1 Computer - 7 Human]*
- Do you believe your opponent was a human or an engine? *[Human / Engine]*
- How would you rate your confidence in identifying the nature of your opponent? *[1 Not at all - 7 Very confident]*
- How strong did you feel like your opponent played? *[1 Weak - 7 Strong]*
- Was there one or more moves of your opponent that made you realise the nature of your opponent? *[Yes - No]*
- If yes, when? What made it recognisable? *[Open Answer]*
- Participant Number
- Condition Number

Table 26. Subjective answers to ’How human-like were the moves of your opponent?’

Human	3 3	6 6	7 6	2 7	2 2	5 6	7 5	7 4	6 7	6 7	2 5	7 7	6 6	7 4	6 7	5 5	3 6	4 7	6 7	1 7	3 6	5 6	6 3	6 4
Maia	2 5	6 6	2 4	4 6	4 5	6 5	6 3	3 5	6 7	5 4	6 7	6 4	7 7	3 6	2 6	6 1	3 7	5 6	7 6	7 7	5 6	4 5	2 6	6 2
Stockfish1500	4 1	3 1	7 5	6 2	6 2	5 2	5 5	1 3	3 4	2 7	1 2	3 5	7 5	3 3	4 2	2 1	6 5	5 4	4 3	2 4	5 2	1 6	2 3	3 5
Stockfish3500	2 2	1 2	5 5	7 3	7 3	5 2	6 4	4 3	2 1	2 2	2 3	7 7	2 5	1 2	6 2	2 7	6 4	2 1	1 6	1 3	3 2	1 1	3 3	3 1

The original questionnaire data is also available within the .csv file.

5 Appendix E: Design choices

- **The duration of the experiment / number of moves in a game**: After pilots, it was decided that a full game of chess was not applicable for this experiment due to time problems.
- **Number of moves in a game**: In another pilot, the participant was only allowed to play 10 moves and then directed to a questionnaire. With the data of this pilot, it was concluded that 10 moves was probably not enough information to represent an accurate representation of the opponent.
- **The repetition of conditions**: To increase the data produced within this study, the conditions were repeated. Having double the participants would also have been an option, however, finding 24 participants who could play chess was already challenging. To reduce the number of participants needed to join, the conditions were repeated once with a break in between.
- **Starting from a certain position**: Starting from a certain position meant that chess players were pulled out of comfort zone into chess situations which could be unknown to them. Playing from a relatively unknown position results in less playing from the top of their head and increases the possibility that the chess player will need to explore the position first, possibly resulting in more eye-tracking data.

6 Appendix F: Eye Tracking data

Extracted Eye Tracking measures from Weblink

A list containing all extracted values from each eye-tracking experiment, in this order:

- RECORDING_SESSION_LABEL
- TRIAL_INDEX
- TRIAL_START_TIME
- TIMESTAMP
- AVERAGE_GAZE_X
- LEFT_GAZE_X
- RIGHT_GAZE_X
- AVERAGE_GAZE_Y
- LEFT_GAZE_Y
- RIGHT_GAZE_Y
- AVERAGE_VELOCITY_X
- LEFT_VELOCITY_X
- RIGHT_VELOCITY_X
- AVERAGE_VELOCITY_Y
- LEFT_VELOCITY_Y
- RIGHT_VELOCITY_Y
- AVERAGE_ACCELERATION_X
- LEFT_ACCELERATION_X
- RIGHT_ACCELERATION_X
- AVERAGE_ACCELERATION_Y
- LEFT_ACCELERATION_Y
- RIGHT_ACCELERATION_Y
- AVERAGE_PUPIL_SIZE
- LEFT_PUPIL_SIZE
- RIGHT_PUPIL_SIZE
- AVERAGE_IN_SACCADE
- LEFT_IN_SACCADE
- RIGHT_IN_SACCADE
- AVERAGE_IN_BLINK
- RIGHT_IN_BLINK
- LEFT_IN_BLINK
- WEBPAGE_SEQUENCE

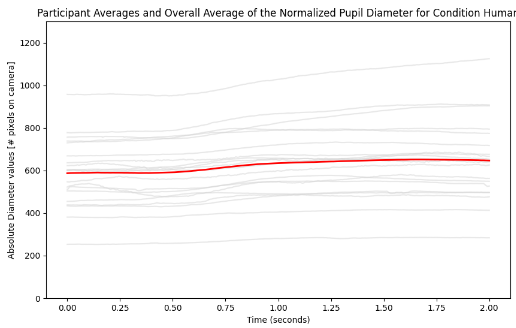
Filtered Eye Tracking data: Overview of all files

- **PXX_EyeTracking_Excel.xls**: All webpage eyetracking parameters shown & all data included
- **PXX_EyeTracking_Filtered.xls**: EyeTracking when a chessboard is shown during the experiment, excluding data like instructions, questionnaire time and waiting for the game to start within LiChess.org

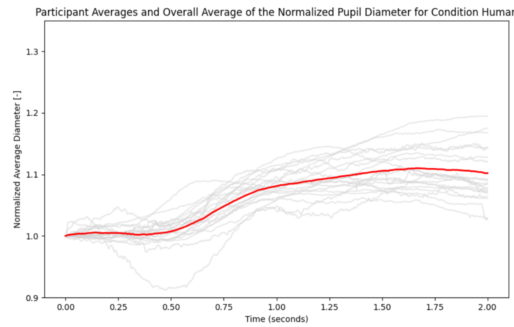
- **PXX_EyeTracking_Filtered_Gaze**$_{n}ew.xls : Excludesallblinks, specificallyperiodsduringwhichhorizontalgazedataonthescree$
  Where XX is the participant number.

Pupil diameter plots, showing average and relative pupil diameter plots per condition

.



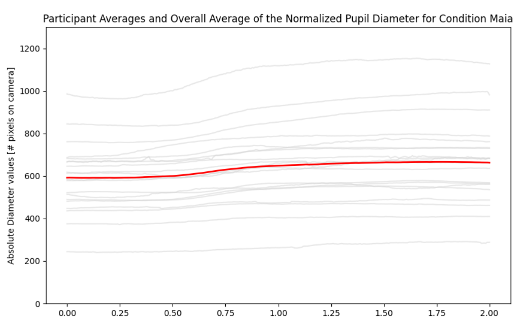
(a) Absolute values of the averages of the pupil diameter.



