# Turing Tests in Chess

## An Experiment Revealing the Role of Human Subjectivity

Eisma, Y.B.; Koerts, R.; de Winter, J.C.F.

**DOI**
10.1016/j.chbr.2024.100496

**Publication date**
2024

**Document Version**
Final published version

**Published in**
Computers in Human Behavior Reports

**Citation (APA)**
Eisma, Y. B., Koerts, R., & de Winter, J. C. F. (2024). Turing Tests in Chess: An Experiment Revealing the Role of Human Subjectivity. *Computers in Human Behavior Reports*, *16*, Article 100496. https://doi.org/10.1016/j.chbr.2024.100496

Computers in Human Behavior Reports 16 (2024) 100496

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

# Turing tests in chess: An experiment revealing the role of human subjectivity

Yke Bauke Eisma, Robin Koerts, Joost de Winter *

*Department of Cognitive Robotics, Delft University of Technology, the Netherlands*

A B S T R A C T

With the growing capabilities of AI, technology is increasingly able to match or even surpass human performance. In the current study, focused on the game of chess, we investigated whether chess players could distinguish whether they were playing against a human or a computer, and how they achieved this. A total of 24 chess players each played eight 5 + 0 Blitz games from different starting positions. They played against (1) a human, (2) Maia, a neural network-based chess engine trained to play in a human-like manner, (3) Stockfish 16, the best chess engine available, downgraded to play at a lower level, and (4) Stockfish 16 at its maximal level. The opponent's move time was fixed at 10 s. During the game, participants verbalized their thoughts, and after each game, they indicated by means of a questionnaire whether they thought they had played against a human or a machine and if there were particular moves that revealed the nature of the opponent. The results showed that Stockfish at the highest level was usually correctly identified as an engine, while Maia was often incorrectly identified as a human. The moves of the downgraded Stockfish were relatively often labeled as 'strange' by the participants. In conclusion, the Turing test, as applied here in a domain where computers can perform superhumanly, is essentially a test of whether the chess computer can devise suboptimal moves that correspond to human moves, and not necessarily a test of computer intelligence.

## 1. Introduction

As AI continues to advance and becomes increasingly capable of closely mimicking human behavior, a pressing question arises: *is this a human or a machine?* This question can appear in various contexts, such as road traffic, where automated driving behavior is becoming almost indistinguishable from human driving (Lambert, 2024), or in text generation, where it is often unclear whether content was produced by a human or in part by a large language model (LLM) (De Winter et al., 2024; Farazouli et al., 2024). Similar challenges are evident in intellectual pursuits like chess, where concerns about cheating have recently surged (Solon, 2024; Yue, 2024).

Deciding whether an interacting agent is a human or a machine, to determine if a machine can mimic human intelligence, is known as a Turing test. In the original Turing test, a human evaluator asks questions via a terminal to both a human participant and a machine (Turing, 1950). The machine passes the Turing test if the human evaluator cannot reliably determine which of the two is the machine. During a Turing test, the human evaluator can apply various strategies, such as using ambiguous language that might be difficult for a computer to understand. Another strategy is to ask the chatbot about recent events or outside weather that the AI might not have knowledge of (e.g., Jannai et al., 2023; Jones & Bergen, 2024).

A common criticism of the Turing test is that, while it is often presented as a test of computer intelligence, computer intelligence is not the same as human intelligence (French, 1990). When the computer exhibits behavior that is *too intelligent*, such as a highly articulate response or a fast and accurate response to an arithmetic problem, it gives itself away as a computer and thereby fails the Turing test (Michie, 1993; Turing, 1950). Humans can act unintelligently at times. In a chess game, for example, a human player will occasionally make a blunder. Similarly, in a Turing test, the human evaluator will look for signs that the opponent's responses lack any mistakes or demonstrate unusually consistent response times, which are not typical of human behavior (e.g., Ciardo et al., 2022).

Bazilinskyy et al. (2021) had human observers evaluate the behavior of a passing self-driving car. Their study revealed that the same behaviors of this car were interpreted differently by different observers. There were certain behaviors, such as hard braking, which some observers interpreted as human-like ('*a computer will always drive defensively and never brake so aggressively*') or computer-like ('*this looks like an autonomous car that only recognized the need to brake at a late moment because of sensor limitations*'). The findings of Bazilinskyy et al. indicate that passing a Turing test is not necessarily about the machine's behavior, but may be more indicative of the human evaluator's expectations concerning typical computer and human abilities. These findings

* Corresponding author. Mekelweg 2, 2628 CD, Delft, the Netherlands.
*E-mail address:* j.c.f.dewinter@tudelft.nl (J. de Winter).

https://doi.org/10.1016/j.chbr.2024.100496
Received 30 June 2024; Received in revised form 23 September 2024; Accepted 25 September 2024
Available online 26 September 2024


**Fig. 1.** (a) The seating position of the participant, with the voice recorder and the Portable Duo eye-tracker. (b) The participant's view during the experiment. (c) The seating position of the experimenter. (d) The view of the experimenter, containing from left to right 1) a monitor displaying the chess GUI (here for condition StockfishH), 2) a laptop where the experimenter plays chess against the participant, and 3) a laptop provided by SR Research, where the experimenter is able to check whether the Portable Duo is collecting eye-gaze data.

are consistent with an earlier study by Warwick and Shah (2015), which described the thought processes of human evaluators in Turing tests involving a chatbot. It was found that the judgments were highly subjective and related to the expectations that the person had. For example, the use of humor by an entity was perceived by some human participants as a computer-initiated attempt to appear human, rather than as expected human behavior. These findings relate to the *mirror hypothesis* described in an article about the Turing test by Sejnowski (2023). He argued that current LLM-based chatbots, such as ChatGPT, primarily reflect the expectations and intelligence of the user. When the user provides a prompt, the LLM will generate an output that matches the tone, style, and content of that prompt. This mirroring behavior gives the impression that the LLM is intelligent, but in fact, the output partly reflects the quality and intelligence of the human input.

The current paper examines whether human chess players can determine whether they are playing against a human or against a computer. The game of chess provides a clearly defined environment with only a limited number of rules and possible moves at a given time. Current chess engines can execute superhuman moves and defeat human players of any chess skill level (computerchess, 2024; Wired, 2023). To make the game against a computer interesting, chess computers are equipped with features that sometimes introduce a bad move, thereby achieving an overall skill level that corresponds to a typical human player. Certain chess engines are created to simulate human play or are trained using human gameplay data, intentionally producing inaccuracies and blunders to match human play (Barrish et al., 2024; McIlroy-Young et al., 2020; Rosemarin & Rosenfeld, 2019). The current study has human participants compete against a human opponent and computer opponents of different strengths and different degrees of intended human-likeness. The underlying question in the present Turing-test study is not whether the computer is intelligent enough to match a human, but rather what level of play and type of moves convince the participants about the nature of their opponent. The aim is to gain insights into the predictors of Turing test outcomes, which in turn may generalize to other applications, such as interactions with AI or robotics in general.

