Anthropic probability and other puzzles affecting the human survival

An informal look at some things the FHI gets up to
Why is the FHI in the philosophy department?

An informal look at some things the FHI gets up to
Why philosophy?

Dealing with areas where the scientific method cannot apply.

Where biases and uncertainties rule the day.

Where the uncertainties are fundamental: everything is open to **justified** questioning.
What is the probability of Heads?

1/2
What is the probability of Heads?

1/3
The Doomsday argument

100 billion humans have lived on Earth
The Doomsday argument

You are... 1 in 200 trillion

100 billion humans have lived on Earth

You should expect humanity to survive... Only if you are extremely unlikely!
The Doomsday argument

Either:
A: People are born every day
B: People are only born on the 1st of January

You, and everyone you know, were born on the 1st of January

Is A or B the most likely?
What are existential risks?

1. Pandemics
2. Synthetic biology
3. Nanotechnology
4. Artificial intelligence
5. Nuclear war
6. Asteroid impact
7. Environmental collapse

(not risky enough)
What are existential risks?

1. Pandemics
2. Synthetic biology
3. Nanotechnology
4. Artificial intelligence
5. Nuclear war
6. Asteroid impact
7. Environmental collapse
AI: Power of Intelligence

Terminator: big muscles, no brain
AI: Power of Intelligence

Who’s the dominant specie?
AI: Power of Intelligence

Pop: 200,000
AI: Power of Intelligence
How bad could it be?

"Prevent human suffering"

How bad could it be?

Kill all humans

"Keep humans safe and happy"

Entomb everyone in underground concrete coffins on heroin drips

Train AIs on human smiles/body language and build on this
PREDICTING AGI
We’re doing it badly
AGI predictions

“significant advance can be made in [machines using language and improving themselves, if a] group of scientists work on it together for a summer.”

(Dartmouth conference, 1956)

“Nonetheless, the dramatic slowdown in [computerised chess playing ability] suggests the boundary may be near.”

(Dreyfus, 1965)
AGI predictions

“[AGI will be developed in 15-25 years]”

(various)

Plan for the Talk

AGI predictions: timelines and philosophy.

What performance should we expect?

What performance do we get?

Singularity Institute’s database of 257 AI predictions (1950-2012).
How predictions in AGI compare

“... leads to the unavoidable conclusion that Krugman isn’t reading real economics anymore...”

John Cochrane
How predictions in AGI compare

“...comments from Chicago economists are the product of a Dark Age of macroeconomics...”

Paul Krugman

“... leads to the unavoidable conclusion that Krugman isn’t reading real economics anymore...”

John Cochrane

Average quarterly GDP adjustments:
±1.7 points
How predictions in AGI compare

“...comments from Chicago economists are the product of a Dark Age of macroeconomics...”

Paul Krugman

“... leads to the unavoidable conclusion that Krugman isn’t reading real economics anymore...”

John Cochrane
Disagreements and Overconfidence

- Life experience
- Evidence!!
- Detailed arguments
- Thought experiments
- Well-formed intuitions
- Biases
- Rationalisations
  - Reasonable conclusion

- Life experience
- Evidence!!
- Detailed arguments
- Thought experiments
- Well-formed intuitions
- Biases
- Rationalisations
  - Biased conclusion
Disagreements and Overconfidence

Genuine expertise?

My objectivity criteria!

Opinions relevant, only if **objectively** better

Genuine expertise?
When are experts good?

<table>
<thead>
<tr>
<th>Good performance</th>
<th>Poor performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static stimuli</td>
<td>Dynamic (changeable) stimuli</td>
</tr>
<tr>
<td>Decisions about things</td>
<td>Decisions about behavior</td>
</tr>
<tr>
<td>Experts agree on stimuli</td>
<td>Experts disagree on stimuli</td>
</tr>
<tr>
<td>More predictable problems</td>
<td>Less predictable problems</td>
</tr>
<tr>
<td>Some errors expected</td>
<td>Few errors expected</td>
</tr>
<tr>
<td>Repetitive tasks</td>
<td>Unique tasks</td>
</tr>
<tr>
<td>Feedback available</td>
<td>Feedback unavailable</td>
</tr>
<tr>
<td>Objective analysis available</td>
<td>Subjective analysis only</td>
</tr>
<tr>
<td>Problem decomposable</td>
<td>Problem not decomposable</td>
</tr>
<tr>
<td>Decision aids common</td>
<td>Decision aids rare</td>
</tr>
</tbody>
</table>

“Competence in experts: The role of task characteristics”
James Shanteau: Organizational Behavior and Human Decision Processes
When are experts good?

<table>
<thead>
<tr>
<th>Good performance</th>
<th>Poor performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static stimuli</td>
<td>Dynamic (changeable) stimuli</td>
</tr>
<tr>
<td>Decisions about things</td>
<td>Decisions about behavior</td>
</tr>
<tr>
<td>Experts agree on stimuli</td>
<td>Experts disagree on stimuli</td>
</tr>
<tr>
<td>More predictable problems</td>
<td>Less predictable problems</td>
</tr>
<tr>
<td>Some errors expected</td>
<td>Few errors expected</td>
</tr>
<tr>
<td>Repetitive tasks</td>
<td>Unique tasks</td>
</tr>
<tr>
<td>Feedback available</td>
<td>Feedback unavailable</td>
</tr>
<tr>
<td>Objective analysis available</td>
<td>Subjective analysis only</td>
</tr>
<tr>
<td>Problem decomposable</td>
<td>Problem not decomposable</td>
</tr>
<tr>
<td>Decision aids common</td>
<td>Decision aids rare</td>
</tr>
</tbody>
</table>
When are experts good?

