

Challenges in Forecasting AI Progress

Miles Brundage
Future of Humanity Institute
University of Oxford



Overview

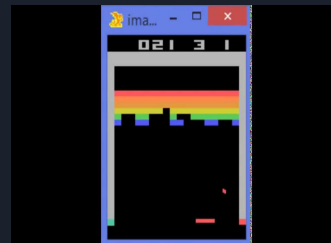
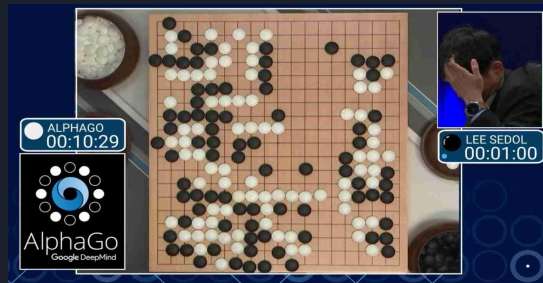
- Historical perspective
- Comparative technological perspective
- Different tools for forecasting AI progress:
 - Expert surveys
 - Trend extrapolation
 - Qualitative analysis
- Where to go from here?



Historical perspective

“An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.” McCarthy et al., 1955

Historical perspective



Historical perspective



Greg Brockman  @gdb · 18h

Looks like there's a prediction market for whether OpenAI Five will win on Aug 5th: twitter.com/metaculus/stat...

Currently at 61% in favor of OpenAI Five :). Highly uncertain!

Metaculus @metaculus

OpenAI's @DOTA2 player system is going head to head against the cream of the (human) gaming crop on Aug 5. Will they triumph?

You can make your prediction here! buff.ly/2L7YBYM ...



5



13



89



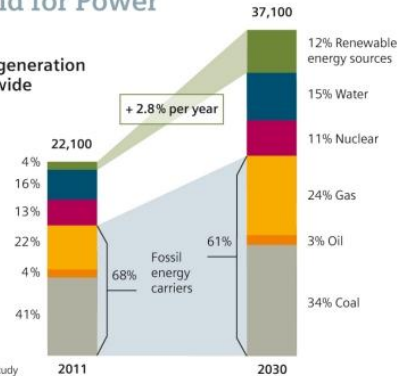
Big uncertainty a few weeks out

Comparative technological perspective

Other areas like energy/climate aren't perfect and make mistakes, but are at least more clear about assumptions, good at tracking data, and prone to follow up to see where they erred

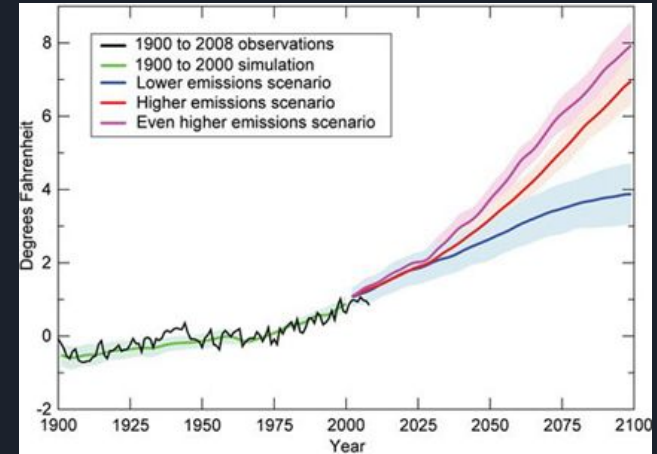
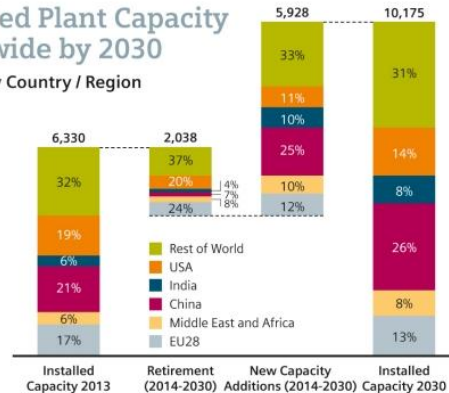
Demand for Power

