Human Decision Making (At Chess) – White Paper

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 "Intrinsic Performance Ratings" (IPRs), hypothesis tests relevant to cheating cases, and other quantitative results.

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, and general problems in rational-choice theory, psychometrics, statistical methodology, and scientific modeling.

Current Composition: Papers presented at professional international conferences. Original software is over 29,000 lines of C++ code for data management and statistical analysis, plus over 10,000 lines of Perl (dup. in Python) to collate the data. Almost 1TB of high-entropy textual data covering events in the entire history of chess, exhaustively (in a quicker mode) since 2015.

1 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 steady since the 1970's, contrary to wide belief in over 100 points of "rating inflation" [Regan-Haworth, 2011; Biswas-Regan 2013].

2. Other evidence that human skill at chess has been improving over time, analogous to athletics, with possible import for other intellectual fields. [Regan-Haworth-Macieja, 2011]

3. 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 more robust than in other performance-judging fields.

4. Human performance (at chess) has intrinsic lower bounds on its variability, represented by the projected error bars. Results partially supporting theories by Kahneman, Tversky, Simon et al. [papers 2014-15 with Biswas et al.]

5. Distributional studies of player performance, including cheating tests and controls for them. Official and unofficial involvement in prominent cheating cases, including sworn testimony [Regan-Haworth 2015; FIDE ACC since 2013].

2 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.

Outside Chess: The model applies in any situation where an agent must choose one of several alternatives, each of which has a prescribed or hindsight-determined value that is not directly perceived:

1. *Multiple-choice tests*, especially scored by value of answer rather than zero for all but one answer.

2. *Trader aptitude*, including insider-trading detection and gauging fidelity to particular advisors.

3. *Other games*, any in which options can be given authoritative (hindsight) numerical evaluations.

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

3 General Scientific Impacts

1. *Big Data aspects:* Several model-design features have been inferred from data. These include discovery that the average error per move made by human players, when plotted against the overall advantage in position for one side, scales up affine-proportionally to it. Log-log scaling flattens this plot and tunes the model. Other examples are a persistent 58%-42% frequency split between two moves tied in value at high depth, and comparison of statistical fitting methods.

2. The model fits large-scale human behavior to a single simple equation.

3. Scaling law analogizes to argue that humans perceive financial differences in proportion to the benefit or overall size of a transaction. Effect is less pronounced in games played by computers.

4. The parameter-fitting task lends itself to a wide range of fitting methods: Bayesian; maximum-likelihood; frequency fitting; clustering; minimizing various fitting scores or distances. The model and data provide multiple criteria for judging closeness of fit. *Providing a natural application that discriminates fitting methods may be the most immediate general import of this work.*

5. Particular human tendencies, such as time-(mis-)management, risk-taking, reliability, and aggression can be quantified and graphed.

6. Psychometric measures and issues, such as separating answer prediction from skill assessment.

4 Research Opportunities

1. Expand the model to represent more features (currently implementing depth-of-calculation).

2. Tune the model to yield robust results on smaller data sets (current).

3. Compare and evaluate different statistical fitting methods in this test-bed (current).

4. Apply model to multiple-choice tests, compare to current psychometric measures.

5. Develop relations to other cognitive models and decision-making applications.

6. Other social-science investigations, and general modeling issues in machine-learning.

7. Applications within chess, such as compiling IPR's from current and historical events, evaluating past controversies, player forecasting and training, tournament format analysis (ongoing).

8. Wider applications include transaction integrity, cross-validation, and personnel assessment.

Web Sites and References: Home page `http://www.cse.buffalo.edu/~regan/` and public anti-cheating site `http://www.cse.buffalo.edu/~regan/chess/fidelity/`, access to data by request.