Skill Rating by Bayesian Inference

Skill Rating by Bayesian Inference

Giuseppe Di Fatta, Guy McC. Haworth and Kenneth W. Regan

***Abstract*— Systems Engineering often involves computer modelling the behaviour of proposed systems and their components. Where a component is human, fallibility must be modelled by a stochastic agent. The identification of a model of decision-making over quantifiable options is investigated using the game-domain of Chess. Bayesian methods are used to infer the distribution of players' skill levels from the moves they play rather than from their competitive results. The approach is used on large sets of games by players across a broad FIDE Elo range, and is in principle applicable to any scenario where high-value decisions are being made under pressure.**

I. Introduction

A PILOT attempts to land in marginal conditions. Multiple agencies work furiously on a major emergency. A student progresses his learning with less than total awareness, motivation or organisation. The combined pressure of events, real-time, partial information, problem complexity, and limitations on human (and computer) resources may cause the human component to perform fallibly, short of the utopian agent in the 'How To' manual. To model such systems effectively, it is necessary to model fallible decision making.

Cognitive psychology has tried to define and explain skilled behavior such as human expertise in problem solving and decision making. Chess players' thinking ([1], [2]) has been studied for a long time and two main models have been provided. One mechanism is based on pattern recognition to access a knowledge database. The second approach is a search strategy through the problem space. The relative importance given to knowledge and quantitative search ([3], [4]) varies in the proposed theories of skilled behavior. Chess has always been a favourite demonstration domain in the fields of cognitive psychology and artificial intelligence as it is a well-documented, familiar, large, complex model-domain, many of whose aspects are subject to quantification. The Elo rating system [5] was adopted by the United States Chess Federation in 1960, by the World Chess Federation (FIDE) in 1970, and by other sports' and games' governing bodies as a suitable way to determine relative strength of participants.

This work presents a method to determine the strength of chess players that is in principle more accurate and more applicable than the system of Elo ratings. It does not require paired comparisons, on which Elo and most other rating systems are based.

The approach was originally proposed in [6], [7] with the aim of correlating computer chess engines' and human players' skills on the same scale for two reasons. First, a scale based on engine capability would be independent of time and player population, and secondly, engines and humans would be more easily compared. The present work evolves the theory and carries out the first experimental analysis.

The approach can be easily adopted in rating skilled behaviour and general types of expertise in other domains. It does not infer a relative strength of players from the outcome of games. Rather, it infers skill ratings directly from the innate quality of the decisions and independently of the competitive nature of the activity. For example, this methodology could be effectively adopted to monitor and evaluate training and education activities.

This work does not assume any specific model for the decision making process of chess players. Instead, the approach is based on the definition of a stochastic agent with parameterised skill level. The empirical evidence of players' skill is given by their chosen moves in chess games, which can be assessed using a heuristic evaluation function. A statistical inference method can be applied to a large set of data in order to map chess players' skills into a model space. In the present work a simple 1-parameter space is considered, in which modelling leads directly to rating and ranking.

Our main experimental results yield sharp, significant and self-consistent differences in the inferred skill level across medium-sized intervals of players in the Elo rating scale.

The rest of this paper is organized as follows: Section II discusses related works in skill rating in general and in Chess in particular. Section III describes the Reference Agent Space (RAS) for modelling fallibility in decision making. Section IV presents the adopted Bayesian inference method for the identification of the model from data. Experimental results in the domain of Chess are presented in Section V. Finally, Section VI summarises the paper and indicates some future research directions.

II. Skill rating

Most rating systems are based on the Bradley-Terry model for paired comparisons [8]. The assumptions of such rating systems are that the strength of a player can be described by a single value (rating) and that expected game results depend only on the difference between the ratings of the two players.

