Assessing Human Error Against a Benchmark of Perfection

Ashton Anderson
University of Toronto

Joint work with Jon Kleinberg and Sendhil Mullainathan
Humans and Machines

One leading narrative for AI: humans versus machines
For any given domain, when will algorithms exceed expert-level human performance?
Humans and Machines

A set of questions around human/AI interaction:

• Relative performance of humans and algorithms
• Algorithms as lenses on human decision-making
• Humans and algorithms working together: pathways for introducing algorithms into complex human systems

Can we use algorithms to characterise and predict human error?
Chess for Decision-Making

Long-standing model system for decision-making

• “The drosophila of artificial intelligence.”
  —John McCarthy, 1960

• “The drosophila of psychology.”
  —Herb Simon and William Chase, 1973

Chess provides data on a sequence of cognitively difficult tasks. When a human player chooses a move, we have data on:

  • The task instance: the chess position itself.
  • The skill of the decision-maker: a chess player’s Elo rating.
  • The time available to make the decision.

Can we use computation to analyze human performance?

  • Characterize human “blunders” (mistakes in choice of move)
  • Chess as the drosophila of machine superintelligence?
A History of Chess Engines

• 1988: First recorded win by computer against human grandmaster under standard tournament conditions.

• 1997: Deep Blue defeats world champion Kasparov in 6-game match.


• 2005: Last recorded win by a human player against a full-strength desktop computer engine under standard tournament conditions.

• 2007: Computers defeat several top players with “pawn odds.”
Chess for Decision-Making

Could use chess engines to evaluate moves [Biswas-Regan 2015]

• Promising, since engines are vastly superior to the world’s best players

• Engines sometimes detect clear-cut errors, but very often a “grey area”: engines and humans disagree, but doesn’t necessarily change the outcome of the game
We use the fact that chess has been solved for positions with at most 7 pieces on the board.

• “Tablebases” record all possible positions with <=7 pieces
• Can determine (game-theoretic) blunders by table look-up
• These positions are still difficult for even the world’s best players

_The Stiller moves are awesome, almost scary, because you know they are the truth, God’s Algorithm; it’s like being revealed the Meaning of Life, but you don’t understand one word._

— Tim Krabbé, commenting on an early tablebase by Lewis Stiller
# Chess for Decision-Making

Data from two sources:

<table>
<thead>
<tr>
<th>Source</th>
<th># Games</th>
<th>Rating</th>
<th>Duration</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>FICS</td>
<td>200M</td>
<td>1200–1800</td>
<td>Minutes</td>
<td>Casual enthusiasts playing online</td>
</tr>
<tr>
<td>GM</td>
<td>1M</td>
<td>2400–2800</td>
<td>Hours</td>
<td>Professional tournaments</td>
</tr>
</tbody>
</table>

Take all <7-piece positions, classify a move as a blunder if and only if it changes the win/loss/draw outcome.
Basic Dependence on Fundamental Dimensions

How does decision quality vary with \{ skill, time, difficulty \}?
Human Error as a Function of Skill

- 1000: Winner of a local scholastic contest
- 1600: Competent amateur
- 2000: Top 1% of players
- 2300: Lowest international title
- 2500: Grandmaster
- 2850: Current world champion
Human Error as a Function of Time

- **Y-axis:** Empirical blunder rate
- **X-axis:** Time remaining (sec.)

Graph shows data across different Elo levels from 1200 to 1800.
Human Error as a Function of Time

![Graph showing the relationship between time spent and blunder rate for different Elo ratings. Each line represents a different Elo rating, with Elo values marked on the right side of the graph.]
A simple measure for the difficulty of a position: the “blunder potential” is the probability of blundering if you choose a move at random.

\[
\text{Blunder potential} = \frac{\text{# possible blunders}}{\text{# legal moves}}
\]

Blunder potential \(= \frac{9}{18} = 0.5\)
Human Error as a Function of Difficulty

Simple, quantal-response model captures how error varies with difficulty: a particular non-blunder is $c$ times more likely than a particular blunder.
Blunder Prediction

Use fundamental dimensions to predict: will the player blunder in a given instance?

- The difficulty of the position
- The skill of the decision-maker (Elo rating)
- The time remaining
- A set of features encoding difficulty deeper in the game tree

Performance using decision-tree algorithms:

- All features: 75%
- Blunder potential alone: 73%
- Elo of player and opponent: 54%
- Time remaining: 52%
Human Error as a Function of Skill

**FICS**

- Empirical blunder rate vs. Skill (Elo rating)

**GM**

- Empirical blunder rate vs. Skill (Elo rating)
Human Error as a Function of Skill

Difficulty is the dominant feature

To the extent this is surprising, connections with fundamental attribution error, and Abelson’s Paradox [Abelson 1985]
Human Error as a Function of Skill

Fix blunder potential: higher-depth blunder potential is the dominant feature.

Fix the exact position: skill and time become predictive.

Difficulty is dominant on average. Is this true point-wise?

• For position $p$, examine blunder rate as a function of skill in $p$
• Call a position skill-monotone if blunder rate is decreasing in $r$
• Natural conjecture: all positions are skill-monotone
Fixing the position

Difficulty is dominant on average. Is this true point-wise?

- For position $p$, examine blunder rate as a function of skill in $p$
- Call a position skill-monotone if blunder rate is decreasing in $r$
- Natural conjecture: all positions are skill-monotone

In fact, we observe a wide variation, including skill-anomalous positions

Connections with U-shaped development
Challenges arising from misleading analogies?
Reflections on Teaching

Contrast:

Traditional organization in textbooks
Adding information about frequency and rate
Reflections on Teaching

High-level goal: create a human-like AI

Understand and model human decision-making qualities at various levels

Can we build an algorithmic teacher from large-scale data on human decisions?
Reflections

Framework for analyzing human error given large numbers of similarly structured instances.

- Compare human performance to computational benchmark (in this case a perfect one)

In chess, difficulty is the dominant predictor of human error

- Similar for other domains?

Opportunities for rich understanding of human decision-making using algorithms