Scoping the Mind With Turing's Chess Machine

University of Bergen

Kenneth W. Regan[1]
University at Buffalo (SUNY)

11 Sept., 2012 (remember...)

[1] Incl. joint work with Guy Haworth and GM Bartlomiej Macieja. Sites:
http://www.cse.buffalo.edu/∼regan/chess/fidelity/ (linked from homepage)
http://www.cse.buffalo.edu/∼regan/chess/ratings/ (not yet publicly linked)
Slides updated 9/19/12 to include graphics and links from webpages that were shown, plus Astana World Rapid 2012 data.

Alan Turing Centenary 1912–2012

- 1948–50: *Turochamp*
  Turing sees chess as way to impart an activity of the human mind to computers.
- 1997: Turing's (and H. Simon's) dream of beating human WC realized.
  But chess engines not "like" the mind?
- 2006: Human WC cheats with computer?
- 2012: Use computer's "mindless" chess analysis to study the human mind itself.
- Small advantage in processing: two ordinary family 4-core PC's.
- Bigger advantage in data: tens of millions of pages, cheaply stored.
- Over 1 milion moves in 50-PV, over 10 million moves in Single-PV.

Jeg og Turing og Norge

- I have a "Turing Number" of 2: in 1983 I played Sir Stuart Milner-Barry who was with Turing at Bletchley Park.
- Princeton '81, then Oxford D.Phil. 1986, met Alonzo Church in 1990, Konrad Zuse in 1994.

Jeg og Turing og Norge

- I have a "Turing Number" of 2: in 1983 I played Sir Stuart Milner-Barry who was with Turing at Bletchley Park.
- Princeton '81, then Oxford D.Phil. 1986, met Alonzo Church in 1990, Konrad Zuse in 1994.
- 40th anniversary of Fischer-Spassky match: I was almost 2200 at age 12 and panelist for TV coverage of games 7 and 16.
- 1975: played in Sandefjord Nordiskturnering (under Norwegian flag!).
- 1980: played in Gausdal and Baerum.
- *Gratia Arnold Eikrem*.

How I got back into chess...

- 1989–2006: said **No** to every request to do computer chess.
- 2005: Mom bought me Fritz 8 as present, later upgraded to Fritz 9. Kibitz-chatted on PlayChess server during San Luis 2005; Yasser Seirawan noted me in video commentary.

How I got back into chess...

- 1989–2006: said **No** to every request to do computer chess.
- 2005: Mom bought me Fritz 8 as present, later upgraded to Fritz 9. Kibitz-chatted on PlayChess server during San Luis 2005; Yasser Seirawan noted me in video commentary.

- 2006 WC match: Kibitzing when cheating scandal broke. Frederic Friedel asked on same channel for help evaluating Danailov's statistical accusations. As cognizant Math/CS expert, felt obliged to help.
- Stayed up late trying to reproduce Danailov's stats on this now-old laptop. No methodology or data logs were given.

How I got back into chess...

- 1989–2006: said **No** to every request to do computer chess.
- 2005: Mom bought me Fritz 8 as present, later upgraded to Fritz 9. Kibitz-chatted on PlayChess server during San Luis 2005; Yasser Seirawan noted me in video commentary.
- 2006 WC match: Kibitzing when cheating scandal broke. Frederic Friedel asked on same channel for help evaluating Danailov's statistical accusations. As cognizant Math/CS expert, felt obliged to help.
- Stayed up late trying to reproduce Danailov's stats on this now-old laptop. No methodology or data logs were given.
- Only Game 2 showed reproduction. Topalov was brilliantly winning at Move 32 but lost on Move 64.

Qualitative and Quantitative Answers

By Thu. Oct. 12, 2006—the eve of the Rapid playoff on the 13th—I had a firm *qualitative* answer:

Yes Kramnik matches Fritz 9 on 29 of the last 32 moves of Game 2, and Rybka 30 of 32, both over 90%! But 21 of those moves are completely forced, and 5 are multi-way ties. That leaves only 6 moves: only 4 "significant" matches and 2 clear mistakes. *That's indistinguishable from random.*

Main Principle: A match on a clear standout move (per computer eval) is much less significant than a match amid many nearly-equal moves.

