Governing the AI Revolution

Allan Dafoe

Yale University

Future of Humanity Institute
University of Oxford

governance.ai

The AI Governance Problem: *the problem of devising global norms, policies, and institutions to best ensure the beneficial development and use of advanced AI.*

Common Misunderstanding 1

Attention to technological risks implies one believes ...the technology is net negative or risks are probable.

Common Misunderstanding 1

Attention to technological risks implies one believes
...the technology is net negative or risks are probable.
...there are risks which attention could mitigate.

Near-term Governance Challenges

Safety in critical systems, such as finance, energy systems, transportation, robotics, autonomous vehicles.

(Consequential) algorithms that **encode values**, such as in hiring, loans, policing, justice, social network. Desiderata: fairness ▸ Hardt, accountability, transparency, efficiency, privacy, ethics.

AI impacts on employment, equality, privacy, democracy...

Some Extreme Challenges from Near-Term AI

- **Mass labor displacement and inequality.** If AI substitutes, rather than complements, labor.

Some Extreme Challenges from Near-Term AI

- **Mass labor displacement and inequality.** If AI substitutes, rather than complements, labor.
- **AI Oligopolies: strategic industry and trade.** If AI industries are natural global monopolies, due to low/zero marginal costs of AI services, incumbent advantage, high fixed costs from AI R&D.

Some Extreme Challenges from Near-Term AI

- **Mass labor displacement and inequality.** If AI substitutes, rather than complements, labor.
- **AI Oligopolies: strategic industry and trade.** If AI industries are natural global monopolies, due to low/zero marginal costs of AI services, incumbent advantage, high fixed costs from AI R&D.
- **Surveillance and Control**: mass surveillance (sensors, digitally-mediated behavior), intimate profiling, tailored persuasion, repression (LAWS).

Some Extreme Challenges from Near-Term AI

- **Mass labor displacement and inequality.** If AI substitutes, rather than complements, labor.
- **AI Oligopolies: strategic industry and trade.** If AI industries are natural global monopolies, due to low/zero marginal costs of AI services, incumbent advantage, high fixed costs from AI R&D.
- **Surveillance and Control**: mass surveillance (sensors, digitally-mediated behavior), intimate profiling, tailored persuasion, repression (LAWS).
- **Strategic (Nuclear) Stability**: autonomous escalation; counterforce vulnerability from AI intel, cyber, drones; autonomous nuclear retaliation (esp w/ hypersonics).
- **Military Advantage**: LAWS, cyber, intel, info operations.

Some Extreme Challenges from Near-Term AI

- **Mass labor displacement and inequality.** If AI substitutes, rather than complements, labor.
- **AI Oligopolies: strategic industry and trade.** If AI industries are natural global monopolies, due to low/zero marginal costs of AI services, incumbent advantage, high fixed costs from AI R&D.
- **Surveillance and Control**: mass surveillance (sensors, digitally-mediated behavior), intimate profiling, tailored persuasion, repression (LAWS).
- **Strategic (Nuclear) Stability**: autonomous escalation; counterforce vulnerability from AI intel, cyber, drones; autonomous nuclear retaliation (esp w/ hypersonics).
- **Military Advantage**: LAWS, cyber, intel, info operations.
- **Accident/Emergent/Other Risks**, from AI-dependent critical systems and transformative capabilities.

Corner-Cutting

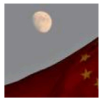
Corner-Cutting

The coordination problem is one thing [we should focus on now]. We want to avoid this harmful race to the finish where corner-cutting starts happening and safety gets cut.... That's going to be a big issue on a global scale, and that's going to be a hard problem when you're talking about ***national governments****.*

Demis Hassabis, January 2017


"
Whoever leads in AI will rule the world
Vladimir Putin
"
RT

Massive Media Reaction



The Next Space Race Is **Artificial Intelligence**
Foreign Policy (blog) - 3 Nov 2017
The Next Space Race Is **Artificial Intelligence** ... forecast in light of the next great technological race: **artificial intelligence**, or **AI**. "Whoever becomes the leader in this sphere will become the ruler of the **world**," **Putin** said. ... on **AI** grows, we need to keep our competitiveness in mind when we put **rules** in place.

Artificial Intelligence, The New Chess Piece Of Geopolitics
Worldcrunch - 12 hours ago
PARIS — **Artificial Intelligence** is now an open topic of geopolitical debate. Russian President Vladimir **Putin** recently said that "whoever becomes the leader in this ... of Tesla and SpaceX, warned of the potential risk of **AI** leading to a third **World** War. ... Negotiating the "**rules** of the game" is a pressing issue.

How To Stop Worrying And Love The Great **AI** War Of 2018
Fast Company - 10 Oct 2017
Inspired by news that Vladimir **Putin** had told Russian students the country that leads in **artificial intelligence** will **rule** the **world**, the Tesla and ...

