Digital Mindprints

Kenneth W. Regan
University at Buffalo (SUNY)

Workshop on "Digital Assistance," Puné, 6 August 2016

Main Question

My original talk thesis from the human point of view:

> Does the use of digital assistants alter our thinking in ways demonstrably different from human agents acting alone?

Main Question

My original talk thesis from the human point of view:

> Does the use of digital assistants alter our thinking in ways demonstrably different from human agents acting alone?

Speaking from the computer's point of view is also valid:

> Does the use of human assistants alter our thinking in ways demonstrably different from computer agents acting alone?

Main Question

My original talk thesis from the human point of view:

> Does the use of digital assistants alter our thinking in ways demonstrably different from human agents acting alone?

Speaking from the computer's point of view is also valid:

> Does the use of human assistants alter our thinking in ways demonstrably different from computer agents acting alone?

- Should "we" cut humans out of the loop in driverless cars?

Main Question

My original talk thesis from the human point of view:

> Does the use of digital assistants alter our thinking in ways demonstrably different from human agents acting alone?

Speaking from the computer's point of view is also valid:

> Does the use of human assistants alter our thinking in ways demonstrably different from computer agents acting alone?

- Should "we" cut humans out of the loop in driverless cars? the military?

Main Question

My original talk thesis from the human point of view:

> Does the use of digital assistants alter our thinking in ways demonstrably different from human agents acting alone?

Speaking from the computer's point of view is also valid:

> Does the use of human assistants alter our thinking in ways demonstrably different from computer agents acting alone?

- Should "we" cut humans out of the loop in driverless cars? the military? financial trading?

Main Question

My original talk thesis from the human point of view:

> Does the use of digital assistants alter our thinking in ways demonstrably different from human agents acting alone?

Speaking from the computer's point of view is also valid:

> Does the use of human assistants alter our thinking in ways demonstrably different from computer agents acting alone?

- Should "we" cut humans out of the loop in driverless cars? the military? financial trading? other daily applications?

Main Question

My original talk thesis from the human point of view:

> Does the use of digital assistants alter our thinking in ways demonstrably different from human agents acting alone?

Speaking from the computer's point of view is also valid:

> Does the use of human assistants alter our thinking in ways demonstrably different from computer agents acting alone?

- Should "we" cut humans out of the loop in driverless cars? the military? financial trading? other daily applications?
- What are important differences in cognitive tendencies?

Main Question

My original talk thesis from the human point of view:

> Does the use of digital assistants alter our thinking in ways demonstrably different from human agents acting alone?

Speaking from the computer's point of view is also valid:

> Does the use of human assistants alter our thinking in ways demonstrably different from computer agents acting alone?

- Should "we" cut humans out of the loop in driverless cars? the military? financial trading? other daily applications?
- What are important differences in cognitive tendencies?
- Dores each side need to do *xenospection*—building a model of the other's characteristic behavior?

Chess For “Mindprints”

- Chess is not a normal daily human activty.

Chess For "Mindprints"

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.

Chess For “Mindprints”

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.
- “The Stress of Chess” (autobio by late US grandmaster Walter Browne).

Chess For "Mindprints"

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.
- "The Stress of Chess" (autobio by late US grandmaster Walter Browne).
- Idea is that from specialized situations with extreme human input we can isolate distinct patterns in cognition.

Chess For "Mindprints"

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.
- "The Stress of Chess" (autobio by late US grandmaster Walter Browne).
- Idea is that from specialized situations with extreme human input we can isolate distinct patterns in cognition.
- Copious data from real competitions not simulations.

Chess For “Mindprints”

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.
- “The Stress of Chess” (autobio by late US grandmaster Walter Browne).
- Idea is that from specialized situations with extreme human input we can isolate distinct patterns in cognition.
- Copious data from real competitions not simulations.
- Computers chess programs (called *engines*) now *outclass* all human players, even when running on commodity hardware.

Chess For “Mindprints”

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.
- “The Stress of Chess” (autobio by late US grandmaster Walter Browne).
- Idea is that from specialized situations with extreme human input we can isolate distinct patterns in cognition.
- Copious data from real competitions not simulations.
- Computers chess programs (called *engines*) now *outclass* all human players, even when running on commodity hardware.
- Positive benefit: it enables authoritative judgment of human decisions

Chess For "Mindprints"

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.
- "The Stress of Chess" (autobio by late US grandmaster Walter Browne).
- Idea is that from specialized situations with extreme human input we can isolate distinct patterns in cognition.
- Copious data from real competitions not simulations.
- Computers chess programs (called *engines*) now *outclass* all human players, even when running on commodity hardware.
- Positive benefit: it enables authoritative judgment of human decisions *intrinsically*, rather than "extrinsically" from results of games.

Chess For "Mindprints"

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.
- "The Stress of Chess" (autobio by late US grandmaster Walter Browne).
- Idea is that from specialized situations with extreme human input we can isolate distinct patterns in cognition.
- Copious data from real competitions not simulations.
- Computers chess programs (called *engines*) now *outclass* all human players, even when running on commodity hardware.
- Positive benefit: it enables authoritative judgment of human decisions *intrinsically*, rather than "extrinsically" from results of games. V. Anand plays fewer than 100 rated games per year—

Chess For "Mindprints"

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.
- "The Stress of Chess" (autobio by late US grandmaster Walter Browne).
- Idea is that from specialized situations with extreme human input we can isolate distinct patterns in cognition.
- Copious data from real competitions not simulations.
- Computers chess programs (called *engines*) now *outclass* all human players, even when running on commodity hardware.
- Positive benefit: it enables authoritative judgment of human decisions *intrinsically*, rather than "extrinsically" from results of games. V. Anand plays fewer than 100 rated games per year—but they furnish over 3,000 relevant decisions by him to analyze.

Chess For "Mindprints"

- Chess is not a normal daily human activty.
- Players in events *rated* by the World Chess Federation (FIDE) : under 500,000 worldwide.
- "The Stress of Chess" (autobio by late US grandmaster Walter Browne).
- Idea is that from specialized situations with extreme human input we can isolate distinct patterns in cognition.
- Copious data from real competitions not simulations.
- Computers chess programs (called *engines*) now *outclass* all human players, even when running on commodity hardware.
- Positive benefit: it enables authoritative judgment of human decisions *intrinsically*, rather than "extrinsically" from results of games. V. Anand plays fewer than 100 rated games per year—but they furnish over 3,000 relevant decisions by him to analyze.
- Negative side: "E-Doping" by human players. . .

Cheating Detection

- My project began with the near-breakup of the 2006 World Championship Match between V. Topalov and V. Kramnik over allegations that Kramnik was getting help from the engine Fritz 9 via Internet cable to his *toilet*, the only part of the stage area off-camera.

Cheating Detection

- My project began with the near-breakup of the 2006 World Championship Match between V. Topalov and V. Kramnik over allegations that Kramnik was getting help from the engine Fritz 9 via Internet cable to his *toilet*, the only part of the stage area off-camera.
- Only evidence given was a letter stating abnormally high agreement between Kramnik's moves and Fritz 9's recommendations. (Fritz 10 beat Kramnik in a 6-game match two months later.)

