GENTLEMEN, STOP YOUR ENGINES!

G. McC. Haworth[1]

Reading, England

ABSTRACT

For fifty years, computer chess has pursued an original goal of Artificial Intelligence, to produce a chess-engine to compete at the highest level. The goal has arguably been achieved, but that success has made it harder to answer questions about the relative playing strengths of man and machine. The proposal here is to approach such questions in a counter-intuitive way, handicapping or *stopping-down* chess engines so that they play less well. The intrinsic lack of man-machine games may be side-stepped by analysing existing games to place computer-engines as accurately as possible on the FIDE ELO scale of human play. Move-sequences may also be assessed for likelihood if computer-assisted cheating is suspected.

1. INTRODUCTION

The recently celebrated Dartmouth Summer Workshop of 1956 (Moor, 2006) coined the term *Artificial Intelligence*. The AI goal most clearly defined was to create a chess engine to compete at the highest level. Moore's law plus new versions and types of chess engine such as FRUIT, RYBKA and ZAPPA, have increased the likelihood that this goal has now been reached. Ironically, recent silicon successes in man-machine play have made this claim harder to verify as there is now a distinct lack of enthusiasm on the human side for such matches, especially extended ones. Past encounters have often been marred by clear blunders[2], highlighting the unsatisfactory nature of determining the competence of *homo sapiens* by the transitory performance of one individual. Ad hoc conditions have increasingly compromised the engines, sometimes in chess-variant matches one or more removes from normal chess matches.

There is a need for a new approach to engine-rating, one which does not rely directly on unachievably large sets of man-machine games. The strategy here is based on an engine E, constrained for reasons of experiment-repeatability to be E_p, i.e., searching for p plies and then for quiescence. E_p is, to borrow a photography term, *stopped down* by *competence factor* c to be engine $E_p(c)$, choosing its move using a stochastic function $f(c)$: $E_p(\infty) \equiv E_p$. Let $EP \equiv \Sigma q_i E_p(c_i)$ be an engine in E_p-space which plays as $E_p(c_i)$ with probability q_i. The objective is to associate engines EP with various players P and levels on the FIDE ELO scale F.

Let engine E, search-depth p and player P (rated at ELO L) be chosen such that E_p manifestly plays better[3] than P. A sample set $S \equiv \{(Q_i, m_i)\}$ of moves m_i from positions Q_i characterises P. A Bayesian inference process defines a transform $T{:}P \rightarrow EP$ of P to E_p-space, analysing S to iterate towards EP (Jeffreys, 1961). ELO(EP) = $L' = L + \delta F_1(E,p,S)$. Match E_p-EP shows to arbitrary accuracy that ELO(E_p) = $L' + \delta F_2$, see Figure 1.

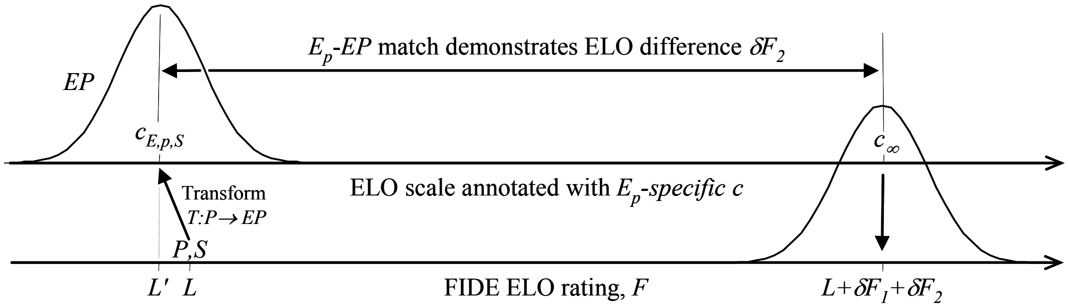

Figure 1. Mapping ELO L players to $E_p(c)$, and comparing EP and E_p.

