Distinguishing Human from Computer Traits at Game Play

Tamal T. Biswas and Kenneth W. Regan

University at Buffalo

Introduction

Alan Turing both proposed the Turing test and sketched a computer chess player. The Turing test, a test of a machine's ability to bridge the gap between humans and machines, leads to the more general question—What differences in *cognitive style* exist between humans and computers?

We have found new ramifications of that question in the context of chess.

Q1: How strong is the impact of first impressions?

Q2: Is playing first always advantageous?

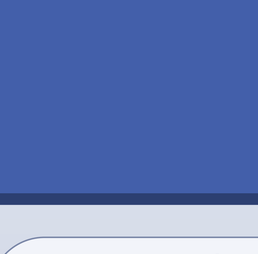
Q3: Are human preferences governed by perception or rational risk-taking?

Q4: What are the effects of human intervention in automated systems?

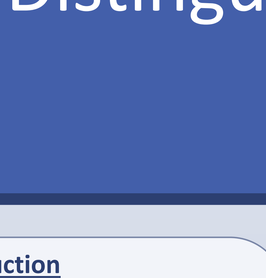
We have also tried to quantify notions such as *depth of thinking* and *complexity of problems* by starting in a domain where they can be clearly formulated, cleanly quantified, and analyzed with large data. Then we aim to transfer the formulations, results, and lessons from interpreting the results to domains of wider interest. Our home domain is competitive chess, in which the items are thousands to millions of positions from recorded games between human players—and also computer players in various kinds of high-level tournaments. Deep analysis of these positions by *computers* reveals individual characteristics of humans and computers.

Terms and Metrics

The following figure shows two chess positions that occurred during the world championship match in 2008 between V. Kramnik and V. Anand.



Kramnik-Anand, 2008 WC match game 8. Position before Kramnik's 29. Nxd4



Position after Anand's 34.... Ne3!! Ouch!

Fig. 1

The best chess programs, including Stockfish, Houdini, Komodo, and Rybka, can beat any human player. They share common measurement units, search in progressively deeper rounds d = 1,2,3..., and can either focus on one best move (Single-PV mode) or give full consideration to multiple moves (Multi-PV mode).

Terms	Notation	Explanation	Range
Unit of Measurements	$v_{i,d}$	centipawns (hundredths of a pawn value)	-999 to +999
Depth	d	Number of plies the chess engine thought in advance.	1 to 19
PV	i	Number of legal moves the chess engine considers for any position.	1 to 50

The following grid demonstrates the evaluation of various legal moves by the engine (Stockfish 5) . Each row represents the evaluation of a legal moves at depths from 1 to 19.

Moves are sorted by the evaluation of moves at the highest depth.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nd2	+118	+076	+085	+098	+032	+000	+005	+005	+006	+021	+024	+013	+026	+017	+017	+022	+000	+000	+000
Bxd7	+069	+069	-028	-028	+002	-041	-011	-023	-022	-022	+016	-008	-003	-014	+000	+000	+000	+000	-001
Qh5	+045	+045	-055	-055	-055	-008	-027	-012	-006	+003	-014	-014	-017	-017	-036	-046	-039	-034	-040
Qg8	+117	+117	-046	-046	-055	-008	-027	-012	-006	+003	-014	-030	-017	-017	-036	-044	-047	-041	-040
Kf1	+030	+030	+030	-006	+052	+009	+001	-032	-005	-044	+000	-034	-044	-034	-057	-043	-057	-036	-042
Rd1	+006	+006	+006	+001	-074	-074	-074	-074	-040	-021	-082	-075	-044	-075	-082	-063	-088	-083	-090
Kh1	+029	+029	+029	-015	+051	-029	+025	+013	-019	-044	-048	-062	-034	-099	-099	-099	-099	-088	-092
Qg5	-077	-077	-077	-077	-127	-094	-144	-086	-108	-117	-150	-165	-165	-092	-102	-098	-104	-104	-117
Qh4	-077	-077	-077	-077	-127	-094	-144	-086	-086	-117	-117	-165	-165	-092	-087	-105	-107	-104	-117
Qh3	-034	-034	-034	-087	-101	-055	-082	-091	-088	-101	-135	-128	-150	-140	-127	-181	-165	-159	-148
Ra1	-004	-004	-004	-040	-107	-107	-107	-094	-098	-109	-119	-095	-098	-087	-103	-110	-149	-119	-170
Rf1	-015	-015	-015	-051	-124	-124	-124	-111	-122	-122	-125	-111	-126	-123	-134	-125	-149	-134	-170
Rb1	-005	-005	-005	-067	-108	-108	-108	-096	-101	-127	-124	-114	-091	-098	-100	-110	-149	-147	-170
h4	+041	+041	+041	-026	-214	-214	-163	-115	-167	-181	-206	-206	-200	-195	-182	-144	-150	-157	-181
h3	+034	+034	+034	-031	-184	-184	-150	-137	-169	-153	-176	-201	-215	-175	-198	-205	-205	-203	-203
Nxd4	-029	-029	-030	-089	-089	+182	+182	+054	+054	+054	+083	+040	+065	+052	+061	-198	-199	-205	-205
Nh4	+100	+055	+009	+079	+079	+079	+026	+056	+070	+071	+059	+047	+057	+027	+027	+027	-115	-180	-218
Ng5	+189	+153	+133	+059	+064	+053	+056	+018	+031	+009	-008	-007	-075	-084	-179	-194	-229	-221	-244
a5	+117	+117	+117	+050	-063	-063	-063	-070	-154	-164	-134	-208	-183	-220	-273	-273	-303	-305	-287
Be2	+049	+001	+029	+001	-052	-100	-100	-100	-073	-058	-186	-181	-240	-243	-263	-308	-271	-271	-318