Cheating Detection

- My project began with the near-breakup of the 2006 World Championship Match between V. Topalov and V. Kramnik over allegations that Kramnik was getting help from the engine Fritz 9 via Internet cable to his *toilet*, the only part of the stage area off-camera.
- Only evidence given was a letter stating abnormally high agreement between Kramnik's moves and Fritz 9's recommendations. (Fritz 10 beat Kramnik in a 6-game match two months later.)
- **How can one evaluate such claims statistically?**

Cheating Detection

- My project began with the near-breakup of the 2006 World Championship Match between V. Topalov and V. Kramnik over allegations that Kramnik was getting help from the engine Fritz 9 via Internet cable to his *toilet*, the only part of the stage area off-camera.
- Only evidence given was a letter stating abnormally high agreement between Kramnik's moves and Fritz 9's recommendations. (Fritz 10 beat Kramnik in a 6-game match two months later.)
- **How can one evaluate such claims statistically?**
- I developed a *predictive analytic* model that provides statistical hypothesis tests.

Cheating Detection

- My project began with the near-breakup of the 2006 World Championship Match between V. Topalov and V. Kramnik over allegations that Kramnik was getting help from the engine Fritz 9 via Internet cable to his *toilet*, the only part of the stage area off-camera.
- Only evidence given was a letter stating abnormally high agreement between Kramnik's moves and Fritz 9's recommendations. (Fritz 10 beat Kramnik in a 6-game match two months later.)
- **How can one evaluate such claims statistically?**
- I developed a *predictive analytic* model that provides statistical hypothesis tests.
- It explained Kramnik's higher concordance by how Topalov had severely limited his options.

Cheating Detection

- My project began with the near-breakup of the 2006 World Championship Match between V. Topalov and V. Kramnik over allegations that Kramnik was getting help from the engine Fritz 9 via Internet cable to his *toilet*, the only part of the stage area off-camera.
- Only evidence given was a letter stating abnormally high agreement between Kramnik's moves and Fritz 9's recommendations. (Fritz 10 beat Kramnik in a 6-game match two months later.)
- **How can one evaluate such claims statistically?**
- I developed a *predictive analytic* model that provides statistical hypothesis tests.
- It explained Kramnik's higher concordance by how Topalov had severely limited his options.
- Since 2013 I've been an official advisor to FIDE's Anti-Cheating Committee (ACC).

Cheating Detection

- My project began with the near-breakup of the 2006 World Championship Match between V. Topalov and V. Kramnik over allegations that Kramnik was getting help from the engine Fritz 9 via Internet cable to his *toilet*, the only part of the stage area off-camera.
- Only evidence given was a letter stating abnormally high agreement between Kramnik's moves and Fritz 9's recommendations. (Fritz 10 beat Kramnik in a 6-game match two months later.)
- **How can one evaluate such claims statistically?**
- I developed a *predictive analytic* model that provides statistical hypothesis tests.
- It explained Kramnik's higher concordance by how Topalov had severely limited his options.
- Since 2013 I've been an official advisor to FIDE's Anti-Cheating Committee (ACC). More cases than I wish would happen. . .

Larger Issue: Skill Assessment “Versus” Prediction

- Skill Assessment: how well people did.

Larger Issue: Skill Assessment "Versus" Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.

Larger Issue: Skill Assessment "Versus" Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?

Larger Issue: Skill Assessment “Versus” Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?
- *Meta:* Is this performance really by this person?

Larger Issue: Skill Assessment "Versus" Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?
- *Meta:* Is this performance really by this person?

- Chess cheating detection needs separating the two in order that it is not merely flagging people for "playing too well."

Larger Issue: Skill Assessment "Versus" Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?
- *Meta:* Is this performance really by this person?

- Chess cheating detection needs separating the two in order that it is not merely flagging people for "playing too well."
- First key: Chess programs revise their move values by increasing *depth* of search.

Larger Issue: Skill Assessment "Versus" Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?
- *Meta:* Is this performance really by this person?

- Chess cheating detection needs separating the two in order that it is not merely flagging people for "playing too well."
- First key: Chess programs revise their move values by increasing *depth* of search.
- My student Tamal Biswas especially demonstrated *strong* impact of lower-depth appearances on human choices.

Larger Issue: Skill Assessment "Versus" Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?
- *Meta:* Is this performance really by this person?

- Chess cheating detection needs separating the two in order that it is not merely flagging people for "playing too well."
- First key: Chess programs revise their move values by increasing *depth* of search.
- My student Tamal Biswas especially demonstrated *strong* impact of lower-depth appearances on human choices.
- So: *predict* using lower depths;

Larger Issue: Skill Assessment "Versus" Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?
- *Meta:* Is this performance really by this person?

- Chess cheating detection needs separating the two in order that it is not merely flagging people for "playing too well."
- First key: Chess programs revise their move values by increasing *depth* of search.
- My student Tamal Biswas especially demonstrated *strong* impact of lower-depth appearances on human choices.
- So: *predict* using lower depths; *assess* on the highest-depth values.

Larger Issue: Skill Assessment "Versus" Prediction

- Skill Assessment: how well people did.
- Prediction: how well people will do.
- Both: how unusual is how well some person did?
- *Meta:* Is this performance really by this person?

- Chess cheating detection needs separating the two in order that it is not merely flagging people for "playing too well."
- First key: Chess programs revise their move values by increasing *depth* of search.
- My student Tamal Biswas especially demonstrated *strong* impact of lower-depth appearances on human choices.
- So: *predict* using lower depths; *assess* on the highest-depth values.
- Second key: human-computer cognitive differences.

Cycling Analogy

- E-Doping means cheating with computer assistance.

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."
2. "Person X cannot make a champion spin and jump and shoot so fast and accurately.

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."
2. "Person X cannot make a champion spin and jump and shoot so fast and accurately. *versus:*

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."
2. "Person X cannot make a champion spin and jump and shoot so fast and accurately. *versus:*
3. "Person X has hematocrit $> 50\%$."

Cycling Analogy

- E-Doping means cheating with computer assistance.
- Jan. 2013: Lance Armstrong (cycling) and Borislav Ivanov (chess) in news at same time.
- Applies to online games in much greater volume than chess.

1. "Person X cannot cycle up *that* hill *that* fast."
2. "Person X cannot make a champion spin and jump and shoot so fast and accurately. *versus:*
3. "Person X has hematocrit $> 50\%$."
4. "Person X made moves highly similar to Code Patch Y."

Elo Rating System

- Named for the Hungarian-American statistician **Arpad Elo**, the system gives every player P a number R_P representing skill.

Elo Rating System

- Named for the Hungarian-American statistician **Arpad Elo**, the system gives every player P a number R_P representing skill.
- Pivotal Idea: a 200-point difference $\approx$ 75% expectation.

Elo Rating System

- Named for the Hungarian-American statistician **Arpad Elo**, the system gives every player P a number R_P representing skill.
- Pivotal Idea: a 200-point difference $\approx$ 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen 2857, 3 others over 2800, Anand 2770.