Electricity generation mix worldwide (in TWh)



Expected Plant Capacity Worldwide by 2030

Capacity by Country / Region (GW)



Siemens; Climate Central



Different tools for forecasting AI progress

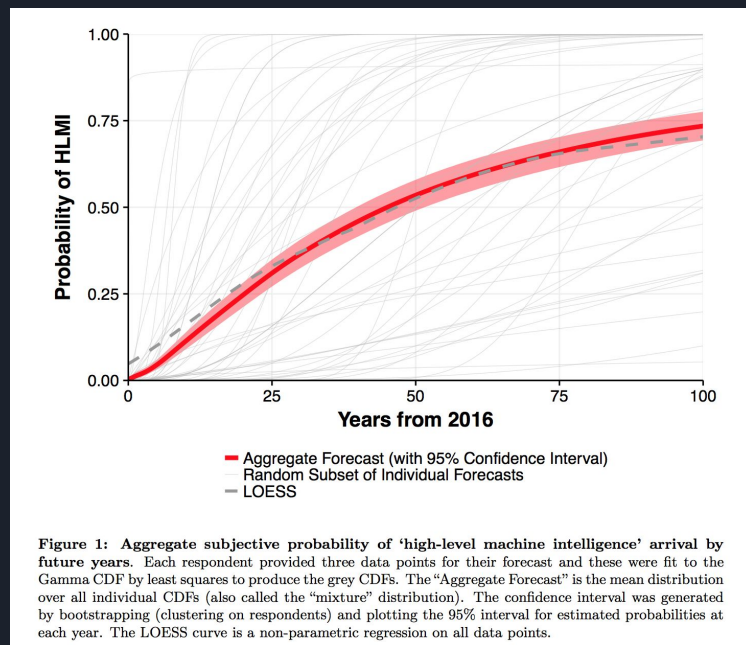
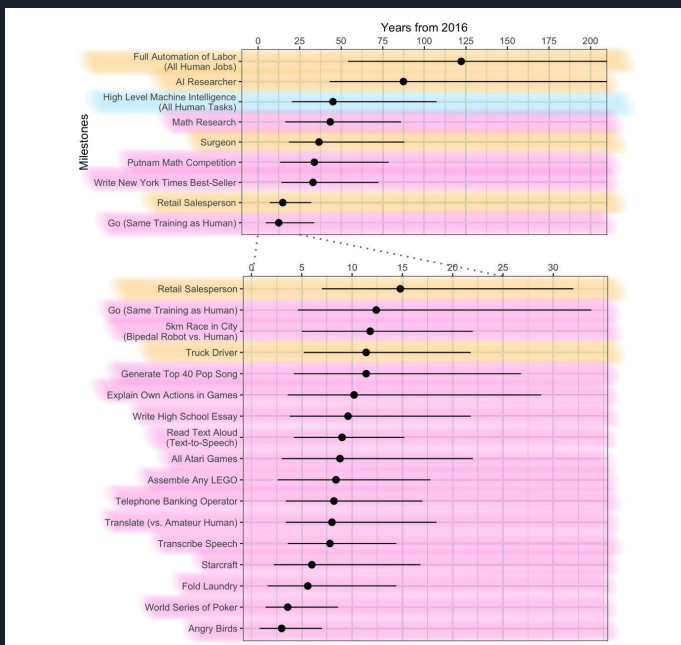
- Different tools
 - Expert surveys
 - Trend extrapolation
 - Qualitative analysis
- Different goals
 - Unconditional forecasts (e.g. “X job automatable in Y year”)
 - Conditional forecasts (e.g. “intelligence explosion” given human-level AI, compute-related acceleration)