---

[1] guy.haworth@bnc.oxon.org
[2] Kasparov (v DEEP BLUE match 2 game 6, DEEP FRITZ and DEEP JUNIOR), Kramnik (v DEEP FRITZ) mated in 1
[3] This is to reduce uncertainty, see Section 3.1, and a recommendation rather than a provably necessary condition.

Section 2 revisits the earlier *Reference Fallible Endgame Player* (RFEP) concept and reviews its use. Section 3 generalises the RFEP to the *Reference Fallible Player* (RFP) concept and warns of the fallibility of the RFP. Section 4 reviews uses of the RFP and section 5 summarises the implied experimental program.

2. THE REFERENCE FALLIBLE ENDGAME PLAYER

The *Reference Fallible Endgame Player*, RFEP (Haworth, 2002) was defined after Jansen (1992a, 1992b, 1993) suggested exploiting opponent-fallibility but did not define fallible opponents to play against.

The RFEP only plays chess when there is an Endgame Table, EGT, available in some metric. This subdomain of chess might be called the *Endgame Table Zone*, ETZ, and is currently restricted to 6^--man chess. Nalimov's Depth to Mate (DTM) EGTs are widely available but RFEPs can use other EGTs. The metrics DTC, DTZ and DTZ_{50}[4] have been seen as more useful and economical, and have been used in computations (Tamplin and Haworth, 2003; Bourzutschky, Tamplin and Haworth, 2005). RFEP R_c is assumed to have a theoretical win and to retain that win[5]. It chooses its moves in the following way:

- at position Q, R_c has n moves m_j to positions Q_j of depth d_j respectively
- $0 \le d_j$ which is arithmetically convenient: $d_1 \le d_2 \le ... \le d_n$, so move m_1 minimises DTx
- R_c is defined as follows: $q_{j,c} \equiv \text{Prob}[R_c \text{ chooses move } m_j] \propto (\kappa + d_j)^{-c}$ with $\kappa > 0$ and $\kappa = 1$ here[6]

The R_c have the following required properties:

- $c = 0$ corresponds to 'zero skill': R_0 assigns the same probability to all available moves
- R_∞ is infallible, always choosing a move minimising depth to the next win-goal
- $R_{-\infty}$ is anti-infallible, always choosing a move which maximises depth to the next win-goal
- for $c > 0$, better moves in a metric-minimising sense are more likely than worse moves
- as c increases, R_c becomes more competent in the sense that post-move E[depth] decreases
- if $d_{j+1} = d_j+1$, as $d_j \to \infty$, $q_{j,c}/q_{j+1,c} \to 1$ monotonically.[7]

A set $S = \{(Q_i, cm_i)\}$ characterises P, who chooses move cm_i in position Q_i. Profiling P in R_c terms:

- let $R = \{R_c \mid c = c_{min}(\delta c)c_{max}\}$ be the defined set[8] of candidate R_c,
- before analysing S, the *a priori* beliefs are {hypothesis$_c \equiv H_c \equiv$ "$\text{Prob}[P \text{ is } R_c \mid R_c \in R] \equiv q_{1,c}$"}
  defining an Aeolian harp of multiple hypotheses to be held up to the wind of evidence
- let $p_{i,c} \equiv \text{Prob}[R_c \text{ chooses move } cm_i \text{ in positions } Q_i]$,
- let $q_{2,c} \equiv \prod p_{i,c} = \text{Prob}[R_c \text{ chooses all moves } cm_i \text{ in positions } Q_i]$, and
- Bayesian inferences implies that the *a posteriori* $\text{Prob}[P \text{ is } R_c \mid S, R_c \in R] \equiv q_{3,c} \equiv \mu.q_{1,c}.q_{2,c}$.

Thus, set S defines a transform $T{:}P \to EP$ of player P into R_c-space: $EP \equiv \Sigma q_{3,c}R_c$ with mean $c_{P,S} \equiv \Sigma c.q_{3,c}$.[9] Haworth and Andrist (2003) reported in Section 6.4 the experimental confirmation that if in fact $P \equiv R_c$, $EP \to P$ in the limit. A conjecture here is that if $P \equiv \Sigma q_c R_c$, then $EP \to P$ in the same way.