Elo Rating System

- Named for the Hungarian-American statistician **Arpad Elo**, the system gives every player P a number R_P representing skill.
- Pivotal Idea: a 200-point difference $\approx$ 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen 2857, 3 others over 2800, Anand 2770. Adult beginner $\approx$ 600, kids $\to$ 100.

Elo Rating System

- Named for the Hungarian-American statistician **Arpad Elo**, the system gives every player P a number R_P representing skill.
- Pivotal Idea: a 200-point difference $\approx$ 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen 2857, 3 others over 2800, Anand 2770. Adult beginner $\approx$ 600, kids $\to$ 100.
- Komodo 10.1 **3230**, Stockfish 7+ **3225**, Houdini 4+ **3175**, Gull 3 **3125**...

Elo Rating System

- Named for the Hungarian-American statistician **Arpad Elo**, the system gives every player P a number R_P representing skill.
- Pivotal Idea: a 200-point difference $\approx$ 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = "Super GM,"; Carlsen 2857, 3 others over 2800, Anand 2770. Adult beginner $\approx$ 600, kids $\rightarrow$ 100.
- Komodo 10.1 **3230**, Stockfish 7+ **3225**, Houdini 4+ **3175**, Gull 3 **3125**... So computers are at "Class 13"—a kind of "Moore's Law."

Elo Rating System

- Named for the Hungarian-American statistician **Arpad Elo**, the system gives every player P a number R_P representing skill.
- Pivotal Idea: a 200-point difference $\approx$ 75% expectation.
- **Class Units**: 2000–2200 = Expert, 2200–2400 = Master, 2400–2600 is typical of International/Senior Master and Grandmaster ranks, 2600–2800 = “Super GM,”; Carlsen 2857, 3 others over 2800, Anand 2770. Adult beginner $\approx$ 600, kids $\to$ 100.
- Komodo 10.1 **3230**, Stockfish 7+ **3225**, Houdini 4+ **3175**, Gull 3 **3125**... So computers are at “Class 13”—a kind of “Moore’s Law.”
- US: “Class A” = 1800–2000, “B” = 1600–1800, “C” = 1400–1600, ...; adult beginner said to be 600; scholastics down to minumum 100 rating.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.

Case Example: April 2015

- The “San Sebastian Open”—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles. . .

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles...
- But my full cheating test showed only a "1.3-sigma" deviation,

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles. . .
- But my full cheating test showed only a "1.3-sigma" deviation, and his IPR was "only" 2455 also within the "2-sigma" range.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles. . .
- But my full cheating test showed only a "1.3-sigma" deviation, and his IPR was "only" 2455 also within the "2-sigma" range.
- Was dead lost against Epishin, lucked out also in previous round,

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles. . .
- But my full cheating test showed only a "1.3-sigma" deviation, and his IPR was "only" 2455 also within the "2-sigma" range.
- Was dead lost against Epishin, lucked out also in previous round,
- World #2 Fabiano Caruana had sensational 7-win streak against the top last Sept.

Case Example: April 2015

- The "San Sebastian Open"—a 9-round, 8-day prize-giving Swiss—had players up to 2600, 24 above 2200, 170 players total.
- Surprise winner: 2115-rated Badr Al-Hajiri of Kuwait.
- Won last 3 games over a 2356, 2412, and GM Vl. Epishin, 2563.
- Loud "whispers" in various circles...
- But my full cheating test showed only a "1.3-sigma" deviation, and his IPR was "only" 2455 also within the "2-sigma" range.
- Was dead lost against Epishin, lucked out also in previous round,
- World #2 Fabiano Caruana had sensational 7-win streak against the top last Sept.—but his IPR was "only" 2900 while his opponents played under 2600.

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make. . .
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make. . .
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds “even out” over hundreds of betting events (for us, moves).

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make. . .
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
    - How often did the favored horses win in a racing week?

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
    - How often did the favored horses win in a racing week?
    - Do basketball teams average "covering their spread"?

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
  - How often did the favored horses win in a racing week?
  - Do basketball teams average "covering their spread"?
  - How often did Player X make the move favored by an engine?

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
  - How often did the favored horses win in a racing week?
  - Do basketball teams average "covering their spread"?
  - How often did Player X make the move favored by an engine?
  - How does his/her "Average Error" compare?

Prediction: Not the Bettor but the Book

- Not a crystal ball to say what move a player will make...
- Though a GM sports-analyst friend tells me there is real-time betting on chess moves in Germany.
- How a bookie sets odds—for the *initial betting line*.
- Accuracy is how well odds "even out" over hundreds of betting events (for us, moves).
- Quantify *aggregate statistics*:
  - How often did the favored horses win in a racing week?
  - Do basketball teams average "covering their spread"?
  - How often did Player X make the move favored by an engine?
  - How does his/her "Average Error" compare?
- Also project *standard deviation* and *confidence intervals*.

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i
3. Parameters: $s, c, \ldots$ denoting skills and levels.
   Trained correspondence to chess Elo rating E

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i
3. Parameters: $s, c, \ldots$ denoting skills and levels.
   Trained correspondence to chess Elo rating E
4. Defines *fallible agent* $P(s, c, \ldots)$.

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i
3. Parameters: $s, c, \ldots$ denoting skills and levels.
   Trained correspondence to chess Elo rating E
4. Defines *fallible agent* $P(s, c, \ldots)$.
5. Main Output: Probabilities $p_{t,i}$ for $P(s, c, \ldots)$ to select option i at time t.

Context: Decision-Making Model at Chess

1. Domain: A set T of decision-making situations t.
   Chess game turns
2. Inputs: Values v_i for every option at turn t.
   Computer values of moves m_i
3. Parameters: $s, c, \dots$ denoting skills and levels.
   Trained correspondence to chess Elo rating E
4. Defines *fallible agent* $P(s, c, \dots)$.
5. Main Output: Probabilities $p_{t,i}$ for $P(s, c, \dots)$ to select option i at time t.
6. Derived Outputs:
   - Aggregate statistics: *move-match* **MM**, *equal-top value* **EV**, *average scaled difference* **ASD**, ...
   - Projected confidence intervals: Bernoulli Trials + $|T|$-adjustment.
   - IPRs similarly reflect errors from the regression.

Reducing Chess to Numbers

- Chess engines all work by *incremental search* in rounds of increasing *depth* $d = 1, 2, 3, \ldots$

Reducing Chess to Numbers

- Chess engines all work by *incremental search* in rounds of increasing *depth* $d = 1, 2, 3, \ldots$
- For each round d and legal move m_i the program outputs a value $v_{i,d}$ in units of 0.01 called *centipawns*, figuratively 100ths of a pawn value (roughly P = 1, N = 3, B = 3+, R = 5, Q = 9).

