Human and Computer Preferences at Chess

Kenneth W. Regan
Department of CSE
University at Buffalo
Amherst, NY 14260 USA
regan@buffalo.edu

Tamal Biswas
Department of CSE
University at Buffalo
Amherst, NY 14260 USA
tamaltan@buffalo.edu

Jason Zhou
The Nichols School
Buffalo, NY 14216 USA

Abstract

Distributional analysis of large data-sets of chess games played by humans and those played by computers shows the following differences in preferences and performance:

(1) The average error per move scales uniformly higher the more advantage is enjoyed by either side, with the effect much sharper for humans than computers;

(2) For almost any degree of advantage or disadvantage, a human player has a significant 2–3% lower scoring expectation if it is his/her turn to move, than when the opponent is to move; the effect is nearly absent for computers.

(3) Humans prefer to drive games into positions with fewer reasonable options and earlier resolutions, even when playing as human-computer *freestyle* tandems.

The question of whether the phenomenon (1) owes more to human perception of relative value, akin to phenomena documented by Kahneman and Tversky, or to rational risk-taking in unbalanced situations, is also addressed. Other regularities of human and computer performances are described with implications for decision-agent domains outside chess.

Keywords. Game playing, Computer chess, Decision making, Statistics, Distributional performance analysis, Human-computer distinguishers.

1 Introduction

What can we learn about human behavior and decision-making agents via large-scale data from competitions? In this paper we use data sets from high-level human and computer chess matches totaling over 3.5 million moves to demonstrate several phenomena. We argue that these phenomena must be allowed for and reckoned with in mainstream machine-learning applications aside from chess. They also show new contrasts between human and computer players. Computer chess-playing programs, called *engines*, are rated markedly higher than all human players even on ordinary personal computer hardware (Banks and others 2013). They are observed to make fewer *blunders* than human players (see (Guid and Bratko 2006; 2011)) even after adjusting for difference in overall playing strength.

Here *blunder* means a mistake of sizable cost, whose cost can be demonstrated over a relatively short horizon. We distinguish this from *aggregate error* as judged by a third party. In our case the third parties are computer chess programs analyzing the position and the played move, and the *error* is the difference in analyzed value from its preferred move when the two differ. We have run the computer analysis to sufficient depth estimated to have strength at least equal to the top human players in our samples, depth significantly greater than used in previous studies. We have replicated our main human data set of 726,120 positions from tournaments played in 2010–2012 on each of four different programs: Komodo 6, Stockfish DD (or 5), Houdini 4, and Rybka 3. The first three finished 1-2-3 in the most recent Thoresen Chess Engine Competition, while Rybka 3 (to version 4.1) was the top program from 2008 to 2011.

All computer chess programs give values in units of *centipawns*, figuratively hundredths of a pawn. They may differ in how they *scale* these units—in particular, our work confirms that versions of Stockfish give values about 1.5 times higher magnitude than most programs, while Rybka, Houdini, and Komodo are close to each other. The value of the computer's best move in a position denotes the overall value v of the position. In all except the small PAL/CSS dataset of games by human-computer tandems and some computer-played sets, for which we used a much slower analysis mode that evaluates all "reasonable" moves fully, we reckoned the value of the played move (when different) as the value v' of the next position from the same player's perspective (contrast (Guid and Bratko 2006; Guid, Pérez, and Bratko 2008; Guid and Bratko 2011)). The proper subtraction $v - v'$ is added to the running error total, and dividing by the total number of positions gives the (*raw*) *average difference* (AD) statistic. Now we can state the three main phenomena demonstrated in this contribution:

1. When the AD $ad(v)$ in positions of value v in human-played games is graphed against v, it grows markedly with $|v|$—indeed the marginal AD is nearly proportional to $|v|$. Ratios of the form $ad(c \cdot v)/ad(v)$ for $c > 1$ are largely independent of the skill level of the human players, but smaller by half in games played by computers.

2. The proportion $p(v)$ of points scored (1 for win, 0.5 for draw) by human players with positions of value v is 2–3% less when those players are to move, than in positions where it is the opponent's turn to move. The effect is absent in computer-played games.

3. Positions in games played by humans—even by human-computer tandems—have a significantly smaller range of reasonable moves than positions played by computers alone.