Jansen's *random player* (1992a) is equivalent to R_0 here. Haworth (2002) analysed two KQKR demonstration games between Walter Browne (KQ) and Ken Thompson's BELLE chess engine. Browne's moves imply an *apparent competence c* of about 19 on the basis of 100 moves, but this value does not transfer to other endgames and has no meaning in other terms. Haworth and Andrist (2003) reported much lower figures for c in KBBKN: Andrist and Haworth (2005) characterised intrinsic endgame difficulty and quantified the degree to which KBBKN is harder than KQKR. Competitive human games rarely spend long in the EGT zone and there is more opportunity to analyse the apparent competence of chess engines not armed with the EGTs.

The quality of play of a losing defender may be measured in the same way (Haworth, 2002).

---

[4] Depths to Conversion, move-count Zeroing move, and DTZ in the context of the 50-move draw-claim rule
[5] There is no loss of generality: R_c*'s* handling of drawing and losing moves when having a win, and its treatment of drawn and lost positions is also defined in Haworth (2002), but is not needed for illustrative purposes here.
[6] as κ increases, the RFEP differentiates less well between better and worse moves.
[7] This is why the more obvious and mathematically simpler $q_{j,c} \propto 1/e^{\wedge}c.d_j$ was not used
[8] After a parallelisable search-evaluation of complexity $O(\nu^p)$, probability computations per move-R_c are merely O(1).
[9] As $c_{min} \to -\infty$, $c_{max} \to +\infty$ and $\delta c \to 0$, $\text{Prob}[P \text{ is } R_c \text{ with } c < x]$ and $c_{P,S}$ converge to limit values.

3. THE REFERENCE FALLIBLE PLAYER

The RFEP concept is now generalised to that of the *Reference Fallible Player* (RFP) as heralded by Haworth (2005). As before, player P provides a sample of play $S = \{(Q_i, cm_i)\}$, not now restricted to the endgame table zone ETZ. Player P, inferior to E_p, may be an individual or community of human players P_L rated or playing in the FIDE ELO range $[L\text{-}\delta L, L\text{+}\delta L]$. P may also be an engine Ei_{pj} or a composite engine $EP \equiv \Sigma q_i E_p(c_i)$.

A chess engine E is constrained to be E_p, searching to p plies depth and for quiescence, to give experimental repeatability. R_c, a handicapped version $E_p(c)$ of E_p, chooses its chess[10] moves as follows:

- at position Q, R_c has n moves m_j to positions Q_j: engine E gives value v_j to position Q_j
- $v_1 \geq v_2 \geq ... \geq v_n$ and move m_1 is, in this sign convention, apparently the best move
  however, some or all of the v_i may be negative – which was not so with the d_i of section 2
- let $w_j = |v_1| + |v_1 - v_j|$, ensuring that $0 \leq w_i$ and that the w_i can play the role of the d_i of section 2
- if (won) position Q is in an endgame for which an EGT is available, $v_j = w_j = d_j$ as in section 2 above
  thus, $R_c \equiv E_p(c)$ is an extension of Section 2's RFEP, choosing moves as before in the ETZ
- persisting with the RFEP's probability function of Section 2, R_c is defined here as follows[11]:
  $q_{j,c} \equiv$ Prob[R_c chooses move m_j] $\propto (\kappa + w_j)^{-c}$, $\kappa > 0$, and $\kappa = 1$ here.

Note that the 'stopping down' process using parameter c is separate from, independent of and applicable to all engines E. The more fallible engine R_c may be welcomed for appearing to be 'more human'. A game of chess between $R1_c$ and $R2_c$ may also be thought of as a game $R1_\infty\text{-}R2_\infty$ in a new variant of chess, *Chess*$_c$.