I was ready to propound all this in match commentary—but got wiped out by the 2006 Buffalo October Storm. Power back on 16th but it was over. So I worked slowly: how to make this principle *quantitative*?

Statistical *Heksebrygg*

The main ingredients of my model:

Statistical *Heksebrygg*

The main ingredients of my model:

1. An equation for the Main Principle: $\Pr(m) =$ a function of the value of the move m in relation to the values of other moves, and the Elo rating E of the player.

Statistical *Heksebrygg*

The main ingredients of my model:

1. An equation for the Main Principle: $\Pr(m)$ = a function of the value of the move m in relation to the values of other moves, and the Elo rating E of the player.
    - From the probabilities you can project the expected number N of agreements with moves preferred by a computer.

Statistical *Heksebrygg*

The main ingredients of my model:

1. An equation for the Main Principle: $\Pr(m)$ = a function of the value of the move m in relation to the values of other moves, and the Elo rating E of the player.

   - From the probabilities you can project the expected number N of agreements with moves preferred by a computer.
   - And you can project the "error" when a player makes a move the computer says has less value. $\longrightarrow$ Average Error (AE).

Statistical *Heksebrygg*

The main ingredients of my model:

1. An equation for the Main Principle: $\Pr(m)$ = a function of the value of the move m in relation to the values of other moves, and the Elo rating E of the player.

   - From the probabilities you can project the expected number N of agreements with moves preferred by a computer.
   - And you can project the "error" when a player makes a move the computer says has less value. ⟶ Average Error (AE).
   - And you can project **95% confidence intervals** for these quantities.

Statistical *Heksebrygg*

The main ingredients of my model:

1. An equation for the Main Principle: $\Pr(m)$ = a function of the value of the move m in relation to the values of other moves, and the Elo rating E of the player.

    - From the probabilities you can project the expected number N of agreements with moves preferred by a computer.
    - And you can project the "error" when a player makes a move the computer says has less value. $\longrightarrow$ Average Error (AE).
    - And you can project **95% confidence intervals** for these quantities.

2. Computer values for millions of moves. Training sets of games with both players within 10 points of each Elo century mark:

    - 2700, 2600, 2500, 2400, 2300, 2200, ..., 1600, ...

Ingredients, Continued

1. Player Skill Parameters fitted by these training sets:
   - *Sensitivity* s: how well you see small differences in value.
   - *Consistency* c: how well you avoid blunders.
   - *Depth of calculation* d. Not yet implemented.

Ingredients, Continued

1. Player Skill Parameters fitted by these training sets:
   - *Sensitivity* s: how well you see small differences in value.
   - *Consistency* c: how well you avoid blunders.
   - *Depth of calculation* d. Not yet implemented.
2. Equation converting parameters to an Intrinsic Performance Rating (IPR):
   - $E = MC^2$.

Ingredients, Continued

1. Player Skill Parameters fitted by these training sets:
   - *Sensitivity* s: how well you see small differences in value.
   - *Consistency* c: how well you avoid blunders.
   - *Depth of calculation* d. Not yet implemented.

2. Equation converting parameters to an Intrinsic Performance Rating (IPR):
   - $E = MC^2$.
   - $Elo = Magnus\ Carlsen^2$.

Ingredients, Continued

1. Player Skill Parameters fitted by these training sets:
    - *Sensitivity* s: how well you see small differences in value.
    - *Consistency* c: how well you avoid blunders.
    - *Depth of calculation* d. Not yet implemented.

2. Equation converting parameters to an Intrinsic Performance Rating (IPR):
    - $E = MC^2$.
    - $Elo = Magnus\ Carlsen^2$.

    - OK, actual equation is more complicated. See technical slides after the Conclusions slide for formulas.
    - Basically it is like “Solitaire Chess" but using *your* games in *real* chess tournaments.