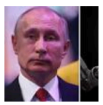
Careers in **artificial intelligence** continue to rise. But what is it?
Q13 FOX - 24 Oct 2017
Careers in **artificial intelligence** continue to rise. ... And Russian President Vladimir **Putin** said whoever controlled **AI** would **rule** the **world**.

Google: China's military plans to 'dominate' **artificial intelligence**
Washington Examiner - 1 Nov 2017
China's military plans to "dominate" the **artificial intelligence** industry in the ... contractors, and the military has not been asking for **AI** systems. ... part of national security strategy around the **world**, not only in China. ... for all humankind," Russian President Vladimir **Putin** said in September, according to RT.

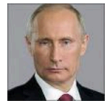
Putin: Leader in **artificial intelligence** will **rule world**
CNBC - 3 Sep 2017
Putin: Leader in **artificial intelligence** will **rule world** ... speaking Friday at a meeting with students, said the development of **AI** raises "colossal ...
'Whoever leads in **AI** will **rule** the **world**': **Putin** to Russian children on ...
International - RT - 1 Sep 2017
View all

Vladimir **Putin** Says Whoever Leads in **Artificial Intelligence** Will ...
Fortune - 4 Sep 2017
Vladimir **Putin** Says Whoever Leads in **Artificial Intelligence** Will **Rule** the **World** ... While some more excitable outlets have reported this as **Putin** saying Russia will use **AI** to take over the **world**, that's not quite what he said.
Putin says the nation that leads in **AI** 'will be the ruler of the **world**'
Highly Cited - The Verge - 4 Sep 2017
View all

EU unsure how to 'make most' of **AI**
EUobserver - 25 Oct 2017
"**Artificial intelligence** is the future, not only for Russia, but for all humankind," **Putin** said last month according to RT, a Russia-funded TV network. ... the leader in this sphere will become the ruler of the **world**," he added. ... a robot can take autonomous decisions, the traditional **rules** will not suffice to give rise ...



Vladimir **Putin** orders students to 'master **AI** so RUSSIA can **RULE** ...
Express.co.uk - 3 Sep 2017
... Vladimir **Putin** has sensationally stated that the nation which wins the **artificial intelligence** (**AI**) arms race will become the rulers of the **world**.



Putin: Whoever **Rules AI Rules** the **World**
TechNewsWorld - 12 Sep 2017
Russian President Vladimir **Putin** last week poked the nest of anxieties over the use of **artificial intelligence** to gain power in a video address to ...

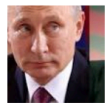
The Data Strategy
AutomatedBuildings.com (press release) - 1 Nov 2017
It takes **Artificial Intelligence** via machine learning to manage and analyze ... ruler of the **world** (via Russian President Vladimir **Putin**)? "Robots are coming for your Job. ... For buildings, **AI** can operate the buildings, but related controls are likely to ... the project team should establish **rules** for the data management that will be ...

Putin and Musk are right: Whoever masters **AI** will run the **world**
CNN - 5 Sep 2017
Gregory C. Allen is an adjunct fellow at the Center for a New American Security. In July 2017, his report "**Artificial Intelligence** and National ...
Vladimir **Putin**: 'Whoever Masters **AI** Will **Rule** The **World**'
Trill (blog) - 7 Sep 2017
View all

Experts: Syrian War Prompting Russians to Expand Unmanned ...
USNI News - 9 Oct 2017
He added work on **artificial intelligence** has become a new priority in systems that are "smarter, larger, ... research and President Vladimir **Putin's** goal that "whoever gets **AI** [**artificial intelligence**] right is going **to rule the world**.

A Sputnik Moment for **Artificial Intelligence** Geopolitics
Council on Foreign Relations (blog) - 7 Sep 2017
Russian President Vladimir **Putin** attends a meeting with Indian ... Whoever assumes leadership in **artificial intelligence** (**AI**) will **rule** the **world**.

For Superpowers, **Artificial Intelligence** Fuels New Global Arms Race
WIRED - 8 Sep 2017
"**Artificial intelligence** is the future, not only for Russia but for all ... last Friday, **Putin** suggested that Russian gains in **AI** could make the **world** ...



Putin Weighs in on **Artificial Intelligence** and Elon Musk Is Alarmed
Big Think - 24 Sep 2017
As Vorozh explained the potential of **artificial intelligence**, **Putin** ... technology, will **rule** the **world**, according to the Russian President. ... In a tweet, Musk cautioned that competition for "**AI** superiority' could result in **World** War 3.

Elon Musk warns battle for **AI** supremacy will spark Third **World** War
The Independent - 6 Sep 2017
Speaking to schoolchildren about **AI** on 1 September, **Putin** declared, "Whoever becomes the leader in this area will **rule** the **world**. ... Kronstadt Group to equip drones with **artificial intelligence**, a similar effort for missiles by the ...
Why Elon Musk needs to be taken seriously on the Russian **AI** threat
In-Depth - The Indian Express - 7 Sep 2017
View all

Those Who Control **AI** Will **Rule** The **World**, **Putin** Warns
Wall Street Pit - 4 Sep 2017
More than 100 robotics and **artificial intelligence** entrepreneurs have signed a letter addressed to the United Nations, which calls for action to ...