The notation of section 2 requires no change for the RFP, and R_c has the same properties:

- a set $S = \{(Q_i, cm_i)\}$ characterises P, who chooses move cm_i in position Q_i,
- let $R = \{R_c \mid c = c_{min}(\delta c)c_{max}\}$ be the defined set of candidate R_c,
- before analysing S, the *a priori* beliefs are {hypothesis$_c \equiv H_c \equiv$ "Prob[P is $R_c \mid R_c \in R$] $\equiv q_{1,c}$"},
- $p_{i,c} \equiv$ Prob[R_c chooses move cm_i from position Q_i],
- Prob[R_c chooses all the moves cm_i] $\equiv q_{2,c} = \Pi\, p_{i,c}$,
- the *a posteriori* Prob[P is R_c] $\equiv q_{3,c} \equiv \mu.q_{1,c}.q_{2,c}$ for some scaling factor μ, and therefore
- $EP \equiv \Sigma q_{3,c} E_p(c)$ and the mean c of EP is $c_{S,E,p} \equiv \Sigma c.q_{3,c}$: n.b., forced moves do not change $c_{S,E,p}$,
- A match $E_p\text{-}E_p(c)$ can determine the ELO difference of the engines $\delta R_{E,p,c}$ to arbitrary accuracy

Note that the opponent's moves do not affect the assessment of the play of P. One or both players in a game may be assessed independently of the other by this method. The concept of the engine $E(c)$ or $E_p(c)$ may be exploited in various ways as described in Section 4. However, it is first worth caveating the fact that E_p is not infallible, and comparing this RFP-approach with the experiment of Guid and Bratko (2006).

3.1 Moving off the gold standard

In section 2, the benchmark engine $E \equiv R_\infty$ is infallible and defines a gold standard of perfect DT*x*-minimaxing play. $c_{P,S}$ is therefore an absolute indicator, unaffected by engine or search-depth, of the apparent competence of P as demonstrated by S: it measures the degree to which S represents less than perfect play. However in section 3, the mean of EP, the E_p-transform of P, is $c_{S,E,p}$: both EP and $c_{S,E,p}$ are affected by E and p as well as by S. Benchmark engine E_p is now fallible, its evaluation of moves $\{m_j\}$ affected by p, its search strategy and its evaluation function. ELO(EP) $\approx$ ELO(P) but 'transform error' δF_1 must be estimated.

$c_{S,E,p}$ is merely a statistical 'distance measure' of how *differently* E_p and P play, as seen by E_p. Thus:

- it is necessary to convert the c-measure into a measure of rating difference between E_p and P, but
- if, at one extreme, E_p makes errors exactly when P moves correctly, $c_{S,E,p}$ is lowered,
- if, at the other extreme, E_p makes exactly the same errors as P, $c_{S,E,p}$ is raised to ∞, and
- if P actually plays better than E_p, $c_{S,E,p}$ decreases as the difference between their play increases.

The uncertainty associated with $c_{S,E,p}$ needs to be understood, but is reduced by reducing the fallibility of the benchmark engine E_p, i.e. by using the best engine E and greatest search-depth p available. The latter need not

[10] The RFP concept in fact applies to any game domain with a set of *moves* to *evaluable* positions.
[11] Other ways of stopping-down $E_{(p)}$ based on actual players' error-patterns may prove to be even more useful here.

be restricted to the search-depth naturally achievable at a 'classic' 40/120 rate of play. Note that, if the errors of E_p and P are uncorrelated, E_p*'s* error-effect is proportional to $|S|^{-½}$, and to the mean size and percentage of errors. Given that E_p will identify $E_p(c_0)$*'s* capability as c_0, an experiment can identify the sensitivity of that identification by adding normal, variance V, 'random anti-noise' to cancel out, as it were, the noise in E_p*'s* position-values. The following experiment will characterise the transform ELO-error δF_1 directly:

- given benchmark engine E_p, consider engines Ei_j, $j < p$: let $T(Ei_j) = EP_{i,j}$
- matches Ei_j-$EP_{i,j}$ directly identify the transform's ELO-error $\delta F_{1,i,j}$ for the player Ei_j
- matches E_p-Ei_j identify the profile of transform-errors δF_1 down the ELO scale relative to E_p[12].