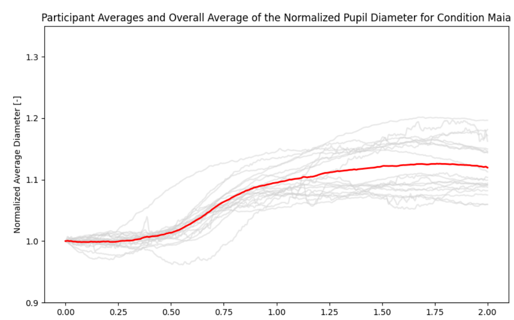
(b) All absolute values divided by their starting value, showing relative change of the averages of the pupil diameter.

Figure 11. Visual representation showing absolute change of the averages of the pupil diameter for 2000 ms after each move for the Human condition. The averages per participant are represented by the grey lines & the average of these averages is shown in red. Measurements containing blinks have been filtered out.

.



(a) Absolute values of the averages of the pupil diameter.

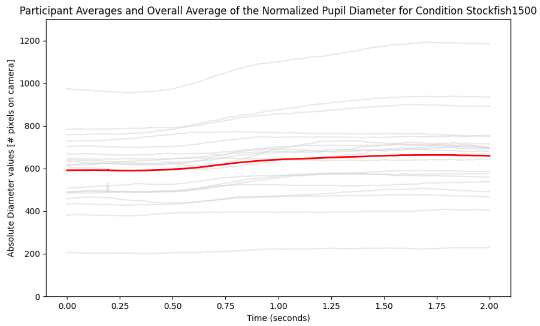


(b) All absolute values divided by their starting value, showing relative change of the averages of the pupil diameter.

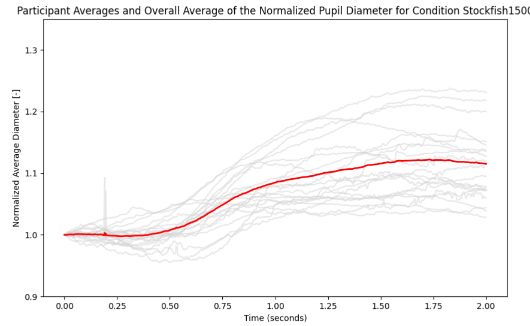
Figure 12. Visual representation showing absolute change of the averages of the pupil diameter for 2000 ms after each move for condition Maia. The averages per participant are represented by the grey lines & the average of these averages is shown in red. Measurements containing blinks have been filtered out.

.

.

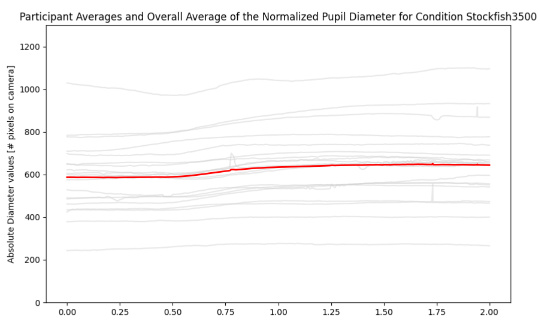


(a) Absolute values of the averages of the pupil diameter.

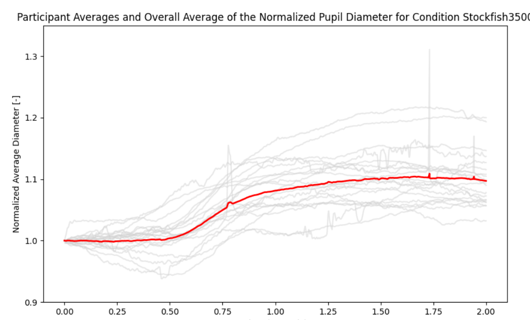


(b) All absolute values divided by their starting value, showing relative change of the averages of the pupil diameter.

Figure 13. Visual representation showing absolute change of the averages of the pupil diameter for 2000 ms after each move for condition Stockfish1500. The averages per participant are represented by the grey lines & the average of these averages is shown in red. Measurements containing blinks have been filtered out.



(a) Absolute values of the averages of the pupil diameter.



(b) All absolute values divided by their starting value, showing relative change of the averages of the pupil diameter.

Figure 14. Visual representation showing absolute change of the averages of the pupil diameter for 2000 ms after each move for condition Stockfish3500. The averages per participant are represented by the grey lines & the average of these averages is shown in red. Measurements containing blinks have been filtered out.

Extra results eye tracking data

To obtain a first estimation of the data, one value for each participant for the pupil data has been obtained within table 27, 28 for when the opponent made a move, 1500ms after the opponent made a move and the change in these two values. This was not included in the report due to no averages being taken, this data being quite sensitive to outliers.

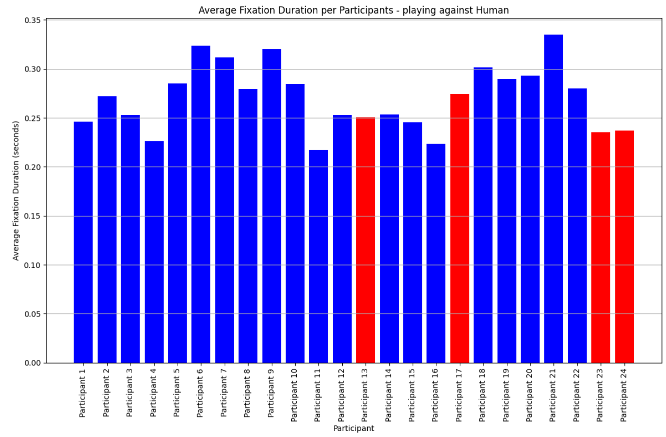


Figure 15. Showing fixation duration per participant when playing against the human condition, strong FIDE rated players shown in red.