Ratings based on pairwise comparisons attempt to estimate skill levels by means of the outcomes of competitive activity involving two or more individuals. Such rating systems for competitive activities are intrinsically relative. Most rating

G. Di Fatta and G. McC. Haworth are with the School of System Engineering, University of Reading, Reading, Berkshire, RG6 6AY, UK. (corresponding author's phone +44 (0) 118-378-8221, fax +44 (0) 118-975-1994 and e-mail G.DiFatta@reading.ac.uk).

K. W. Regan is with the Department of Computer Science and Engineering, University at Buffalo, The State University of New York, 201 Bell Hall, Buffalo, NY 14260-2000, USA. He also holds the title of International Master from the World Chess Federation (FIDE).

systems fall into this category, including the most prevalent, the Elo system [5].

More generally, the direct and persistent measurement of skill is required in non-competitive domains of complex decision making where professional standards must be maintained despite the pressures of events, time constraints, partial information, problem complexity and ability.

Skill rating must take into account the fallible nature of human decisions. We consider the situation where human beings take decisions under certain constraints, such as time bounds, imprecise information and psychological conditioning. In this case, the decisions may appear to be the result of a stochastic process informed by knowledge and experience. Skill rating, then, must evaluate the quality of decisions in terms of such a stochastic process.

A. Skill Rating in Chess

The Elo system [5], perhaps the best known rating system, was originally created for Chess and later adopted in games including Scrabble and Go, and sports such as bowling, golf, table tennis, football and basketball.

Within a pool of players, Elo differences are meaningful, but Elos from different pools of players are not comparable as Elos have no absolute meaning, being also affected by the Elos being imported/exported to/from the pool by players entering/leaving[1]. They are determined from the results of games and not by the innate quality of the moves played: they therefore measure performance rather than underlying skill. There have been criticisms of the Elo approach [11] and improvements [10], [12], [13], [14] have been proposed. However, they are still results-based and affected over time by the changing player population. As such, these approaches cannot accurately determine:

- inflationary trends in ratings, changing the quality of play at a specific Elo figure,
- the relative skill of players in a specific part of the game, e.g. the 'opening' or 'endgame',
- the relative skill of contemporary players in different leagues,
- the relative skill of players of different eras,
- whether a match or game was won by good play or lost by bad play,
- whether a player is playing abnormally, suspiciously well and perhaps cheating.

In contrast, a few systems have been proposed to assess, rank and rate absolute chess skill [6], [15], [7], [16]. Authors in [15] and in [16] use only the 'error' of move-decisions to calculate mean-error: they do not use the full move-context of the decisions. Neither have fully addressed the issues of statistical confidence. The author in [6] built on the opponent fallibility work in [17], [18], [19], [20], [21] by defining a Reference Agent Space based on an infallible chess engine for the Endgame Zone (EZ), defined as that part of Chess for which Endgame Tables (EGTs) have been computed. An EGT provides the value, win/draw/loss, of a position and its depth to win if decisive: an engine E may play chess infallibly in EZ by simply extracting the data required. The author in [7] proposed an extension of the RAS concept to the whole of Chess.

This work evolves the theory of [7] and carries out the first experimental analysis based on it.

[1]The FIDE and USCF Elo for human players and SSDF Elo [9] for computers scales are said to have been affected by both deflationary and inflationary forces [10].

III. Reference Agent Space

A. Reference Chess Engine

Rational decision making requires a definite set of alternative actions and knowledge of the utilities of the outcomes of each possible action. Chess players make decisions according to individual judgment under time pressure. A player's skill is a measurement of their ability to make choices as close as possible to the optimal ones. In order to assess this, we ideally need the 'best move' benchmark but this is only available via the EGTs in the Endgame Zone. In the general case of the whole game this is clearly not feasible. However, significant advances have been made in the last decades in terms of chess engines' playing strength [9].

