REVISTA Mundi
Engenharia, Tecnologia e Gestão
INSTITUTO FEDERAL

ON THE NEED TO INVESTIGATE CHESS ENGINES FOR EDUCATIONAL AND PSYCHOLOGICAL PURPOSES

SOBRE LA NECESIDAD DE INVESTIGAR LOS MOTORES DE AJEDREZ CON FINES EDUCATIVOS Y PSICOLÓGICOS

William M. Bart[1]

Abstract: The purpose of this study was to explore why and how chess researchers should investigate chess engines for educational and psychological purposes. The motivation for the study of chess engines for educational and psychological purposes is that such study will facilitate chess training and learning. A chess engine is software that can play chess. Examples of chess engines are STOCKFISH 14 and ALPHAZERO. Chess engines tend to be used for entertainment, recreation, and sport. But chess engines have the potential to be used for educational and psychological purposes in settings such as schools, colleges, universities, senior settings, and correctional facilities. However, certain issues regarding chess engines need to be investigated to allow chess engines to be more fully used for educational and psychological purposes. For example, there are chess engines that are assumed to be at certain levels of chess skill (i.e., Elo ratings), but the assumed Elo ratings for the chess engines have not been validated. Empirical research is needed to validate the assumed Elo ratings for various chess engines. Also, some chess engines are able to identify chess blunders. The blunders committed by chess engines and by human chess players should be analyzed and interpreted. Such blunder analysis can then inform efforts to improve chess teaching and practice for human chess players at various levels of chess skill.

Keywords: Chess engines. Elo ratings. Chess blunders. Education. Chess training.

Resumen: El propósito de este estudio era explorar por qué y cómo los investigadores de ajedrez deberían investigar los motores de ajedrez con fines educativos y psicológicos. La motivación para el estudio de los motores de ajedrez con fines educativos y psicológicos es que dicho estudio facilitará el entrenamiento y el aprendizaje del ajedrez. Un motor de ajedrez es un software que puede jugar al ajedrez. Ejemplos de motores de ajedrez son Stockfish 14 y AlphaZero. Los motores de ajedrez tienden a ser utilizados para el entretenimiento, la recreación y el deporte. Sin embargo, los motores de ajedrez tienen el potencial de ser utilizados con fines educativos y psicológicos en entornos como escuelas, colegios, universidades, centros de la tercera edad y centros penitenciarios. Sin embargo, es necesario investigar ciertas cuestiones relativas a los motores de ajedrez para que éstos puedan utilizarse más plenamente con fines educativos y psicológicos. Por ejemplo, hay motores de ajedrez que se supone que tienen ciertos niveles de habilidad ajedrecística (es decir, clasificaciones Elo), pero las clasificaciones Elo asumidas para los motores de ajedrez no han sido validadas. Se necesita una investigación empírica para validar las clasificaciones Elo asumidas para varios motores de ajedrez. Además, algunos motores de ajedrez son capaces de identificar errores de ajedrez. Los errores cometidos por los motores de ajedrez y por los ajedrecistas humanos deben ser analizados e interpretados. Este análisis de los errores puede servir de base para mejorar la enseñanza y la práctica del ajedrez por parte de los ajedrecistas humanos en los distintos niveles de habilidad ajedrecística.

Palabras-clave: Máquina de ajedrez. Clasificaciones Elo. Errores en el ajedrez. Educación. Entrenamiento de ajedrez.

[1] Ph.D., University of Chicago, director of the Thinking Lab at the University of Minnesota, bartx001@umn.edu

1 INTRODUCTION

The purpose of this study was to explore the reason why chess researchers should investigate chess engines for educational and psychological purposes and how they may do such investigation. The motivation for the study of chess engines for educational and psychological purposes is that such study will facilitate chess training and learning.

This report begins with a definition of a chess engine. Chess engines are chess software that can play chess. FRITZ 18 (CHESSBASE.COM, 2021) and STOCKFISH 14 (STOCKFISH team, 2021) are examples of chess engines.

Most research on chess engines has been on high-level chess engines such as STOCKFISH 14 and ALPHAZERO (DEEPMIND, 2018) that have Elo ratings in excess of 3400. But there is a problem with high-level chess engines. High-level chess engines have chess skill that far exceeds the chess skill of all human chess players, even the present World Chess Champion, Magnus Carlson, with the Elo rating of 2856 as of December, 2021 (INTERNATIONAL CHESS FEDERATION, 2021). Most research on chess engines is research on very high-skill chess engines such as ALPHAZERO (SILVER, et al., 2018). The application of high-powered chess engines to the study of human players is neither evident nor clear, because most human players have much lower levels of chess skill than high-level chess engines.