## 2. Methods

### 2.1. Participants

A total of 24 participants took part in the experiment, 23 of whom were male and one was female.

The participants ranged in age from 14 to 59 years ($M = 26.3$, $SD = 9.5$). Based on a questionnaire administered before the experiment, participants reported playing more chess against human opponents ($M = 5.5$, $SD = 1.2$) than against chess engines ($M = 1.5$, $SD = 0.7$), where 1 indicated a few times a year and 7 indicated daily.

Twenty-three out of 24 participants reported their online or over-the-board chess ratings, which we converted to Lichess-Blitz-equivalent ratings using online conversion tables (ChessGoals, 2024). If a participant provided multiple ratings (e.g., Blitz and Rapid), the mean Lichess-equivalent rating was used. For one participant who did not report any ratings, a Lichess-Blitz-equivalent rating of 1350 was estimated by applying a quadratic fit, based on the Lichess-Blitz-equivalent ratings of the other 23 participants and their mean win rate loss per move (calculated using Stockfish 16.1 with a depth of 20), to the participant's own win rate loss per move. The mean Lichess-Blitz-equivalent rating of the 24 participants was 1564 ($SD = 430$).

Participants were asked in a questionnaire administered before the experiment: "*How would you try to recognize the difference between an engine and a human opponent?*" The participants' responses indicated that

an engine can be recognized by unusual mistakes or odd moves, such as unnecessary sacrifices or highly precise tactics. They also pointed out that engines use time differently, such as responding quickly to complex situations, whereas humans take longer to consider their options in difficult positions. Furthermore, it was reported that humans play with more intuition, while engines tend to focus on finding the best move, even if it breaks general principles.

The experiment received approval from the Delft Human Research Ethics Committee, and each participant provided written informed consent.

### 2.2. Apparatus

Fig. 1 shows the experimental setup. The experiment used a 17.3-inch monitor of the laptop model ROG Zephyrus S17 GX701 GX701LXS-XS78 with a total display area of 383 × 215 mm and a screen resolution of 1920 × 1080 pixels. The root of the screen of the laptop was placed 71 cm from the edge of the table. The participants' verbal utterances were recorded using a digital voice recorder. Eye movements were recorded using the SR Research EyeLink Portable Duo (SR Research, 2024). The Portable Duo was placed 65 cm from the table and captured eye movements at a frequency of 1000 Hz.

### 2.3. Software

The experiment was conducted using WebLink, a screen recording software solution which records eye movements, as well as browser navigation actions, mouse clicks, and mouse positions (SR Research, 2023). Lichess.org (2023a) was chosen because it can be easily configured to start a game from a given chess position. We used Lichess's ZEN mode, which removes all non-essential elements on the webpage. To prevent cheating, Lichess bans the use of chess engines. This is why, for this experiment, the account 'TuringTest001' was set to a BOT status. This function exists for people playing with the help of engines.

### 2.4. Experiment task

Participants were told that the proportion of human and engine opponents might vary. Each chess game was limited to 5 min per player, with no seconds added to the clock for a move that was played. The participant could spend their time as they liked. The experimenter, however, was tasked with playing a move in 10 s. The participants were neither encouraged nor discouraged from winning.

### 2.5. Independent variable

The independent variable was the opponent the participant played against. There were four types of opponents, henceforth called conditions:

- **Human**: The same chess player was used in all experiments. His Chess.com rapid rating was 944, which corresponds to an 1136 Lichess Blitz rating.
- **Maia**: Maia is an adapted version of AlphaZero, developed to predict human chess moves with high accuracy (McIlroy-Young et al., 2020). It is trained on actual human chess games rather than self-play games. Maia employs deep neural networks without using tree search. It is specifically trained to mimic the moves of players of different skill levels. Nine models of skill levels have been trained, each using 12 million games from the open-source chess platform Lichess. In this experiment, the algorithm Maia1 was chosen to replicate a skill level similar to that of a human rated around 1100.
- **StockfishL**: Stockfish 16 was used, a free and open-source chess engine. At the time of the experiment, Stockfish 16 was the strongest available chess engine (Stockfish, 2023a). Its estimated ELO rating was over 3600 (computerchess, 2024), substantially higher than the highest recorded ELO rating of a human ever (Magnus Carlsen, 2882, May 2014; 2700chess, 2024). Stockfish has the option to lower its skill level, which this condition represents. The level of Stockfish was set to a low value of 4 out of 20, to replicate a more human skill level.
- **StockfishH**: In this condition, Stockfish was set to its highest level of 20 out of 20.

### 2.6. Experimental procedures

When the participant arrived, they were welcomed by the experimenter and offered an informed consent form. After signing, participants completed a short questionnaire to gather demographic information, chess playing frequency, and perceptions of differences between human and engine play styles in chess. Next, they were given verbal instructions. It was mentioned that the task was not to win, but to recognize the nature of the opponent while playing chess. Additionally, the participant was asked to think aloud in English and was given some examples of what to talk about: how they are trying to figure out whether the opponent is an engine or human, which moves they expect from the opponent, and which moves they are considering. Furthermore, the participant was instructed to sit as still as possible after the eye-tracker had been calibrated.

In the WebLink environment, instructions similar to the verbal instructions were repeated on screen at the start of the experiment:

- ***Welcome.*** *Today, you will be playing chess from 8 pre-selected positions.*
- ***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.*
- ***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.*
- ***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 the same proportion of human & engine opponents, as they might vary.*

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 eye-tracker to register their eye movements on the screen. During the experiment, the experimenter could check whether the eye gaze was being registered. If this were not the case, the experimenter would ask the participant to move in a specific way; the request was made in a soft voice to avoid disturbing the voice recording.

After calibration, the WebLink application directed the participant to the Lichess website, where the participant received a game invitation from the experimenter. Once the participant accepted the game, the experimenter activated the voice recorder. The participant began by verbalizing their thoughts while analyzing the position to get comfortable with thinking aloud. After completing their analysis, the participant made the first move. The game proceeded until checkmate, a time-out on either side, or the participant's resignation.

High-level engines occasionally play optimal moves in less than a second, potentially giving away their nature. To increase difficulty in recognizing the opponent for the participant, the experimenter made a chess move every 10 s. To prevent mix-ups by the experimenter, the three different chess engines were managed through separate interfaces.

When the game was finished, the experimenter deactivated the voice recorder and made sure the participant was guided towards Google Forms, where the participant answered six questions about the game just played (see Dependent variables section).