Before presenting the data for these results, we discuss potential wider significance and motivations. The first phenomenon was already reported by (Regan, Macieja, and Haworth 2011) but for unstructured data without controls—in particular, without investigating possible dependence on the player's skill rating. They interpreted it as owing to human psychological perception of differences in value between moves in marginal proportion to the overall absolute value of the position, with reference to human consumer behavior demonstrated by Kahneman and Tversky (Kahneman and Tversky 1981). Stedile (Stedile 2013) argued instead that it can be explained rationally with regard to risk taking. The graph of $p(v)$ approximately fits a logistic curve $(1/(1 + e^{-cv})$ where c depends on the scaling) with slope greatest near $v = 0$, so that the marginal impact of an error on the points expectation is greatest when the game is balanced. The argument is that players rationally take fewer risks in close games, while the strong computer analyzers expose the riskier played moves as errors to a greater degree than human opponents; hence more error is recorded by them from the riskier played moves in unbalanced positions (v away from 0).

Can these two hypotheses—psychological perception versus rational risk-taking—be teased apart? Stedile did not claim to resolve the issue, nor did either (Stedile 2013) or (Regan, Macieja, and Haworth 2011) distinguish according to the difference in ratings between players in the games. Human chess players are rated on the well-established and well-documented Elo rating scale, which gives numbers ranging from 600–1200 for "bright beginner," 1200–1800 for regular "club players," 1800–2200 for expert players, 2200–2500 for master players (many of whom frequent international tournaments), and 2500–2800+ for typical holders of the Grandmaster title to the world's best players (fewer than 50 above 2700 as of this writing). The rating system is designed so that a difference d in rating gives expectation $p(d) = 1/(1 + 10^{d/400})$ to the weaker player, which means about 36% when he/she is 100 points lower rated, 30% for 150 points lower, 24% for 200 points lower, 15% for 300 lower, and so on.[1] Our figures emphasize 150-point rating differences and 30/50/70% expectation.

The second phenomenon is robust up and down the value curve: the human player to move scores 2–3% worse for positions of constant value v to him/her than when the opponent is to move. The difference is somewhat greater when weaker players are to move, but is not significantly affected by the difference in ratings. We seem to be left with a pessimistic, even cynical, explanation: the player to move has the first opportunity to commit a game-clarifying *blunder*. Indeed the frequency of blunders matches the 2–3% observed difference. The effect is *absent* in our computer data. This demonstrated aspect of human error highlights a generic advantage of using computerized decision agents. Natural human optimism would view having the turn to move as greater value of opportunity.

The third phenomenon addresses the question of a generic human tendency to try to force earlier skirmishes in unresolved situations, when better value may be had by strategies that keep a wider range of good options in reserve. Compare the hotter inclinations of Kirk or McCoy versus the cooler counsels of Spock in the first halves of many "Star Trek" episodes. In our situation the computers playing the role of "Spock" are proven to be stronger players than the humans. It is thus all the more significant that in so-called "Freestyle" tournaments where humans freely use (often multiple) computers for analysis but have the final say in choice of moves, the pattern of forcing play is close to that of human games, and significantly stronger than in games played by computers alone. Again with reference to computerized decision agents, we regard this as a "natural" demonstration of factors that must be specifically addressed in order for the agents to best please their human masters.

2 Data

The "Human" dataset comes entirely from games played in 208 tournaments in the years 2010 through 2012. It includes all round-robin chess tournaments rated "Category 9" and higher by the World Chess Federation (FIDE) that were held in those years. Each "category" is an interval of 25 points in the average rating of the players. Category 9 denotes averages from 2451 to 2475, while the strongest category in the set (and of any tournament ever played before 1/30/14) is 22, denoting 2776–2800. The set also includes several countries' national or regional championships that were held as "Open" rather than invitational round-robin events, which hence provide a reasonable sample of players at lower rating levels. It does not have any team competitions or events played at time controls faster than those licensed by FIDE for standard Elo ratings.

The total of 10,317 games includes 726,120 analyzed moves, excluding turns 1–8 of all games. Moves inside repeating sequences were also eliminated, leaving 701,619 relevant moves. Of these:

- 108,663 are by players rated 2700 and above;
- 43,456 are by players rated 2300 and below;
- 59,532 are in games between players rated 150 or more points apart.