Let E_{10} source $S1 = \{(Q_{1i}, cm_{1i})\}$ and E_{16} source $S2 = \{(Q_{2i}, cm_{2i})\}$. The 'distance measure' is asymmetric: $c_{S1,E,16} \neq c_{S2,E,10}$ because E_{10}*'s* and E_{16}*'s* perspectives of the 'distance' between their play differ.

3.2 A review of the Guid-Bratko experiment

The Guid and Bratko experiment (2006) ranked World Champions by comparison with CRAFTY searching for 12 plies and for quiescence. Their work differs from that proposed here in the following respects:

- their aim was to rank human players, and not to align carbon and silicon players on one scale,
- their distance measure was Average$\{|w_{chosen} - w_1|\}$ in terms of the notation here
  - not considering, as here, P*'s* complete set of move-choices,
  - giving the same no-penalty reward for 'optimal' moves, even if forced or highly obvious,
  - penalising equally the choice of a non-optimal move in positions Q_1 and Q_2 where, e.g.:
    Q_1 has moves to positions/values $\{0, 1\}$; Q_2 has moves to values $\{0, 1, 1, ... , 1\}$,
- their benchmark engine CRAFTY(12) is inferior to the human players it was assessing
  - the possibly conservative constraint here is that E_p is superior to P, constraining $\{P\}$
- no constraints as to their approach's applicability were stated, apparently leaving CRAFTY(12) to:
  - falsely rate (Beal, 1999) CRAFTY(12+n) as worse than CRAFTY(12), $n > 0$,
  - therefore, falsely rate 'zero error' CRAFTY(12) as the ultimate chess player, and
  - perhaps rate CRAFTY(n), n = 11, 10, ... as better than the World Champions.
- there was no discussion of the uncertainty introduced by CRAFTY(12)'s fallibility:
  - the 'stylistic similarity' correlation of players' and CRAFTY(12)'s errors affects the results.

Their experiment may be valid but was unsupported by theory. It was counter-intuitive, as it used the worst playing agent as the benchmark. It therefore attracted criticism by Riis (2006) and others on both counts. In a response, Guid, Pérez and Bratko (2007) proffered further statistics to support their original data but again did not define theoretically the applicability of their approach. They did model a fallible benchmark agent correctly ranking two less fallible agents, showing their approach to be more than intuitively applicable. However, this extreme scenario from McKinnon (2007) further highlights the fact that there are limitations:

- in a set of positions Q_j, two moves are available, leading to positions valued at 0 and 1,
- however, benchmark engine B values these moves' destinations exactly wrongly at 1 and 0,
- player P_i chooses the correct moves with frequency p_i, $p_1 < p_2 < ... < p_n$,
- player P_i appears to B to be choosing the wrong moves with frequency p_i,
- therefore, B ranks the players in the opposite order to that which is correct,
- in terms of the Guid et al (2007) model[13], $P_C = 0$, $N = 1$ and mirror-like, $P' = 1 - P$,
- more generally in their model, $dP'/dP = [(N + 1)P_C - 1]/N$, negative for $P_C < 1/(N + 1)$, i.e.,
  if E_p*'s* move-choices are bad enough, players are ranked in an exactly inverted order.

Two experiments have been proposed in Section 3.1 to compensate for a similar absence here of the theoretical analysis of the impact of E_p's fallibility on the c and on the ELO of the EP:

- a test of the robustness of the computed c by adding 'noise' to E_p's valuations $\{v_i\}$, and
- a comparison of the ELO rating of Ei_j and $EP_{i,j}$, its transform in E_p-space.

[12] This pattern is analogous to that established by GPS reference stations for use with Differential or Wide-Area GPS.