Given a reference chess engine E, a chess position analysis results in a list of recommended 'best' moves and their heuristic values in pawn units. The value associated with a move corresponds to the estimated advantage of the position that the move will lead to. For time constraints and for the exponential nature of the computational complexity, the analysis of the chess engine can only be performed up to a given maximum depth d (number of plies) and the number mv of alternative variants is limited as well. For the purpose of this work we consider the reference chess engine

$$E \equiv E(d, mv).$$

The analysis of a position p is a function f_E which provides a list $S_{E,p} = \{(m_i, v_i)\}$ of candidate moves m_i and their estimated values v_i ($1 \leq i \leq mv$):

$$f_E : p \rightarrow S_{E,p}.$$

Let $M_{E,p} = \{m_i\}$ be the set of candidate moves in $S_{E,p}$.

In contrast to the EZ scenario, three main factors introduce an approximation in the evaluation of a chess position in terms of candidate moves and relative values. They are the limited search depth and span (parameters d and mv) and the heuristic nature of E's position evaluation function. The influence of this approximation in our analysis needs to be properly addressed and will be the scope of further investigation.

B. Stochastic Reference Agent

To model human players' fallibility we associate a likelihood function L with engine E to create a stochastic chess engine $E(c)$. E always makes a move it sees as best in the position, but $E(c)$ can choose any move in the top mv moves with a non-zero probability defined by the function L. The requirements on c and L are that:

- $E(0)$ plays at random, giving each move an equal chance of being chosen,
- the greater c, the better $E(c)$ in the sense that the expected value after the move is better, and
- c is notionally in $[-\infty, \infty]$ but is initialised in practical computations as in $[c_{min}, c_{max}]$, where $c_{min} \geq 0$.

We define the likelihood of the move $m_i \in M_{E,p}$ being chosen by a stochastic chess engine $E(c)$ as

$$L[E(c), (p, m_i)] = (v_{max} - v_i + K)^{-c} \tag{1}$$

where $(m_i, v_i) \in S_{E,p}$, $v_{max} = max_j\{v_j\}$ and K is a constant > 0.[2]

The probability of $E(c)$ selecting the move m_i is simply given by normalizing its likelihood, viz.:

$$Prob[E(c), (p, m)] = \\ = \begin{cases} \frac{L[E(c),(p,m)]}{\sum_{m_j \in M_{E,p}} L[E(c),(p,m_j)]}, & \text{if } m \in M_{E,p}; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Given a parametric model of a fallible player, i.e. the stochastic engine $E(c)$, we can use the evidence of the moves made by a chess player and fit the model to the data.

In general, we can define a Reference Agent Space (RAS) based on one or more parameters. Given data from games of a benchmarked player BP, we can define a mapping $M : BP \rightarrow M(BP) \in RAS$.

The ultimate aim is to compare the decisions of BP and M(BP), choosing a space RAS and an agent so that the behaviours of BP and M(BP) are as close as possible. Further investigation will be devoted to the choice of the likelihood function and the generalization to a multi-dimensional RAS.

In the next section we describe the Bayesian inference method to determine a probability distribution of the parameter c. In the 1-dimensional case, the expected value of the parameter c defines a skill rating system based on the actual quality of the decisions made. The method can be applied to a player in a single game, a player in a set of games or to a set of related players in many games.

IV. Inference of the parametric model

Let us consider the event $e = (p, m)$, where m is a move made in position p. The posterior probability of the parameter c of $E(c)$, given the evidence of the move m in the position p, depends on the *a priori* probability $Prob[E(c)]$ and the conditional probability of the event e given $E(c)$.

Bayes' theorem states that:

$$Prob[E(c)|e] = \frac{Prob[E(c)] \cdot Prob[e|E(c)]}{\sum_c Prob[E(c)] \cdot Prob[e|E(c)]}. \tag{3}$$

Let us consider a set of position-move events

$$\mathfrak{E} = \{(p_i, m_i)\},$$

where $i = 1 \ldots N$, p_i is the position prior to the move m_i.

[2]In the experimental tests the constant K has been set to 0.1.