Are smart robots going to steal jobs from your children or can they ...
Courier Mail - 13 Oct 2017
Theirs is a clever new **world**, a time when algorithms **rule** and deep ... But beyond their classroom is a **world** in which **artificial intelligence** (**AI**) is being ..... Musk was responding to Russian President Vladimir **Putin** who said the ...

The rise of **AI** is sparking an international arms race
Vox - 13 Sep 2017
"**Artificial intelligence** is the future not only of Russia but of all of ... When **Putin** says the leader in **AI** will become the leader in the **world**, I think ...

Putin seems unconvinced **AI** won't 'eat us'
RT - 21 Sep 2017
Vladimir **Putin** may secretly be on the side of Elon Musk in their indirect debate over the threat posed by **artificial intelligence** (**AI**). ... a group of kids about who would **rule** the **world** in the future, said it will be whichever country ...
Putin reveals fears that robots with **artificial intelligence** will one day ...
Daily Mail - 24 Sep 2017
View all

Will **artificially intelligent** weapons kill the laws of war?
Bulletin of the Atomic Scientists - 18 Sep 2017
So **Putin** says he will share Russian **AI** with the rest of the **world**. ... future of armed conflict to integrate equal levels of **artificial intelligence** (**AI**) ... targets should be attacked, what the **rules** of engagement should be, and so on.

Country to lead the development of **Artificial Intelligence** will ...
Firstpost - 3 Sep 2017
Russian President Vladimir **Putin** says that whoever reaches a breakthrough in developing **artificial intelligence** will come to dominate the **world**. ... Meanwhile, China will launch a series of **artificial intelligence** (**AI**) projects and ... Catalonia declares independence from Spain as direct Madrid **rule** looms; PM ...

Kasparov on **Putin** and the Future of **Artificial Intelligence**
NBCNews.com - 6 May 2017
"The question of **artificial intelligence** still remained unanswered." ... development and Russia's alleged cyber attacks on the free **world** go hand in hand. ... "So we're trying to play by the **rules** and they use our own technology, ... According to Kasparov, part of the problem for **AI** rests in the fact that many ...

Elon Musk warns that **AI** arms race could spark WWIII
New York Post - 4 Sep 2017
... Tesla founder spoke about his fears after Vladimir **Putin** claimed that the nation which controls **artificial intelligence** will come **to rule the world** ...

Artificial intelligence — the arms race we may not be able to control
The Hill - 21 Sep 2017
... in this sphere will become ruler of the **world**," said Vladimir **Putin**. The sphere the President of Russia is referring to is **artificial intelligence** (**AI**) ...

Only 13% of Americans are scared that robots will take their jobs, a ...
CNBC - 5 Sep 2017
Responding to Russian president Vladimir **Putin's** warning this week that "whoever leads in **artificial intelligence** will **rule** the **world**," Tesla CEO Elon Musk said that competition for **AI** will "most likely cause" **World** War 3.

National Strategies

CIFAR

Pan-Canadian Artificial Intelligence Strategy Overview

ENGLISH.GOV.CN
THE STATE COUNCIL
THE PEOPLE'S REPUBLIC OF CHINA

the WHITE HOUSE
PRESIDENT BARACK OBAMA
BRIEFING ROOM | ISSUES | THE ADMINISTRATION | 1600 PENN

HOME · BLOG

The Administration's Report on the Future of Artificial Intelligence

OCTOBER 12, 2016 AT 6:02 AM ET BY ED FELTEN AND TERAH LYONS

HOME | STATE COUNCIL | PREMIER | NEWS | POLICI

HOME >> POLICIES >> LATEST RELEASES

China issues guideline on artificial intelligence development

China's Technology Transfer Strategy:

How Chinese Investments in Emerging Technology Enable A Strategic Competitor to Access the Crown Jewels of U.S. Innovation

Eric Schmidt says America needs to 'get its act together' in AI competition with China

'Trust me, these Chinese people are good'

by James Vincent | @jjvincent | Nov 1, 2017, 1:22pm EDT

Epistemic Calibration

"*Prediction is very difficult, especially about the future.*"
-attributed to Niels Bohr, and others...

Epistemic Calibration

"*Prediction is very difficult, especially about the future.*"
-attributed to Niels Bohr, and others...

Failure Mode 1: Overconfidence that some specific possibility, X, will happen.

Epistemic Calibration

"*Prediction is very difficult, especially about the future.*"
-attributed to Niels Bohr, and others...

Failure Mode 1: Overconfidence that some specific possibility, X, will happen.

Failure Mode 2: Overconfidence that X will not happen.

Epistemic Calibration

"*Prediction is very difficult, especially about the future.*"
-attributed to Niels Bohr, and others...