Reducing Chess to Numbers

- Chess engines all work by *incremental search* in rounds of increasing *depth* $d = 1, 2, 3, \ldots$
- For each round d and legal move m_i the program outputs a value $v_{i,d}$ in units of 0.01 called *centipawns*, figuratively 100ths of a pawn value (roughly P = 1, N = 3, B = 3+, R = 5, Q = 9).
- Values by Stockfish 6 in key Kramnik-Anand WC 2008 position:

Move	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nd2	103	093	087	093	027	028	000	000	056	-007	039	028	037	020	014	017	000	006	000
Bxd7	048	034	-033	-033	-013	-042	-039	-050	-025	-010	001	000	-009	-027	-018	000	000	000	000
Qg8	114	114	-037	-037	-014	-014	-022	-068	-008	-056	-042	-004	-032	000	-014	-025	-045	-045	-050
...			...			...			...			...			...			...	
Nxd4	-056	-056	-113	-071	-071	-145	-020	-006	077	052	066	040	050	051	-181	-181	-181	-213	-213

Reducing Chess to Numbers

- Chess engines all work by *incremental search* in rounds of increasing *depth* $d = 1, 2, 3, \ldots$
- For each round d and legal move m_i the program outputs a value $v_{i,d}$ in units of 0.01 called *centipawns*, figuratively 100ths of a pawn value (roughly P = 1, N = 3, B = 3+, R = 5, Q = 9).
- Values by Stockfish 6 in key Kramnik-Anand WC 2008 position:

Move	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nd2	103	093	087	093	027	028	000	000	056	-007	039	028	037	020	014	017	000	006	000
Bxd7	048	034	-033	-033	-013	-042	-039	-050	-025	-010	001	000	-009	-027	-018	000	000	000	000
Qg8	114	114	-037	-037	-014	-014	-022	-068	-008	-056	-042	-004	-032	000	-014	-025	-045	-045	-050
...			...			...			...			...			...			...	
Nxd4	-056	-056	-113	-071	-071	-145	-020	-006	077	052	066	040	050	051	-181	-181	-181	-213	-213

- Note that two moves have “equal-top value” (EV).

Reducing Chess to Numbers

- Chess engines all work by *incremental search* in rounds of increasing *depth* $d = 1, 2, 3, \ldots$
- For each round d and legal move m_i the program outputs a value $v_{i,d}$ in units of 0.01 called *centipawns*, figuratively 100ths of a pawn value (roughly P = 1, N = 3, B = 3+, R = 5, Q = 9).
- Values by Stockfish 6 in key Kramnik-Anand WC 2008 position:

Move	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nd2	103	093	087	093	027	028	000	000	056	-007	039	028	037	020	014	017	000	006	000
Bxd7	048	034	-033	-033	-013	-042	-039	-050	-025	-010	001	000	-009	-027	-018	000	000	000	000
Qg8	114	114	-037	-037	-014	-014	-022	-068	-008	-056	-042	-004	-032	000	-014	-025	-045	-045	-050
...			...			...			...			...			...			...	
Nxd4	-056	-056	-113	-071	-071	-145	-020	-006	077	052	066	040	050	051	-181	-181	-181	-213	-213

- Note that two moves have "equal-top value" (EV).
- This happens in 8–10% of positions.

Reducing Chess to Numbers

- Chess engines all work by *incremental search* in rounds of increasing *depth* $d = 1, 2, 3, \ldots$
- For each round d and legal move m_i the program outputs a value $v_{i,d}$ in units of 0.01 called *centipawns*, figuratively 100ths of a pawn value (roughly P = 1, N = 3, B = 3+, R = 5, Q = 9).
- Values by Stockfish 6 in key Kramnik-Anand WC 2008 position:

Move	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Nd2	103	093	087	093	027	028	000	000	056	-007	039	028	037	020	014	017	000	006	000
Bxd7	048	034	-033	-033	-013	-042	-039	-050	-025	-010	001	000	-009	-027	-018	000	000	000	000
Qg8	114	114	-037	-037	-014	-014	-022	-068	-008	-056	-042	-004	-032	000	-014	-025	-045	-045	-050
...			...			...			...			...			...			...	
Nxd4	-056	-056	-113	-071	-071	-145	-020	-006	077	052	066	040	050	051	-181	-181	-181	-213	-213

- Note that two moves have "equal-top value" (EV).
- This happens in 8–10% of positions.
- *These values are* (currently) *the only chess-specific inputs.*

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...
- The **Main Equation:**

$$\frac{h(p_i)}{h(p_1)} = 1 - x_i$$

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...
- The **Main Equation:**

$$\frac{h(p_i)}{h(p_1)} = 1 - x_i = \exp(-\left(\frac{\delta(v_1, v_i)}{s}\right)^c),$$

- Here $\delta(v_1, v_i)$ scales $v_1 - v_i$ in regard to $|v_1|$.

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...
- The **Main Equation:**

$$\frac{h(p_i)}{h(p_1)} = 1 - x_i = \exp(-\left(\frac{\delta(v_1, v_i)}{s}\right)^c),$$

- Here $\delta(v_1, v_i)$ scales $v_1 - v_i$ in regard to $|v_1|$.
- Ratio not difference on LHS so x_i on RHS has 0-to-1 scale.

How the Model Operates

- Let v_1, v_i be values of the best move m_1 and ith-best move m_i.
- Given $s, c, \dots$, the model computes $x_i = g_{s,c}(v_1, v_i) =$ the **perceived inferiority** of m_i by $P(s, c, \dots)$.
- Besides g, the model picks a function $h(p_i)$ on probabilities.
- Could be $h(p) = p$ (bad), log (good enough?), $H(p_i)$, *logit*...
- The **Main Equation:**

$$\frac{h(p_i)}{h(p_1)} = 1 - x_i = \exp(-\left(\frac{\delta(v_1, v_i)}{s}\right)^c),$$

- Here $\delta(v_1, v_i)$ scales $v_1 - v_i$ in regard to $|v_1|$.
- Ratio not difference on LHS so x_i on RHS has 0-to-1 scale.
- Given $(x_1, \dots, x_i, \dots, x_\ell)$, fit subject to $\sum_i p_i = 1$ to find p_1. Other p_i follow by $p_i = h^{-1}(h(p_1)(1 - x_i))$.

The Data: Old and New

- **Old:** Over 3 million moves of **Multi-PV** data: $>$ 250 GB.

The Data: Old and New

- **Old:** Over 3 million moves of **Multi-PV** data: $>$ 250 GB.
- Over 40 million moves of **Single-PV** data: $>$ 50 GB

The Data: Old and New

- **Old:** Over 3 million moves of **Multi-PV** data: $>$ 250 GB.
- Over 40 million moves of **Single-PV** data: $>$ 50 GB
- $=$ 150 million pages of text data at 2k/page.

The Data: Old and New

- **Old:** Over 3 million moves of **Multi-PV** data: $>$ 250 GB.
- Over 40 million moves of **Single-PV** data: $>$ 50 GB
- $=$ 150 million pages of text data at 2k/page.
- All taken on two quad-core home-style PC's plus a laptop using the GUI. This involved **retaining hashed move values** between game turns—which is the normal playing mode and only GUI option.

The Data: Old and New

- **Old:** Over 3 million moves of **Multi-PV** data: $>$ 250 GB.
- Over 40 million moves of **Single-PV** data: $>$ 50 GB
- $=$ 150 million pages of text data at 2k/page.
- All taken on two quad-core home-style PC's plus a laptop using the GUI. This involved **retaining hashed move values** between game turns—which is the normal playing mode and only GUI option.
- **New—using UB's Center for Computational Research:** Every published high-level game since 2014 in **Single-PV** mode.

