Human Decision Making (At Chess) – Research Overview

Kenneth W. Regan, University at Buffalo CSE

What It Is: A predictive-analytic model for human decision making. The domain is move-choice at chess. The only chess-specific content is numerical values given to the various move choices by strong chess computer programs run to high depth. These represent hindsight utility values for the choices. The goal is to determine how frequently a fallible (human) agent can find the foresight that produces good hindsight results.

What It Does: The general problem solved is "Converting Utilities Into Probabilities." The model outputs, for each possible move, the probability that an agent having certain skill parameters will select it. This enables forecasts of certain aggregate statistics, importantly with projected confidence intervals. Then actual player scores on these statistics provide both performance ratings and hypothesis tests.

What It's Good For: Skill assessment, prediction of results, cheating detection, player feedback and training, evaluation of other skill rating systems, probing human behavior tendencies (itemized below), and general problems in rational-choice theory, psychometrics, statistical methodology, and scientific modeling (itemized below).

1 Current Makeup and Resources

1. Four jointly authored research papers, all presented at professional international conferences, only one of them specific to computer games. One draft paper online, others in preparation.
2. Over 10,000 lines of C++ code for data management and statistical analysis, plus 2,000 lines of Perl to collate the data.
3. Computer analysis of over 1 million game turns evaluating all move options (in so-called Multi-PV mode with "multi" = 50), plus over 10 million game turns in Single-PV mode as a scientific control. The latter includes all major events in chess history except intra-national team leagues.
4. At 2k/page, over 30 million pages of textual raw data. This is human decision-making data taken under real competition, with no legal restrictions or need for human-subject waivers. All taken on ordinary quad-core desktop computers, ongoing.
5. Co-authors, collegial advisors, student assistants, seminars, interfaces with neighboring applications in predictive analytics and machine learning.
6. The model itself has *no* chess-specific ingredients, besides the data with numerical values of move options.

2 How It Operates

1. Given values $v(\cdot)$ for all relevant move options in a set of positions/game-turns t, and depending on fittable parameters representing the skill profile of a player P, the model outputs probabilities $p_{t,i}$ for P to choose each move m_i available at turn t.

2. These in turn yield forecasts for aggregate statistics including the following:

   (a) Number of agreements with a prescribed sequence of moves m_{t,k_t} for each turn t, such as the sequence of moves favored by a certain chess program suspected in cheating.

   (b) Expected total difference in value from such a sequence, defined as

   $$E[e] = \sum_t \sum_i p_{t,i} |v(m_i) - v(m_{t,k_t})|.$$

   (c) Probabilities of achieving various game objectives, based also on the skill profile of one's opponent.

3. Importantly, the model also projects variances and confidence intervals for the aggregate statistics.

4. Practical tests of the projected intervals show that they are accurate to within 15% for test (a) and 40% for test (b). The difference represents modeling error. Tightening especially the latter is a current goal. Both tests have been used in chess cheating cases, the former in sworn written testimony.

5. The model also does skill assessment based directly on the computer-judged quality of the moves made, rather than the results of games as with the current Elo Rating System used by the World Chess Federation and national bodies.

6. This eliminates "luck" factors from opponents' miscues and provides more-robust samples, insofar as a typical top-player load of 50–100 games per year will yield 1,500–3,000 relevant moves.

7. Can also quantify and assess derived skill factors, such as "challenge created," which means the projected $E[e]$ over all moves by a player's opponents.

3 Results So Far

1. Training data shows a strong linear correlation between the computed "Intrinsic Performance Ratings" (IPR's) and the players' official ratings on the Elo system used by the World Chess federation (FIDE). In Elo, a beginner may be rated 1200, a club regular 1600, a master 2200, and the world champion near 2800. This correspondence has been mostly steady since the 1970's, contrary to wide belief in over 100 points of "rating inflation" [Regan-Haworth, 2011] Watchword: The four-dozen players on the 2700+ Elo list *deserve their ratings*.

2. An exhaustive study of the Elo 2600 level shows 16 points of inflation between the 1980's and now, with 0 and a potentially-arguable 32 within the error bars. The difference can be explained by players in the 1980's having benefit of an extra 30–60 minutes of thinking time per game and adjournments after turn 40. (Recent, unpublished)

3. IPR's extend to years before the 1971 inception of Elo at world level. The IPR's of the best players at a given time increase steadily from about 2000 before Morphy through 2500 around 1900 to 2700 after WW I and 2800 regularly only in the post-Fischer era.