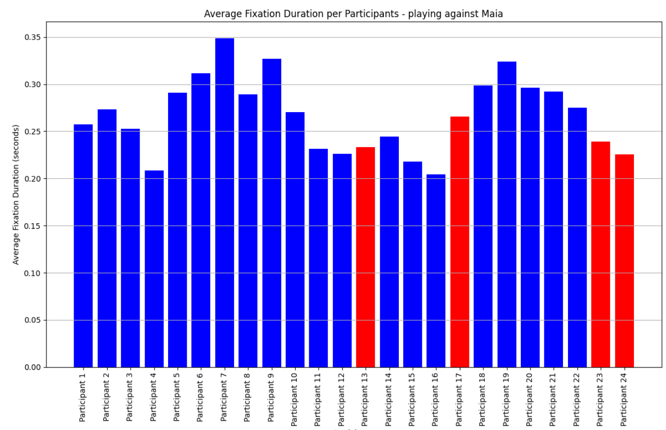


Figure 16. Showing fixation duration per participant when playing against Maia, strong FIDE rated players shown in red.

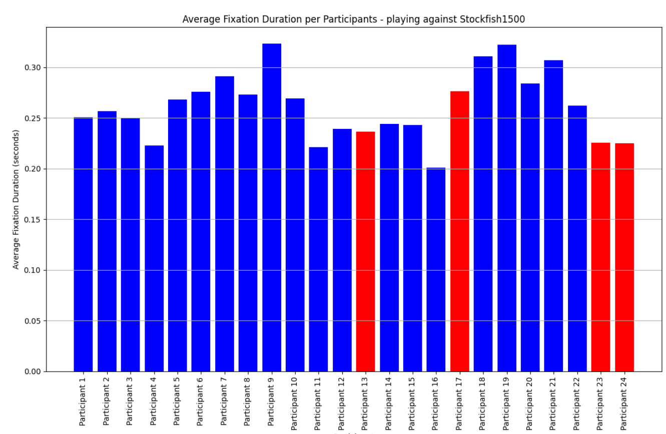


Figure 17. Showing fixation duration per participant when playing against Stockfish1500, strong FIDE rated players shown in red.

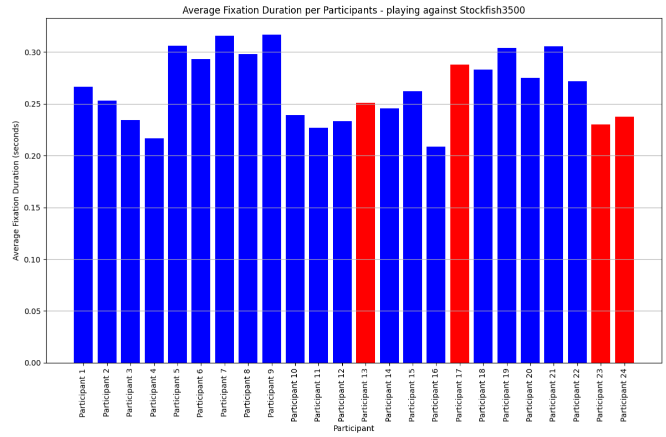


Figure 18. Showing fixation duration per participant when playing against Stockfish1500, strong FIDE rated players shown in red.

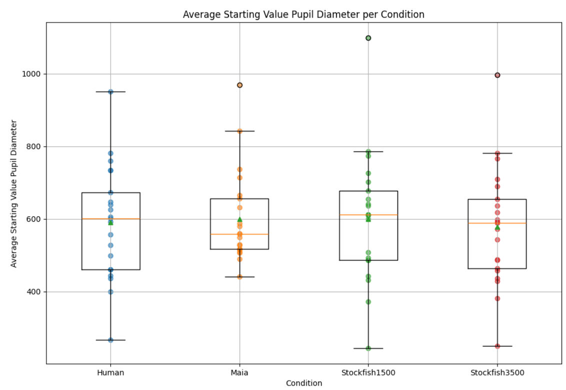


Figure 19. Average pupil diameter at timestamps where the opponent moves. One data point is equal to one participant, shown per condition.

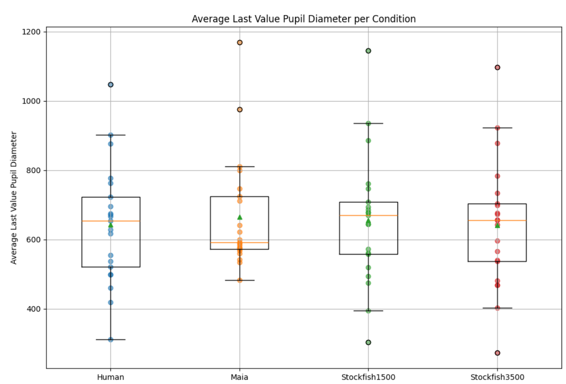


Figure 20. Average pupil diameter 1500 ms (fixed) after timestamps where the opponent moves. One data point is equal to one participant, shown per condition.

Condition	Mean	SD
Human	590.090	151.808
Maia	599.956	124.886
Stockfish1500	599.582	173.333
Stockfish3500	577.954	158.593

Table 27. Average Starting Value Pupil Diameter per participant

Condition	Mean	SD
Human	642.651	168.996
Maia	665.465	160.028
Stockfish1500	654.841	181.631
Stockfish3500	640.803	181.207

Table 28. Average Last Value Pupil Diameter per participant

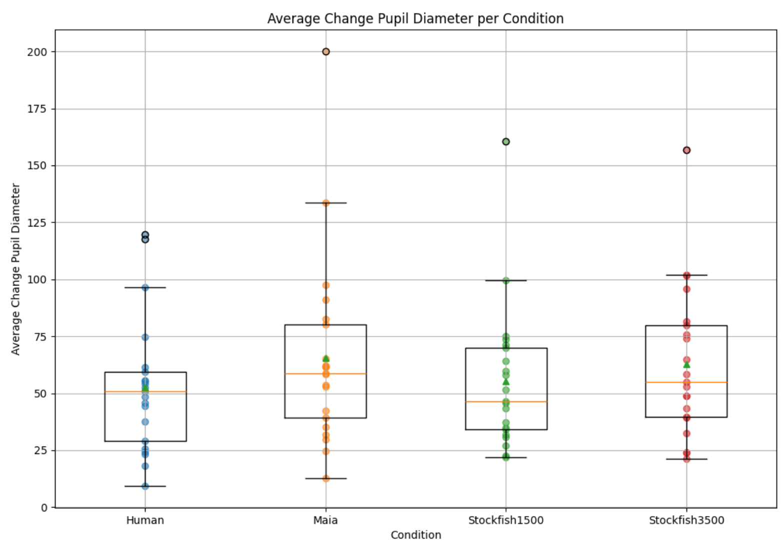


Figure 21. Average absolute change within pupil diameter when the opponent makes a move. Here, pupil diameter is obtained for each move played & the absolute change in pupil diameter after 1500ms (fixed) per participant is shown per condition.