The order in which the conditions were presented was determined before the experiment, using a complete counterbalancing method. This was possible because there were 24 participants and 24 possible combinations of the four conditions. Each participant played a total of eight games, encountering each of the four conditions twice. This was done in

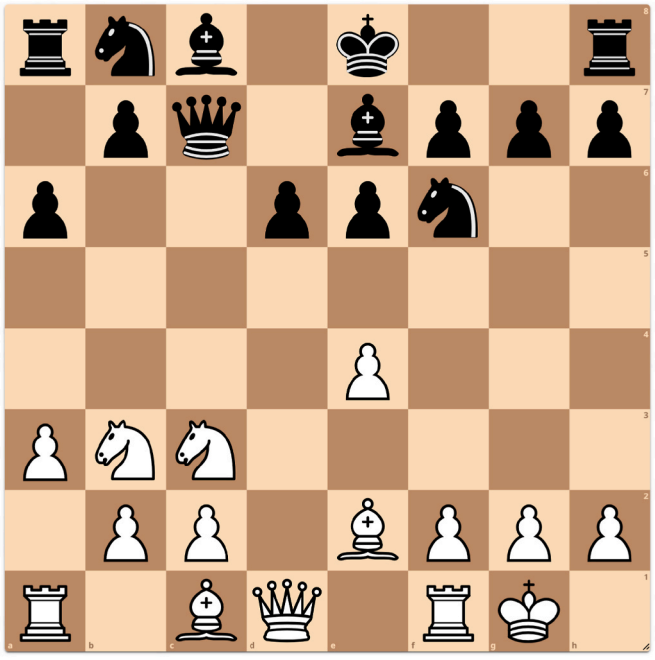

**Fig. 2.** An example of a starting position.

two blocks of four counterbalanced conditions. After playing four games, participants took a break. The entire experiment, including instructions, calibration, and breaks, took approximately 2 h to complete.

### 2.7. Positions played from

All positions played from were games played on Lichess by the highest Rapid-rated player at the time of preparing the experiment, named GM Drvitman. A PGN (Portable Game Notation) file was downloaded containing Rapid games. A PGN file is a plain text format used to record chess games, including moves and clock data. All games were evaluated by Stockfish 16 from white's perspective.

The downloaded games were evaluated after nine moves and filtered based on the following criteria:

- The position should be rated approximately equal by Stockfish 16. Any evaluation between −1.0 and 1.0 was considered sufficiently equal for this experiment.
- Each side must have moved at least two pawns.
- Each side should have developed at least two of the four minor pieces (knights and bishops).
- The position had to be positional, not tactical. Thus, positions containing any forced moves were eliminated, providing the participant with several reasonable options for their first move.

All selected positions were recreated in Lichess Studies (see Fig. 2 for an example of a starting position). In total, 240 positions were selected to supply games for a maximum of 30 participants playing 8 games each. The reason all starting positions were different was to ensure that the human opponent always encountered a new starting position. This prevented him from memorizing previous positions or learning from past mistakes.

### 2.8. Dependent variables

After the experiment, the evaluation of each encountered position, recreated from the PGN files, was calculated using Stockfish 16.1 with a depth of 20. The engine's evaluation is presented in the form of a score that can range from minus infinity (winning for black) to plus infinity (winning for white). For easier interpretation, this score was converted to a win rate percentage as follows (Lichess.org, 2023b):

$$Win\ rate = 50 + 50\left(\frac{2}{1 + e^{-0.00368208 \cdot Evaluation}} - 1\right)$$

Accordingly, the win rate describes the position evaluation, as calculated by Stockfish, on a bounded scale from 0 to 100%. This win rate has been calibrated based on Stockfish analyses of games played by strong players (2300+ Elo) in Rapid time control but is a simplification and does not account for draw probabilities. Since most participants had a rating lower than 2300, the win rate should not be interpreted literally as the probability of winning but rather as a value that indicates how favorable the position was for the participant.

The following dependent variables were subsequently extracted per participant per game:

A) **Number of half-moves.** The number of half-moves played in the game. For example, if the participant had played 10 moves and the opponent also had played 10 moves in a given game, the number of half-moves was 20.
B) **Participant's result.** The result of the participant, where 0 stands for a loss, 0.5 for a draw, and 1 for a win.
C) **Participant's win rate, end of game.** The win rate of the participant according to the engine, when the game had ended. Note that this win rate does not necessarily correspond to the game result. For example, it may be that for the opponent (playing with black), a checkmate combination was available, putting black in a winning position (the participant's win rate will then be 0% for this game), but black lost on time because the experimenter adopted a thinking time of 10 s.
D) **Win rate loss, participant's move.** This variable indicates how much the participant's (i.e., white player's) win rate decreased due to the participant's move compared to the best move as identified by Stockfish 16.1. This score was then averaged over all the participant's moves in the game.
E) **Win rate loss, opponent's move.** This variable indicates how much the opponent's (i.e., black player's) win rate decreased compared to the best move as identified by Stockfish 16.1. This score was then averaged over all the opponent's moves in the game.
F) **Participant's move time.** The average time per move (i.e., thinking time) for the participant. The move time was extracted from the clock times in the saved PGN files and was available in whole seconds.

In addition to the above-mentioned performance-related measures, six self-reported variables were extracted from the questionnaire that was completed after each game and the transcribed think-aloud statements.

A) **Q1. Human-likeness.** The participant's response to the question: *How human-like were the moves of your opponent?* (1: computer, 7: human).
B) **Q2. Engine or Human.** The participant's response to the question: *Do you believe your opponent was a human or an engine?* (coded as: 0: engine, 1: human).
C) **Q3. Confidence.** The participant's response to the question: *How would you rate your confidence in identifying the nature of your opponent?* (1: not confident at all, 7: extremely confident).
D) **Q4. Opponent strength.** The participant's response to the question: *How strong did you feel like your opponent played?* (1: weak, 7: strong).
E) **The number of words spoken.** This was determined by automatically transcribing the recorded think-aloud results using OpenAI's Whisper large-v2 model (OpenAI, 2024; transcription