An issue in judging statistical significance for results on this dataset is correlation for moves in the same game. Suppose Black has an advantage in the range 0.21–0.30 (in the standard "centipawn" units of chess engines) at move 10 and again at move 40. The result of that game will thereby count twice in the histogram for that range. Hence the two turns are correlated. However, since they are 30 game turns apart

[1] That observed scores are a few percentage points higher for the lower player is ascribed to uncertainty in ratings by (Glickman 1999). Whether such a "globbing" effect is present due to uncertainty over value seems harder to pin down, even though both map to points expectations $p(\cdot)$ via logistic curves that are identical up to scale units.)

they are largely independent as situations with that advantage.

We take the conservative policy of regarding all moves from the same game as correlated, and use this to justify the following rough approach: Our target minimum sample size is 41,000 moves. The average number of analyzed moves per game rounds to 70, which ignoring White's frequency of playing one more move than Black gives 35 by each player. From these, one move for each player is typically eliminated, leaving 34 each. Since the samples will never or rarely involve both players from the same game, a sample size of 41,000 expects to involve at least 1,200 games. Now 1,200 is a common poll-size target giving roughly a $\pm 3\%$ two-sigma margin-of-error for frequencies even as far from 0.5 as 0.25 or 0.75. The difference between two independent such samples allows about a 2% difference to count as significant. For samples of about 160,000 moves this becomes 1%, and for halves of the whole data set it is about 0.7%. Note also that each move involves two situations: one for the player to move and one for the player not-to-move, so the overall space has over 1.4 million entries.

These moves were analyzed in the regular playing mode (called Single-PV mode) of four computer chess programs, to the indicated fixed search depths, at which head-to-head matches have indicated approximately equal strength each within 50–100 points of the 2700 mark.

- Rybka 3, depth "13+3."[2]
- Houdini 4, depth 17.
- Stockfish DD (5), depth 19—emphasized in figures.
- Komodo 6, depth 17.

The "Computer" dataset comprises games played by major chess programs running on standard-issue quad-core personal computers, run by the "Computer Engine Grand Tournament" (CEGT) website (http://www.husvankempen.de/nunn/). This is the only site known to publish games played at full standard time controls (40 moves in 120 minutes, the next 20 moves in 60 minutes, then 20 minutes plus a 10-second bonus per move for the rest of the game, one core per engine), like those in effect for FIDE world championship events. These games were all analyzed in Single-PV mode using the previous version 4 of Stockfish.

The PAL/CSS "Freestyle" dataset comprises 3,226 games played in the series of eight tournaments of human-computer tandems sponsored in 2005–2008 by the PAL Group of Abu Dhabi and the German magazine *Computer-Schach und Spiele*. All of these games were analyzed with Stockfish 4 in the same mode as for CEGT.

In addition the round-robin finals of the 5th, 6th, and 8th tournaments, comprising 45, 45, and 39 games, respectively, were analyzed in 50-PV mode with Rybka 3 to depth 13. Insofar as chess positions rarely have more than 50 valid moves, let alone reasonable ones, this mode produced equal-depth analysis of all reasonable moves, as required for the predictive model of (Regan and Haworth 2011). These tournaments were played in 2007–2008, before the August 2008 release of Rybka 3. In the same timeframe, CEGT ran 50-game matches that included a round-robin of 7 of the very strongest programs which were in common use by PAL/CSS participants, including the previous version 2.32a of Rybka. We synthesized one 42-game tournament by taking the first and last game of each match, and another by taking the 25th and 26th games. These were also analyzed by Rybka 3 in the 50-PV mode.

[2]Rybka regards the bottom 4 ply of its search as one depth level.

3 Average Error Versus Overall Value

As stated above, chess engines give values in common units called *centipawns*, which are written either as whole numbers or with two decimal places. For the most part they do not use finer units internally, and until recently the Stockfish engine rounded units of $1/197$ to the nearest multiple of $8/197$ during its search, not just for display. Hence the values are more discrete than continuous.

Figure 1: AD for human games measured by 4 engines, and CEGT computer games measured by Stockfish 4.