.

Condition	Mean	SD
Human	52.561	29.016
Maia	65.509	40.470
Stockfish1500	55.259	30.847
Stockfish3500	62.849	32.047

Table 29. Average Absolute Change in Pupil Diameter per Condition

.

7 Appendix G: Voice data

Total number of spoken sentences recognised by Whisper can be found in Table 30.

Table 30. Total Spoken Sentences per condition

Speaker	Total Spoken Sentences
Human	3688
Maia	3104
Stockfish1500	3582
Stockfish3500	3335

7.1 System Instructions and Prompts used to find specific sentences within ChatGPT-4-turbo

Category sentences containing a comment about mentioning the move of the opponent is either a computer or a human:

- **System Instruction**: "You are an assistant that extracts valuable information from all of the sentences spoken from a think aloud experiment and puts sentences in one of two categories."
  "If the sentence specifically mentions the participant thinks the opponent is a 'computer', 'engine', 'unhuman', 'machine' or 'pc', place the sentence into the first category. Exclude sentences which talk about plans to discover the opponent being an engine, only include sentences reflecting on the previous engine moves."
  "If the sentence specifically mentions the participant thinks the opponent is a 'human' or 'person', place the sentence into the second category. Exclude sentences which talk about plans to discover the opponent being a human, only include sentences reflecting on the previous human moves."
  "If the sentence does not contain anything about the move of the opponent, discard the sentence and do not output the sentence."
  "Output as a json with the keys 'Sentences containing engine moves' and 'Sentences containing human moves'."

- **Prompt**: "Analyse all sentences once from this think aloud experiment. Output the two categories."
  Sentences said during the think aloud experiment: **content_list**

Category sentences containing a comment about mentioning the move of the opponent is either a computer or a human:

- **System Instruction**: "You are an assistant that extracts valuable information from all of the sentences spoken from a think aloud experiment and puts sentences in one of two categories."
  "If the sentence specifically mentions a move is 'unexpected', 'not logical', 'surprising', 'weird' or 'strange', place the sentence into the first category. Exclude 'dumb' or 'not smart' moves. Exclude sentences which talk about the next moves, only include sentences on the previous moves. Include comments like 'okay, that is what he does'."
  "If the sentence specifically mentions a move is 'expected', 'logical' or 'makes sense', place the sentence into the second category. Exclude sentences which talk about the next moves, only include sentences on the previous moves. Include comments like 'okay, that is not what he does'."
  "If the sentence does not contain anything about the move of the opponent, discard the sentence and do not output the sentence."
  "Output as a json with the keys 'Sentences containing unexpected or not logical moves' and 'Sentences containing expected or logical moves'."

- **Prompt**: "Analyse all sentences once from this think aloud experiment. Output the two categories. "
"Analyse all sentences once from this think aloud experiment. Output the two categories."
Sentences said during the think aloud experiment: **content_list**

7.2 System Instructions and Prompts used to find new catevories within ChatGPT-4-turbo, while all sentences of one category have been used as input

- **System Instruction**: "You are an assistant that extracts valuable information from all of the sentences spoken from a think aloud experiment and tries to find differences between the conditions of the experiment."
"List the differences between the sentences spoken from the conditions. "

- **Prompt**: "Analyse all sentences once from this think aloud experiment. What differences between the conditions or between the spoken data looking at the conditions you can find:"
f"Sentences said during the think aloud experiment from the human condition:contentlist1"
f"Sentences said during the think aloud experiment from the maia condition:contentlist2"
f"Sentences said during the think aloud experiment from the stockfish1500 condition:contentlist3"
f"Sentences said during the think aloud experiment from the stockfish3500 condition:contentlist4"

When changing the content_list different for each condition, the answer from ChatGPT remained the exact same for each condition.

The answer obtained, exactly the same for each condition:

Based on the analysis of the sentences spoken during the think aloud experiment, the following categories can be identified:

- **Strategic Planning and Position Analysis**: This category includes sentences where the player is analyzing the position on the board, planning their next moves, considering the strengths and weaknesses of both sides, and strategizing their approach to the game. This is the most frequently occurring category, as it encompasses the core of the think aloud process during a chess game.

- **Piece Movement and Attack Plans**: Sentences in this category involve specific plans for moving pieces, attacking the opponent's pieces, or defending their own pieces. This includes considerations of specific squares to move to, pieces to capture, and threats to create or respond to.

- **Evaluation of Opponent's Moves and Intentions**: This category includes sentences where the player is trying to predict or evaluate the opponent's moves and strategies. It involves trying to understand the opponent's plan, anticipating future moves, and adjusting their own strategy accordingly.

- **Reflections on Mistakes and Missed Opportunities**: Sentences that fall into this category involve the player reflecting on their own mistakes or missed opportunities. This could include realizing a better move after the fact, recognizing a blunder, or considering how a different strategy might have led to a better outcome.

- **General Observations and Comments**: This category includes general observations about the game, comments on the position that don't necessarily lead to a specific plan or strategy, and other miscellaneous thoughts that the player has while playing.

From the analysis, the most frequently occurring category is "Strategic Planning and Position Analysis," followed by "Piece Movement and Attack Plans," "Evaluation of Opponent's Moves and Intentions," "Reflections on Mistakes and Missed Opportunities," and finally "General Observations and Comments."