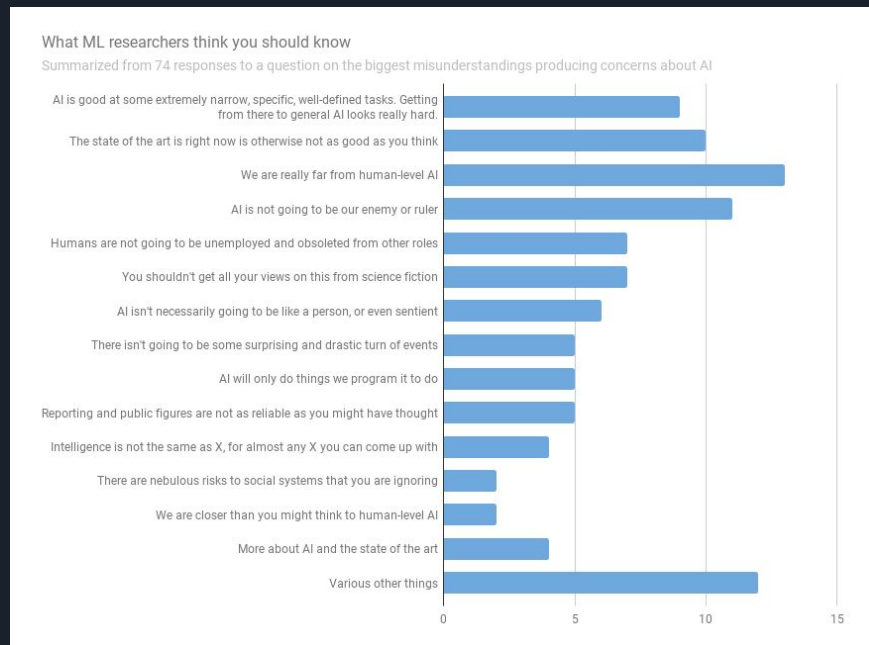
Expert surveys

- What it is
 - Asking people who (hopefully) know!
- Challenges
 - Disagreement (differential weighting doesn't help that much)
 - How informed are these views? Expertise on present vs. future