We iteratively apply the Bayesian rule in (3) to the set of events in $\mathfrak{E}$, where the *a priori* probability at step i ($i > 1$) is the posterior probability at step $i - 1$.

$$Prob[E(c)|e_i] = \\ = \frac{Prob[E(c)|e_{i-1}] \cdot Prob[e_i|E(c)]}{\sum_c Prob[E(c)|e_{i-1}] \cdot Prob[e_i|E(c)]} \tag{4}$$

where $e_i = (p_i, m_i)$ and $e_{i-1} = (p_{i-1}, m_{i-1})$.

In all tests we have set the initial *a priori* probabilities $Prob[E(c)] \equiv Prob[E(c)|e_0]$ to be a 'know nothing' constant.

The inference process produces an *a posteriori* probability distribution of the stochastic reference agent $E(c)$. The mean value of the parameter c measures the innate quality of the moves and can be used as a player's skill rating. Moreover, the variance of the probability distribution, as in [12], [14], provides a measure of the uncertainty of the rating.

A. Adaptive algorithm

We have adopted an adaptive detection of the range of c for a more efficient computation of the probability distribution. The parameters c_{min}, c_{max} and δ_c define a finite set of discrete values of the parameter c:

$$c_i = c_{min} + i \cdot \delta_c, \tag{5}$$

where $0 \leq i \leq (\frac{c_{max} - c_{min}}{\delta_c})$.

The three parameters are adjusted during execution to allow a refinement (better precision) of the values of c. An iterative process starts from a wide range $[c_{min}, c_{max}]$ with a coarse precision ($\delta_c = 0.1$). At each iteration step, the range is narrowed and the precision increased (δ_c is decreased). This results in a more efficient computation in terms of runtime and memory requirements.

V. Experimental analysis

A. Experimental setup

For the experimental analysis the following resources have been used:

- publicly available data in Portable Game Notation (PGN),
- a publicly available, reputable and widely used engine with relatively high 'engine Elo' with control of search-depth (d) and top moves (mv) accurately evaluated,
- engine input/output via the Universal Chess Interface (UCI) [22].

The decisions of a number of players of different skill levels have been analysed. From sources including the ChessBase Mega database [23] with millions of played games, we have extracted thousands of games in which both players had Elo ratings within 10 points of some Elo figure, e.g. games of players rated between 2390 and 2410. During a preprocessing phase positions were acquired from the games, ignoring the first 12 moves by each side (assumed to be 'out

of book'). The Portable Game Notation (PGN) was converted to an indexed set of events $\{e_k = (p_k, m_k)\}$, each a move m_k from a position p_k. Positions were analysed using the chess engine TOGA II 1.3.1 [24], which despite being a free product is considered one of the top ten engines in playing strength, and fully competitive with other chess engines used in studies [15], [16].

Each analysis was carried out to depth $d = 10$ plies. The TOGA II engine was configured to determine and report the top 10 moves (mv) it found in each position. Both values were chosen as compromises between computing speed and comprehensiveness of the data. Depth 10 is not considered sufficient to outplay the stronger players in our samples, but apparently it suffices to ferret out most of their inaccuracies.

Finally, the Bayesian inference process described in section IV has been applied to the preprocessed data to generate the probability distribution of the parameter c.

B. Composite reference Elo players

Games were grouped according to the Elo rating of the players. Each group contains games between players with a similar Elo rating ($ELO_{min} \leq ELO(player) \leq ELO_{max}$). The total number of games and the total number of move-events we have included in the datasets of composite reference players are given in Table I.

TABLE I
CHESS GAMES DATASETS

Dataset	Elo_{min}	Elo_{max}	Games	Events	Time period
E2100	2090	2110	217	12751	1994-1998
E2200	2190	2210	569	29611	1971-1998
E2300	2290	2310	568	30070	1971-2005
E2400	2390	2410	603	31077	1971-2006
E2500	2490	2510	636	30168	1995-2006
E2600	2590	2610	615	30084	1995-2006
E2700	2690	2710	225	13796	1991-2006

We have applied the Bayesian inference method described in section IV to each dataset of Table I. The probability distributions of the parameter c is shown in Figure 1. A summary of these distributions is provided in Table II in terms of the mean, standard deviation and 95% credibility region (CR) for c.