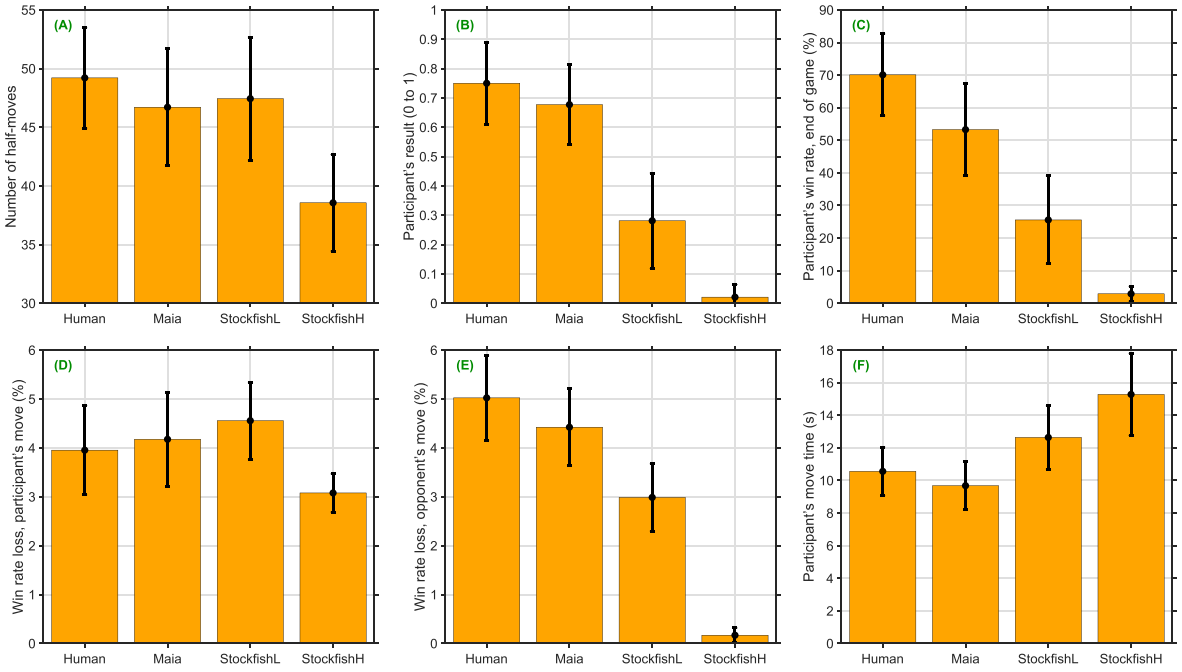


**Fig. 3.** Means and 95% confidence intervals of chess performance-related variables across the four conditions. All values are determined per game and then averaged over the 2 games per participant per condition. The means and confidence intervals are thus calculated over 24 data points corresponding to the 24 participants.

performed in May 2024). Audio recordings were not available for 4 out of 8 games for 1 of 24 participants, and for 1 out of 8 games of 3 participants.

F) **Surprise keyword count.** This was determined by automatically extracting the number of surprise-related keywords from the transcripts. The following keywords were defined for this purpose: strange, weird, odd, unnatural, unusual, dubious, awkward, unorthodox, questionable, random, puzzling, not logical, illogical, bizarre, unconventional, unexpected, peculiar, seems off, not a typical, did not expect, surprising, and surprise. The underlying reason for doing this analysis is that Maia is programmed to exhibit human-like inaccurate moves, while StockfishL sometimes makes an artificially bad move that is not necessarily human-like, and thus can be perceived as strange by a human.

The above variables were determined per game per participant and then averaged over the 2 games per condition. Thus, for each dependent variable, a 24 participants × 4 matrix with scores was obtained per dependent variable. The variables were plotted in bar plots with 95%

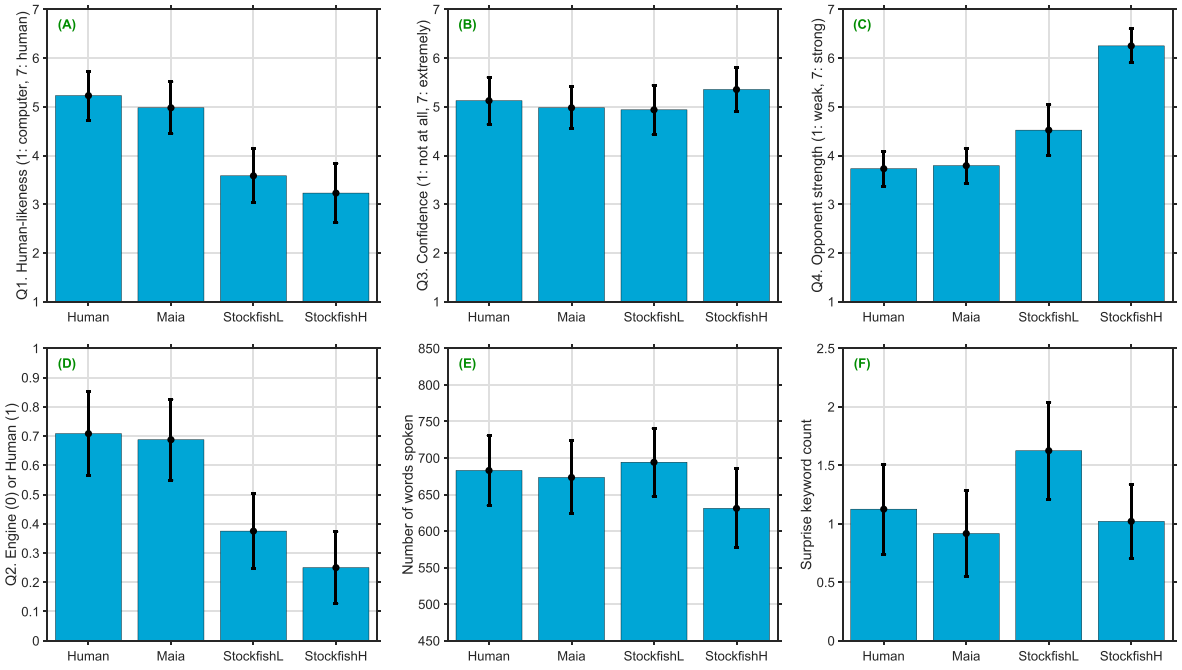


**Fig. 4.** Means and 95% confidence intervals for within-subject designs of variables derived from the post-game questionnaire and think-aloud data across the four conditions. All values are determined per game and then averaged over the 2 games per participant per condition. The means and confidence intervals are thus calculated over 24 data points corresponding to the 24 participants.

**Table 1**
GPT-4o-based summaries of questionnaire responses for the question: "Was there one or more moves of your opponent that made you realize the nature of your opponent? If yes, when? What made it recognisable?"

| | Participant thinks opponent is Engine | Participant thinks opponent is Human |
|---|---|---|
| **Human** | 14 (9 available responses analyzed)<br>**GPT-4o:** The respondents believe their chess opponent was an engine due to a combination of unusual or suboptimal moves, such as waiting moves, simple mistakes, and decisions that felt more characteristic of a computer's logic rather than human intuition. | 34 (19 available responses analyzed)<br>**GPT-4o:** The respondents believe their chess opponent was a human due to the presence of mistakes, missed tactics, and natural but imperfect play that an engine would likely avoid. |
| **Maia** | 15 (10 available responses analyzed)<br>**GPT-4o:** The respondents believe their chess opponent was an engine due to a combination of highly strategic and unnatural moves, minimal mistakes, and sudden shifts in play quality that are characteristic of computer algorithms rather than human intuition. | 33 (24 available responses analyzed)<br>**GPT-4o:** The respondents believe their chess opponent was a human due to various blunders, missed opportunities, and strategic errors that are uncharacteristic of chess engines, such as giving away pieces, missing checkmates, and making moves that seemed to lack deeper calculation or clear purpose. |
| **StockfishL** | 30 (20 available responses analyzed)<br>**GPT-4o:** The respondents believe their chess opponent was an engine due to a combination of highly calculated, cold, and unnatural moves, missed obvious checkmates, strange blunders, and random or illogical piece movements that are atypical for human players. | 18 (13 available responses analyzed)<br>**GPT-4o:** The respondents believe their chess opponent was a human due to various mistakes and suboptimal decisions, such as missing checkmate opportunities, making imprecise moves, and not capturing pieces when possible, which they think a computer would not make. |
| **StockfishH** | 36 (23 available responses analyzed)<br>**GPT-4o:** The respondents believe their chess opponent was an engine due to the execution of highly sophisticated, non-human-like moves, such as avoiding obvious captures, making complex tactical decisions, and demonstrating deep positional understanding and consistency. | 12 (6 available responses analyzed)<br>**GPT-4o:** The respondents believe their chess opponent was a human due to the opponent's unpredictable and less optimal moves, adherence to predicted strategies, and natural, intuitive play that deviated from the precise and structured approach typical of a chess engine. |