Some IPRs—Historical and *Actuel*

- Magnus Carlsen:
    - 2983 at London 2011 (Kramnik 2857, Aronian 2838, Nakamura only 2452).
    - 2855 at Biel 2012.
- Bobby Fischer:
    - 2921 over all 3 Candidates' Matches in 1971.
    - 2650 vs. Spassky in 1972 (Spassky 2643).
    - 2724 vs. Spassky in 1992 (Spassky 2659).
- Hou Yifan: 2971 vs. Humpy Koneru (2683) in Nov. 2011.
- Paul Morphy: 2344 in 59 most impt. games, 2124 vs. Anderssen.
- Capablanca: 2936 at New York 1927.
- Alekhine: 2812 in 1927 WC match over Capa (2730).
- Simen Agdestein: 2586 (wtd.) at Hoogevens 1988.

Sebastien Feller Cheating Case

- Khanty-Mansiysk Olympiad 2010: Feller played 9 games (6-1-2, board 5 gold).
- Cyril Marzolo confessed 4/2012 to cheating most moves of 4 games. On those 71 moves:
    - Predicted match% to Rybka 3 depth 13: 60.1% ± 10.7%
    - Actual: 71.8%, z-score 2.18 (Barely significant: rumor says he used Firebird engine.)
    - AE test more significant: $z = 3.37$ sigmas.
    - IPR on those moves: 3240.
- On the other 5 games: actual $<$ predicted, IPR $=$ 2547.
- Paris Intl. Ch., July 2010: 3.15 sigmas over 197 moves, IPR 3030.
- Biel MTO, July 2010: **no** significant deviation, alleged cheating on last-round game only.

What is a Scientific Control?

- If I say odds are 2,000-to-1 against Feller's performance being "by chance," then I should be able to show 2,000 other players who did not match the computer as much.
- (Show "Control" site on Internet. "Opens" page is still private, but top of it was NY Times graphic.)
- But note—if I have many more performances, say over **20,000**, then I should expect to see higher match % by non-cheating players! "Littlewood's Law"
- (Show master MM% list. Still sensitive. . .)
- To be sure, stats must combine with other evidence.
- (show "Parable of the Golfers" page)

Wider Human Implications

- Aside from cheating, what does this tell us about humanity?
- Millions of pages of data on human decision-making. What patterns emerge?
- Data come from actual competitions, unlike studies based on simulated circumstances.

1. Perception Proportional to Benefit

How strongly do you perceive a difference of 10 kronor, if:

- You are buying lunch and a drink in a pub. (100 Kr)
- You are buying dinner in a restaurant. (400 Kr)
- You are buying an I-pod. (1000 Kr)
- You are buying a car. (100,000 Kr)

For the car, maybe you don't care. In other cases, would you be equally thrifty?

> *If you spend the way you play chess, you care maybe* 4× *as much in the pub!*

(Show pages from Net, or show next two slides.)

Average Error

Eval for PTM:	Error(.cp)	/ #moves	= AE
-1.00 -- -0.91:	2370.72	/ 14312	= 0.1656
-0.90 -- -0.81:	2537.31	/ 16929	= 0.1499
-0.80 -- -0.71:	2357.24	/ 17982	= 0.1311
-0.70 -- -0.61:	2794.65	/ 23956	= 0.1167
-0.60 -- -0.51:	3525.21	/ 32718	= 0.1077
-0.50 -- -0.41:	3155.00	/ 33945	= 0.0929
-0.40 -- -0.31:	4203.85	/ 50242	= 0.0837
-0.30 -- -0.21:	4990.28	/ 65310	= 0.0764
-0.20 -- -0.11:	6346.10	/ 89116	= 0.0712
-0.10 -- -0.01:	5745.90	/ 84775	= 0.0678
0.00 -- 0.00:	7931.69	/ 95112	= 0.0834
0.01 -- 0.10:	4927.55	/ 87933	= 0.0560
0.11 -- 0.20:	6025.43	/ 97595	= 0.0617
0.21 -- 0.30:	5215.15	/ 75272	= 0.0693
0.31 -- 0.40:	4605.31	/ 59469	= 0.0774
0.41 -- 0.50:	3392.78	/ 40222	= 0.0844
0.51 -- 0.60:	3510.60	/ 38036	= 0.0923
0.61 -- 0.70:	2728.45	/ 27891	= 0.0978
0.71 -- 0.80:	1999.12	/ 20280	= 0.0986
0.81 -- 0.90:	1956.12	/ 18954	= 0.1032
0.91 -- 1.00:	1685.87	/ 15973	= 0.1055