Research on lower-skill chess engines is suggested as it has potential to reveal interesting findings in the study of chess. Lower-level chess engines have greater utility for educational and psychological programs as they can have chess skills that approximate the chess skills of human chess players.

Following is an example of how useful a lower-level chess engine can be. Let us assume that a chess engine X has an approximate Elo rating of 1100.Then one could assert that a human chess player has an Elo rating of at least 1100 if the human chess player can defeat chess engine X playing both with the White pieces and then with the Black pieces.

Lower-level chess engines have thus the potential to measure the chess skills of humans and thus to serve as tests of chess skills. Lower-level chess

engines can serve as measures of human chess skill as meter sticks can serve as measures of human height.

If lower-level chess engines have utility in educational and psychological programs, where are they? Many lower-level chess engines can be found in chess-related websites, such as CHESS.COM (2021) and LICHESS.ORG (2021). PLAYMAGNUS.COM (2021) has 29 chess engines, most of which are weakened versions of one powerful chess engine, GLAURUNG (LINUSLINKS, 2021). CHESS.COM has 25 chess engines, most of which are weakened versions of STOCKFISH 14. LICHESS.ORG has 8 chess engines, most of which are weakened versions of STOCKFISH 14. All of these 52 chess engines are freely available to chess researchers.

Many of these chess engines are given Elo ratings. For example, chess engine 6 is said to have an Elo rating of 1000. There is no evidence provided to support that assertion. In fact, there is no empirical evidence in published research to support the Elo rating of any of the chess engines that are weakened versions of either GLAURUNG or STOCKFISH 14.

To validate the Elo ratings of the 51 chess engines that are weakened versions of either GLAURUNG or STOCKFISH, empirical research with human chess players with comparable Elo ratings is required. One way to validate the assumed Elo rating of a chess engine is to have the chess engine play games against either a human chess player with the same but validated Elo rating or a chess engine with the same but validated Elo rating. The games should result in mostly draws with perhaps a few wins and losses. However, a relatively simple statistical analysis of the results should support the hypothesis that the two chess players have a similar probability of winning or drawing when playing against each other. With the use of such games and the subsequent analysis of the results, the assumed Elo ratings of chess engines could be validated.

At present, there is no published evidence that the various available chess engines have the Elo ratings ascribed to them. From a measurement point of view, we can only assume that the chess engines at each of those public websites form ordinal scales. However, even that assumption could be subjected to empirical analysis.

One way to test the hypothesis that one chess engine to be simply called A has a higher Elo rating than other chess engine to be simply called B is to have A play chess games against B with the two chess engines alternating piece colors in games. If chess engine A defeats chess engine B in all of the games, then one may conclude that chess engine A has a higher Elo rating than chess engine B. However, such a series of games and the subsequent statistical analysis of game results provide no information as to the validity of any specific Elo rating for either chess engine. If one chess engine defeats another chess engine in a series of, for example, 18 games, then one can only assert that the former chess engine has an Elo rating that is superior to the Elo rating of the latter chess engine.

2 A STUDY WITH CHESS ENGINES

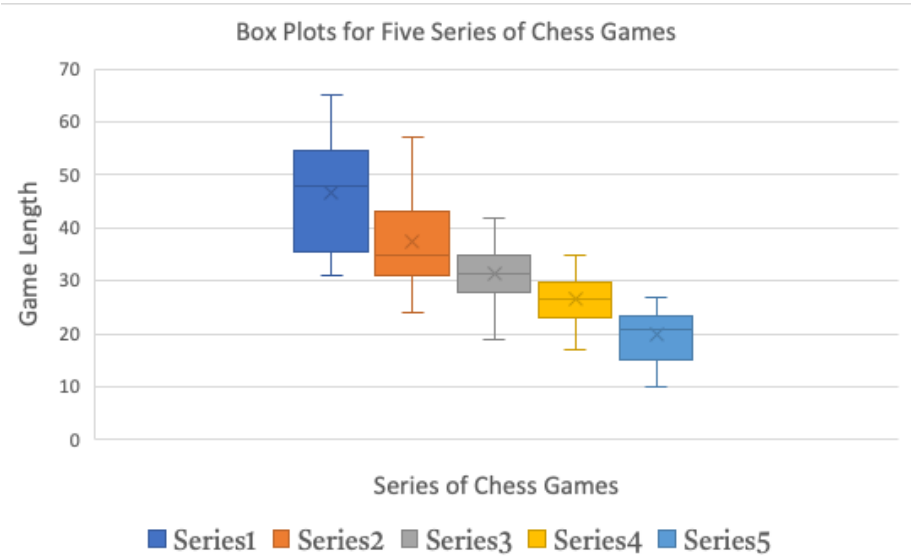
Research involving chess engines with levels of chess skill that are typical of average or below-average human chess players is rare. One example is a recent study by BART, RITTER, and RITTER (2021) that produced some interesting findings. That study will be reviewed not only for its findings but also what it suggests for future studies with chess engines with levels of chess skill that are comparable to levels of chess skill of a wide range of human chess players.