<table>
<thead>
<tr>
<th>Good performance</th>
<th>Poor performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static stimuli</td>
<td>Dynamic (changeable) stimuli</td>
</tr>
<tr>
<td>Decisions about things</td>
<td>Decisions about behavior</td>
</tr>
<tr>
<td>Experts agree on stimuli</td>
<td>Experts disagree on stimuli</td>
</tr>
<tr>
<td>More predictable problems</td>
<td>Less predictable problems</td>
</tr>
<tr>
<td>Some errors expected</td>
<td>Few errors expected</td>
</tr>
<tr>
<td>Repetitive tasks</td>
<td>Unique tasks</td>
</tr>
<tr>
<td>Feedback available</td>
<td>Feedback unavailable</td>
</tr>
<tr>
<td>Objective analysis available</td>
<td>Subjective analysis only</td>
</tr>
<tr>
<td>Problem decomposable</td>
<td>Problem not decomposable</td>
</tr>
<tr>
<td>Decision aids common</td>
<td>Decision aids rare</td>
</tr>
</tbody>
</table>

- Experts agree on stimuli
- Experts disagree on stimuli
- Problem decomposable
- Problem not decomposable
Grind is easy, insight hard

How long will it take to produce the next Michael Bay ‘blockbuster’?

When will someone solve the Riemann hypothesis?

Moore’s law, hence AGI:
By year XXXX, computers will have Y (a level comparable with the human brain!), then AGI.
The evidence: AGI predictions

The Singularity Institute collected a database of 257 AGI-related predictions (online, in research journals, news articles, etc...) 1950-2012.

95 are timeline to AGI predictions.

“By golly, I predict that we will have human-level AGI by year XXXX!” A Renown Expert
The evidence: AGI predictions

The Singularity Institute collected a database of 257 AGI-related predictions (online, in research journals, news articles, etc...) 1950-2012

95 are timeline to AGI predictions.

I transformed each one into a median date of AGI estimate.
The evidence: AGI predictions

The Singularity Institute collected a database of 257 AGI-related predictions (online, in research journals, news articles, etc...) 1950-2012

95 are timeline to AGI predictions.

I transformed each one into a median date of AGI estimate.

We also assessed the expertise of the predictor.
When, oh when, will we have AGI?

![Graph showing predicted dates for AGI arrival]
One AGI to see before you die

“Maes-Garreau law”
One AGI to see before you die
Tomorrow never gets any closer...

15-25 years time: not soon, not too far

No evidence that experts have any predictive advantage (but that does not imply your own guess is better!)
Spread your wings of uncertainty

Experts

Non-experts

Wrong, deluded

Confused, mistaken

AGI at 2040!

Pretty likely.

Very certain.

Date prediction made

Predicted AGI arrival
Spread your wings of uncertainty

Experts
Non-experts

Wrong, deluded
Confused, mistaken

AGI at 2040... Pretty likely...
Spread your wings of uncertainty

Experts

Non-experts

AGI at 2040...

Approximately...
Current best timeline prediction

Whole brain emulations (Uploads)

Fix a brain, slice it up, scan it, construct a model, instantiate it on a computer.

Very decomposed.

Justified grind.

Clear assumptions and scenarios.

Integrates new data (partial feedback).

Multiple pathways.
Current best timeline prediction

Whole brain emulations (Uploads)
What *can* we say about AGI?

• Timeline predictions are pretty poor.

• Other types of predictions (such as plans for how to build AGIs) have similar problems.

• But we can get good ideas about AGI from...

...Philosophy!
Ugh, philosophy – what’s it good for?

Gödel’s theorem proves AGI is impossible!

I don’t think it does...

Does too! My argument is sound!

Does not! The argument is not convincing!

Does too!

Does not!
Ugh, philosophy – what’s it good for?

Philosophers are also very overconfident!

Their arguments need more caveats, uncertainty, and decomposition...
Ugh, philosophy – what’s it good for?

1. Gödel’s theorem applies to certain formal systems.
2. Those formal systems *could* be model for likely AGI designs.
3. Hence there *may* be a problem with self-reference in AGI.
4. AGI programmers should be aware of this.
5. But, *in my expert opinion*, that problem will still be insoluble.

Agree with 1-3, partially with 4, disagree with 5.

Let’s discuss some more...
Ugh, philosophy – what’s it good for?

A few minor philosophical results:

- Occam’s razor
- Church-Turing thesis
- Decision theory
- Formal logic
- Scientific method
Ugh, philosophy – what’s it good for?

Example of improved philosophical arguments:

Dreyfus: Computers can’t cope with ambiguity...
...using current [1965] AI approaches.

Gozzi: “Identifying the computer with a brain may be putting together things that don't belong [...] computing isn’t thinking”.

AGIs may be nothing like human brains. We may go astray thinking that they are.
Current best philosophical prediction

Simplified “Omohundro-Yudkowsky thesis”:

Behaving dangerously is a generic behaviour for high-intelligence AGIs.

Economic model: simplified model of what AGI will be.
Current best philosophical prediction

Simplified “Omohundro-Yudkowsky thesis”, refined and narrowed:

Many AGI designs have the potential for unexpected dangerous behaviour.

AGI programmers should demonstrate to (moderate) sceptics that their design is safe.

Is the thesis wrong, in your opinion?
Conclusions

• Our own opinions are not strong evidence

• Philosophy has some useful things to say

• AGI timeline predictions are problematic

• It’s very hard to know where to begin with existential risks – but we have to begin