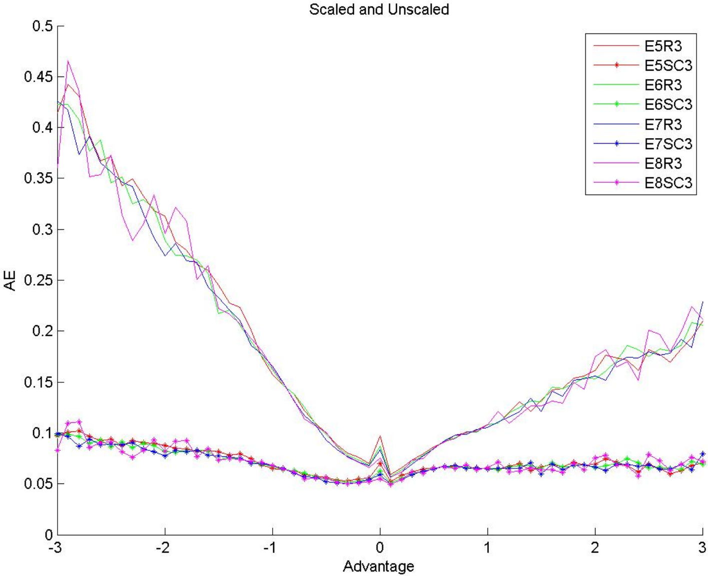
Table covers all Cat. 11 and higher tournaments played in 2000—2009.

Read: In 65,310 positions the player to move was judged 21 to 30 cp behind, and made a (raw, unscaled) “error” of 7.64 cp per move.

Scripts miss some non-immediate repetitions, hence 0.00 eval set aside.

Raw figures say players make 60-90% more error when half a pawn ahead or behind than when the game is even.

Is this a “real” verdict on skill in these cases? We think not. Instead we deduce a proportionality law.


Scaled and Unscaled
E5R3
E5SC3
E6R3
E6SC3
E7R3
E7SC3
E8R3
E8SC3
AE
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-3
-2
-1
0
1
2
3
Advantage

2. Is Savielly Tartakover Right?

The winner is the player who makes the next-to-last blunder.

- We like to think chess is about Deep Strategy.
- This helps, but is it statistically dominated by blunders?
- Recent Examples:
  - USA-Russia and USA-China matches at 2012 Olympiad.
  - Gelfand-Anand 2012 Rapid playoff.
- My Average Error (AE) stat shows a tight linear fit to Elo rating.
- Full investigation will need ANOVA (analysis of variance).

3. Procrastination...

- (Show graph of AE climbing to Move 40, then falling.)
- Aug. 2012 *New In Chess*, Kramnik-Grischuk, Moscow Tal Mem.
    - King's Indian: 12. Bf3!? then 13. Bg2 N (novelty)
    - "Grischuk was already in some time pressure."

3. Procrastination...

- (Show graph of AE climbing to Move 40, then falling.)
- Aug. 2012 *New In Chess*, Kramnik-Grischuk, Moscow Tal Mem.
  - King's Indian: 12. Bf3!? then 13. Bg2 N (novelty)
  - "Grischuk was already in some time pressure."