In the study by BART, RITTER, and RITTER (2021), there were six participants: PLAY MAGNUS Age 10, 9, 8, 7, and 6 and CHESS.COM Level 10 from an earlier version of CHESS.COM that was comparable to chess engine 19 in the present version of CHESS.COM. The procedure was the following: CHESS.COM Level 10 played 18 games against each of the five artificial opponents, PLAY MAGNUS Age 10, 9, 8, 7, and 6, with 9 games
as White and 9 games as Black. Five series of 18 games were played: (1) Series 1, CHESS.COM Level 10 vs PLAY MAGNUS Age 10; (2) Series 2, CHESS.COM Level 10 vs PLAY MAGNUS Age 9; (3) Series 3, CHESS.COM Level 10 vs PLAY MAGNUS Age 8; (4) Series 4, CHESS.COM Level 10 vs

PLAY MAGNUS Age 7; and (5) Series 5, CHESS.COM Level 10 vs PLAY MAGNUS Age 6. The number of moves in each game was recorded.

The games between the artificial chess opponents involved two computers. One computer was used to operate the CHESS.COM Level 10 chess engine and another computer was used to operate the appropriate PLAY MAGNUS chess engine. The same chess position would be displayed on both computers. When one chess engine made a move on one computer that may be called computer A, that move would be entered on the other computer that may be called computer B. After the chess position on computer B is changed as a result of the chess move being entered into computer B, the chess engine operating on computer B would make a chess move on the chess position displayed on computer B and that chess move would be entered into computer A and change the chess position displayed on computer A. That would lead to the chess engine operating on computer A to make a chess move on the game board displayed on computer A and that chess move would be entered into computer B that would prompt a chess move by the chess engine operating on computer B to be made on the chess position displayed on computer B and so forth. This procedure involving two computers was the procedure that was used in playing the games between the artificial chess participants.

Figure 1 - Box plots of game lengths of five series of chess games.



Source: From the author.

Figure 1 displays the box plots of the game lengths in the five series of games with game length being the number of moves in a game. There are certain interesting features in Figure 1 that are quite evident. One, the range of game lengths decreases from Series 1 to Series 5. In other words, the game lengths decrease in their range as the levels of chess skill of the chess engines become more similar. Two, game lengths tended to decrease as the levels of the chess skills of the chess engines become more similar.

A statistical analysis of the data resulted in certain interesting findings. One, the chess engine CHESS.COM Level 10 won all 90 games. Two, the length of a game in terms of the number of moves tended to decrease as the age level of the PLAY MAGNUS artificial player decreased. Three, the variance regarding the number of moves in a game decreased as the age level of the PLAY MAGNUS artificial player decreased. Four, the chess engines PLAY MAGNUS Age 10-6 formed an ordinal scale.

The major finding of the study was arguably the following. Using chess engines from CHESS.COM and PLAY MAGNUS.COM, BART, RITTER, and RITTER (2021) found that the length of a chess game in terms of number of moves in a game is related to the disparity in chess skill for the chess engines (artificial chess players). The BART, RITTER, & RITTER (2021) provides an example of empirical research that can occur with lower-level chess engines.