[13] $P_C \equiv$ Prob[B plays optimally], $1/N \equiv$ Prob[B and player P choose the same wrong move],
$P' \equiv$ Prob[P appears to B to play optimally], $P \equiv$ Prob[P actually plays optimally],

4. USES OF THE REFERENCE FALLIBLE PLAYER $E_p(c)$

4.1 Assessing positions with $E_p \equiv E_p(\infty)$

The large-scale deposition of position assessments by $\{E_p\}$ in a standard format provides a set of reference data which will be valuable in itself, and may suggest stopping-down $E_{(p)}$ in a better way. Any engine E participating in this production will have the extra status of being a Reference Engine.

4.2 The Player

It is good practice to spar against a suitably stretching but not overpowering opponent, often today a chess-engine. The competence-factor c allows an engine E to be tuned on a continuous scale which is preferable to a discrete set of choices such as FRITZ's {*very easy, easy, ... , really hard*}: the scale extends from 'grandmaster' to 'incompetent'. As described below, the c-scale can be correlated with the FIDE ELO scale for human players, and this leads to inferences of engine-rating and assessments of the likelihood of cheating.

4.3 The Analyser

The Analyser E_p proceeds as in Section 3 from an *a priori* belief that the observed player P is $E_p(c)$ with probability $q_{1,c}$. These $q_{1,c}$ may be independent of P in the 'know nothing' situation, or informed by some prior knowledge of P such as their ELO rating. The analyser recalculates $q_{3,c} \equiv \text{Prob}[P \text{ is } R_c]$ after each move using the rule of Bayesian inference. Guid and Bratko (2006) sensibly recommend not considering the first 'opening book' moves or moves where the advantage, say 2.00+, is already decisive. At any time, $c_{E,p,S} \equiv \Sigma c.q_{3,c}$ is the *apparent competence* of P as seen by engine E_p.

In fact, different measures of apparent competence may conveniently be developed for a range of search-depths p at the same time, say $p \in [8, 18]$. Beal (1999) proved that increasing search-depth p improves the quality of play, a fact previously assumed on empirical grounds only. Thus, it is to be expected that, as p decreases, E_p*'s* assessment $c_{E,p,S}$ of player P will first increase as E_p*'s* quality of play reduces to that of P. It should then decrease as E_p*'s* increasing errors are seen by E_p as P*'s* increasing errors. This behaviour provides a way of confirming that E_p does indeed play better than P as required here.

Let the transform of P_L in E_p-space be EP_L. As highlighted in Section 3.1, ELO(EP_L) is $L'=L+\delta F_1$, approximated by $L''=L+\delta F_1'$ where $\delta F_1'$ is an estimate of δF_1 based on the $\delta F_{1,i,j}$. If engine EP_L plays on the web, its ELO may be adjusted to L''' but this is susceptible to many social sources of error.

Given two players *P1* and *P2*, it is possible to compare their E_p-profiles and compare the c-distances of their play from that of E_p. P1 and P2 might be the benchmark engine E at different depths, subsets of P_L before and after say 1980, or if E_p is good enough, different World Champions. The *a priori* beliefs about *P1* and *P2* should be the same, or midway between profiles corresponding to their ELO ratings.

4.4 The Imitator

The Imitator E_p analyses its opponent as in Section 4.2, identifies their apparent competence c_O, and then itself plays as $E_p(c)$, with c bearing some defined relationship to c_O.

4.5 Estimating the FIDE ELO rating of engines E_p and E

Let the best estimate of the FIDE ELO rating of $E_p(c_{L,p})$ be L'' as above. Let E be E_p, or E at a 'classic' 40/120 tempo on some platform. A match between E and $E_p(c_{L,p})$ will determine an ELO-superiority δF_L with a precision proportional to the square-root of the number of games played. An estimate of the FIDE ELO rating[14] of E_p is therefore $L''+\delta F_L$. Given that an $E_p(c_{L,p})$ engine may be determined for several ELO levels L, several estimates $\{L''+\delta F_L\}$ of ELO(E_p) may be made and compared. Without the benefit of experimental evidence, the author's expectation is that the better E is and the greater p is, the more accurate these estimates will be. Also, the values L chosen should be well below the eventual FIDE ELO rating estimate of E_p so that the evaluation-errors of E_p are not significant compared to the errors of players at level L.