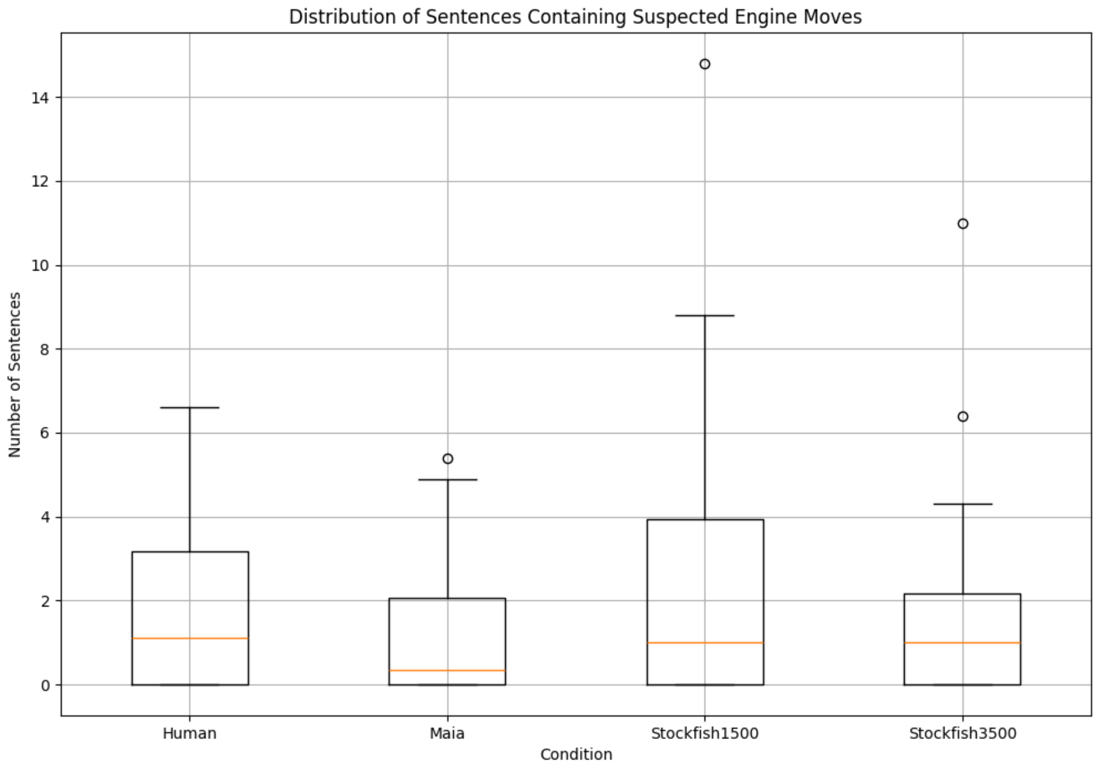


Figure 22. Visual representation of the number of sentences containing a comment about an expected computer opponent. Spoken data transcribed to text with Whisper & data analysed with GPT-4. This shows the average of 10 GPT-4 analysis.

7.3 Boxplots Results Voice Data

Figures 22, 23, 24 and 25 visually display the same results given in Tables 11, 12, 10 and 9 respectively.

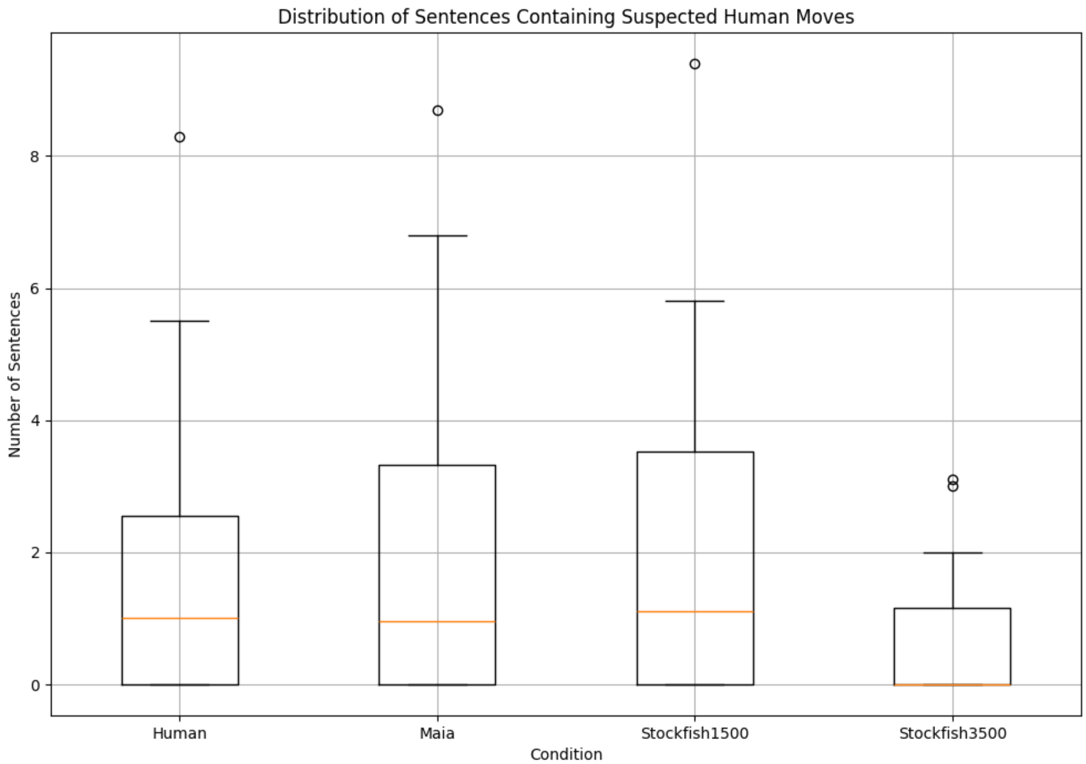


Figure 23. Visual representation of the number of sentences containing a comment about an expected human opponent. Spoken data transcribed to text with Whisper & data analysed with GPT-4. This shows the average of 10 GPT-4 analysis.

.