Research with chess engines with human-like chess skills has certain advantages. One, the participants in empirical studies involving lower-level chess engines are likely freely available, as the participants are likely chess engines freely available on chess-related websites. As a result, empirical research with lower-level chess engines as participants is likely relatively inexpensive and convenient. Two, there is no need for a review by an Institutional Review Board concerned with the ethical treatment of human participants, because the participants are artificial. Institutional Review Board reviews of studies involving human participants can take weeks to complete. As a result, empirical research with lower-level chess engines can proceed on schedule; whereas, comparable research with human participants can often experience delays in study proposal approval. Three, empirical research

involving chess engines with human-like chess skills tend to be more readily replicable than empirical research involving human participants. Replication of results is of major concern in contemporary science. These advantages support the view that empirical research with lower-level chess engines is an efficient form of inquiry.

Invariably when chess engines play chess games, there will be cases where one chess engine loses against another chess engine. But why did one chess engine lose to another chess engine? The loss could be due to a blunder committed by the losing chess engine. The analysis of blunders requires the use of another component of chess engines and that component is the evaluation function.

3 RESEARCH WITH THE EVALUATION FUNCTION

Every chess engine has an evaluation function. An evaluation function takes a chess position as input and produces a number that is an estimate of who is winning and by how much. For example, an evaluation function value of +1.5 indicates that White is winning by approximately 1.5 pawns. An evaluation function value of - 2.75 indicates that Black is winning by 2.75 pawns.

Evaluation function values are used to evaluate the quality of chess moves and are important in the analysis of chess games. The analysis of chess games with the use of chess engines is a rather complex, technical domain as the analysis of chess games can vary depending upon the chess engine and evaluation function being used in the analysis.

This topic also includes the issue of how to represent a chess position or the chess game to be analyzed. The Forsyth-Edwards Notation (FEN) is a standard notation used to describe a particular chess position in a chess game (CHESS TERMS, 2022a). The Portable Game Notation (PGN) is a notation used to describe an entire chess game (CHESS TERMS, 2022b). Using FEN, chess positions can be stored and analyzed by a chess engine. Using PGN, whole chess games can be stored and then analyzed by a chess engine.

The topic of chess game analysis is indeed quite complex. As a simple example of how an analysis of a chess game can proceed, please consider the following case. First, go to the Analysis Board on CHESS.COM, https://www.chess.com/analysis, on a computer. Second, click on the Analysis button on the top right side of the screen. Third, please make the moves for the Fool's Mate and notice the evaluation function values presented on the right side of the screen. For the move f4 for White, the top evaluation function value is -0.37 indicating a slight advantage for Black. For the move e6 for Black, the top evaluation function value is +0.04 indicating an equal position. For the next move g4 for White, the top evaluation function value is -M1 indicating that g4 is a horrendous move leading to a checkmate with the next move for Black, Qh4. The chess engine being used in this case was an advanced version of STOCKFISH with NNUE. Using the Analysis Board for CHESS.COM, the moves of any chess game can be entered and analyzed for their quality by a powerful STOCKFISH chess engine and a formidable evaluation function software.

The evaluation function used by STOCKFISH 14 is NNUE, Efficiently Updatable Neural Networks, which is an example of artificial intelligence (STOCKFISH team, 2020). This evaluation function is available on CHESS.COM to assess the quality of moves. Using NNUE, moves that are blunders, mistakes, inaccuracies, and best moves are identified. In the case that was previously cited, the second move by White, g4, is an example of a profound blunder.

4 BLUNDER ANALYSIS

Both CHESS.COM and LICHESS.ORG use NNUE, an advanced evaluation function associated with STOCKFISH 14 to analyze game moves and positions. That advanced evaluation function can identify blunders, mistakes, inaccuracies, and best moves. STOCKFISH 14 using NNUE can

identify a blunder and a best move but cannot explain why a move is a blunder or why a best move is a best move.

As an example of blunder analysis, let us consider what could occur in a chess training intervention in which human participants periodically play artificial chess opponents. Let us also consider what would result if the human participant makes the basic moves for a Fool's Mate. To illustrate this situation, please first go to CHESS.COM and click the Play Computer button on the right side of the screen. Then, go to the right side of the screen and select an artificial opponent with an intermediate Elo rating such as Isabel with an Elo rating of 1600. Next, click the Choose button on the bottom right side of the screen. Then, click the White King button and the Friendly button on the right side of the screen and then click the Play button also on the right side of the screen. The first move for White is f4 followed by the move d5 for Black. The second move for White is g4 followed by the move e5 for Black. The move g4 is given an evaluation function value of -2.52 that indicates that that move is a blunder (i.e., a serious mistake). The third move for White is a3 that is quickly followed by the move Qh4# (checkmate) for Black. On the right side of the screen, there is an indication of 1 Mistake and 1 Blunder in the game. In this game, the move a3 for White was both a blunder and a mistake for White that resulted in the game winning move of Qh4# for Black.