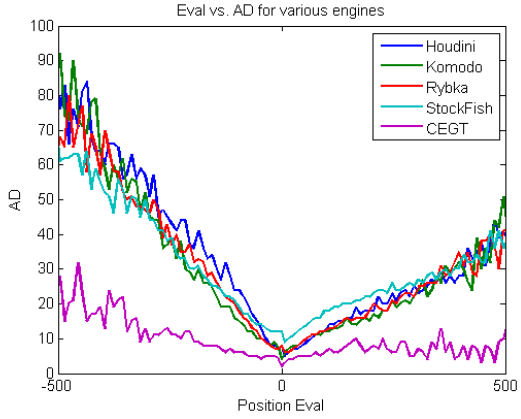


Figure 1 shows that the phenomenon of increasing human error in unbalanced games is recorded similarly by all four analyzing engines, and the error is immediately steeper when players of all kinds are behind than ahead, by any amount. When computer players are ahead, however, the error curve is nearly flat.

4 Scoring Expectation Versus Value, By Player to Move

As observed also in (Stedile 2013), the graph of percentage score by all players to move in positions of value v closely fits a logistic curve, for all four engines analyzing the human games. In particular, the graph for Stockfish DD fits

$$\frac{0.9837}{1 + 1.03457e^{-.0078v}} \tag{1}$$

with average error 0.0155. When forced to fit $\frac{1}{2}(1 + \tanh(ax))$ it gives $a = 0.0038$ with slightly higher error 0.0179. The distortion between this and the constants near unity in (1) appears mostly due to the discrepancy from 50%

expectation when the player is to move, which is detectable also in the graphs in (Stedile 2013) but not remarked there.

Figure 2 shows this discrepancy clearly for human games, and shows a robust 2–3% lower expectation for players to move compared to when the opponent is to move across positions of all values. As argued above about sample size the differences are significant. However, the effect is completely absent for games between computers, as shown by Figure 3.

Figure 2: Points expectation versus position evaluation by Stockfish DD, for human player to move versus not to move.

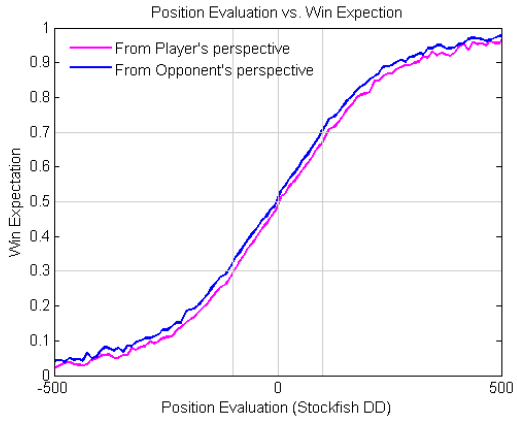
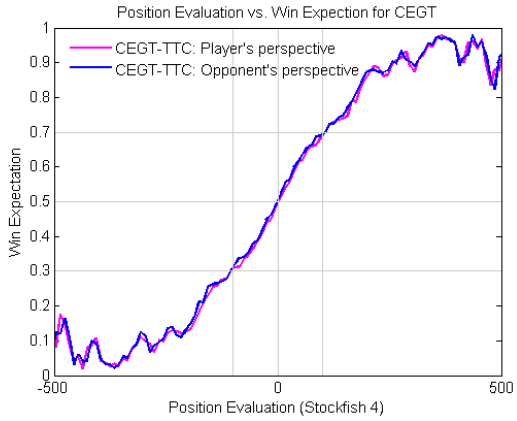


Figure 3: Expectation vs. eval for computer players, to move and not to move.



Figures 4 and 5 show that the effect does not depend greatly on whether the player is high-rated or lower-rated, nor whether the player is facing a higher or lower rated opponent. The observed frequency near 4% of large mistakes in our tabular data supports the explanation that the player to move has the first opportunity to blunder, while most blunders simplify the game into a clear loss. The near absence of crass mistakes by computers makes the effect disappear.

Combining player-to-move and opponent-to-move turns makes a perfectly symmetrical curve. Figure 6 shows that the stage of the game matters little until well past move 40.

Figure 4: Expectation vs. eval for humans rated over 2700 and under 2300, to move and not to move.