Failure Mode 1: Overconfidence that some specific possibility, X, will happen.

Failure Mode 2: Overconfidence that X will not happen.

Failure Mode 3: Given uncertainty, dismiss value of studying X.

Epistemic Calibration

"*Prediction is very difficult, especially about the future.*"
-attributed to Niels Bohr, and others...

Failure Mode 1: Overconfidence that some specific possibility, X, will happen.

Failure Mode 2: Overconfidence that X will not happen.

Failure Mode 3: Given uncertainty, dismiss value of studying X.

Lesson: Accept uncertainty and distributional beliefs. Uncertainty does not imply futility.

Narrow Transformative Capabilities

Narrow Transformative Capabilities

Most likely where: data rich, can simulate environment, narrow domains, ripe technical problem, and/or high stakes.

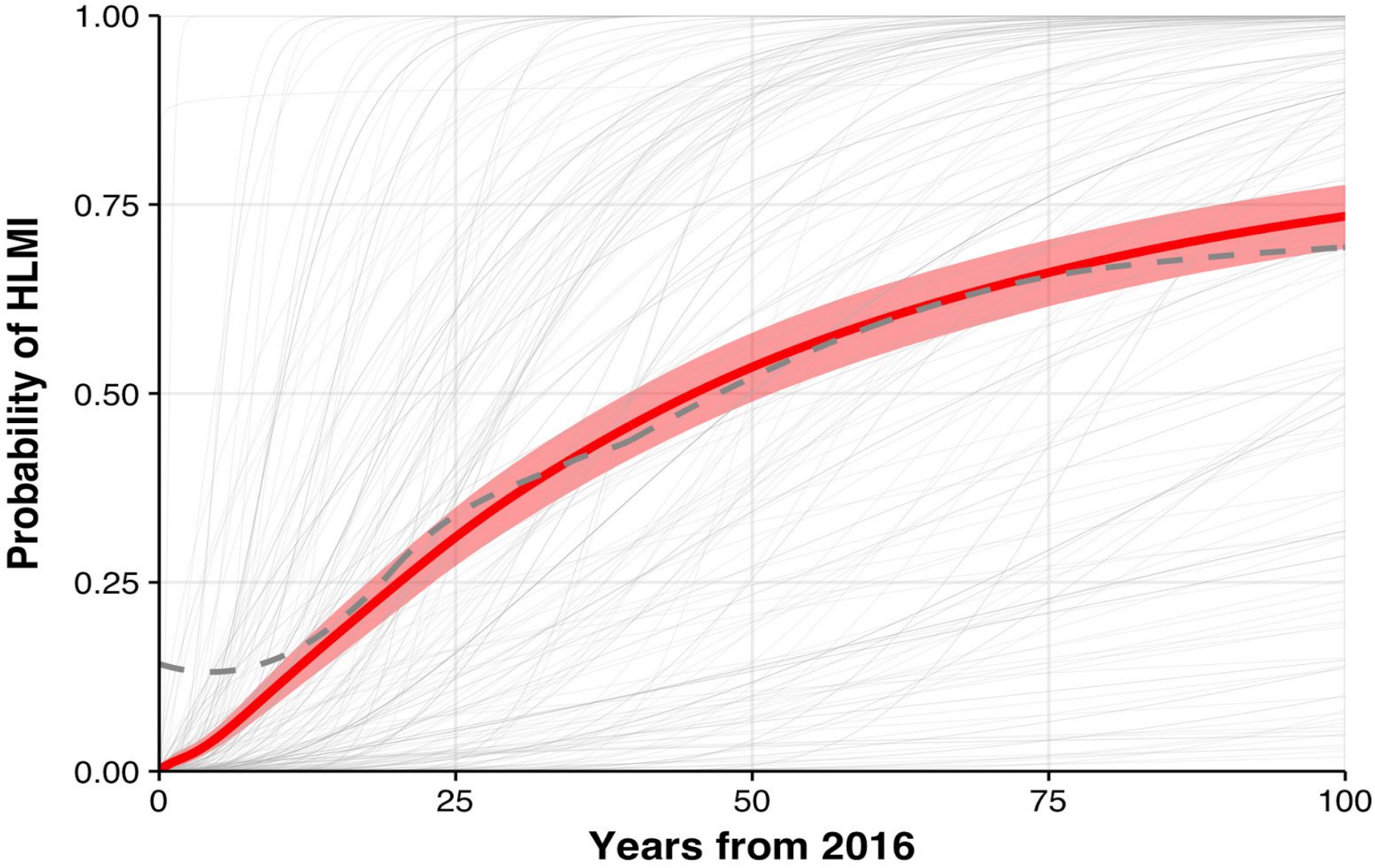
Narrow Transformative Capabilities

Most likely where: data rich, can simulate environment, narrow domains, ripe technical problem, and/or high stakes.