There are several purposes of blunder analysis. One, such blunder analysis done on games between human chess players and on games between human chess players and artificial chess players could identify the nature of blunders committed by human chess players at various skill levels. Two, blunder analysis could provide guidance regarding the instruction that is needed for human chess players who commit blunders at various parts of the game: opening, middle game, or endgame.

5 FINAL CONSIDERATIONS

BART and his colleagues are playing games involving chess engines comparable to the games played by BART, RITTER, and RITTER (2021) to

identify blunders, when they occur (e.g., opening, middle game, or endgame), and what they are (e.g., hanging a piece). BART and his colleagues are expecting that lower-level chess engines tend to commit blunders earlier in games than chess engines with higher chess skill. Preliminary analysis of blunders committed by lower-level chess engines indicate that many of the blunders occurred in the later phases of the opening or the middle game and that many of the blunders involved hanging a piece (i.e., exposing an unprotected piece for possible capture). A more complete analysis of the blunder data will reveal a more complete picture of chess blunders and what they are.

Empirical research with lower-level chess engines has the capacity to reveal interesting findings regarding chess and chess play. For example, lower-level chess engines can be used as measures of chess skill for human chess players. However, many lower-level chess engines have hypothetical Elo ratings and that warrants the implementation of empirical research to validate the hypothetical Elo ratings ascribed to the various chess engines to be used in subsequent empirical research.

Another line of inquiry to be used with lower-level chess engines is blunder analysis. Blunder analysis of games played by human chess players could inform chess trainers as to what types of instruction are best suited to correct blunders committed by their chess students and how best to improve the chess skills of their chess students. Studies of lower-level chess engines including blunder analysis have the potential to improve chess programs for educational and psychological purposes.

From a consideration of these ideas, certain recommendations are presented. One, use chess technology to teach chess. Two, use chess engines to research chess. Three, do blunder analysis to understand blunders in chess games and to improve chess training programs. The scientific study of chess engines has substantial potential to contribute to theory and practice in education and psychology.

REFERENCES

BART, W., RITTER, J., & RITTER, N. An empirical study of artificial participants: A factor affecting game length in chess. **Journal of Technological Advancements**, v. 1, n. 1, p. 1-11, 2021. DOI: 10.4018/JTA.20210101.oa2

CHESS.COM. Play vs, 2021. https://www.chess.com/play/computer

CHESSBASE.COM. Fritz 18: **Evolving genius**, 2021, October 14.

CHESS TERMS. **Forsyth-Edwards Notation (FEN)**, June, 2022a. https://www.chess.com/terms/fen-chess

CHESS TERMS. **Portable game notation (PGN)**, June, 2022b. https://www.chess.com/terms/chess-pgn https://en.CHESSBASE.com/post/fritz-18-evolving-genius

DEEPMIND. **AlphaZero: Shedding new light on chess, shogi, and Go**, 2018, December 6. https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

LINUSLINKS. **Glaurung – UCI chess engine**, 2021. https://www.linuxlinks.com/glaurung/

INTERNATIONAL CHESS FEDERATION. **Top 100 chess players**, 2021, December. https://ratings.fide.com

LICHESS.ORG. **Play with the computer**, 2021. https://lichess.org/

PLAY MAGNUS.COM. **Play Magnus, 2021**. https://www.playmagnus.com/en/play

SILVER, D., HUBERT, T., SCHRITTWIESER, J., ANTONOGLOU, I., LAI, M., GUEZ, A., LANCTOT, M., SIFRE, L., KUMARAN, D., GRAEPEL, T., LILLICRAP, T., SIMONYAN, K., & HASSABIS, D. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. **Science**, v. 362, n. 6419, p. 1140–114, 2018. doi:10.1126/science.aar6404 PMID:30523106

STOCKFISH team. **Introducing NNUE evaluation**, 2020, August 7. https://stockfishchess.org/blog/2020/introducing-nnue-evaluation/

STOCKFISH team. **Stockfish 14**, 2021, July 2. https://stockfishchess.org/blog/2021/STOCKFISH-14/

Edição especial – Xadrez, Ciência & Tecnologia
Enviado em: 24 dez. 2021
Aceito em: 03 jul. 2022
Editores responsáveis: Valério Brusamolin/ Mateus das Neves Gomes