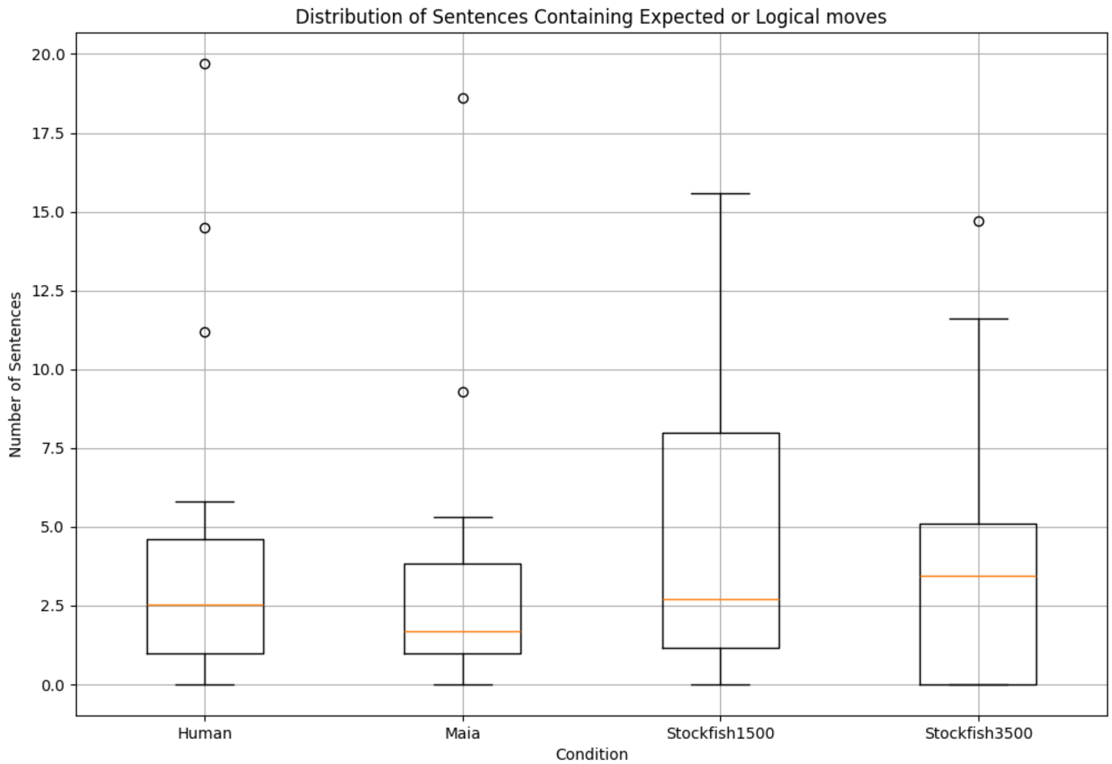


Figure 24. Visual representation of the number of sentences containing a comment about an expected or logical move of the opponent. Spoken data transcribed to text with Whisper & data analysed with GPT-4. This shows the average of 10 GPT-4 analysis.

.

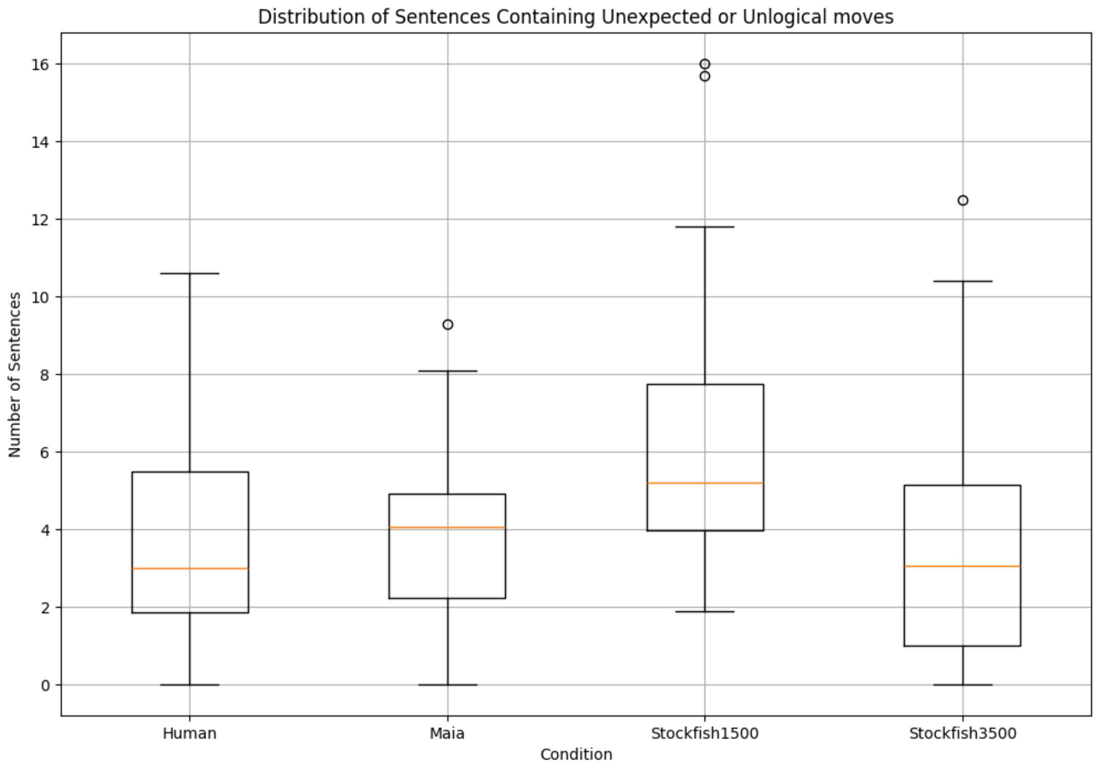


Figure 25. Visual representation of the number of sentences containing a comment about an unexpected or illogical move of the opponent. Spoken data transcribed to text with Whisper & data analysed with GPT-4. This shows the average of 10 GPT-4 analysis.

.

8 Appendix H: Examples hand edited text file

To serve as a benchmark, ten files have been read and edited by hand. Visible in figure 26 is one example of the ten.

Can I begin? Yes, go. Okay, let's start. Well, first of all, very much bounce to the center. I see my pieces are nicely developed. But it's somewhat of a position I'm not really sure what my next move will be from the start. So, it's fairly complicated. What comes to mind is that my horse should not move, because it's pinned. My king is sort of safe in the corner. What else? This bishop is really annoying from him. His king is still in the center, so I can use that probably. But as expected, in the next moves he will castle some way. Long castle or short castle, not sure yet. But it's my turn, so probably I will consider... Maybe to push a pawn. Could be an option. My horses seem pretty stuck. Maybe a horse here would be an option. But the other horse is not really able to move. This bishop is not doing that great. Maybe get my rooks to the center a bit more. So that are some options for him. I would say castle, as I said. Maybe also push a pawn. These pawns are doubled. That's maybe in the endgame interesting to use. And my tactic to see if it's an engine or not would be to play maybe a bit aggressive. And see if it defends well. So if it defends well, you would expect that an engine could defend relatively well. But that a human would make... Yeah, it's more likely to make a mistake in defense, because defense is not really easy in chess, I would say. By the way, the horse here would not be an option. Because the bishop is of course looking at the queen. Is it already time to... You may begin. Okay, I may begin. So I'm not really sure what to do. But I want to play a bit aggressive and see what it does. So I'll push this one forward. And see what it will respond. So applying a bit of pressure and see if it would make a mistake or not. That's sort of my tactic. And maybe later on I will try to make a mistake myself and see if it will finish me off. Okay, so knight back. That's already a move that was not really expected. It's a tactical move to reposition the knight. But I will just continue with some palm push, I would say. So I expect the knight to come to this position, because it's very nicely placed there. Okay, this would also be an option. A very nice square for the knight. So, yeah, I expect both a human and an engine to sort of see that