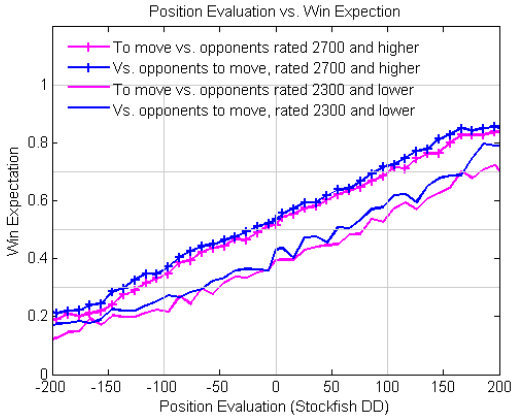
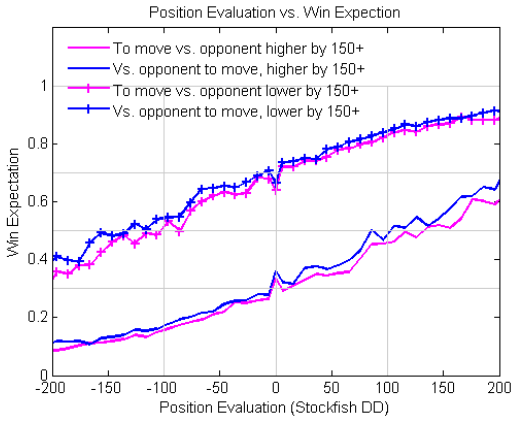


Figure 5: Expectation vs. eval in games with opponent rated 150 higher/lower, player to move and not to move.



5 Psychological Versus Rational Preferences

When a player is rated significantly higher or lower than the opponent, the curve of expectation versus position value naturally shifts higher or lower, as shown in Figure 7. Being 150 points higher rated appears to offset over a 1.00 disadvantage for Stockfish, though as noted above the value is closer to 0.60–0.70 for the other engines.

This enables a partial test of the hypothesis in (Stedile 2013) that the error phenomenon in Section 3 results from rational risk taking. Differences over 100 points are anecdotally psychologically felt as "being out-rated" in tournament games, and conventional wisdom does advise the lower-rated player to carry the fight and try to "scare" the stronger, rather than sit tight and try not to lose. Were this advice to show up as taking more risk where the win-probability curve is less steep, we might expect the graph of AD versus evaluation to shift over similarly by ± 1.00. However, Figure 8 shows that the bottom of the curve is at 0 regardless of whether the opponent's rating is lower or higher. It also shows that players who are out-rated make notably higher error when they are ahead, while there may be too little data to tell significant differences when the players are behind.

Figure 6: Expectation vs. eval within game intervals, using Stockfish DD.

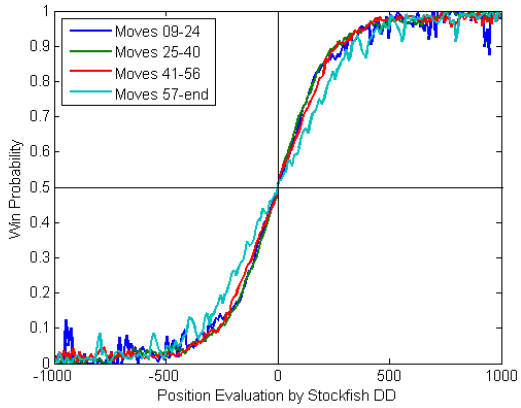
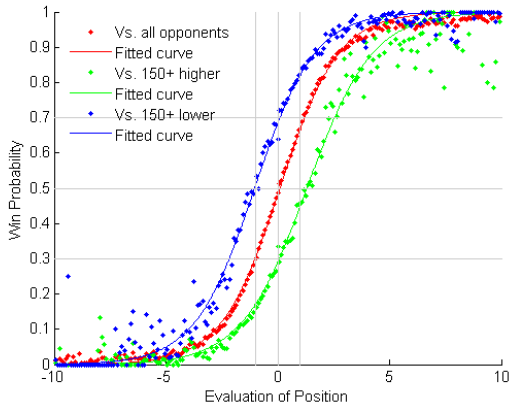


Figure 7: Shift in expectation versus eval when opponent is rated 150 points higher/lower.



6 Computer-Freestyle Comparison

Anecdotal accounts of the PAL/CSS Freestyle tournaments (Cowen 2013) relate that the best human teams used multiple computer programs, and did not blindly play the first move recommended by any of them. Instead the human players looked for moves consistent with their conceived strategies and tactical plans. We ran Multi-PV analysis of the last three 10-team round-robin finals (the 7th PAL/CSS championship was a Swiss System event without a final) of the best teams. The TCEC competition had a six-program semifinal of 90 games total, advancing two to the final. The TCEC data gives a comparison to any of the PAL/CSS finals.