[14] Existing CCRL, CEGT, CSS, SCCT and SSDF ELO ratings cannot be interpreted as FIDE ELO ratings.

4.6 Use of a suite of engines Ei_j

Different engines Ei_j have different search and evaluation algorithms, and will not be unanimous on position-values or even choice of best move. The deployment of engines $\{Ei_j\}$ as analysers of a game would produce a range of perspectives about the balance of advantage, the best moves, and the loci of apparent competence $c_{Ei,j,S}$ of the two players. This would contribute to a more engaging commentary for the audience.

4.7 Analysing suspected cheating

Advancing technology has increasingly made engine-assisted cheating in chess an issue (Friedel, 2001). Recently, the manager of Topalov claimed that Kramnik had been cheating in their World Championship match, this on the basis that the percentage of moves which coincided with the choices of FRITZ. It is notable that the claim did not specify the version, settings or search-depth if any of the FRITZ used to identify the moves. Nor were the percentages compared with those of Topalov himself. Regan (2007) correctly points to a lack of detail, rigour and scientific method in statements to date about suspected cheating.

Worse, merely counting the number of times that engine E or E_p agrees with player P is unsatisfactory. A move-choice may be forced or obvious for both E_p and P, or not, and an assessment of coincidence needs to consider the full move-context of that choice as in Section 3. Tracking $|v_{best} - v_{chosen}|$ (Guid and Bratko, 2006) is an improvement but does not use the available information fully, as stated above.

The proposal here is to use the Analysis process, together with an *a-priori* E_p-profile EP of player P, to calculate the probability of EP making a sequence of *P's* moves, and track *apparent competence c* after each move. A sharp increase in apparent competence indicates sustained move-agreement between P and E_p. The pattern of such variations in apparent competence may be compared for players of the pre- and post-engine eras.

5. SUMMARY

This paper proposes an approach to the assessment of human play and the ranking of chess-engines on the FIDE ELO scale. The principle is to use games already played rather than require new games to be played. A possibly notional player P may be profiled in terms of a chess engine E_p, the recommended precaution being that E_p is a superior player to P. Standard statistical-confidence intervals apply to all results but:

- a set of players rated at FIDE ELO L may be profiled in E_p terms,
- the effect of the fallibility of E_p on that profiling process may be assessed
- engines $E_p(c)$ and in particular, engines E_p and E may be rated in FIDE ELO terms
- for each engine E_p, c may be calibrated against the FIDE ELO scale
- given $T(P)=EP$, Prob[P and E_p agree over n moves] can be calculated if cheating is suspected
- the likelihood that P has cheated may therefore be considered with better quantitative input
- sparring partners $R_c \equiv E_p(c)$ may be supplied at any level of difficulty
- an engine-opponent can tune itself dynamically to the apparent competence of its opponent

The intention is that this theoretical proposal is tested by an experimental programme as follows:

- some engines E are modified to be a set RE of Reference Engines, outputting in common format,
- the conjecture '$P \equiv \Sigma q_i E_p(c_i) \Rightarrow EP \equiv P$' is investigated,
- quorate and appropriate samples S of actual play are defined for various FIDE ELO levels L,
- the samples S are analysed by the engines in RE at various attainable depths p,
- the notional player P at FIDE ELO level L is profiled as a composite engine EP of E_p-engines,
- this profile's accuracy is tested by available $\delta F_{1,i,j}$ and by variations of S and E_p's $\{v_i\}$,
- matches E_p-EP and E_p-$E_p(c)$ are conducted to estimate the FIDE ELO level of E_p,
- if there are E_p appearing to be better than the World Champions, then
  - the World Champions may be ranked in terms of apparent competence c,
  - the robustness of those assessments will define what confidence can be placed on the ranking,
  - "Who was the 'best player?" may be addressed separately, partially informed by the $\{c\}$.