The expected value $\bar{c}$ measures the average quality of moves played in $\mathfrak{C}$ and for convenience we refer to it as the 'apparent skill'. The standard deviation σ_c measures the uncertainty of the apparent skill level, caused by the varying performance of the player and the finiteness of the data.

This experiment shows that the Bayesian inference approach is able to detect different skill levels among players with different Elo ratings. This also shows that the proposed skill rating system, which is based on the quality of the moves, is consistent with the FIDE Elo system.

As expected, the standard deviation and the width of the credibility region depend on the amount of training data. The 2100 and 2700 Elo datasets contain less data then the others and show slightly higher standard deviation. In the next section we analyse the effect of the amount of training data more in detail.

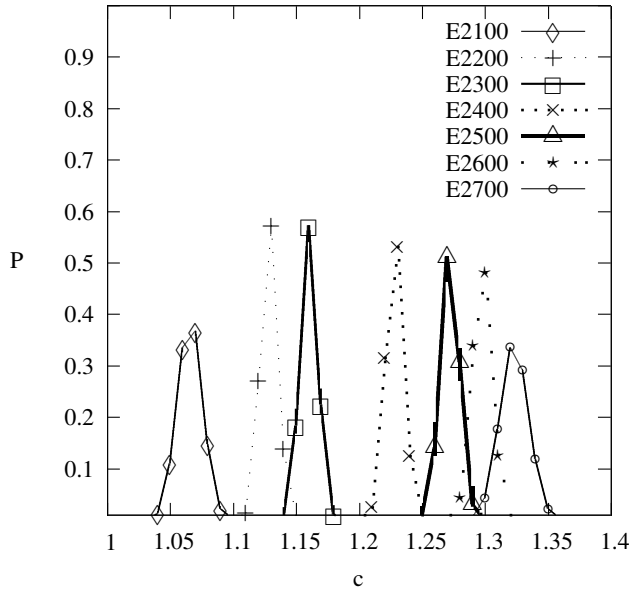


Fig. 1. The probability of the model $E(c)$ for composite reference Elo players

TABLE II
STATISTICS ON THE PROBABILITY DISTRIBUTIONS OF THE APPARENT SKILL

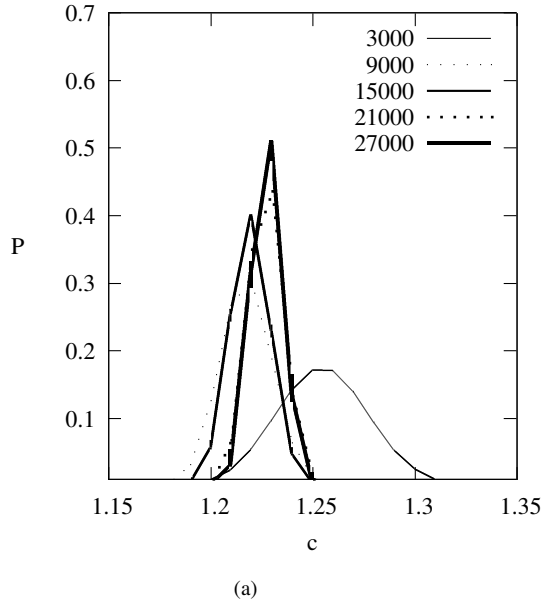
Dataset	$\bar{c}$	σ_c	CR_{min}	CR_{max}
E2100	1.0660	0.00997	1.04	1.10
E2200	1.1285	0.00678	1.11	1.15
E2300	1.1605	0.00694	1.14	1.18
E2400	1.2277	0.00711	1.21	1.25
E2500	1.2722	0.00747	1.25	1.29
E2600	1.2971	0.00770	1.27	1.33
E2700	1.3233	0.01142	1.29	1.35