The Data: Old and New

- **Old:** Over 3 million moves of **Multi-PV** data: $>$ 250 GB.
- Over 40 million moves of **Single-PV** data: $>$ 50 GB
- = 150 million pages of text data at 2k/page.
- All taken on two quad-core home-style PC's plus a laptop using the GUI. This involved **retaining hashed move values** between game turns—which is the normal playing mode and only GUI option.
- **New—using UB's Center for Computational Research:** Every published high-level game since 2014 in **Single-PV** mode.
- **Master training sets** of 3 million moves by players of Elo ratings 1050, 1100, 1150, . . . (stepping by 50) . . . , 2700, 2750, 2800, all in **Multi-PV mode**.

The Data: Old and New

- **Old:** Over 3 million moves of **Multi-PV** data: $>$ 250 GB.
- Over 40 million moves of **Single-PV** data: $>$ 50 GB
- = 150 million pages of text data at 2k/page.
- All taken on two quad-core home-style PC's plus a laptop using the GUI. This involved **retaining hashed move values** between game turns—which is the normal playing mode and only GUI option.
- **New—using UB's Center for Computational Research:** Every published high-level game since 2014 in **Single-PV** mode.
- **Master training sets** of 3 million moves by players of Elo ratings 1050, 1100, 1150, . . . (stepping by 50) . . . , 2700, 2750, 2800, all in **Multi-PV mode**.
- Taken with Komodo 10 and Stockfish 7, *all years since 1971* except 2000,2050,. . . ,2500 years 2006–2015 only.

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- Now does "**Solitaire Chess**" for Chess Life magazine:

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- 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.

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- 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?

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- 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.

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- 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.
- Goal is **natural** scoring and distribution evaluation for multiple-choice tests, especially with partial-credit answers.

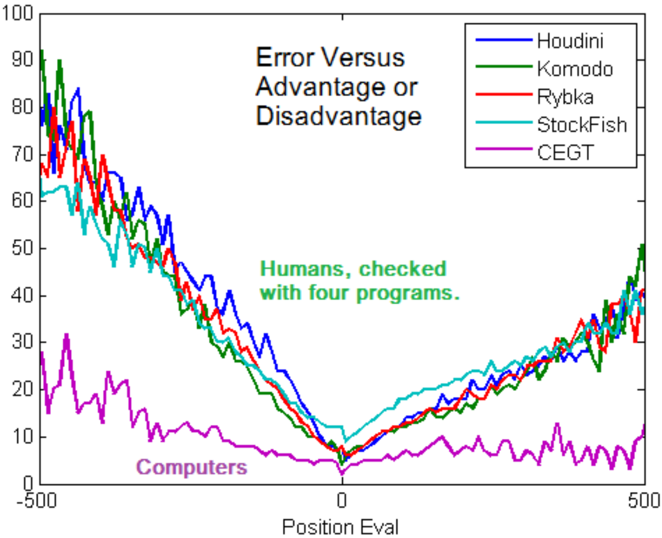
The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- 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.
- Goal is **natural** scoring and distribution evaluation for multiple-choice tests, especially with partial-credit answers.
- Connect to parameters in **Item-Response Theory** (IRT) test-taking models.

The Turing Pandolfini?

- Bruce Pandolfini — played by Ben Kingsley in "Searching for Bobby Fischer."
- 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.
- Goal is **natural** scoring and distribution evaluation for multiple-choice tests, especially with partial-credit answers.
- Connect to parameters in **Item-Response Theory** (IRT) test-taking models. IRT does both skill and prediction.

Human Versus Computer Phenomena



The Imbalance-Error Phenomenon

- Regard as real, or only as perception (and so correct for it)?

The Imbalance-Error Phenomenon

- Regard as real, or only as perception (and so correct for it)?
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \textit{with} \quad d\mu = \frac{1}{1+cx}\,dx$$

balances well, with c in the range 1.2–1.6 or so.

The Imbalance-Error Phenomenon

- Regard as real, or only as perception (and so correct for it)?
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \textit{with} \quad d\mu = \frac{1}{1+cx}\,dx$$

balances well, with c in the range 1.2–1.6 or so.

- A mix of three factors?

The Imbalance-Error Phenomenon

- Regard as real, or only as perception (and so correct for it)?
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \text{with} \quad d\mu = \frac{1}{1+cx}\,dx$$

balances well, with c in the range 1.2–1.6 or so.

- A mix of three factors?

(A) Human perception of value as proportional to stakes, *per* Ariely-Kahneman-Tversky.

The Imbalance-Error Phenomenon

- Regard as real, or only as perception (and so correct for it)?
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \text{with} \quad d\mu = \frac{1}{1+cx}\,dx$$

balances well, with c in the range 1.2–1.6 or so.

- A mix of three factors?

(A) Human perception of value as proportional to stakes, *per* Ariely-Kahneman-Tversky.

(B) Rationally playing less *catenaccio* when marginal impact of evaluation on win probability is minimal. (Leo Stedile, working under Mark Braverman)

The Imbalance-Error Phenomenon

- Regard as real, or only as perception (and so correct for it)?
- The *metric correction*

$$\int_{e-\delta}^{e} d\mu \quad \text{with} \quad d\mu = \frac{1}{1+cx}\,dx$$

balances well, with c in the range 1.2–1.6 or so.

- A mix of three factors?

(A) Human perception of value as proportional to stakes, *per* Ariely-Kahneman-Tversky.

(B) Rationally playing less *catenaccio* when marginal impact of evaluation on win probability is minimal. (Leo Stedile, working under Mark Braverman)

(C) Greater volatility intrinsic to chess as game progresses.

A. Perception Proportional to Benefit

How strongly do you perceive a difference of 500 rupees, if:

- You are buying lunch and a drink in a pub.
- You are buying dinner in a restaurant.
- You are buying an I-pad.
- You are buying a car.

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* $4\times$ *as much in the pub!*

B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

- Here a gives pretty steep slope near 0, $a \approx 4.5$ for Rybka and Houdini.

B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

- Here a gives pretty steep slope near 0, $a \approx 4.5$ for Rybka and Houdini.
- *How to test apart from cause A?*

B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

- Here a gives pretty steep slope near 0, $a \approx 4.5$ for Rybka and Houdini.
- *How to test apart from cause A?*
- Expect eval-error curve to shift in games between unequally-rated players.