- Finance. Operations/logistics.
- Engineering, science, math.
- Cyber.
- Surveillance.
- Profiling (lie detection, emotion detection, psychological insight, DNA). Personal assistants/advertising.
- Social network mapping and manipulation.

Survey of NIPS/ICML about HLMI (Grace et al 2017)

Survey of NIPS/ICML about HLMI (Grace et al 2017)



Indicators

Relevant: indicators for altered circumstances: transformative capabilities or altered probabilities. Eg economic, security, technical.
Informative: close to necessary and/or sufficient condition. (Not game-able.)
Precise: perhaps not "AGI"
Leading Indicators: eg not "HLMI: better than all humans at all tasks"

Technical Landscape

- Rapid & Broad Progress?
- Species and Properties
- Other Strategic Tech
- Measuring Inputs, Capabilities, Performance
- AI Production Function
- Forecasting and Indicators
- Safety Production Function

Role for AI Researchers in AI Governance

- Engage (eg with FHI) on Technical Landscape.
- Improve public understanding of AI.
- Develop international network for science diplomacy.
- Consider social impact of work.
- Focus on technical problems especially pertinent to social problems.

(h/t Brundage)

AI Politics

Sample of projects:

- Public opinion and the public as relevant political actor
- Government - AI industry relations
- China's AI landscape and opportunities for coordination
- Case studies: nuclear power, cryptography, space race
- Strategic properties of tech: offense/defense, destructiveness, first-mover, power volatility and observability, dual-use, strategic gradient, cooperation promoting/inhibiting.
- AI race mitigation, through modeling dynamics
- Deployment Problem: "what if AGI breakthrough tomorrow..."
- Unipolar versus multipolar outcomes

(and many more)

AI Governance

What potential global governance systems, including norms, policies, laws, processes, and institutions, can best ensure the beneficial development and use of advanced AI systems?

- Institutional, constitutional, and procedural design of an AI governance body (or bodies)
- How to incentivize creation of an AI governance regime or organization
- Mechanisms for increased cooperation and coordination
- Case studies: Baruch plan and related, CERN, ITER, cyber, medical trials.
- AI verification and agreements
- AI IGO for Common Good
- World preferences and values

Summary

- (1) Impending global governance challenges.
- (2) Warrants attention, even given uncertainty.
- (3) Transformative possibilities.
- (4) Lots of ways you can contribute.

Research Landscape

Technical Landscape

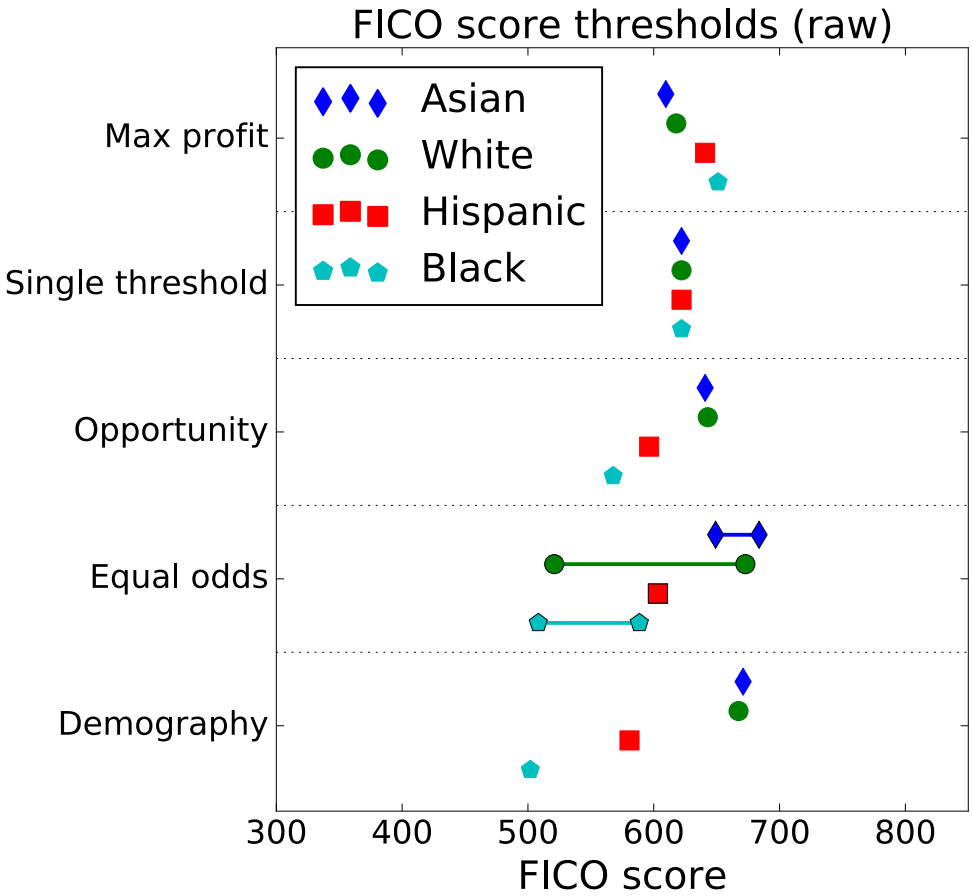
- Rapid & Broad Progress; Species and Properties; Other Tech
- Measuring Inputs, Capabilities, Performance; AI Production Function; Forecasting AI
- Safety Production Function