C. Training analysis

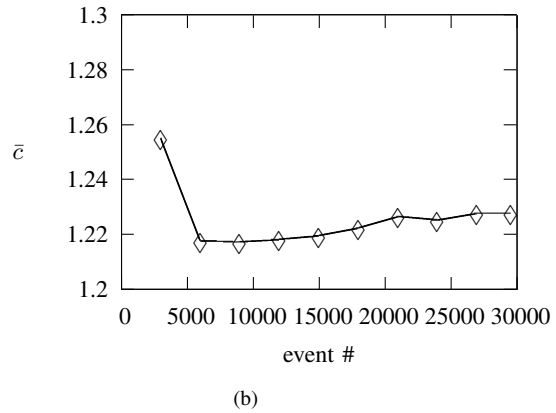
In order to verify the learning process of the probability distribution of the apparent c, we have taken snapshots at different iteration steps (i.e. number of positions). Figure 2 shows the analysis that has been carried out on the 2400 Elo data. The curves in Figure 2(a) show the evolution of the probability distribution during the refinement of the Bayesian inference process. The expected value $\bar{c}$ (Fig. 2(b)) quickly converges and the standard deviation (Fig. 2(c)) decreases as the inference process includes more data. The asymptotic value of the standard deviation is a measure of the intrinsic uncertainty of the skill level.

D. Players with similar Elo ratings

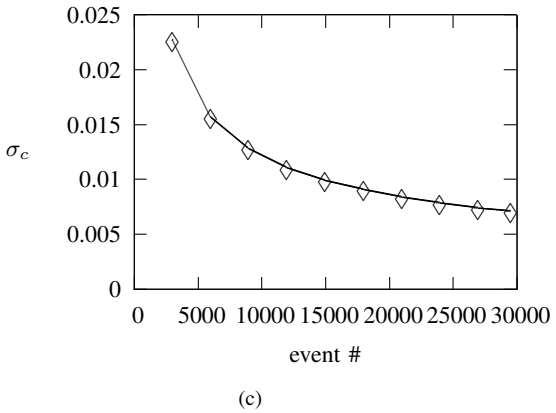
In this section we present the experimental test aimed at investigating differences of the apparent skill in single games



(a)



(b)



(c)

Fig. 2. Training evolution for the dataset 2400 Elo: a) *probability distribution at different iteration steps*, b) *expected value*, c) *standard deviation*.

between (real) players with a similar Elo rating. Note that such a difference cannot be detected by the Elo system in principle. The Elo rating captures an average performance of a player in terms of game outcomes and not in terms of the quality of the moves.

Given a set of games $\{\mathbf{G}\}$ among players with similar Elo rating (e.g. E2400), from each game we have extracted two lists of events $\mathfrak{E} = \{(p, m)\}$, one for each player, and associated each list with the outcome $(1, \frac{1}{2}, 0)$ to generate three sets of events $\mathbf{L_0}$, $\mathbf{L_1}$ and $\mathbf{L}_{\frac{1}{2}}$. The set $\mathbf{L_1}$ contains sets of move-events which have been made by players who won, $\mathbf{L_0}$ those made by players who lost and $\mathbf{L}_{\frac{1}{2}}$ those made by players who drew.

$$\{\mathbf{G}\} \Rightarrow \begin{cases} \{(\mathfrak{E}_{0k}, \mathfrak{E}_{1k})\} & \Rightarrow \begin{cases} \mathbf{L_0} = \{\mathfrak{E}_{0k}\}, & k = 1 \dots n_1 \\ \mathbf{L_1} = \{\mathfrak{E}_{1k}\}, & k = 1 \dots n_1 \end{cases} \\ \{(\mathfrak{E}_{\frac{1}{2}k}, \mathfrak{E}_{\frac{1}{2}k})\} & \Rightarrow \mathbf{L}_{\frac{1}{2}} = \{\mathfrak{E}_{\frac{1}{2}k}\}, k = 1 \dots n_2 \end{cases}$$

We have applied the Bayesian inference process to each set of events $\mathfrak{E}_{rk}$ ($r \in \{0, 1, \frac{1}{2}\}$) to obtain a probability distribution of $E(c)$. In this case, the apparent skill $\bar{c}$ measures the quality of moves played by a 'single' player during a 'single' game, with a consequent expected high uncertainty because of the limited amount of data on which the inference is carried out.