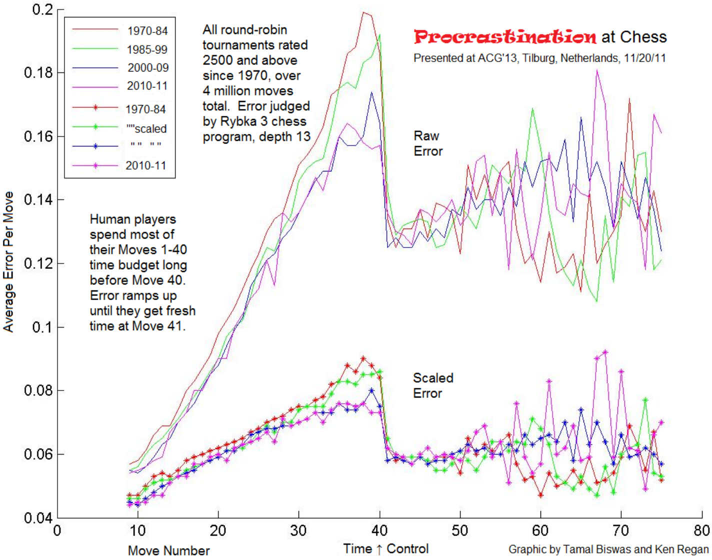
- IPR for Astana World Blitz 2012 (cat. 19, 2715 avg.) time control 3' + 2"/move: 2135 ±49. Difference of -580.
- IPR for Amber Rapid 2010+2011 (cat. 20+21, 2758 avg.) time control 25' + 10"/move: 2549 ±57. Difference of -210.
- IPR for Astana World Rapid 2012: (cat. 19, 2715 avg.) time control <u>15'</u> + 10"/move: 2394 ±62. Difference of -320.

3. Procrastination...

- (Show graph of AE climbing to Move 40, then falling.)
- Aug. 2012 *New In Chess*, Kramnik-Grischuk, Moscow Tal Mem.
  - King's Indian: 12. Bf3!? then 13. Bg2 N (novelty)
  - "Grischuk was already in some time pressure."

- IPR for Astana World Blitz 2012 (cat. 19, 2715 avg.) time control 3' + 2"/move: 2135 ±49. Difference of -580.
- IPR for Amber Rapid 2010+2011 (cat. 20+21, 2758 avg.) time control 25' + 10"/move: 2549 ±57. Difference of -210.
- IPR for Astana World Rapid 2012: (cat. 19, 2715 avg.) time control 15' + 10"/move: 2394 ±62. Difference of -320.

- *Can players be coached to play like the young Anand?*



All round-robin tournaments rated 2500 and above since 1970, over 4 million moves total. Error judged by Rybka 3 chess program, depth 13

Human players spend most of their Moves 1-40 time budget long before Move 40. Error ramps up until they get fresh time at Move 41.

Procrastination at Chess

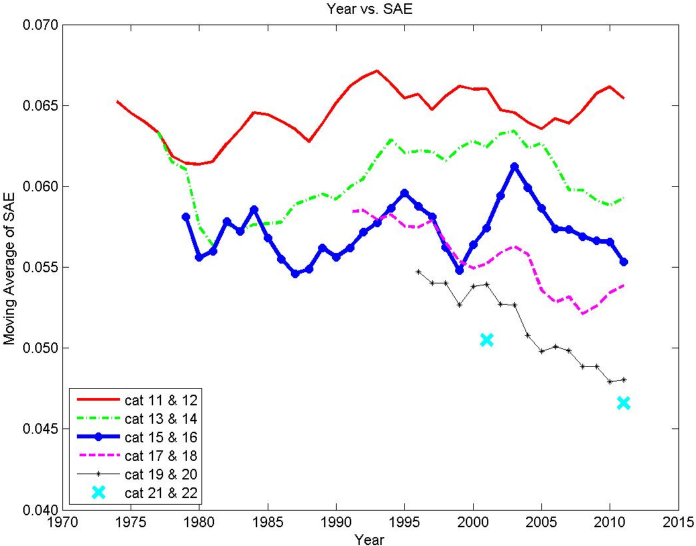
Presented at ACG'13, Tilburg, Netherlands, 11/20/11