*Note.* The number in each cell indicates how many times the participant indicated that the opponent was an engine (left column) or a human (right column), corresponding to Fig. 4D, with the number of comments provided and analyzed shown in parentheses. The maximum number is 48 (24 participants × 2 games per participant per condition).

confidence intervals. For the six performance-related variables, the 95% confidence interval was determined for each condition separately; this was done due to the nonhomogeneous variances between conditions (for example, StockfishH resulted in a win rate for the participant of almost 0%, a clear floor effect). For the self-report variables, a confidence interval for within-subject designs was determined according to Morey (2008).

The final part of the questionnaire completed after each game consisted of the following two questions: Q5 "*Was there one or more moves of your opponent that made you realize the nature of your opponent?*" (Response options: Yes; No, more a general feeling; Other), with a follow-up question: Q6 "*If yes, when? What made it recognisable?*". The responses to Q6 were analyzed using OpenAI's GPT-4o (model: gpt-4o-2024-05-13). We chose to use a large language model because we aimed to summarize the text results without the current researchers potentially introducing bias (Tabone & De Winter, 2023). Although large language models can also introduce bias due to the data on which they were trained and the way the models have been fine-tuned, using GPT seemed like a suitable approach to summarize the responses in a reproducible manner, where the model remained blind to the experimental condition under which the data was collected.

The following prompt was used: "*Summarize the participants' statements below in 1 sentence, focusing on why the respondents believe their chess opponent was an engine. Do not mention individual statement numbers.*" Here, we submitted the text responses when the participant thought the opponent was an engine, based on Q2. The same process was repeated with the prompt "*Summarize the participants' statements below in 1 sentence, focusing on why the respondents believe their chess opponent was a human. Do not mention individual statement numbers.*" for comments where the participant indicated in Q2 that they thought the opponent was a human.

Finally, we determined correlation coefficients between the Lichess-Blitz-equivalent rating and the responses to Q2 and Q4, as well as the average move time. This was done to investigate whether stronger players were better able to recognize the nature of the opponent and adjust their chess strategy accordingly. A Pearson product-moment correlation coefficient was chosen, which for Q2, a question with dichotomous response options, is equivalent to the so-called point-biserial correlation coefficient.

## 3. Results

Fig. 3 provides an overview of the chess-performance-related results for each of the four conditions. It can be seen that participants mainly lost or were in a losing position at the end of the game when they played against StockfishH, and to a lesser extent when they played against StockfishL, while they mostly won against the human and Maia opponents (Fig. 3B & C). The superhuman playing strength of StockfishH explains why the games in this condition ended in fewer moves compared to the other conditions (Fig. 3A). When participants played against StockfishH, they generally thought longer (Fig. 3F) and made slightly better moves compared to the other conditions (Fig. 3D).

Fig. 4 shows the means and 95% confidence intervals obtained from the questionnaire completed after each game (Fig. 4A–D) and the think-aloud analysis (Fig. 4E & F). A clear trend is evident that participants perceived StockfishH, and to a lesser extent StockfishL, as stronger opponents than the human experimenter and Maia (Fig. 4C). Accordingly, participants believed that StockfishL and StockfishH played like a computer (Fig. 4A) and that these were chess engines (Fig. 4D). The confidence in this judgment was equivalent across the different conditions, with a slight tendency to judge StockfishH with more certainty as being either human or computer (Fig. 4B). When playing against StockfishH, slightly fewer words were spoken compared to the other conditions (Fig. 4E), which might be explained by the fact that these games were shorter in duration (see Fig. 3A). Finally, it is evident that surprise-related keywords were spoken more frequently when playing against StockfishL than in the other conditions.

Table 1 provides an overview of the characteristics of the given reasons that led participants to realize the nature of the opponent. Participants correctly identified the human as human due to observed mistakes, blunders, and inaccuracies. The same applies to Maia, which was often perceived as a human due to blunders and other mistakes, or short-sighted thinking. StockfishH was often correctly recognized as an engine because it operated with high precision and calculation in a manner that did not seem human. StockfishL was identified as a computer due to its 'cold calculation' and strange moves that did not appear to serve an immediate purpose.

Regarding the incorrect assessments in Table 1, where participants thought the human opponent was a computer, the given reasons included that simple mistakes were made; participants thought it was a

**Table 2**
Pearson product-moment correlation coefficients between Lichess-equivalent ratings and responses to selected questions as well as participants' mean move times ($n = 24$).

| | Human | Maia | StockfishL | StockfishH |
|---|---|---|---|---|
| Q2. Engine (0) or Human (1) | −0.12 | 0.38 | 0.09 | −0.37 |
| Q4. Opponent strength (1: weak, 7: strong) | −0.43[a] | −0.54[b] | 0.00 | 0.35 |
| Participant's mean move time (s) | −0.41[a] | −0.22 | −0.26 | 0.27 |

*Note.* The correlation between the rating (a continuous variable) and the response to Q2 (a dichotomous variable) is also known as a point-biserial correlation coefficient.
[a] $p < 0.05$.
[b] $p < 0.01$.

low-level engine (equivalent to why participants correctly identified StockfishL as an engine). Additionally, participants incorrectly identified StockfishL as human due to the mistakes made by the engine or its passive play.

Finally, we investigated whether the overall playing strength of the participant, operationalized by their Lichess-equivalent Blitz rating, was predictive of whether they could distinguish between a human and a machine (Q2) and how strong they perceived their opponent to be (Q4). For this, the correlation coefficient was used between the player's rating and their response to the questions, as shown in Table 2. Additionally, the correlation between the rating and the participant's average move time is shown.

Although the correlations are, in many cases, not statistically significantly different from zero, there appears to be a divergence, where stronger players more frequently correctly identified StockfishH as an engine and incorrectly identified Maia as a human (Q2). Furthermore, stronger players perceived StockfishH as playing stronger, and both the human experimenter and Maia as playing weaker (Q4). Additionally, it seems that stronger players used their time differently by playing faster against the human (whom they could defeat) while taking more time per move when playing against StockfishH, an engine that defeats them.