We have computed first order statistics of the apparent skill $\bar{c}$ over the three sets $\mathbf{L_0}$, $\mathbf{L_1}$ and $\mathbf{L}_{\frac{1}{2}}$. In this test we have used 602 games from the dataset E2400 ($n_1 = 313$ and $n_2 = 578$). The average apparent skill $\mu_{\bar{c}}$ over all 1204 $\mathfrak{E}_{rk}$, regardless of the result r, is 1.2109. The results over each set $\mathbf{L_0}$, $\mathbf{L_1}$ and $\mathbf{L}_{\frac{1}{2}}$ are shown in Table III.

TABLE III
ANALYSIS OF THE OPPONENT PLAYERS IN THE DATASET E2400

set	n_1/n_2	$\mu_{\bar{c}}$	$\sigma_{\bar{c}}$
$\mathbf{L_0}$	313	1.1493	0.0686
$\mathbf{L_1}$	313	1.2302	0.0623
$\mathbf{L}_{\frac{1}{2}}$	578	1.2339	0.0460

In spite of the small number of events in a single game the Bayesian approach is able to detect a meaningful difference between the players of a single game, who have similar Elo ratings. On average, players who have won the game have a higher apparent skill $\bar{c}$ than their opponents who have lost. Players who have drawn have even higher apparent skill. This can be explained considering that in drawn games both opponents have played well with no or irrelevant errors. The intrinsic quality of the game is in general higher. When a player has reached a significant advantage during the game, they may prefer to play an easy and safe strategy. They can even afford to make small errors provided the outcome is ensured. In this case there is a lack of motivation to play high quality and difficult strategies.

VI. Conclusions

The proposed approach has demonstrated the viability of rating skill in Chess by benchmarking against chess-engines, which themselves continue to improve in playing ability. This work does not assume any specific model for the decision making process of chess players. The approach is rather based on the definition of a stochastic agent and a mapping of the apparent player's skill into a Reference Agent Space based on the empirical evidence of the chosen moves.

A Bayesian inference method has been successfully applied to a large set of data. The experimental analysis has provided evidence of the accuracy of the method in estimating the skill level for players regardless of the outcome of the games and of the opponent rating. It has been shown that the probability distribution of the apparent skill is able to discriminate players' performance in different Elo ranges. The apparent skill can also discriminate the quality of the moves of opponent players with similar Elo rating during a single game.

In the demonstration domain of Chess, skill rating based on the proposed approach may be used to:

- assess a player on the basis of a single game, match, or set of games,
- assess composite reference Elo players, a composite of a number of players with similar Elo rating,
- calculate the prior probability that a player of Elo e will make a sequence of moves,
- create 'likelihood evidence' as to whether someone is being illegally informed by computer,
- assess a players learning and skill, rather than their performance, over time, and
- compare players skill, even though those players are from different eras.

Given that the 'apparent error' of the player analysed is in part affected by the errors of the reference engines, we intend to test the statistical robustness of the results by comparing the models derived from the varying position-valuations at different depths of search. Furthermore, by analysing the pattern of human error, it should be possible to derive likelihood functions that better fit the model to the data. Further work will also address the generalization of the Stochastic Reference Agent into a multi-dimensional Reference Agent Space.

In principle, the proposed methodology can be effectively adopted in other domains and we intend to do so. It can be used, for example, to analyse in real-time the likely abilities of students, skilled workers in defined-process scenarios, and high-value professional decision-making.