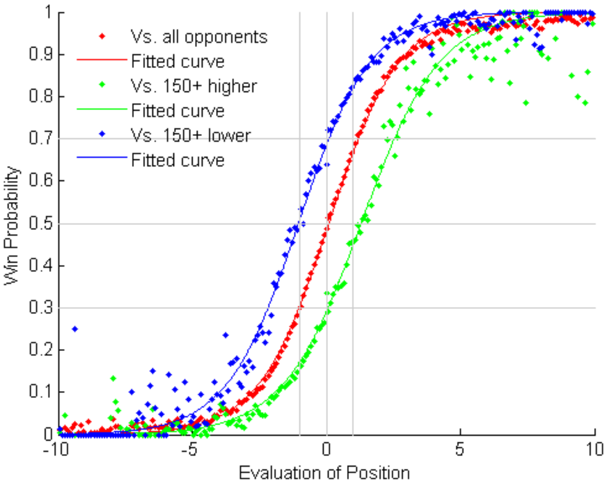
B. Rational Risk-Taking

- Expectation curves according to position evaluation v are sigmoidal, indeed close to a hyperbolic tangent

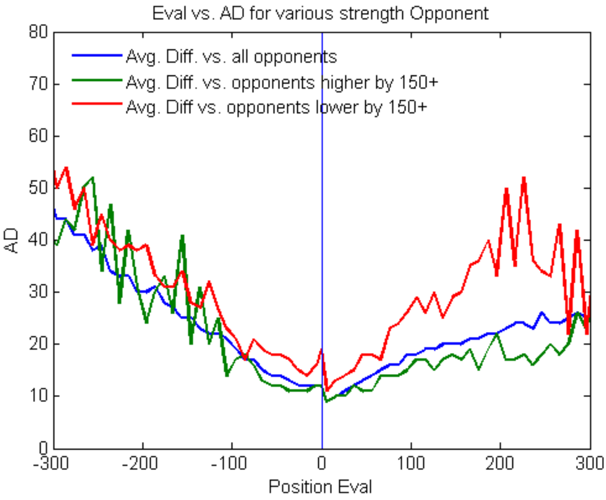
$$E = \frac{e^{av} - e^{-av}}{e^{av} + e^{-av}}.$$

- Here a gives pretty steep slope near 0, $a \approx 4.5$ for Rybka and Houdini.
- *How to test apart from cause A?*
- Expect eval-error curve to shift in games between unequally-rated players.
- Results so far show no shift—

Human Versus Computer Phenomena



Eval-Error Curve With Unequal Players



C. Complexity and Depth

- How to measure the *complexity* of a chess position?

C. Complexity and Depth

- How to measure the *complexity* of a chess position?
- Guid and Bratko [2006] suggested the number of times the computer changes its preferred move as the search depeens.

C. Complexity and Depth

- How to measure the *complexity* of a chess position?
- Guid and Bratko [2006] suggested the number of times the computer changes its preferred move as the search depeens.
- We use the full information about "swings" in move values as shown above.

C. Complexity and Depth

- How to measure the *complexity* of a chess position?
- Guid and Bratko [2006] suggested the number of times the computer changes its preferred move as the search depeens.
- We use the full information about "swings" in move values as shown above.
- Related, we isolate the depth at which a player's errors are exposed by the computer.

C. Complexity and Depth

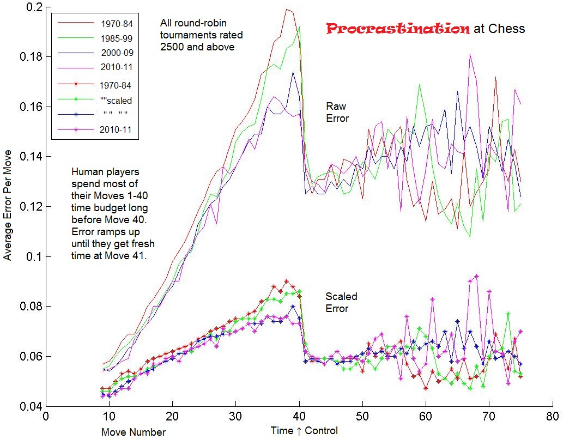
- How to measure the *complexity* of a chess position?
- Guid and Bratko [2006] suggested the number of times the computer changes its preferred move as the search depeens.
- We use the full information about "swings" in move values as shown above.
- Related, we isolate the depth at which a player's errors are exposed by the computer.
- This turns out to correlate regularly with skill.

C. Complexity and Depth

- How to measure the *complexity* of a chess position?
- Guid and Bratko [2006] suggested the number of times the computer changes its preferred move as the search depeens.
- We use the full information about "swings" in move values as shown above.
- Related, we isolate the depth at which a player's errors are exposed by the computer.
- This turns out to correlate regularly with skill.
- [show animations from https://rjlipton.wordpress.com/2015/10/06/depth-of-satisficing/]

Procrastination...

Chess players tend to use up most of a ≈ 2-hour time budget early on, leaving little time for moves 30 to 40 when a fresh budget of time comes. Note ramped-up error until turn 41. (Anand was an exception.)



Are We Reliable?

- One blunder in 200 moves can “ruin” a tournament.

Are We Reliable?

- One blunder in 200 moves can “ruin” a tournament.
- But we were reliable 99.5% of the time.

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.

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 expectation based on reading Nicholas Nassim Taleb's book *The Black Swan*.

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 expectation based on reading Nicholas Nassim Taleb's book *The Black Swan*.

- But we are even more reliable if we can use a computer...
- This is shown by computing *Intrinsic Performance Ratings* (IPRs) for humans and computers separately,

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 expectation based on reading Nicholas Nassim Taleb's book *The Black Swan*.

- But we are even more reliable if we can use a computer...
- This is shown by computing *Intrinsic Performance Ratings* (IPRs) for humans and computers separately, then for them playing in tandem in so-called "Freestyle" tournaments.

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 expectation based on reading Nicholas Nassim Taleb's book *The Black Swan*.

- But we are even more reliable if we can use a computer...
- This is shown by computing *Intrinsic Performance Ratings* (IPRs) for humans and computers separately, then for them playing in tandem in so-called "Freestyle" tournaments. (Where 'cheating' is OK!)

Computer and Freestyle IPRs

Analyzed Ratings of Computer Engine Grand Tournament (on commodity PCs) and PAL/CSS Freestyle in 2007–08, plus the Thoresen Chess Engines Competition (16-core) Nov–Dec. 2013.

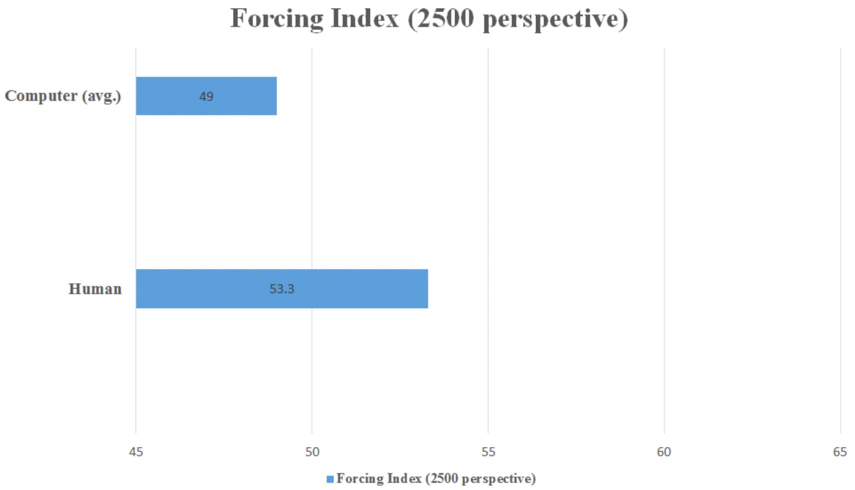
Event	Rating	2σ range	#gm	#moves
CEGT g1,50	3009	2962–3056	42	4,212
CEGT g25,26	2963	2921–3006	42	5,277
PAL/CSS 5ch	3102	3051–3153	45	3,352
PAL/CSS 6ch	3086	3038–3134	45	3,065
PAL/CSS 8ch	3128	3083–3174	39	3,057
TCEC 2013	3083	3062–3105	90	11,024