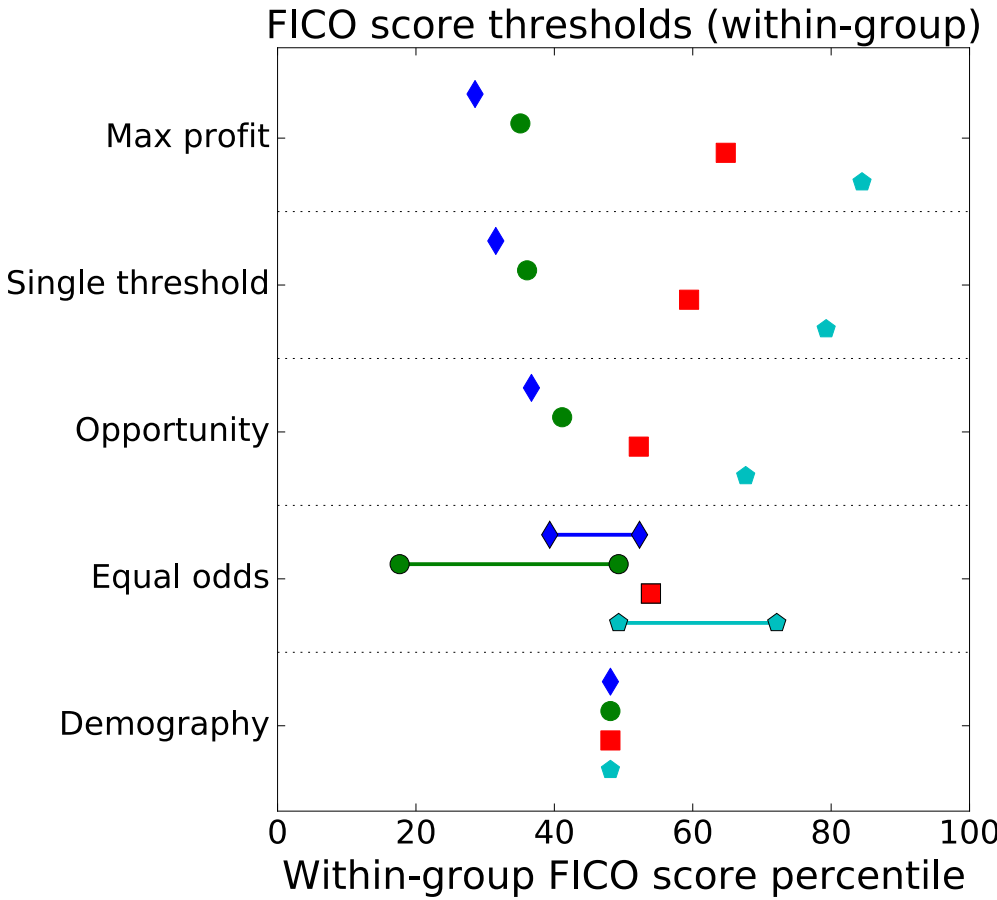
AI Politics

- Domestic & Mass Politics; Unemployment & Inequality; Public Opinion; Authoritarian Control
- IPE, Strategic Trade
- International Security; AI Race; Norms, Treaties, Int'l Control

AI Governance

- Values and Principles
- Institutions and Mechanisms


FICO score thresholds (raw)
Asian
White
Hispanic
Black
Max profit
Single threshold
Opportunity
Equal odds
Demography
300
400
500
600
700
800
FICO score


FICO score thresholds (within-group)
Max profit
Single threshold
Opportunity
Equal odds
Demography
0
20
40
60
80
100
Within-group FICO score percentile

AI Safety Production Function

How difficult is it to build a safe, aligned advanced AI? How subtle are the challenges? What is the **Performance-Safety Tradeoff?**

- Not that hard. OR: a necessary part of building functional AI, so a challenge engineers will confront in due time.
- Don't know. Maybe +10% to +1000% development costs (in dollars, time, compute, talent,...).
- "not a hard problem if we have two years once we have the system. It is almost impossible if we don't."
- Near impossible. Like building a rocket and spacecraft for a moon-landing, without ever having done a test launch.

Dimensions of AI Safety Production Function

How difficult is it to agree on and build a safe, aligned advanced AI system?

- **Extent of Externalities of Risks**: All risks internal/local vs substantial systemic/global/catastrophic risks.

Dimensions of AI Safety Production Function

How difficult is it to agree on and build a safe, aligned advanced AI system?