Additionally, the Lichess-Blitz-equivalent rating criterion demonstrated criterion validity, as it was predictive of the participant's win rate at the end of the game (depicted in Fig. 3C) averaged over the four conditions and the average win rate loss (depicted in Fig. 3D), with $r = 0.70$ ($p < 0.001$, $n = 24$) and $r = -0.81$ ($p < 0.001$, $n = 24$), respectively.

## 4. Discussion

In this study, 24 chess players each played eight Blitz games against an unknown opponent. The opponents included a human player in two games, Stockfish at a low level in two games, Stockfish at its maximum superhuman level in two games, and Maia, a chess engine designed to play in a human-like manner in the remaining two games. The order of these conditions was counterbalanced.

The original Turing Test, as described by Turing (1950), has been interpreted by some as a measure of computer intelligence. When a chess computer first defeated a world champion, this achievement was seen by some as an instance of passing the Turing Test (Krol, 1999). However, it can be argued that Turing did not intend his test to be taken so literally, but rather to provoke thought and challenge the views of philosophers, mathematicians, and scientists who were skeptical of the cognitive potential of computers (Gonçalves, 2023).

With the increasing capabilities of computing and neural networks, the Turing Test remains relevant today. However, in certain domains, including chess, computers now perform at levels far beyond human abilities. As we demonstrated, when a computer exhibits superhuman chess skills, it actually fails the Turing Test, even when its time per move is forced to 10 s. In 75% of the games, participants accurately identified StockfishH as a computer. Stronger players were more capable in this task, likely due to their deeper understanding of the game. Alternatively, this could also be attributed to more pragmatic factors, namely, that it is unlikely for a human to outperform a top-level human player. A thought experiment can illustrate this point: Suppose the participant is an extremely strong player, such as the current world chess champion. In this scenario, the participant would find it highly unlikely that his opponent is human, simply because there are no human players stronger than him.

A relatively simple way to make a computer play chess at a lower level, thereby giving humans a greater chance of winning, is to have the computer play at a lower level, or as pointed out by Shannon (1950, p. 272), "*the strength of the play can be easily adjusted by changing the depth of calculation and by omitting or adding terms to the evaluation function*". The method applied in StockfishL is to occasionally make a lower-ranked move (Stockfish, 2023b), resulting in the engine displaying a mix of superhuman skills with occasional illogical inaccuracies. This comes across as somewhat human-like, as shown by our results (Fig. 4A & D), but human players are still able to recognize that it is not a human because the suboptimal moves made often do not correspond to typical human mistakes. During the experiment, participants remarked that the moves of StockfishL were sometimes strange and inexplicable.

Maia is a neural network-based chess engine trained on a large number of human games to replicate the move choices of players, including typical human decisions and mistakes (McIlroy-Young et al., 2020). This approach represents a fundamentally different way of making mistakes compared to StockfishL. Our participants generally believed that Maia was a human player (Fig. 4D). Moreover, stronger players were more likely to hold this belief (see Table 2), which means that stronger players were more often *incorrect*. Previous research shows that chess experts distinguish themselves from beginners by their ability to quickly recognize and analyze complex positions due to well-organized knowledge, also referred to as chunks and templates (e.g., Chase & Simon, 1973; Gobet & Jansen, 2006). Analogously, experts should also be able to recognize unusual, i.e., non-human, moves more easily. A plausible explanation for the stronger players' judgments of Maia is that these participants were better at recognizing human-like chess moves, yet did not realize that an engine could also play chess like a human.

In summary, this study yields some interesting observations. A superhuman level of performance is often correctly recognized by participants as non-human, while players often incorrectly identify Maia, a computer agent that has been trained to perform in a human-like manner, as human. This means that Maia succeeds in deceiving players, a finding that resembles studies on ChatGPT, a chatbot trained on human text that appears to (almost) pass a Turing test (Buz et al., 2024; Jones & Bergen, 2024; Kovács, 2024). Furthermore, we showed that an unusual move or blunder can be characterized by different participants as either a human error or an attempt by a computer to play at a human-like level. This finding corresponds to the studies by Bazilinskyy et al. (2021) and Warwick and Shah (2015), who showed, in the contexts of automated driving and chatbots, respectively, that human judgments in Turing tests are not necessarily dependent on the agent's output or behavior but rather on people's expectations regarding human or computer-based performance.

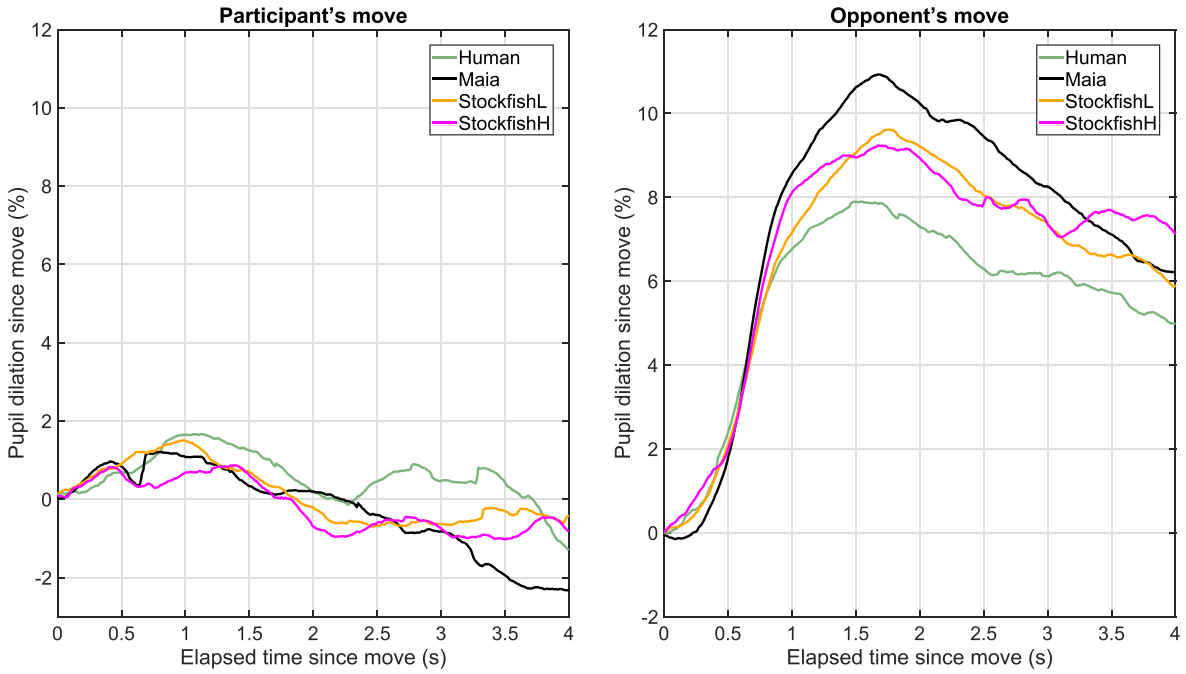


**Fig. 5.** Average pupil diameter change as a function of elapsed time, relative to the moment the move was made, for moves by the participant (left) and by the opponent (right), split into the four experimental conditions. This figure is based on 21 of the 24 participants (for 3 of the 24 participants, the eye-tracking data could not be unambiguously linked to the move data). The figure is based on a total of 6303 moves; for 1087 moves of the 21 participants, no pupil diameter data was available (these counts are combined for all conditions, and both participant and opponent moves).