Figure 9 measures that the two CEGT and three PAL/CSS data sources are respectively close to each other, that personal computer engines under similar playing conditions were significantly stronger in 2013 than in 2007–08, and that the human-computer tandems were significantly ahead of the engines playing alone even without aggregating the events together. The 2-sigma confidence intervals are the empirically-tested "adjusted" ones of (Regan and Haworth 2011); we show them to four digits although the rating values themselves should be rounded to the nearest 5 or 10.

Figure 8: AD versus eval when opponent is rated lower/equal/higher.

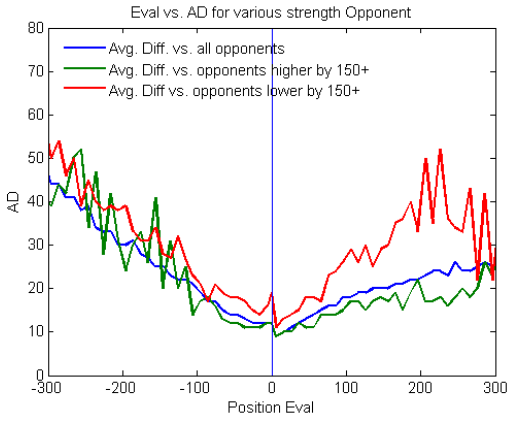


Figure 9: Analyzed Ratings of CEGT and PALCSS in 2007–08, plus TCEC Nov. 2013.

Event	Rating	2σ range	#gm	#moves
CEGT g1,50	3009	2962–3056	42	4,212
CEGT g25,26	2963	2921–3006	42	5,277
PAL/CSS 5ch	3102	3051–3153	45	3,352
PAL/CSS 6ch	3086	3038–3134	45	3,065
PAL/CSS 8ch	3128	3083–3174	39	3,057
TCEC 2013	3083	3062–3105	90	11,024

Notice that the CEGT games have significantly more moves, especially the set drawn from games 25 and 26 of each match. (There was no "memory" during the matches.) To equalize this factor, we eliminated turns after turn 60.

Figure 10: Aggregate comparisons, with and without move-60 game cutoff.

Sample set	Rating	2σ range	#gm	#moves
CEGT all	2985	2954–3016	84	9,489
PAL/CSS all	3106	3078–3133	129	9,474
TCEC 2013	3083	3062–3105	90	11,024
CEGT to60	3056	3023–3088	84	7,010
PAL/CSS to60	3112	3084–3141	129	8,744
TCEC to60	3096	3072–3120	90	8,184

Imposing the move-60 cutoff significantly affects the analyzed rating of the CEGT games, but not the others. The commonly-voiced assertion that the computer-human tandems played at a higher level, which was largely borne out by results in the qualifying stages of the PAL/CSS tournaments (see (Cowen 2013)), survives the use of the cutoff but with under 3-sigma significance. The TCEC results are enough to indicate that the difference does not survive to today, while unfortunately no Freestyle events of comparable level and prize funds have been held between 2008 and a tournament begun by InfinityChess.com in February 2014 through April 10. Our purpose here is to argue a larger human effect on the *style* of play.

Relatively few of the PAL/CSS games extended beyond move 60, and only 730 moves total were eliminated from those games (5.66 per game), compared to 2,479 moves being eliminated from the CEGT data (29.51 per game). This already hints that the PAL/CSS games were driven to quicker conclusions than the games by 2007–08's best computer programs playing alone.

The comparison is made fully quantitative by employing the model of (Regan and Haworth 2011). Given parameter settings denoting a rating such as 2500 or 3050, the model computes prior estimates for expected performance by players of that rating on any given set of analyzed positions. The performance estimates are given as aggregate statistics, including the projected frequency f_1 of choosing the computer's first move, and the projected average difference ad.

Most in particular, the projection of f_1 itself acts as an index of how *forcing* the position is. Here *forcing* means that there is only one move to prevent speedy defeat, or to preserve one's advantage. When such positions are analyzed by strong computers, the "forced" move is given a significantly higher value than any other. The model then projects significantly higher probability for players—of any sufficient skill—to choose that move, compared to *non-forcing* positions which have alternative moves of near-optimal value. Thus a higher aggregate f_1 means that the set of positions in the sample are more forcing.