Expert surveys



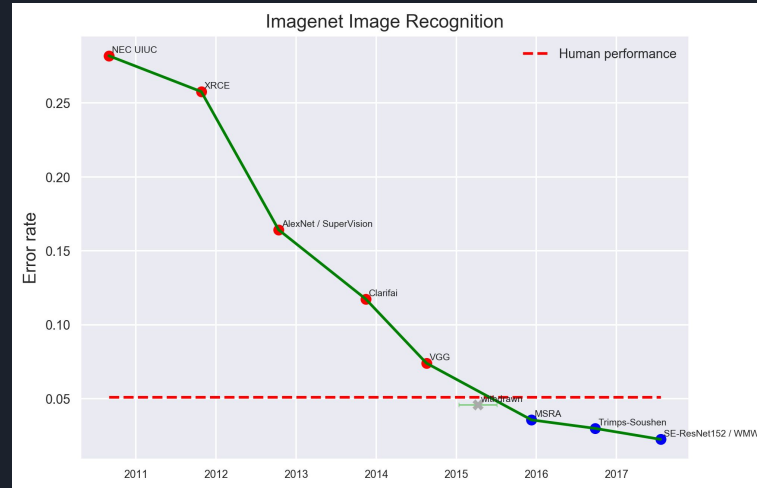
Expert surveys



AI Impacts, 2016

Trend extrapolation

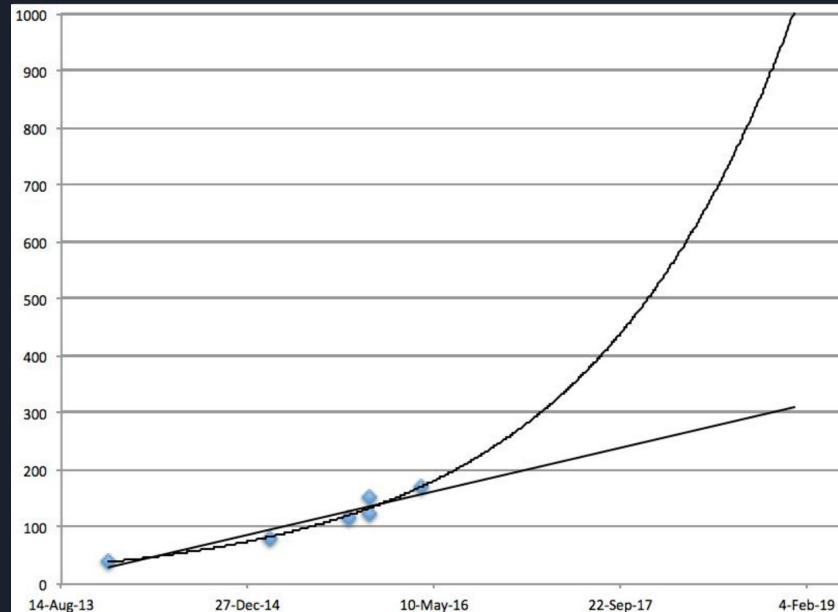
- Challenges
 - Which trends matter? Robustness often not captured
 - Which trends are missing?
 - When will the trends break?
 - Do we have enough data?