The game of chess provides a delimited domain to test the implications of superhuman intelligence, as there is only a limited number of moves possible at any given moment, thereby restricting the degree of information transfer. When we extrapolate the current findings to broader applications of AI, such as chatbots or humanoid robots, several thoughts arise. If these AI systems want to pass as human, they must avoid giving themselves away by displaying their superhuman abilities, and they must also refrain from conspicuously failing or stumbling. To be perceived as human, AI agents must exhibit human-like behaviors, which can be achieved by training their neural networks to replicate typical human actions and responses. Additionally, human expectations play a role, making it more likely for a person to perceive an AI system as a human when the AI's imperfections are not interpreted as a stereotypical computer gimmick. In a future era of superhuman AI, passing the Turing test, as explained, is no longer a test of computer intelligence, but primarily a test of whether a computer with superhuman intelligence can adapt itself in such a way that it mimics human behavior.

A follow-up question that can be asked is whether AI should actually behave in a human-like manner or rather in a superhuman manner. The answer to this question will depend on the user's objectives. For example, humans might want to rely on superhuman intelligence for the evaluation or preparation of games, in order to improve their skill (Gaessler & Piezunka, 2023; Shin et al., 2023). However, humans may want to play against a human-like AI when the goal is to learn to exploit human mistakes. This question is also relevant for other domains, such as automated driving. Current automated vehicles are trained to drive in a human-like manner to meet the expectations of passengers and other road users. However, such an approach is not necessarily time-optimal; it might be more efficient to remove all traffic lights and let cars coordinate among themselves which gaps in the traffic they accept (Tonguz, 2018). In short, whether it is desirable for a computer to pass the Turing test, i.e., to mimic human behavior, is a subject of further discussion. It should also be noted that human players prefer to play against each other rather than against an engine due to the psychological dimension involved (Kulikov, 2020).

### 4.1. Limitations and recommendations

One limitation of the current study is the small sample size of 24 chess players. Another limitation is the disparity in skill levels among the participants (mean rating of 1564), the human opponent (rating of 1136), Maia (trained on games of players rated around 1100), StockfishL (rated slightly higher than the human participants), and StockfishH (playing at a level surpassing the human world champion). These skill differences may have led participants to identify stronger opponents as machines and weaker opponents as humans. While the varying skill levels produced interesting results, they also made it difficult to statistically distinguish between human-like playing styles and skill levels. Future research could perform Turing tests in a more controlled condition, by using participants, experimenters, and engines with comparable Elo ratings.

In the past, chess engines primarily relied on brute force tree searching algorithms (minimax algorithm, alpha-beta pruning), combined with handcrafted evaluations (e.g., Kasparov, 2017). This has changed in recent years with the integration of neural networks into Stockfish, allowing for better evaluation of candidate moves. Chess engines such as Leela Chess Zero (Lc0; e.g., Jenner et al., 2024) and Maia fully utilize neural networks to evaluate positions, with the latter being trained on human games, including typical human errors. An interesting topic for further research could be to investigate, from a psychological perspective, what typical human mistakes are and *why* players of varying strengths make these mistakes. Additionally, it would be interesting to explore the limits of Maia: in what cases does Maia, despite being trained on human games, make a different prediction than a human player would? This difference might be related to the player's personality or current state (e.g., nervousness, limited time on the clock). To address such questions, game positions could be extracted from online databases where Maia, Stockfish, and humans made different moves, and these positions could be presented to human chess players for in-depth reflection.

Another limitation of the current study is that we devoted considerable attention to collecting eye-tracking data to gain insight into the

visual-cognitive mechanisms underlying the execution of Turing tests. However, when facing a stronger opponent (such as StockfishH), participants often found themselves in more losing positions and spent more time thinking about their moves. These differences made it challenging to compare eye movements between the four conditions. Furthermore, we employed an eye-tracker in a configuration without a stabilizing head support, which negatively affected eye-data availability and accuracy. Nevertheless, upon exploring the data, we identified an interesting phenomenon in pupil diameter. It appears that directly after the opponent made a move, the participant's pupil dilated (Fig. 5, right panel). This pattern matches the increased pupil diameter observed when participants are confronted with cognitive tasks such as multiplication problems (De Winter et al., 2021). Further research on pupil diameter changes in chess is recommended to better understand fluctuations in cognitive load during play.

### 4.2. Conclusion

This study found that participants could correctly identify StockfishH (the most powerful chess engine available) as a computer opponent in 75% of the games. This was likely because of its sophisticated moves, which seemed non-human, along with the fact that participants typically lost the game. More skilled players were better at recognizing StockfishH as a computer, likely due to their ability to detect high-level play. StockfishL, a weaker version of Stockfish, was also often identified as a computer due to its combination of logical and sometimes illogical moves. On the other hand, Maia, a neural network-based chess engine designed to imitate human-like mistakes, was often mistaken for a human opponent, particularly by stronger players. In short, when AI plays at a level that includes human-like errors, it can trick expert players into thinking it is human.

In conclusion, the Turing test in chess relies not just on how well an AI plays but also on how convincingly it mimics human errors. A machine that displays superhuman ability too clearly is easy to recognize as non-human. The experiment also showed that participants' judgments are shaped by their expectations of human versus machine behavior. For example, a certain mistake could be deemed computer-like (because the player thinks only a computer would be programmed to make it) or human-like (because the player believes computers do not usually make mistakes). In other words, passing a Turing test is not just about AI intelligence; it is also about how we perceive intelligence.

These findings could apply to areas beyond chess. For example, in the development of chatbots or robots, AI may need to display human-like flaws to be perceived as a human. The experiment also raised the question of whether AI should aim to mimic human behavior or use its superhuman potential.

## CRediT authorship contribution statement

**Yke Bauke Eisma:** Writing – review & editing, Supervision, Methodology, Data curation, Conceptualization. **Robin Koerts:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Joost de Winter:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

MATLAB scripts and raw data are available at https://doi.org/10.4121/25142e2b-9c97-4002-8fc2-c9a4eac17cb8.

## References


2700chess. (2024). Live chess ratings. https://2700chess.com/records.

Barrish, D., Kroon, S., & Van der Merwe, B. (2024). Making superhuman AI more human in chess. In M. Hartisch, C. H. Hsueh, & J. Schaeffer (Eds.), *Advances in computer games. ACG 2023* (pp. 3–14). Cham: Springer. https://doi.org/10.1007/978-3-031-54968-7_1

Bazilinskyy, P., Sakuma, T., & De Winter, J. (2021). What driving style makes pedestrians think a passing vehicle is driving automatically? *Applied Ergonomics, 95*, Article 103428. https://doi.org/10.1016/j.apergo.2021.103428

Buz, T., Frost, B., Genchev, N., Schneider, M., Kaffee, L.-A., & De Melo, G. (2024). Investigating wit, creativity, and detectability of Large Language Models in domain-specific writing style adaptation of Reddit's showerthoughts. *arXiv*. https://doi.org/10.48550/arXiv.2405.01660

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

ChessGoals. (2024). Rating comparisons. https://chessgoals.com/rating-comparison.