Figures 11 and 12 show the projections for parameters denoting Elo 2500 and Elo 3050. (The actual printed values are 2499 and 3052, but differences in the last digit are not significant, while the (s, c) pairs are individually close to the "central fit" diagonal identified in (Regan and Haworth 2011).) Elo 3050 is chosen as midway between the analyzed ratings for the CEGT and PAL/CSS players themselves in the above tables. Again the recent TCEC entries are included only to witness that the CEGT comparison in 2007–2008 retains its point currently.

Figure 11: Agreement frequency projections, with and without move-60 game cutoff.

For 2500 ($s = 0.10$, $c = 0.50$)				
Sample set	f_1	2σ range	#gm	#moves
CEGT all	50.0%	49.1–51.0%	84	9,489
PAL/CSS all	54.5%	53.5–55.4%	129	9,474
TCEC all	48.0%	47.1–48.8%	90	11,024
CEGT to60	51.7%	50.6–52.8%	84	7,010
PAL/CSS to60	54.9%	53.9–55.9%	129	8,744
TCEC to60	48.9%	47.9–50.0%	90	8,184

The projections for both skill levels show a tangibly significant difference in the forcing nature of the games, even after the move-60 cutoff is applied to make the position sets more comparable. We conclude that human oversight of the computer programs drove the games to earlier crises than computers playing alone have judged to do.

7 Conclusions

We have demonstrated several novel phenomena from direct analysis of the quality of game decisions made by human and computer players. These phenomena establish stylistic differences in perception and preferences. The comparison in Section 5 supports the hypothesis that humans perceive differences according to relative rather than absolute valuations, while only the latter affect choices made by computers (at least when they are not behind in the game). There is a significant human disadvantage in the onus to choose a move first. This is legion with betting in poker but perhaps a surprise to find so robustly in chess, and tempers the optimism humanly associated with going first. Section 6 shows a significant preference in computer players for "wait-and-see," while this is overruled when humans use the same computer programs to inform rather than execute their decisions. It is possible, however, that human-computer cooperation produces better results than either acting separately.[3]

Figure 12: Agreement frequency projections, with and without move-60 game cutoff.

For 3050 ($s = 0.05$, $c = 0.55$)				
Sample set	f_1	2σ range	#gm	#moves
CEGT all	59.5%	58.5–60.4%	84	9,489
PAL/CSS all	65.0%	64.1–65.9%	129	9,474
TCEC all	57.1%	56.2–57.9%	90	11,024
CEGT to60	62.2%	61.2–63.3%	84	7,010
PAL/CSS to60	65.7%	64.8–66.7%	129	8,744
TCEC to60	58.9%	57.9–59.9%	90	8,184

References

Banks, G., et al. 2013. CCRL, computerchess.org.uk/ccrl/.

Cowen, T. 2013. *Average is Over*. Dutton Adult.

Glickman, M. E. 1999. Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics* 48:377–394.

Guid, M., and Bratko, I. 2006. Computer analysis of world chess champions. *ICGA Journal* 29(2):65–73.

Guid, M., and Bratko, I. 2011. Using heuristic-search based engines for estimating human skill at chess. *ICGA Journal* 34(2):71–81.

Guid, M.; Pérez, A.; and Bratko, I. 2008. How trustworthy is Crafty's analysis of world chess champions? *ICGA Journal* 31(3):131–144.

Kahneman, D., and Tversky, A. 1981. The framing of decisions and the psychology of choice. *Science* 211.

Regan, K., and Haworth, G. 2011. Intrinsic chess ratings. In *Proceedings of AAAI 2011, San Francisco*.

Regan, K.; Macieja, B.; and Haworth, G. 2011. Understanding distributions of chess performances. In *Proceedings of the 13th ICGA Conference on Advances in Computer Games*. Tilburg, Netherlands.

Stedile, L. 2013. Analysis of risk in chess. Junior Paper, Princeton University, advised by Mark Braverman.

[3] Preliminary analysis of the InfinityChess competition supports the conclusion that humans make the games more forcing, but in possible contrast to the PAL/CSS results, not the overall quality increase compared to today's programs acting alone..