Metrics Defined

Terms	Formulation	Explanation
Scaled delta or Error	$\delta_{i,d} = \int_{x=v_{i,d}}^{x=v_d^*} \lvert\log(1+ax)\rvert dx$	Scaled difference of value $v_{i,d}$ for any move m_i at depth d from the evaluation of the best move at that depth v_d^*. Constant a is engine specific.
Swing	$sw(m_i) = \sum_{d=1}^{D}(\delta_{i,d} - \delta_{i,D})$	Sum of scaled differences in value between depth d and highest depth D.
Generalized Kendall tau coefficient	$\tau_{X,Y} = \frac{\sum_{i,j}\mu(x_i,x_j)\mu(y_i,y_j)}{\lvert\lvert\mu_X\rvert\rvert \cdot \lvert\lvert\mu_Y\rvert\rvert}$	Used for measure rank aggregation between ordered sequence $X = (x_i)$, $Y = (y_i)$
Complexity	$\kappa(\pi) = 1 - \frac{1}{D-1}\sum_{d=1}^{D-1}\tau_{L_d,L_{d+1}}$	Complexity of position π for d ranging from 1 to $D-1$.
Difficulty	$Diff(\pi) = \kappa(\pi)\cdot\rho(\pi)$	The measure $\rho(\pi)$ is the depth where 50% of the total swings occurs.

Dataset

First Dataset: All recorded games in standard round-robin tournaments in 2006-2009 between players each within 10 *Elo-points* of a "milepost" value. The mileposts used were Elo 2200, 2300, 2400, 2500, 2600, 2700.

Second Dataset: All 900 games of the 2013 World Blitz (WB) championship. The average rating of the 60 WB players was 2611.

"Human" Dataset: Games played in 2010-2012 (208 human tournaments

Computer Dataset: Comprises games played by major chess programs in the Computer Engine Grand Tournaments (CEGT) and Thoresen Chess Engines Competition (TCEC).

Results

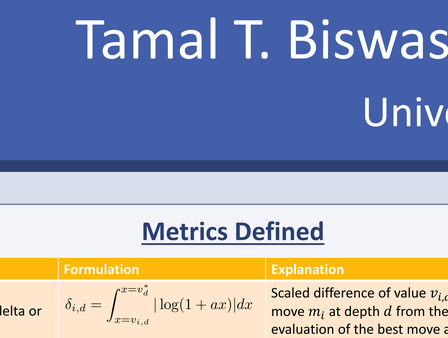


Fig. 2

Observation: High swing moves are 'tricky' to find. The players often choose inferior moves. The phenomenon is consistent with players of any Elo ratings, where higher rated players are slightly less tricked by the swing values.

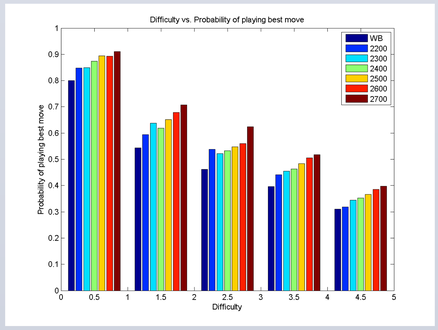


Fig. 3

Observation: Players of all calibers could find the best move when the position is easy, but less than 50% of the time when the difficulty lies between 4 and 5.

Is Playing First Always Advantageous ?

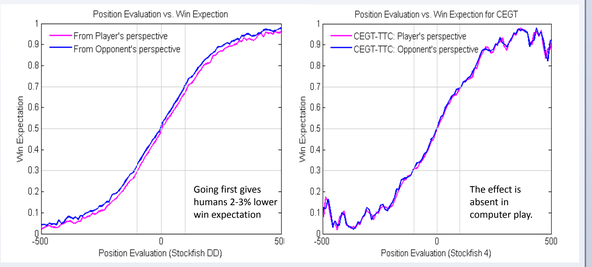


Fig. 4

Psychologically we think so.

All other things being equal, we want it to be our move.