EFF AI Progress Measurement
Project

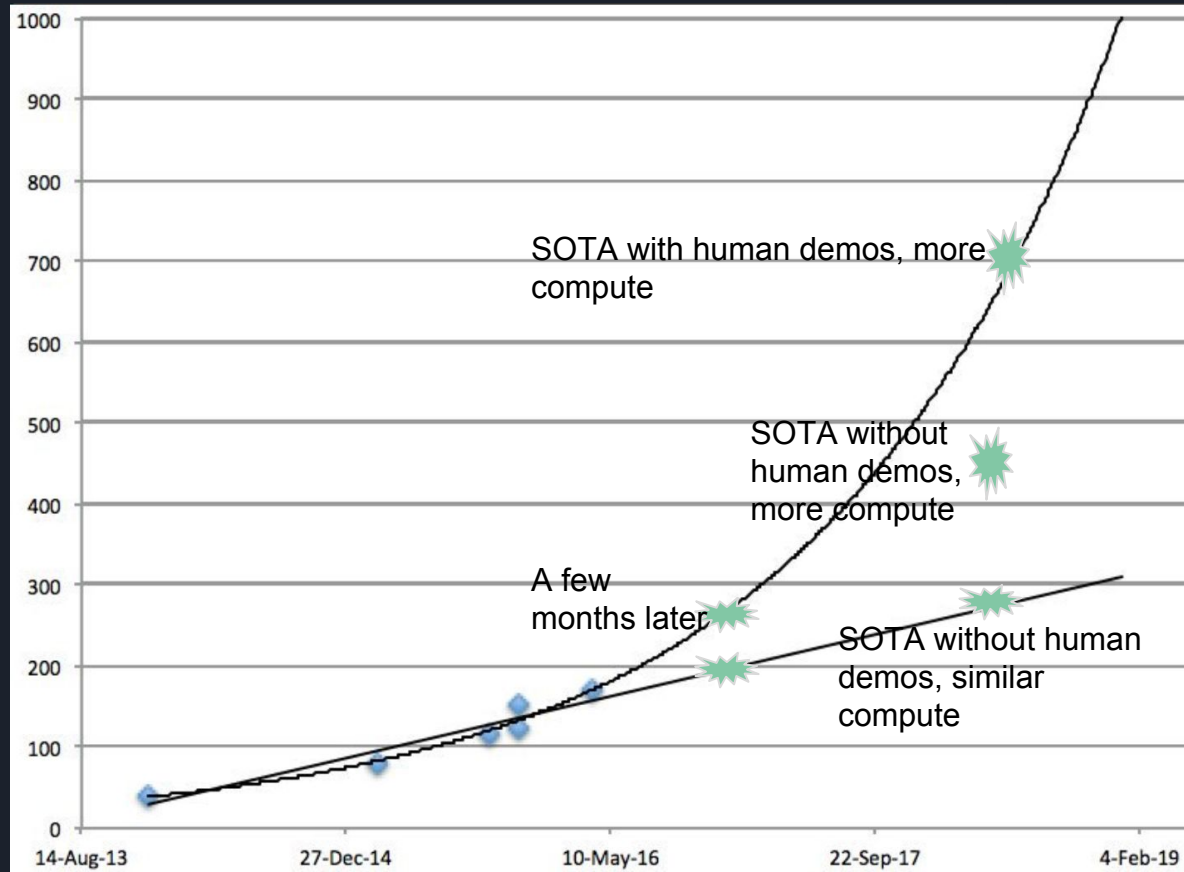
Trend extrapolation

- Sometimes works pretty well!

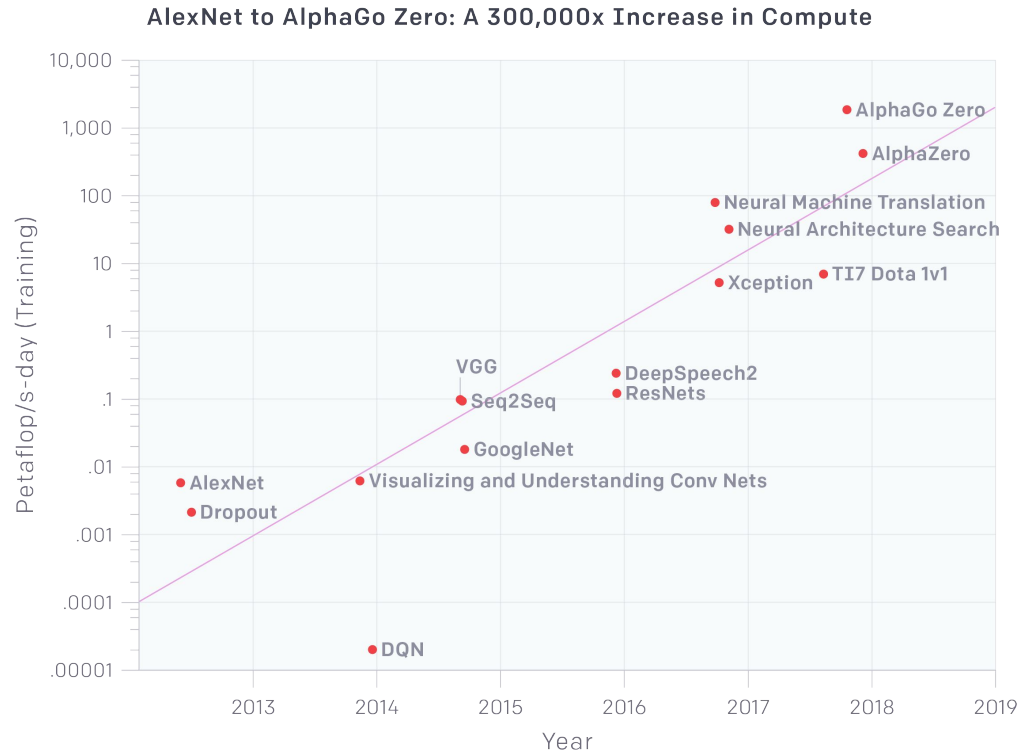


Median
human-normalized scores
on 57 Atari games;
Brundage, April 23, 2016

Trend extrapolation



Trend extrapolation



Amodei and
Hernandez,
2018



Problem with this framework...

- There are many drivers of progress
- There is insufficient data, and what data we have isn't well organized

Problem with this framework...

- costs
- There are many ~~drivers~~ of progress. Which do you factor out?