Graphic by Tamal Biswas and Ken Regan

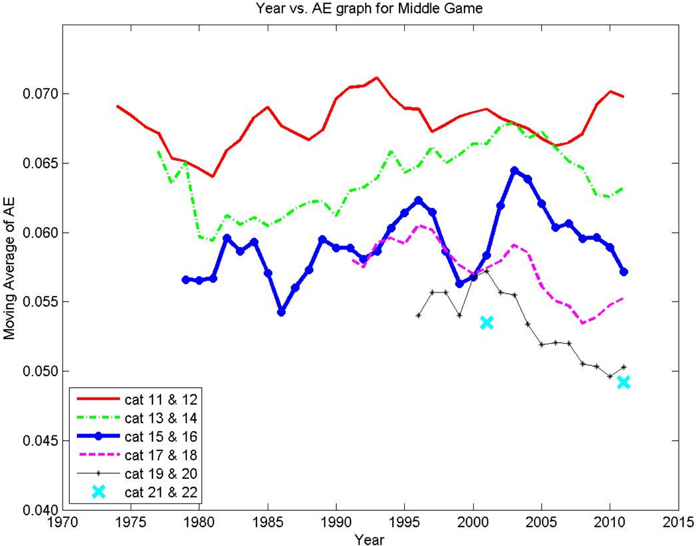
4. Human Skill Increasing Over Time?

- In 1970s, two 2700+ players: Fischer and Karpov. In 1981: none!
- Sep. 2012 list, 44 2700+ players. Rating Inflation?
- My results:
- 1976–1979 vs. 1991–1994 vs. 2006–2009: Little or no difference in IPR at all rating levels.
- 2600 level, 1971–present:
  - Can argue 30-pt. IPR difference between 1980's and now.
  - Difference measured at 16 pts. using 4-yr. moving averages, 10-year blocks.
  - Explainable by faster time controls, no adjournments?
- Single-PV AE stat in **all** Cat 11+ RRs since 1971 hints at mild **deflation**.
- Moves 17–32 show similar results. Hence not just due to better opening prep?
- Increasing skill consistent with Olympics results.

Error Mostly Constant Per Rating Level



Also Constant For Moves 17–32 Only



5. Variance in Performance, and Motivation?

- Let's say I am 2400 facing 2600 player.
- My expectation is 25%. Maybe:
  - 60% win for stronger player.
  - 30% draw.
  - 10% chance of win for me.
- In **12-game match**, maybe under 1% chance of winning if we are random.
- But my model's intrinsic error bars are often 200 points wide over 9–12 games.
- Suggests to take event not game as the unit.
- How can we be motivated for events?
- (Show examples, e.g. this about Svetozar Gligoric.)

6. Are We Reliable?

- One blunder in 200 moves can "ruin" a tournament.
- But we were reliable 99.5% of the time.
- Exponential $g(s,c)$ curve fits better than inverse-poly ones.
- Contrary to my "Black Swan" expectation.
- But we are even more reliable if we can use a computer...
- (Show PAL/CSS Freestyle stats if time...)

7. Not Just About Chess?

- *Only chess aspect of entire work is the evaluations coming from chess engines.*
- No special chess-knowledge, no "style" (except as reflected in fitted s, c, d).
- General Problem: **Converting Utilities Into Probabilities** for *fallible agents*.
- Framework applies to multiple-choice tests, now prevalent in online courses.
- Alternative to current psychometric measures?
- Issue: Idea of "best move" at chess is the same for all human players, but "best move" in sports may depend on natural talent.

Conclusions

- Lots more to do!
- Can use helpers!
    - Run data with other engines, such as Stockfish.
    - Run more tournaments.
    - Run to higher depths—how much does that matter?
- Spread word about general-scientific aspects; fight gullibility and paranoia over cheating.
- Deter cheating too.
- Learn more about human decision making.
- Thus the Turing Tour comes back to the human mind.
- **Thank you very much for the invitation.**

Addendum: Some Technical Slides

Let $\Pr_E(m_i)$ stand for the probability that a player of Elo skill rating E will choose move m_i in a given position.