- **Extent of Externalities of Risks**: All risks internal/local vs substantial systemic/global/catastrophic risks.
- **Difficulty**: Not too hard (+10%), hard (+100%) or extremely hard (+1000%), relative to development costs.

Dimensions of AI Safety Production Function

How difficult is it to agree on and build a safe, aligned advanced AI system?

- **Extent of Externalities of Risks**: All risks internal/local vs substantial systemic/global/catastrophic risks.
- **Difficulty**: Not too hard (+10%), hard (+100%) or extremely hard (+1000%), relative to development costs.
- **Parallelizability**: Safety work can be done in parallel vs only at the end (eg testing) or start (architecture dependent).

Dimensions of AI Safety Production Function

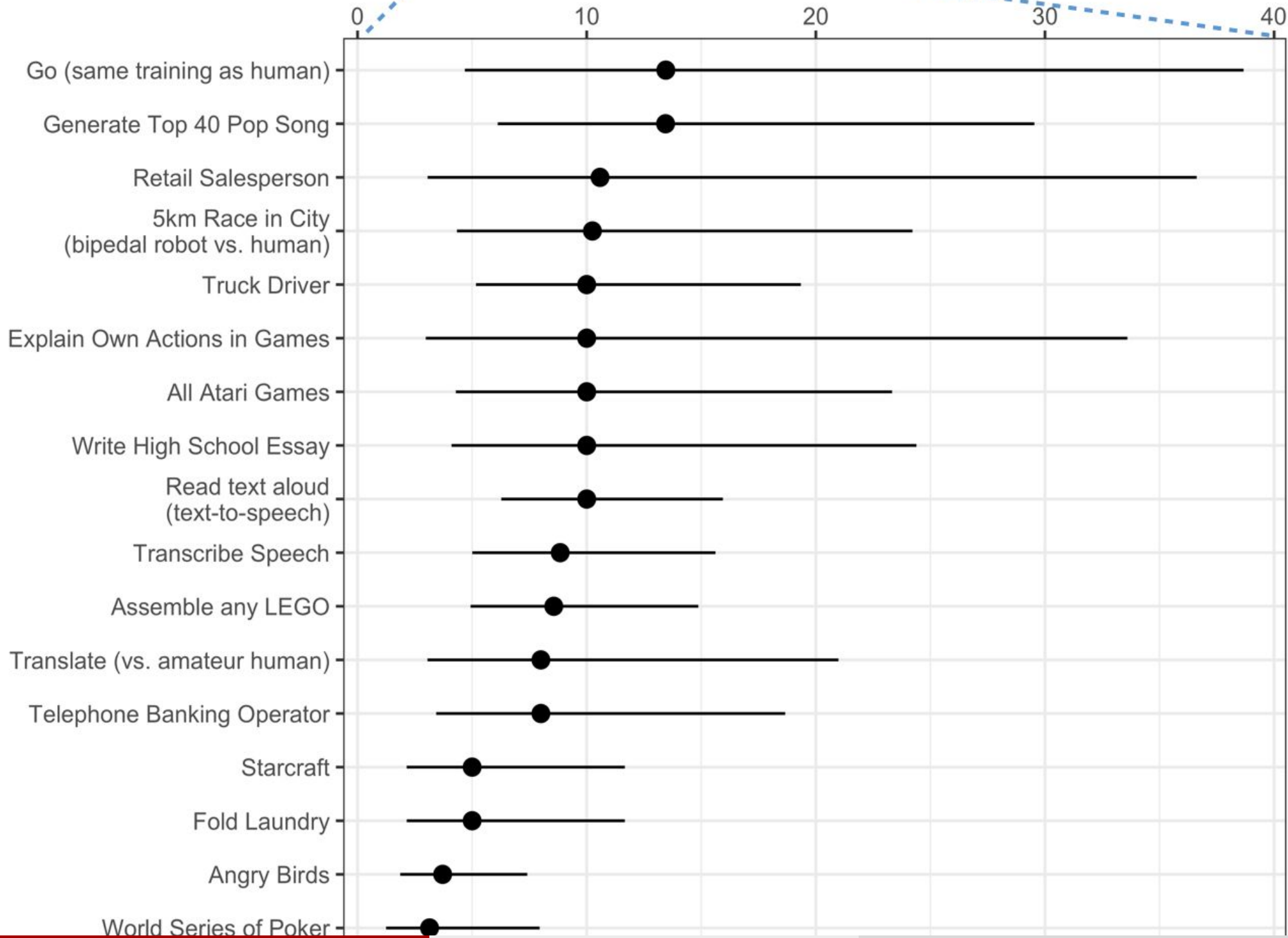
How difficult is it to agree on and build a safe, aligned advanced AI system?

- **Extent of Externalities of Risks**: All risks internal/local vs substantial systemic/global/catastrophic risks.
- **Difficulty**: Not too hard (+10%), hard (+100%) or extremely hard (+1000%), relative to development costs.
- **Parallelizability**: Safety work can be done in parallel vs only at the end (eg testing) or start (architecture dependent).
- **Observability/Provability**: Safe systems can be demonstrably/provably safe or not.