Acknowledgments

In addition to the references' authors, particularly those who have actively contributed to our work, we thank Paul Janota, Michael Scheidl, David McClain and Ian Bland for help along the way. We also thank the referees for their comments.

References

[1] A. De Groot, *Het denken van den schaker*. Amsterdam, Noord Hollandsche, 1946.

[2] ——, *Thought and choice in chess (2nd ed.) (Revised translation of De Groot, 1946)*. The Hague: Mouton Publishers, 1978.

[3] F. Gobet, "Chess players' thinking revisited," *Swiss Journal of Psychology*, vol. 57, pp. 18–32, 1998.

[4] F. Gobet and N. Charness, *Expertise in chess, Chess and games. Cambridge handbook on expertise and expert performance*. Cambridge, MA: Cambridge University Press, 2006.

[5] A. Elo, *The Rating of Chessplayers, Past and Present*. Arco, ISBN 0-668-04721-6, 1978.

[6] G.McC. Haworth, "Reference fallible endgame play," *ICGA Journal*, vol. 26-2, pp. 81–91, 2002.

[7] ——, "Gentlemen, stop your engines!" *ICGA Journal*, vol. 30-3, pp. 150–156, 2007.

[8] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs. I. The method of paired comparisons," *Biometrika*, vol. 39, pp. 324–345, 1952.

[9] The SSDF rating list. [Online]. Available: http://ssdf.bosjo.net/list.htm

[10] M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," *Applied Statistics*, vol. 48, pp. 377–394, 1999.

[11] J. Beasley, *The Mathematics of Games*. Dover ISBN 0-4864-4976-9, 2006.

[12] R. Herbrich, T. Minka, and T. Graepel, "$TrueSkill^{TM}$: A Bayesian skill rating system," in *Advances in Neural Information Processing Systems (NIPS 2006)*. MIT Press, 2007, pp. 569–576.

[13] P. Dangauthier, R. Herbrich, T. Minka, and T. Graepel, "$TrueSkill$ Through Time: Revisiting the History of Chess," in *Advances in Neural Information Processing Systems (NIPS 2007)*. MIT Press, 2008, pp. 931–938.

[14] R. Coulom, "Whole-History Rating: A Bayesian rating system for players of time-varying strength," in *Proceedings of the Conference on Computers and Games*, Beijing, China, 2008.

[15] M. Guid and I. Bratko, "Computer analysis of world chess champions," *ICGA Journal*, vol. 29-2, pp. 65–73, 2006.

[16] C. Sullivan. Comparison of great players, 2008. [Online]. Available: http://www.truechess.com/web/champs.html

[17] P. Jansen, "Problematic positions and speculative play," *Computers, Chess and Cognition (Eds. T.A. Marsland and J. Schaeffer), Springer-Verlag, New York.*, pp. 9–32, 1990.

[18] ——, *Using Knowledge about the Opponent in Game-Tree Search*. Ph.D. thesis, Carnegie-Mellon University, Pittsburgh, 1992.

[19] ——, "KQKR: Awareness of a fallible opponent," *ICCA Journal*, vol. 15-3, pp. 111–131, 1992.

[20] ——, "KQKR: Assessing the utility of heuristics," *ICCA Journal*, vol. 15-4, pp. 179–191, 1992.

[21] ——, "KQKR: Speculatively thwarting a human opponent," *ICCA Journal*, vol. 16-1, pp. 3–17, 1993.

[22] S. Meyer-Kahlen. Definition of the Universal Chess Interface. [Online]. Available: http://tinyurl.com/65hxat

[23] ChessBase GMBH, Mexikoring 35, D22297 Hamburg, Germany. Chessbase player database. [Online]. Available: http://www.chessbase.com

[24] T. Gaksch. Toga II 1.3.1 Chess Engine. [Online]. Available: http://www.superchessengine.com/togaii.htm