But "to err is human"—and not so much computer…

Human-Computer Error Phenomenon

Figure 7 at upper right plots the raw error against the overall value of the position. It shows a steep change in humans that is absent in computers.

Two hypotheses to explain this:

1. We humans perceive differences in value in proportion to the total value involved.
2. We play tightest when the slope of the point expectation curve is the highest, meaning the marginal cost of the error is more.

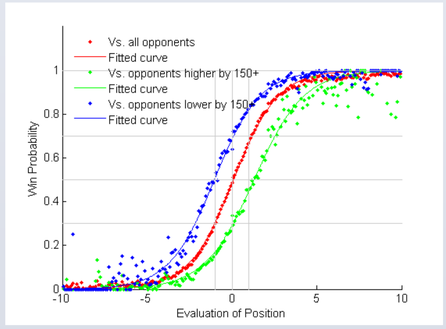


Fig. 5

Differences over 100 points are anecdotally psychologically felt as "being out-rated" in tournament games, and conventional wisdom would advise the lower rated player to carry the fight and "scare" the stronger.

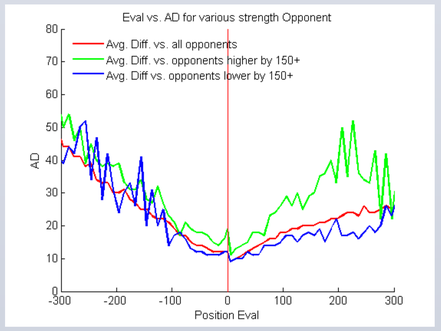


Fig. 6

However, this plot shows the bottom to be 0 regardless of the opponent's rating, which lends support to hypothesis 1 about human preferences.

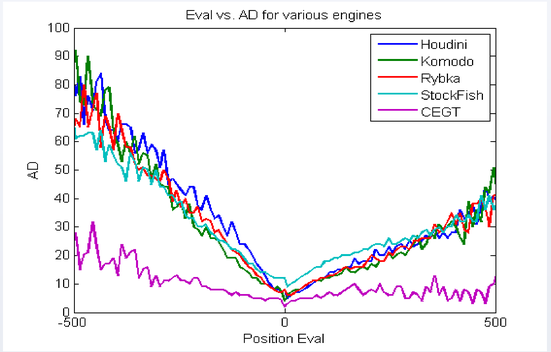


Fig. 7

Human, Computer, or Human-Computer Tandem

Human impetuousness compared to computers shows in other ways. The PAL/CSS Freestyle tournaments of 2005-2008 allowed human-computer teams. We compare them to the CEGT computer-only tournaments played in the same year and the 2013 TCEC computer championship.

In the following table, freestyle tournaments showed higher projection of f_1 (frequency of choosing the engine's first move) which itself is an indicator of how forcing the position is. The phenomenon is consistent for projections for both skill levels (Elo 2500 and Elo 3050), and even after the move-60 cutoff is applied. We conclude that in freestyle chess, human oversight of the computer programs drove the games to earlier crises than computers playing alone have judged to do.

Dataset			For 2500		For 3050		Count	
Sample set	Rating	2σ range	f_1	2σ range	f_1	2σ range	#gms	#moves
CEGT all	2985	2954–3016	50.0%	49.1–51.0%	59.5%	58.5–60.4%	84	9,489
PAL/CSS all	3106	3078–3133	54.5%	53.5–55.4%	65.0%	64.1–65.9%	129	9,474
TCEC 2013	3083	3062–3105	48.0%	47.1–48.8%	57.1%	56.2–57.9%	90	11.024
CEGT to60	3056	3023–3088	51.7%	50.6–52.8%	62.2%	61.2–63.3%	84	7.010
PAL/CSS to60	3112	3084–3141	54.9%	53.9–55.9%	65.7%	64.8–66.7%	129	8,744
TCEC to60	3096	3072–3120	48.9%	47.9–50.0%	58.9%	57.9–59.9%	90	8,184

Conclusions

- Demonstrated several novel phenomena from direct analysis of the quality of game decisions made by human and computer players.
- Established stylistic differences in perception and preferences.
- There is a significant human disadvantage in the onus to choose a move first. For computers no such disadvantage.
- Humans perceive differences according to relative rather than absolute valuations, while only the latter affect choices made by computers.
- Showed a significant computer preference for "wait-and-see," while this is overruled when humans use the same computer in tandem.
- However, human-computer cooperation produced better results in 2005-2008 than humans or computers acting separately.

References

- BISWAS, T., AND REGAN, K. Quantifying depth and complexity of thinking and knowledge. In *Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART)*, Lisbon, Portugal (2015).
- REGAN, K., BISWAS, T., AND ZHOU, J. Human and computer preferences at chess. In *Proceedings of the 8th Multidisciplinary Workshop on Advances in Preference Handling (Mpref)*, Quebec City, (2014).