Dimensions of AI Safety Production Function

How difficult is it to agree on and build a safe, aligned advanced AI system?

- **Extent of Externalities of Risks**: All risks internal/local vs substantial systemic/global/catastrophic risks.
- **Difficulty**: Not too hard (+10%), hard (+100%) or extremely hard (+1000%), relative to development costs.
- **Parallelizability**: Safety work can be done in parallel vs only at the end (eg testing) or start (architecture dependent).
- **Observability/Provability**: Safe systems can be demonstrably/provably safe or not.
- **Common perspective**: Stakeholders agree what it means for the system to be safe vs not.


0
10
20
30
40
Go (same training as human)
Generate Top 40 Pop Song
Retail Salesperson
5km Race in City
(bipedal robot vs. human)
Truck Driver
Explain Own Actions in Games
All Atari Games
Write High School Essay
Read text aloud
(text-to-speech)
Transcribe Speech
Assemble any LEGO
Translate (vs. amateur human)
Telephone Banking Operator
Starcraft
Fold Laundry
Angry Birds
World Series of Poker

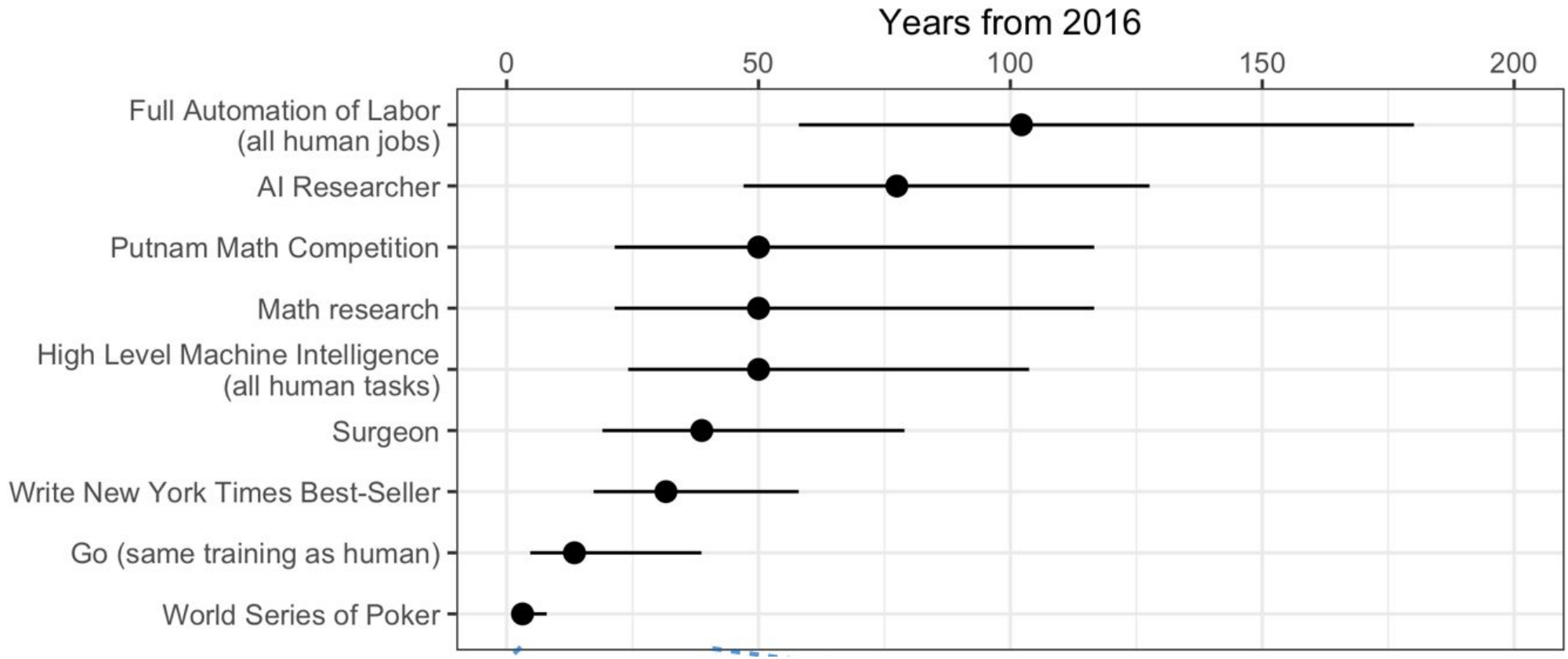

Years from 2016
0
50
100
150
200
Full Automation of Labor
(all human jobs)
AI Researcher
Putnam Math Competition
Math research
High Level Machine Intelligence
(all human tasks)
Surgeon
Write New York Times Best-Seller
Go (same training as human)
World Series of Poker

Figure: Median Probabilities Assigned to HLMI Outcomes