4. These points argue that human skill at chess has been improving over time, analogous to athletics, with possible import for other intellectual fields.

5. IPR's show the effect on performance of faster time controls and other tournament formats and conditions. The correspondence to Elo makes the meaning of these computed ratings universally understood, compared for instance to rating musicians.

6. Human performance (at chess) has intrinsic lower bounds on its variability, represented by the projected error bars. Preliminary indications are that actual performance varies more on a per-event than a per-game or per-move basis.

7. Distributional studies of player performance, including cheating tests and controls for them. Official and unofficial involvement in prominent cheating cases.

4 General Human Implications

1. A logarithmic scaling law: Define v to be the value of the best move in absolute terms and u the value of the played move. When the average raw error $e = v - u$ by human players over many positions with the same v is plotted against v, e scales upward with $|v|$ asymmetrically about 0. When e is replaced by the scaled error

$$e' = \int_{x=u}^{x=v} d\mu(x) \qquad \text{with normalized metric} \qquad d\mu(x) = \frac{a}{a+x}dx,$$

the plot of e' versus v becomes approximately flat and symmetric, for values of a near 1.0. When u, v are on the same side of 0, this makes

$$e' = |\ln(1 + |v|) - \ln(1 + |u|)|.$$

The effect is less pronounced in games played by computers.

2. Interpretation: human players perceive differences in value between moves in proportion to the overall value or imbalance in the position. Analogy may be that we perceive financial differences in proportion to the cost or benefit or overall size of a transaction.

3. Human reliability may fit "thin-tail" rather than "fat-tail" curves. The model's main equation relates the probability p_i of a move with scaled error e' to the probability p_1 of the best move by equations of the form

$$R(p_i, p_1) = g_c(\frac{e'}{s}) \tag{1}$$

where s, c are fittable parameters. Here $g_c(x)$ is a family of curves parameterized by c that satisfy $g_c(0) = 1$ and $g_c(x) \to 0$ as $x \to \infty$. Thus far inverse-exponential curves such as e^{-x^c} are observed to fit better than inverse-polynomial curves such as $1/(1+x^c)$ or $1/(1+x)^c$, especially in the tail. Although one blunder can proverbially ruin one's tournament, large errors by strong players are rare in relation to all moves played.

4. Human time mis-management ("procrastination") graphically displayed by plotting e or e' against the turn number, ramping up sharply to the standard Move 40 time control, then dropping suddenly as the players are given a fresh budget of time.

5. An effort to quantify human *depth of calculation* via a third fittable parameter d is underway. A strangely-steady observed 58%-42% frequency split between moves with equal high-depth values may indicate how preference responds to skin-deep factors.

6. This will also separate *prediction* from *skill assessment* in the model.

5 Applications

Within Chess: Skill assessment—overall and in various kinds of positions such as endgames, tactical play, positional play, attack, defense, ahead, behind. Player training. Forecasting the capacity to improve one's skill. Stability of the rating system. True growth curve of player strength—the belief that Elo ratings "lag" for developing players has led to artificial imposition of "rating bonuses"; controversy about them may be intrinsically resolvable. Objectively-assessed effects of faster time controls and playing two or more games in a day on quality of play. Cheating testing. Standard representation of computer analysis and comparison of different chess programs. Historical ratings and hard data for light on past controversies.

Outside Chess: The model can already apply in any situation where an agent must choose one of several alternatives, each of which has a prescribed or hindsight-determined value that the agent does not directly perceive. Some examples of what is generally a simple, bellwether "bounded-rationality" application:

1. *Multiple-choice tests:* The raw count of unique right answers is less sensitive than a measure that distinguishes between the values of sub-optimal answers. For example, multiple-choice questions may each have a best answer, a nearly-equivalent answer, an inferior answer that still shows substantial comprehension, a superficial answer, and an answer that catches irrelevant personal biases. My model can accommodate prescribed numerical values on the choices, plus weights on the importance of the questions themselves, and generate reliable skill assessment with uniquely-specified criteria and values from actual performance training data. This kind of upgraded psychometric scale may be especially useful for online courses.

2. *Trader aptitude, including insider-trading detection:* The hindsight values can be the actual values of trades after some specified time period. This requires determining the al-

ternatives available to a trader at any given time. Given those values, again the regression and rating mechanism is out-of-the-box.

