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 this 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 rationality-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, cleaned 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.

![Chess Game](image)

The best chess programs, including Stockfish, Houdini, Komodo, and Rybka, can beat any human player. They share common measurement units, search in progressively deeper depth 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).

![Chess Metrics](image)

### Metrics Defined

<table>
<thead>
<tr>
<th>Metric</th>
<th>Formulation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaled delta or Error</td>
<td>( \delta_{i,j} = \sum_{s=0}^{D} v_{i,s} )</td>
<td>Scaled difference of value ( v_{i,s} ) for any move ( m_{i,s} ) at depth ( s ) from the evaluation of the best move at that depth ( \delta_{i} ). Constant ( i ) is an engine-specific.</td>
</tr>
<tr>
<td>Gain</td>
<td>( g_{i} = \sum_{s=0}^{D} \sum_{m_{i,s}} v_{i,s} - \delta_{i} )</td>
<td>Sum of scaled differences in value between depth ( s ) and highest-depth ( D ).</td>
</tr>
<tr>
<td>Kendall tau coefficient</td>
<td>( \tau(X,Y) = \frac{P_{+/+} - P_{-/-}}{\sqrt{(P_{+/+} + P_{-/-})(P_{+/0} + P_{0/-})}} )</td>
<td>Used to measure rank aggregation between ordered sequence ( X ) and ( Y ).</td>
</tr>
<tr>
<td>Complexity</td>
<td>( C(i) = \left( \frac{1}{2} \right) \sum_{i=1}^{N} \sum_{j=1}^{N} \left</td>
<td>\frac{1}{D} \sum_{s=0}^{D} \frac{v_{i,s} - v_{j,s}}{N} \right</td>
</tr>
<tr>
<td>Difficulty</td>
<td>( D(i) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left</td>
<td>\frac{1}{D} \sum_{s=0}^{D} \frac{v_{i,s} - v_{j,s}}{N} \right</td>
</tr>
</tbody>
</table>

### Dataset

First recorded games in standard round-robin tournaments in 2006-2009 between players each within 10 Elo-points of a “milepost” value. The milepost 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-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 tighter when the slope of the point expectation curve is the highest, meaning the marginal cost of the error is more.

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 use of 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 “win-and-run,” while this is overridden 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