- Too Simple:

  $$\Pr_E(m_i) \sim g(E, val(m_i)).$$

  Doesn't take values of the other moves into account.
- Cogent answer—let m_1 be the engine's top-valued move:

  $$\frac{\Pr_E(m_i)}{\Pr_E(m_1)} \sim g(E, val(m_1) - val(m_i)).$$

  That and $\sum_i \Pr(E, m_i) = 1$ minimally give Main Principle.
- Needs Multi-PV analysis—already beyond Guid-Bratko work.
- Single-PV data on millions of moves shows other improvements.

Better, and Best?

Need a general function f and a function $\delta(i)$ giving a *scaled-down* difference in value from m_1 to m_i.

$$\frac{f(\Pr_E(m_i))}{f(\Pr_E(m_1))} = g(E, \delta(i)).$$

Implemented with $f = \log$ and log-log scaling, as guided by the data.

Best model? Let *weights* w_d at different *engine depths* d reflect a player's depth of calculation. Apply above equation to evals at each depth d to define $\Pr_E(m_i, d)$. Then define:

$$\Pr_E(m_i) = \sum_d w_d \cdot \Pr_E(m_i, d).$$

This accounts for moves that *swing* in value and idea that weaker players prefer weaker moves. **In Process Now.**

Why Desire Probabilities?

- Allows to *predict* the # N of agreements with any sequence of moves m_*^t over game turns t, not just computer's first choices:

$$N = \sum_t \Pr_E(m_*^t).$$

- and it gives confidence intervals for N.
- Also predicts *aggregate error* (AE, scaled) by

$$e = \sum_t \sum_i \delta(i) \cdot \Pr_E(m_i^t).$$

  Comparing e with the *actual* error e' by a player over the same turns leads to a "virtual Elo rating" E' for those moves.

- IPR $\equiv$ "Intrinsic Performance Rating."

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- 25th in line for throne of Monaco.
- Now does "**Solitaire Chess**" for Chess Life magazine:
  - Reader covers gamescore, tries to guess each move by one side.
  - E.g. score 6 pts. if you found 15.Re1, 4 pts. for 15.h3, 1 pt. for premature 15.Ng5.
  - Add points at end: say 150=GM, 140=IM, 120=Master, 80 = 1800 player, etc.
- Is it scientific?
- With my formulas, yes—using *your* games in *real* tournaments.

Judgment By Your Peers

Training Sets: Multi-PV analyze games with both players rated:

- 2690–2710, in 2006–2009 and 1991–1994
- 2590–2610, "" "", extended to 2580–2620 in 1976–1979
- 2490–2510, all three times
- 2390–2410, (lower sets have over 20,000 moves)
- 2290–2310, (all sets elim. moves 1–8, moves in repetitions,
- 2190–2210, (and moves with one side $>$ 3 pawns ahead)
- Down to 1590–1610 for years 2006–2009 only.
- 2600-level set done for **all years since 1971**.

Training the Parameters

- Formula $g(E;\delta)$ is really

$$g(s,c;\delta) = \frac{1}{e^{x^c}} \quad \text{where} \quad x = \frac{\delta}{s}.$$

- s for *Sensitivity*: smaller $s \equiv$ better ability to sense small differences in value.
- c for *Consistency*: higher c reduces probability of high-δ moves (i.e., blunders).
- Full model will have parameter d for depth of calculation.

Fitting and Fighting Parameters

- For each Elo E training set, find (s, c) giving best fit.
- Can use many different fitting methods...
    - Can compare methods...
    - Whole separate topic...
    - Max-Likelihood does *poorly*.
- Often s and c trade off badly, but $E' \sim e(s, c)$ condenses into one Elo.
- Strong linear fit—suggests Elo mainly influenced by error.