These experiments may also provide a basis for improving both position-evaluation and the FIDE ELO scale itself. The author invites collaboration with engine-authors and others to expedite this programme.

6. ACKNOWLEDGEMENTS

I thank Rafael Andrist, Paul Chatfield, Giuseppe Di Fatta, Peter Grindrod, Ken McKinnon, Ken Regan, Søren Riis, Virginie Ruiz and the referees for their stimulating and useful dialogue on this topic.

7. REFERENCES

Andrist, R.B. and Haworth, G.M^{c}C. (2005). Deeper model endgame analysis. *Theoretical Computer Science*, Vol. 349 Issue 2, pp. 158-167. ISSN 0304-3975.

Beal, D.F. (1999). *The Nature of Minimax Search*. Ph.D. thesis, University of Maastricht.

Bourzutschky, M.S., Tamplin, J.A. and Haworth, G.M^{c}C. (2005). Chess endgames: 6-man data and strategy. *Theoretical Computer Science*, Vol. 349 Issue 2, pp. 140-157.

Friedel, F. (2001). Cheating in Chess. *Advances in Computer Games 9* (ed. H.J.van den Herik amd B. Monien), pp.327-346. Universiteit Maastricht.

Guid M. and Bratko, I. (2006). Computer Analysis of World Chess Champions. *ICGA Journal*, Vol. 29, No. 2, pp.65-73. ISSN 1389-6911.

Guid, M., Pérez, A. and Bratko, I. (2007). How Trustworthy is Crafty's Analysis of Chess Champions? Computer Games Workshop 2007 (CGW 2007), (eds. H.J. van den Herik, J.W.H.M. Uiterwijk, M.H.M. Winands, and M.P.D. Schadd), pp. 15-26.

Haworth, G.M^{c}C. (2002). Reference Fallible Endgame Play. *ICGA Journal*, Vol. 26, No. 2, pp. 81-91.

Haworth, G.M^{c}C. (2005). Chess Endgame News. *ICGA Journal*, Vol. 28, No. 4, p.243.

Haworth, G.M^{c}C. and Andrist, R.B. (2003). Model Endgame Analysis. *Advances in Computer Games 10*, Graz, Austria (eds. H.J.van den Herik, H. Iida and E.A. Heinz), pp. 65-79. Kluwer Academic Publishers, Norwell, MA. ISBN 1-4020-7709-2.

Jansen, P.J. (1992a). KQKR: Awareness of a Fallible Opponent. ICCA Journal, Vol. 15, No. 3, pp. 111-131.

Jansen, P.J. (1992b). KQKR: Assessing the Utility of Heuristics. ICCA Journal, Vol. 15, No. 4, pp. 179-191.

Jansen, P.J. (1993). KQKR: Speculatively Thwarting a Human Opponent. ICCA Journal, Vol. 16, No. 1, pp. 3-17.

Jeffreys, H. (1961). *Theory of Probability*. 3rd Edition, esp. Ch. 4, p193. Oxford University Press. Most recently republished (1998) with ISBN 0-1985-0368-7.

M^{c}Kinnon, K. (2007). Private Communication.

Moor, J. (2006). http://www.dartmouth.edu/~ai50/homepage.html. The ai@50 Conference.

Regan, K.W. (2007). http://www.cse.buffalo.edu/~regan/chess/fidelity/ Measuring Fidelity to a Computer Agent.

Riis, S. (2006). http://www.chessbase.com/newsdetail.asp?newsid=3465. Critique of Guid and Bratko (2006).

Tamplin, J.A. and Haworth, G.M^{c}C. (2003). Chess Endgames: Data and Strategy. *Advances in Computer Games 10*, Graz, Austria (eds. H.J.van den Herik, H. Iida and E.A. Heinz), pp. 81-96. Kluwer Academic Publishers, Norwell, MA. ISBN 1-4020-7709-2.