Description	Example
r_d <i>Data</i> : All kinds of data (unsupervised, supervised, queries, measurements).	A self-driving car needs on-line traffic information.
r_k <i>Knowledge</i> : Rules, constraints, bias, utility functions, etc., that are required.	A spam filter requires the cost matrix from the user.
r_s <i>Software</i> : Main algorithm, associated libraries, operating system, etc.	A planner uses a SAT solver.
r_h <i>Hardware</i> : Computer hardware, sensors, actuators, motors, batteries, etc.	A drone needs a 3D radar for operation.
r_m <i>Manipulation</i> : Manual (human-operated) intervention through assistance	A robot needs to be manually re-calibrated.
r_c <i>Computation</i> : Computational resources (CPU, GPU usage) of all the components	A nearest neighbor classifier computes all distances.
r_n <i>Network</i> : Communication resources (Internet, swarm synchronisation, distribution).	An automated delivery system connects all drones.
r_t <i>Time</i> : Calendar (physical) time needed: waiting/night times, iteration cycles.	A PA requires cyclical data (weeks) to find patterns.

Table 1: Resources that are frequently needed by AI systems.

Problem with this framework...

Missing data...

	Sarsa	Best Linear	DQN best	NatureDQN	Gorila	DQN noop & hs	DUEL noop & hs	DDQN tuned hs	PRIOR _{hs & noop}	P. DUEL _{hs & noop}	AC3 _{LSTM, FF & FFid}	DDQN _{Pop-Art noop}	AC3 _{CTS}	SARSA _{e & FEB}	TRPO _{hash}	DQN _{CTS & PixelCNN}	C51 _{noop}	ES FF _{(1h) noop}	RAINBOW	REACTOR
r_d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
r_k	○	○	×	✓	×	○	×	○	○	○	○	○	×	×	○	○	○	×	✓	✓
r_s	×	×	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×
r_h	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
r_m	×	✓	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
r_c	○	○	○	○	○	○	○	○	○	○	✓	○	○	○	○	○	○	○	✓	✓
r_n	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
r_t	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ψ	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	○	✓	✓	✓	✓

Table 3: Same as Table 2 for the ALE papers (from EFF [11] and [15, 20]).

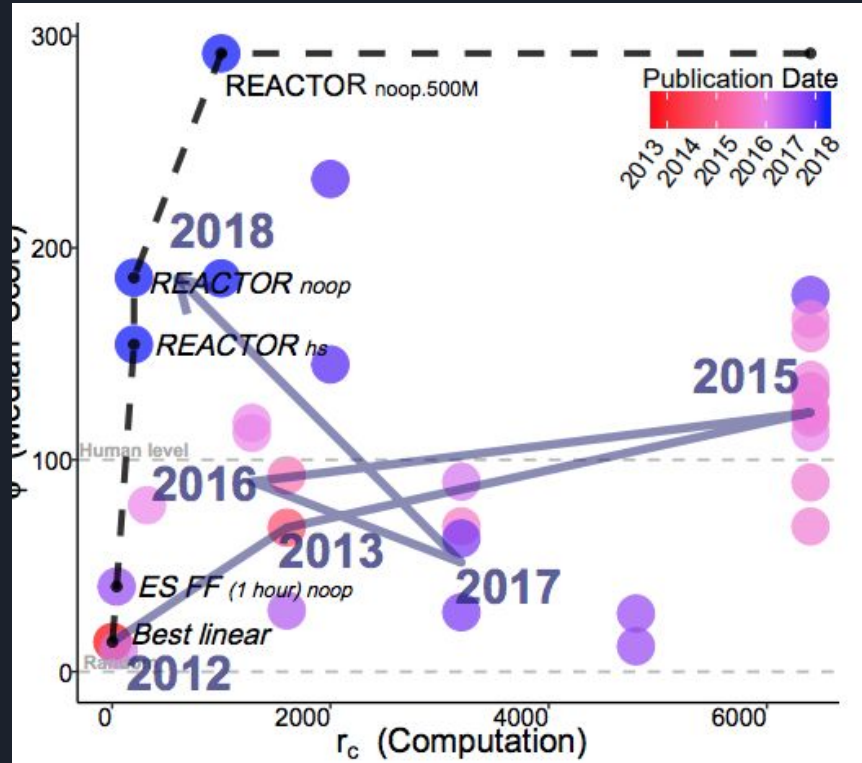
Ibid

Problem with this framework...