Computer and Freestyle IPRs—To Move 60

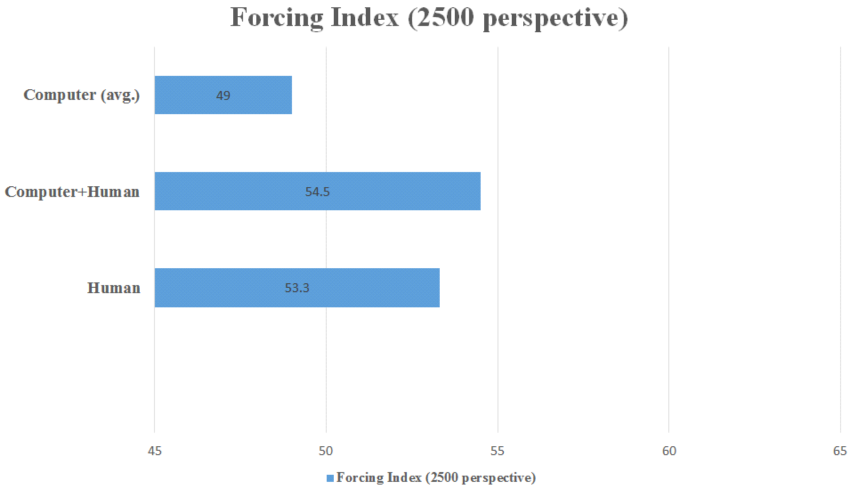
Computer games can go very long in dead drawn positions. TCEC uses a cutoff but CEGT did not. Human-led games tend to climax (well) before Move 60. This comparison halves the difference to CEGT, otherwise similar:

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

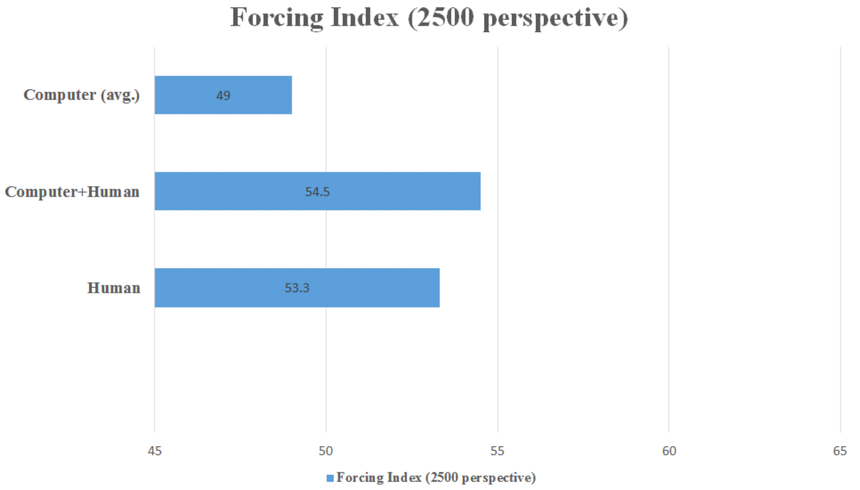
Degrees of Forcing Play



Add Human-Computer Tandems

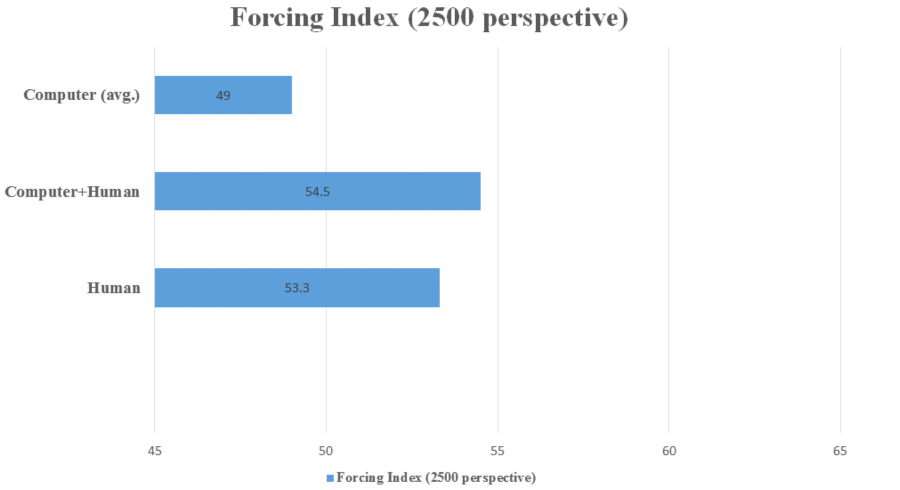


Add Human-Computer Tandems



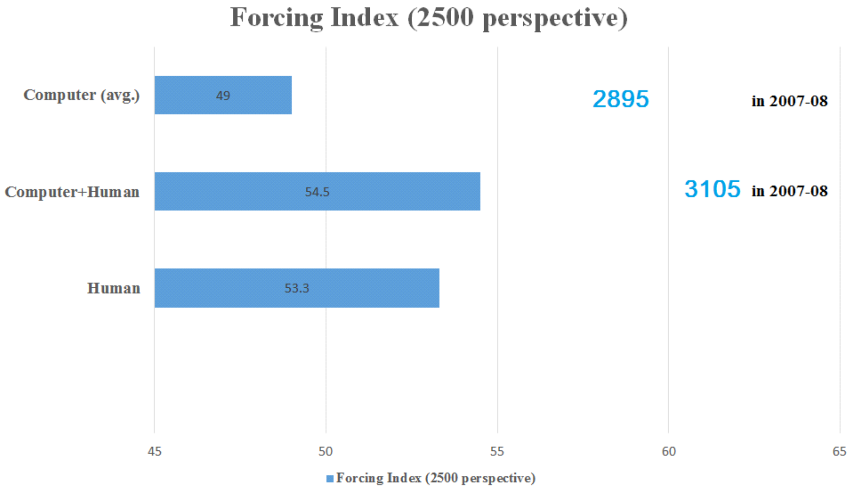
Evidently the humans called the shots.

Add Human-Computer Tandems



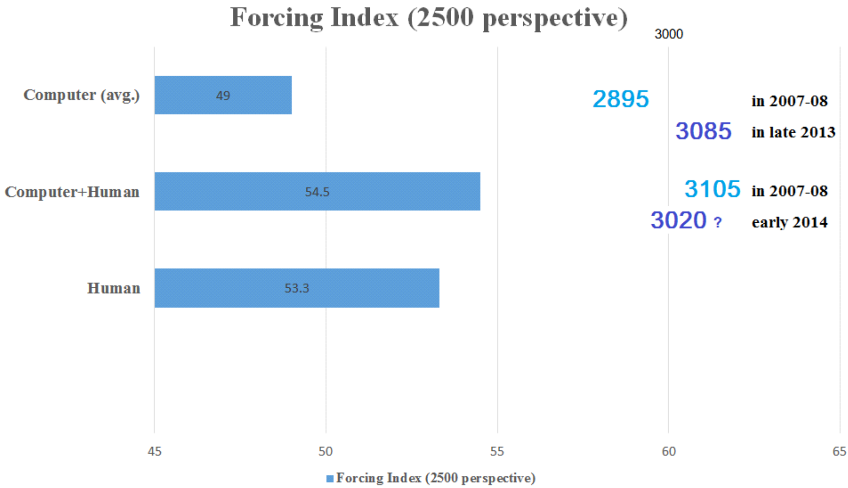
Evidently the humans called the shots. But how did they play?