3. *Fidelity to particular advisors:* A decision-maker's expected rate of correspondence to a series of advised options can be estimated from the hindsight values of those options, and then compared with the actual choices. Among other things, this can determine whether there is agreement with a particular advisor beyond what the ultimate value of the advice would expect.

4. *Other games:* Rating systems for any game in which options can be given authoritative (hindsight) numerical evaluations.

5. *Fraud detection* in general, based on the chess-cheating methodology.

6 General Scientific Implications

1. *Big Data aspects:* On size alone, 60 GB may be just the "Austin Powers Mini-Me" version of big-data, but the point is that aggregation of results from millions of moves has informed scientific choices made in the model. Three examples are the scaling law, the 58%-42% split for tied moves, and comparison of fitting methods described below.

2. *Integrating Disparate Data Points:* Different chess positions provide disparate "spreads" of options and their values. Sometimes there is one obvious "forced" move, sometimes there are many nearly-equal alternatives, sometimes crisis is at hand, and sometimes a clear path is present but hard to spot because opportunity is far off. It is challenging to find a single simple equation such as (1) that handles these cases equally well, without bias.

3. The degree to which large-scale human behavior can be captured by a single simple equation is the purest scientific issue.

4. *Comparison of Fitting Methods:* The parameter-fitting task (s, c now, soon d) lends itself freely to a wide range of fitting methods, with multiple criteria for judging closeness of fit. These include Bayesian methods (implemented by co-author Guy Haworth), maximum-likelihood (ML), my own "percentiling" method which puts the same frequency framework on the disparate data points, clustering similar data points, and minimizing various fitting scores or distances. The application obeys the conditions under which ML gives the limit of Bayesian iteration on large data, but thus far ML does verifiably poorly. *Providing a natural application that discriminates fitting methods may be the most immediate general import of this work.*

5. Particular phenomena in human decision-making behavior, including those described above.

6. Ancillary matters: (I) Chess programs use simple tabulation (Zobrist) hashing with little or no probing, and have been susceptible to rare but large faults from hash collisions.

A potential "holy grail" is a concrete statistical test for distinguishing between pseudo-random and truly-random initialization of the hashing scheme. (II) Test how successful evolutionary-algorithm heuristics can be on the canonically hard task of proving wins in (endgame) positions. (III) Test "fingerprints" of chess programs, perhaps yielding a useful distance metric, with application to quantifying originality.

7 Game Plan and Research Opportunities

1. Implement the depth d parameter, along with an expanded data format (current).
2. Tune the model to yield robust results on smaller data sets, where "robustness" means goodness of fit and agreement among certain fitting methods and model settings. The training sets for the 2011 papers achieved robust results with sizes about 10,000–20,000 moves. Tweaks made in early 2012 usually give robust results for single-tournament sets of 1,500–4,500 moves. The goal is robustness for single-player-in-single-event performances of about 150–450 moves (next, through Spring 2013).
3. Compare and evaluate different statistical fitting methods in this test-bed (Winter 2013, partial draft paper).
4. Apply model to multiple-choice tests and compare to current psychometric measures (topical for Spring 2013).
5. Develop relations to other cognitive models and decision-making applications (topical for Spring 2013).
6. Draw social-science conclusions, such as human risk-taking behavior (current undergraduate project at Princeton) and utility-difference perception (topical for Spring 2013).
7. Applications within chess, such as compiling IPR's from current and historical events, shedding light on past controversies, player forecasting, showing effects of tournament conditions, and uses in training (ongoing).
8. Potential wider applications, including how decision-making differs between honest and fraudulent circumstances, and risk analysis.

8 Reference Links

http://www.cse.buffalo.edu/~regan/
http://www.cse.buffalo.edu/~regan/publications.html#chess
http://www.cse.buffalo.edu/~regan/Talks/
http://www.cse.buffalo.edu/~regan/chess/fidelity/
www.nytimes.com/2012/03/20/science/a-computer-program-to-detect-possible-cheating-in-chess.html
http://rjlipton.wordpress.com/2012/05/31/chess-knightmare-and-turings-dream/
http://rjlipton.wordpress.com/2011/10/12/empirical-humility/
http://rjlipton.wordpress.com/2012/03/30/when-is-a-law-natural/