What counts as progress?

Only Pareto improvements?

Any performance increase?



Ibid



A better (and harder) way of doing trend extrapolation

- Disentangle factors driving performance
- Model the relationship between these factors and performance
- Extrapolate specific factors and the resulting performance
- When data isn't available or organized: organize what's available, ask for what isn't published, and create it yourself with experiments



Qualitative analysis

- Roughly: model-based rather than model-free expert knowledge
 - Moravec's Paradox (what's easy for humans is hard for machines, and vice versa)
 - Perception and manipulation, social intelligence, and creative intelligence are hard (Frey and Osborne, 2013)
 - Machines are good at “prediction” tasks, broadly construed (Agrawal et al., 2018)
- Challenges
 - Some of the same problems as expert surveys - what knowledge do we really have re: the future? What if people disagree about underlying hardness assumptions?

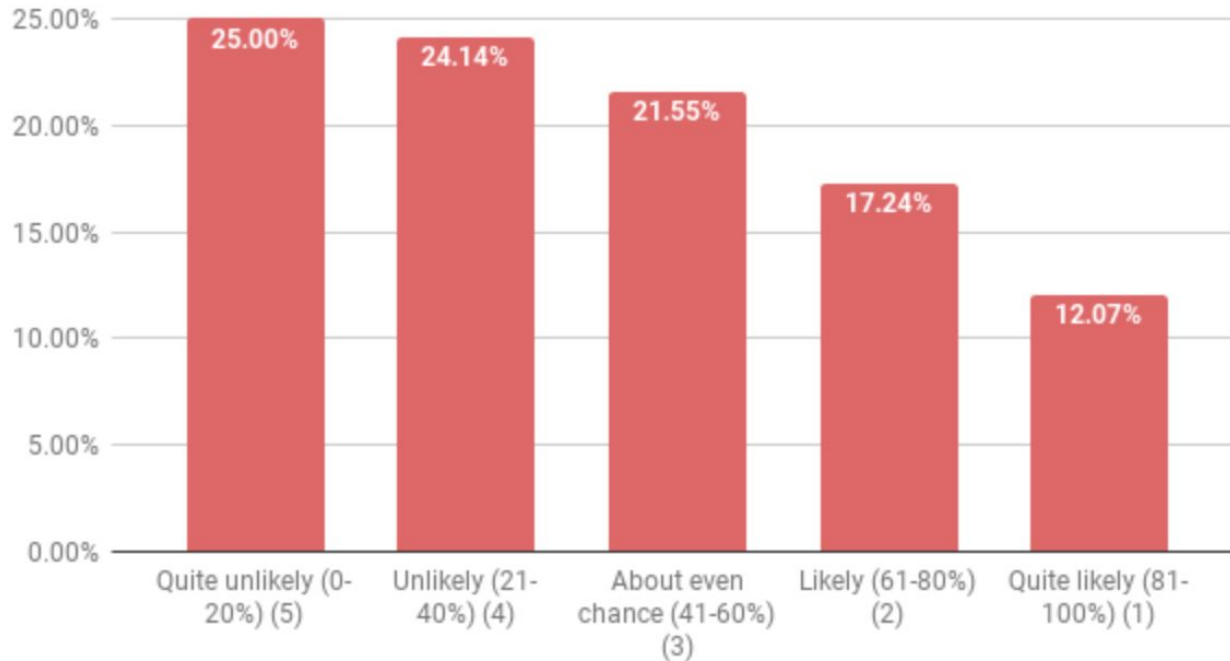


Qualitative analysis

- Qualitative analysis domains
 - Automatability of tasks given economic constraints, commercial incentives, etc.
 - Unconditional likelihood of a task being automated based on first principles
 - Conditional expectations re: future acceleration given certain milestones