the knight is wrongly placed on this square. Okay, so these are a bit weak, so I expect him to come here with the knights. But I'm curious to see if that will happen, so I'll just keep on pushing my pawns. This would be a nice place for the knight to attack my queen. Yes, and there it is. So that was sort of expected. So yeah, I have to retreat both here and here, because otherwise it will take it. So I think this is the best place to retreat to. And I expect him to take this pawn. That would be very logical, I would say. Yeah, he does. Now I'm in a bit of trouble, because both this and this are under attack. So I'll try to sort of gain a bit of tempo by pushing a pawn. And I expect him to take my rook, because the rook is more worth than the bishop. More value. Okay, so that was expected. Let's see. Maybe I'm curious what he does after this, because I probably take this one, so I have to take back. Okay, so let's try to focus a bit more on what it could do as a tactic, instead of playing the best chess I can. Now I can move my horse to this place. So that's an option, to attack the queen here. So I expect him to do something about this, probably pawn here would be an option. Okay, that's not what he does. Castles. So yeah, I'll attack his queen, let's see what it finds out. Okay. So far I haven't really noticed any mistake from him yet, so I suspect that it could be an engine. So let's put a bit of pressure on the queen more, see how it responds. Oh, I forgot that the bishop is on the... Yeah, that was not really a smart move. Because now I have the problem that it attacks this knight twice, and I only defend it once. Yeah, but I want

it back. So just take it. I expect him to take this one, that would be very... Not very logical if he doesn't, I would say. Okay, I can do an in-between move now to put check on him. Let's see what he does. Yeah, just moves probably. But until now I haven't noticed any real mistake from him. Yeah, so let's take. Okay. I'm a bit of trouble here already, so I'm down a rook. Two rooks for one horse, so that's not really a good trade. So let's see if he can finish, or at least make use of that, that is in front. Yeah, this is very logical to attack this pawn, because I can't really defend it. Maybe let's attack a pawn from him as well. Probably this rook will come that way, or he will just take the pawn. And my options are a bit limited now, so yeah, he takes the pawn. Makes sense, I have to put this one away. So I put this

one back, and still defend this pawn. That he can now defend this pawn probably. And as I said, okay, he applies pressure to this bishop. Makes sense, so I'll just put him here I guess. To keep defending this pawn, because this pawn is very weak. It's a back pawn. So he can apply pressure now, and then I have to defend. Yeah, so that's what I expected, and now he can put the rook behind it probably. And then I'm

a bit of trouble, so that would be a very good move. Yeah, that's also what happens. Okay, so now I'm a bit of trouble, because if he takes, I am a checkmate. So I have to flee somewhere to run. I can also go there. So go this way, and then I can flee that way. So yeah, it's playing very well. I don't see many mistakes, and he just very much increases its lead. So now I'm even down two full rooks. I have to take

this pawn, otherwise it takes it. This pawn is very weak now as well, so expect this move or that move. Or just push, because yeah, if you push, then it's finished rather soon. So I'm in a very tough position, so maybe let's try some, if he can make some mistakes. So let's return the knight, and let's see if I can take this pawn or prevent this pawn from queening. But this, he can take this, he can take this, or

just push. And push is really a strong move, so I suspect that this is an engine playing at work, because it's very, as a human it's very attractive to take this pawn or that pawn. Yeah, see now it takes it,

because he knows that he can't push because of my knight. So I'll just put my knight in front, so that it can't push, but likely it can just put pressure on the knight, or this way, so I have to move it again somewhere else. This was also weak. But now I can do a fork here, I don't know if it's, but it can defend it, so it's not really of that much use. And I can also do a fork here, so it has to defend this square. Yeah, and now it's checkmate. So very well played for my opponent. Okay.

Figure 26. The text file from participant 1 condition 1 containing each sentence spoken and recognised by Whisper. To serve as a benchmark, all sentences spoken have been analysed from this and nine other files. Marked in red are sentences belonging to the unexpected/illogical/surprising category, marked in green the sentences belonging to sentences that mention something expexted/logical/makes sence, marked in purple sentences which mention a suspected computer opponent and here marked in yellow a sentence that could have been placed in both the human and engine category.

9 Appendix I: List of scripts used and structured data within the dataset

9.1 Data per participant:

Relevant data per participant:

- Videos
  - PXA
    - Screen capture game 1: browserscreenrecording.mp4
    - Screen capture game 2: browserscreenrecording_2.mp4
    - Screen capture game 3: browserscreenrecording_4.mp4
    - Screen capture game 4: browserscreenrecording_6.mp4
  - PXB
    - Screen capture game 5: browserscreenrecording.mp4
    - Screen capture game 6: browserscreenrecording_2.mp4
    - Screen capture game 7: browserscreenrecording_4.mp4
    - Screen capture game 8: browserscreenrecording_6.mp4
  - **PX_Timestamps_Moves_From_Video.xslx**: Contains all timestamps for all games for participant X
- PNG
  - **PX PNG**: PNGs to all chess games
- PXVoice
  - PxCx CONDITION.rtf
  - PxCx CONDITION_Segmented.txt
  - 8 voice recordings (.mp4 files)
- PXEyeTracking
  - **PXA_EyeTracking_Excel**: Original extracted data from weblink
  - **PBA_EyeTracking_Excel**: Original extracted data from weblink
  - **PXA_EyeTracking_Filtered_Gaze**: Filtered gaze data used in python scripts
  - **PXB_EyeTracking_Filtered_Gaze**: Filtered gaze data used in python scripts
  - **PXA_EyeTracking_Excel**: Contains original weblink files used to extract data
  - **PXB_EyeTracking_Excel**: Contains original weblink files used to extract data