Ciardo, F., De Tommaso, D., & Wykowska, A. (2022). Human-like behavioral variability blurs the distinction between a human and a machine in a nonverbal Turing test. *Science Robotics, 7*, Article eabo1241. https://doi.org/10.1126/scirobotics.abo1241

computerchess. *CCRL 40/15 rating list*, (2024). https://computerchess.org.uk/ccrl/4040/index.html.

De Winter, J., Hancock, P. A., & Eisma, Y. B. (2024). ChatGPT and academic work: New psychological phenomena. *ResearchGate*. https://www.researchgate.net/publication/375742431_ChatGPT_and_academic_work_New_psychological_phenomena.

De Winter, J. C. F., Petermeijer, S. M., Kooijman, L., & Dodou, D. (2021). Replicating five pupillometry studies of Eckhard Hess. *International Journal of Psychophysiology, 165*, 145–205. https://doi.org/10.1016/j.ijpsycho.2021.03.003

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education, 49*, 363–375. https://doi.org/10.1080/02602938.2023.2241676

French, R. M. (1990). Subcognition and the limits of the Turing test. *Mind, 99*, 53–65. https://doi.org/10.7551/mitpress/6928.003.0028

Gaessler, F., & Piezunka, H. (2023). Training with AI: Evidence from chess computers. *Strategic Management Journal, 44*, 2724–2750. https://doi.org/10.1002/smj.3512

Gobet, F., & Jansen, P. J. (2006). Training in chess: A scientific approach. In T. Redman (Ed.), *Chess and education: Selected essays from the Koltanowski conference* (pp. 81–97). Dallas: TX: Chess Program at the University of Texas at Dallas.

Gonçalves, B. (2023). Can machines think? The controversy that led to the Turing test. *AI & Society, 38*, 2499–2509. https://doi.org/10.1007/s00146-021-01318-6

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). Human or not? A gamified approach to the Turing test. *arXiv*. https://doi.org/10.48550/arXiv.2305.20010

Jenner, E., Kapur, S., Georgiev, V., Allen, C., Emmons, S., & Russell, S. (2024). Evidence of learned look-ahead in a chess-playing neural network. *arXiv*. https://doi.org/10.48550/arXiv.2406.00877

Jones, C. R., & Bergen, B. K. (2024). People cannot distinguish GPT-4 from a human in a Turing test. *arXiv*. https://doi.org/10.48550/arXiv.2405.08007

Kasparov, G. (2017). *Deep thinking: Where machine intelligence ends and human creativity begins*. New York: PublicAffairs. Hachette Book Group.

Kovács, B. (2024). The Turing test of online reviews: Can we tell the difference between human-written and GPT-4-written online reviews? *Marketing Letters*. https://doi.org/10.1007/s11002-024-09729-3

Krol, M. (1999). Have we witnessed a real-life Turing test? *Computer, 32*, 27–30. https://doi.org/10.1109/2.751325

Kulikov, V. (2020). Preferential engagement and what can we learn from online chess? *Minds and Machines, 30*, 617–636. https://doi.org/10.1007/s11023-020-09550-7

Lambert, F. (2024). Tesla Full Self-Driving Beta v12 finally rolls out – get your minds blown. https://electrek.co/2024/03/16/tesla-full-self-driving-beta-v12-finally-rolls-out.

Lichess.org. (2023a). lichess.org. https://lichess.org.

Lichess.org. (2023b). Lichess accuracy metric. https://lichess.org/page/accuracy.

McIlroy-Young, R., Sen, S., Kleinberg, J., & Anderson, A. (2020). Aligning superhuman AI with human behavior: Chess as a model system. *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1677–1687). Virtual Event, CA. https://doi.org/10.1145/3394486.3403219

Michie, D. (1993). Turing's test and conscious thought. *Artificial Intelligence, 60*, 1–22. https://doi.org/10.1016/0004-3702(93)90032-7

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61–64. https://doi.org/10.20982/tqmp.04.2.p061

OpenAI. (2024). Speech to text. https://platform.openai.com/docs/guides/speech-to-text.

Rosemarin, H., & Rosenfeld, A. (2019). Playing chess at a human desired level and style. In *Proceedings of the 7th international conference on human-agent interaction* (pp. 76–80). Kyoto, Japan. https://doi.org/10.1145/3349537.3351904.

Sejnowski, T. J. (2023). Large language models and the reverse Turing test. *Neural Computation, 35*, 309–342. https://doi.org/10.1162/neco_a_01563

Shannon, C. E. (1950). XXII. Programming a computer for playing chess. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, 41*, 256–275. https://doi.org/10.1080/14786445008521796

Shin, M., Kim, J., Van Opheusden, B., & Griffiths, T. L. (2023). Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proceedings of the National Academy of Sciences, 120*, Article e2214840120. https://doi.org/10.1073/pnas.2214840120

Solon, N. (2024). Suspicions, statistics and standoffs. *New in Chess, 2*, 12–24.
SR Research. (2023). Weblink. https://www.sr-research.com/weblink.
SR Research. (2024). EyeLink portable Duo. https://www.sr-research.com/eyelink-portable-duo.
Stockfish. (2023a). Stockfish 16. https://stockfishchess.org/blog/2023/stockfish-16.
Stockfish. (2023b). Stockfish wiki. UCI & commands. https://disservin.github.io/stockfish-docs/stockfish-wiki/UCI-&-Commands.html.
Tabone, W., & De Winter, J. (2023). Using ChatGPT for human–computer interaction research: A primer. *Royal Society Open Science, 10*, Article 231053. https://doi.org/10.1098/rsos.231053
Tonguz, O. K. (2018). Red light, green light—no light: Tomorrow's communicative cars could take turns at intersections. *IEEE Spectrum, 55*, 24–29. https://doi.org/10.1109/MSPEC.2018.8482420
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460. https://doi.org/10.1093/mind/LIX.236.433
Warwick, K., & Shah, H. (2015). Human misidentification in Turing tests. *Journal of Experimental & Theoretical Artificial Intelligence, 27*, 123–135. https://doi.org/10.1080/0952813X.2014.921734
Wired. (2023, December 8). Why AI chess bots are virtually unbeatable (ft. GothamChess). WIRED [Video]. YouTube https://www.youtube.com/watch?v=CdFLEfRr3Qk.
Yue, A. (2024). Cheating allegations surface in high level chess. *The Science Survey*. https://thesciencesurvey.com/news/2024/03/05/cheating-allegations-surface-in-high-level-chess.