Qualitative analysis

Perceived chance of intelligence explosion argument being broadly correct

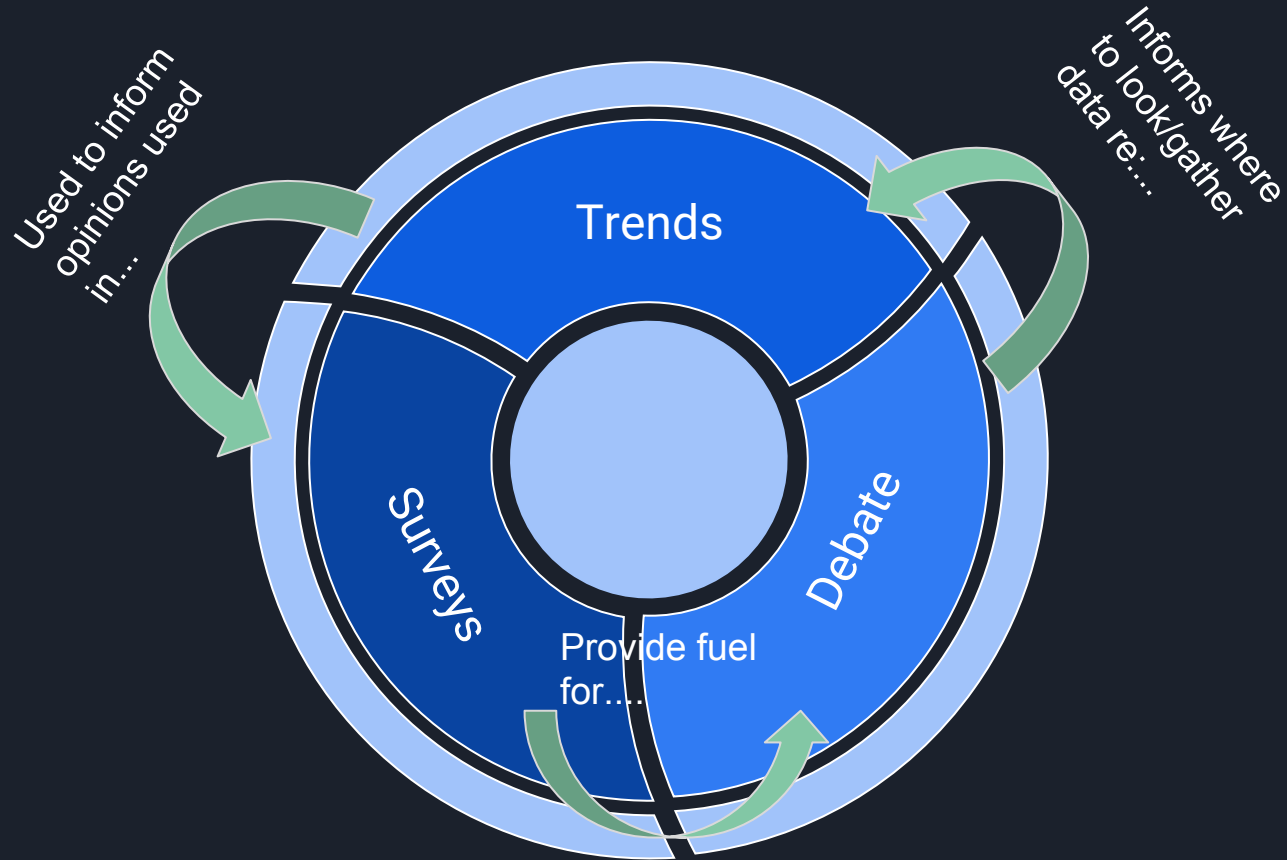




Where to go from here?

- Ongoing: projects tackling the data problem
 - EFF AI Progress Measurement Project
 - AI Index
- Ongoing: better expert surveys
- Ongoing: debate on qualitative issues
- How to integrate the three?

Where to go from here?





Thanks!

miles.brundage@philosophy.ox.ac.uk