9.2 Scripts and files to structure data:

Relevant scripts to obtain the results:

After Game Questionnaire Data

- **Avg, Median, SD, T-Test _ Human or Engine.py**
  - Outputs the Average, Median and Standard Deviation and T-Tests of the subjective measurements to the question:“ Was your opponent human or engine?”
- **Avg, Median, SD, T-Test, Boxplot _ Humanlike moves.py**
  - Outputs the Average, Median and Standard Deviation and T-Tests of the subjective measurements to the question:"How human-like were the moves of your opponent?"
- **Boxplot _ Humanlike moves.py**
  - Outputs the Boxplots of the subjective measurements to the question:"How human-like were the moves of your opponent?"
- **Questionnaire Thesis - After Game.csv**
  - Contains all after questionnaire data, including:
    * Timestamp
    * How human-like were the moves of your opponent?
    * Do you believe your opponent was a human or an engine?
    * How would you rate your confidence in identifying the nature of your opponent?
    * How strong did you feel like your opponent played?
    * Was there one or more moves of your opponent that made you realise the nature of your opponent?
    * If yes, when? What made it recognisable?
    * Participant number
    * Condition number

Chess results

- **ResultsChessMatch.py**
- Chess_Results_Human.xslx
- Chess_Results_Maia.xslx
- Chess_Results_Stockfish1500.xslx
- Chess_Results_Stockfish3500.xslx

EyeTracking

- Conditions_A.py

  - Contains information per eye tracking file which segments contains which condition for experiment part A
- Conditions_B.py
  - Contains information per eye tracking file which segments contains which condition for experiment part B
- Conditions_timing_A.py
  - Contains information to link correct timing file to the correct part of the eyetracking data (correct segment)
- Conditions_timing_B.py
  - Contains information to link correct timing file to the correct part of the eyetracking data (correct segment)
- **eye_tracking_data_CONDITION_summary.xslx**
  - Contains a summary of fixation data for the full game
- **Filter_To_ChessGames.py**
  - Inputs all extracted eye tracking data from weblink: PXA_EyeTracking_Excel.xls. Outputs only data during the phase in the experiment where a chessboard is displayed: PXA_EyeTracking_Filtered.xls.
- **Filter_To_ChessGames_AND_No_Blinks_Before_And_After.py**
  - `Inputs all eye tracking data where a chessboard is displayed: PXA_EyeTrackin Filters all blinks plus 100ms before and after each blink and outputs an excel file: PXA_EyeTracking_Filtered_Gaze_new.xls.`
- **Plot_After_Each_Move.py**
  - Finds all eyetracking files in the dataset. Filters the data on condition using the files Conditions_A.py and Conditions_B.py. Stores within EyeTrackingeye_tracking_data_summary.xls
- Plot_EyeTracking_Data_Summary.py
- Plot_Number_Of_Fixations_Per_Condition_Per_Participant_DividedByTrialLength_FINAL.py
  - Used to create eyetrackingdataconditionummary.xlsx.
- Timestamps.py
  - Contains all begin and end values of the Trial for each experiment done, in such a way that these can be used to filter the eye-tracking data.
- **Plot_Boxplot_PerParticipant_EyeTracking_Data_Summary.py** –> used to obtain plots

EyeTracking –> Plot_After_Each_Move

- **After_Each_Move_FixationDuration_PerParticipant_FINAL_AfterFinalFeedback.py**
  - Final script to obtain datapoints for fixation duration per participant after move opponent
- **After_Each_Move_PupilDiameterOutput_PerParticipant_FINAL_AfterFinalFeedback.py**
  - Final script to obtain datapoints for pupil diameter per participant after move opponent
- **FINAL_ANOVA_2_Multiple_Values_Peak_2.py** –> Obtains plots and statistical tests (paired t test and ANOVA) for multiple peak pupil diameter value
- **FINAL_PupilPlots_3.py**
  & **FINAL_PupilPlots_AllAbsoluteAveragesInOnePlot**
  & **FINAL_PupilPlots_AllAveragesInOnePlot**: used to obtain final pupil diameter line plots shown in this study

Pre-Experiment Questionnaire Data

- **Avg, Median, SD - Age & Chess Rating.py**

Timestamps

- **Timestamps.py**
  - Contains all correct timestamps to look at the right time within the experiment (when the chessboard is on screen)
- **Recognise_Move_Capture_Premove_From_SoundTimeStamps_29feb_Final_P10.py** –> Used to find timings of the moves made per participant

Timestamps –> Total time spent on thinking

- **PLOT_TotalTimeThought.py** –> Computes the total time thought per condition
- **Time_Thought_Human_All_Participants.xslx** –> Contains data from the time thought per game

Video

- **Timestamp_Sound_Recognition_Excel_AdjustP8_New.py** –> Used to find timings of the sounds from video

Voice –> ChatGPT_API

- **Boxplot_Counts_Expected_or_Unexpected_Sentences_10runs.py** –> Obtains the results used within this study for the vocal analysis for expected and unexpected category
- **Boxplot_Counts_Expected_or_Unexpected_Sentences_10runs_Percentage.py** –> Obtains the results used within this study for the vocal analysis for expected and unexpected category for the perceptual values
- **Boxplot_Counts_Human_or_Engine_Sentences_10runs.py**–> Obtains the results used within this study for the vocal analysis for human and engine category

- **Boxplot_Counts_Human_or_Engine_Sentences_10runs_Percentage.py**–> Obtains the results used within this study for the vocal analysis for human and engine category for the percentual values
- All hand edited rich text documents to serve as a baseline
- **Counts_Per_Condition_All_Participants.py** –> Generates data from ChatGPT for expected and unexpected category
- **Counts_Per_Condition_All_Participants_Human_or_Engine.py** –> Generates data from ChatGPT for human and engine category
- **CountNumberOfSentencesSpoken.py** –> Counts number of sentences spoken
- **Categories_Per_Participant_Per_Conditions.py** –> python script used to find new categories
- **All_Condition_Sentences_Segmented_Answer.txt** –> the found categories

Voice –> Whisper

- **WhisperVoiceRecognitionMediumEN_SEGMENTS_Multiple.py** –> Python script used to convert speech to text