2007–08 Freestyle Performance



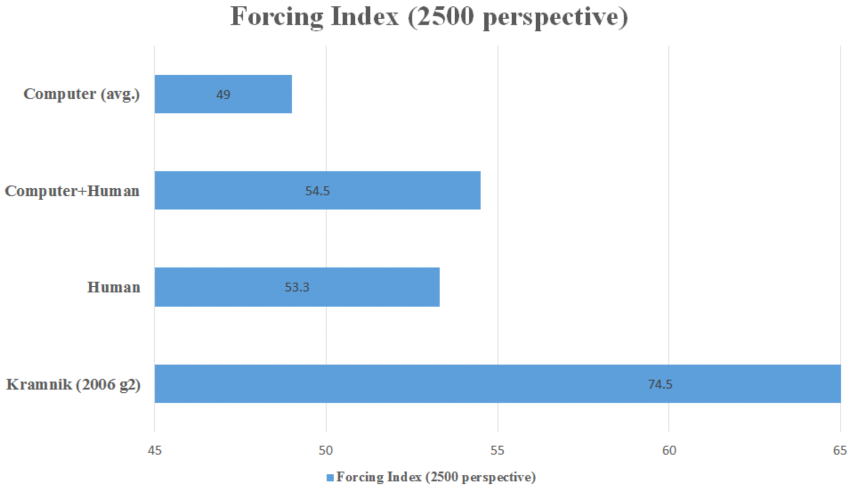
Adding 210 Elo was significant. Forcing but good teamwork.

2014 Freestyle Tournament Performance



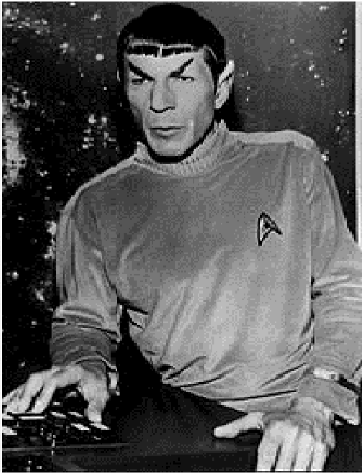
Tandems had marginally better W-L, but quality not clear...

Add Topalov Forcing Kramnik

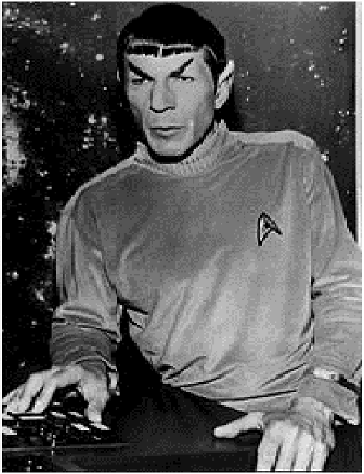


Last bar goes way off the chart

Like “Spock” to our “Kirk”



Like “Spock” to our “Kirk”

"It is logical to cultivate multiple options."

Summary For Us and PDAs

Summary For Us and PDAs

1. PDAs pick up every little difference: “**Forest** *and* **Trees**”

Summary For Us and PDAs

1. PDAs pick up every little difference: **"Forest *and* Trees"**
2. We should avoid overconfidence. . .

Summary For Us and PDAs

1. PDAs pick up every little difference: **“Forest *and* Trees”**
2. We should avoid overconfidence. . . and take counsel when “down.”

Summary For Us and PDAs

1. PDAs pick up every little difference: **“Forest *and* Trees”**
2. We should avoid overconfidence. . . and take counsel when “down.”
3. **Look before we Leap. . .**

Summary For Us and PDAs

1. PDAs pick up every little difference: **"Forest *and* Trees"**
2. We should avoid overconfidence. . . and take counsel when "down."
3. **Look before we Leap. . . Don't rush in. . .**

Summary For Us and PDAs

1. PDAs pick up every little difference: **"Forest *and* Trees"**
2. We should avoid overconfidence... and take counsel when "down."
3. **Look before we Leap... Don't rush in... Measure risks.**

Summary For Us and PDAs

1. PDAs pick up every little difference: **"Forest *and* Trees"**
2. We should avoid overconfidence... and take counsel when "down."
3. **Look before we Leap... Don't rush in... Measure risks.**
4. Even at a purely calculational pursuit like chess, *our brains still contribute.*

Summary For Us and PDAs

1. PDAs pick up every little difference: **"Forest *and* Trees"**
2. We should avoid overconfidence. . . and take counsel when "down."
3. **Look before we Leap. . . Don't rush in. . . Measure risks.**
4. Even at a purely calculational pursuit like chess, *our brains still contribute*. (2014: maybe)

Summary For Us and PDAs

1. PDAs pick up every little difference: **"Forest *and* Trees"**
2. We should avoid overconfidence... and take counsel when "down."
3. **Look before we Leap... Don't rush in... Measure risks.**
4. Even at a purely calculational pursuit like chess, *our brains still contribute*. (2014: maybe)
5. **Main takeaway:**

Summary For Us and PDAs

1. PDAs pick up every little difference: **"Forest *and* Trees"**
2. We should avoid overconfidence... and take counsel when "down."
3. **Look before we Leap... Don't rush in... Measure risks.**
4. Even at a purely calculational pursuit like chess, *our brains still contribute*. (2014: maybe)
5. **Main takeaway:**

> **It should be natural to program digital assistants so they enhance our freedom rather than constrain it.**

Conclusions and Future Activity

- Lots more potential for research and connections...
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.

Conclusions and Future Activity

- Lots more potential for research and connections...
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.
- Detect and deter cheating too—generally.

Conclusions and Future Activity

- Lots more potential for research and connections...
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.
- Detect and deter cheating too—generally.
- Have just created a major upgrade to the model which allows treating *depth* as a skill attribute and which produces 10x sharper fits (at 100x slower pace, however).

Conclusions and Future Activity

- Lots more potential for research and connections...
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.
- Detect and deter cheating too—generally.
- Have just created a major upgrade to the model which allows treating *depth* as a skill attribute and which produces 10x sharper fits (at 100x slower pace, however).
- Learn more about human decision making.

Conclusions and Future Activity

- Lots more potential for research and connections...
- Spread word about general-scientific aspects, including public outreach over what isn't (and is) cheating.
- Detect and deter cheating too—generally.
- Have just created a major upgrade to the model which allows treating *depth* as a skill attribute and which produces 10x sharper fits (at 100x slower pace, however).
- Learn more about human decision making.
- **Thank you very much for the invitation.**