AI and Existential Risk

Levan Bokeria

Presentation for the Foundational Models Discussion Group

2024-02-05

Foreword

- Slides ~= endorsement
- Postpone discussing people and institutes (like “the letter” and who signed it)
- Discuss only the substance of AI x-risk:
  - Some of it!

The AI Safety landscape:

	Short Term	Long Term
Accident	e.g. Self-Driving Car Crashes	Oh boy
Misuse	e.g Deep Fakes	e.g. AI-Enabled Dictatorship

Robert Miles, “Intro to AI Safety” https://www.youtube.com/watch?v=pYXy-A4siMw&t=604s

Definitions

X-risk:

- “one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential” (Bostrom 2002).

Artificial General Intelligence (AGI):

- An AI system capable of solving tasks of at least as much generality as humans.
- “highly autonomous systems that outperform humans at most economically valuable work” (cite, OpenAI: https://openai.com/charter)

Artificial Superintelligence (ASI):

- ”Any intellect that greatly exceeds cognitive performance of humans in virtually all domains of interest” (Bostrom, “Superintelligence”, 2014)

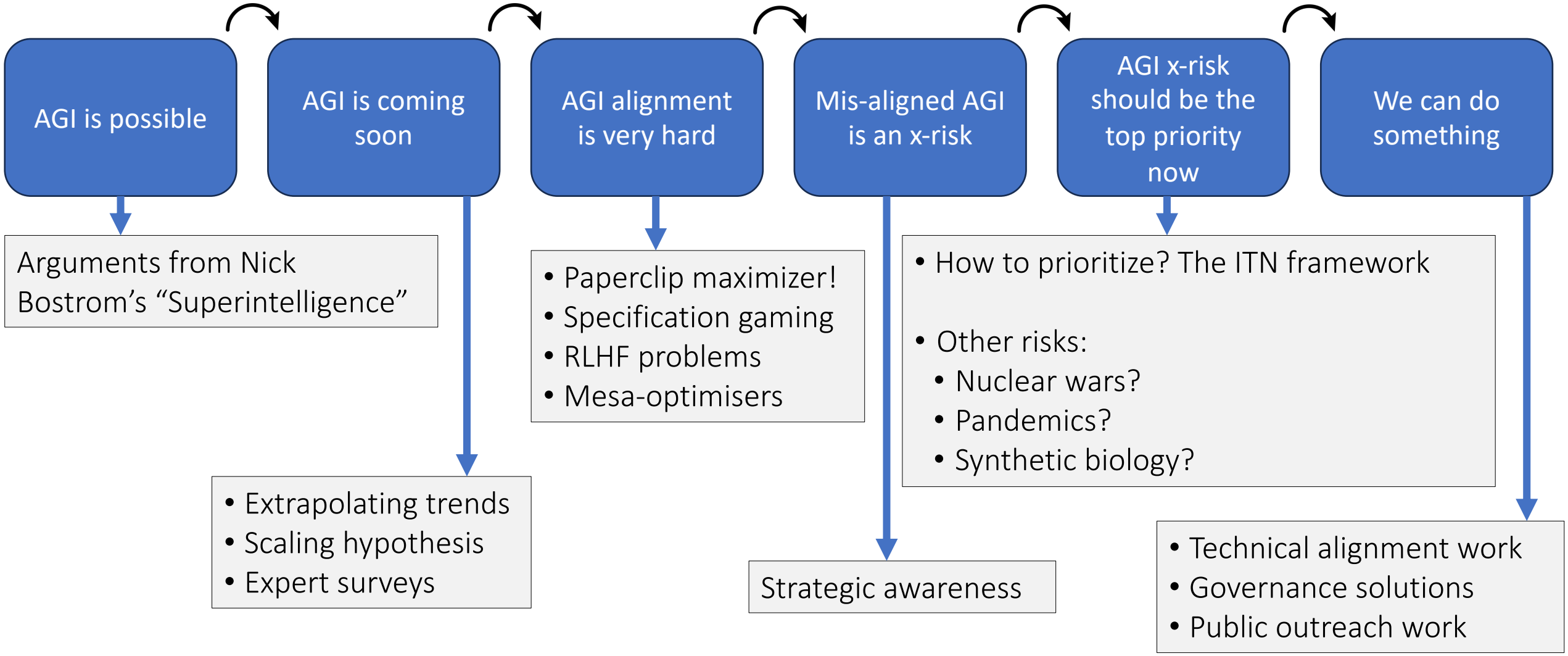
Intelligence:

- Whatever it is that lets an agent choose effective actions to achieve its goals (Robert Miles, “Intro to AI Safety”)

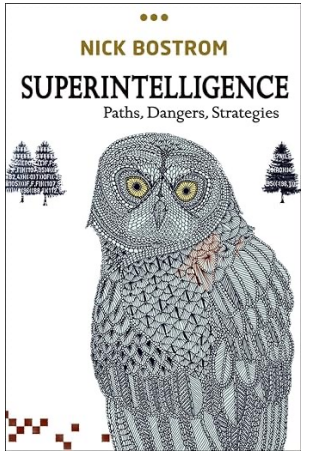
Alignment:

- An AI is “intent aligned” if it is trying to do, or “impact aligned” if it is succeeding in doing what a human person or institution wants it to do (Critch, 2020).

General outline of the (strongest) argument:



AGI is possible



- Recapitulate evolution
- Human brain as a template – whole brain emulation
- Turing (1950) – A child machine:

> Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain.[3]

- Brain-Computer interfaces
- Genetically engineer to higher biological abilities

- Counterarguments:

Article | Open access | Published: 17 June 2020

Why general artificial intelligence will not be realized

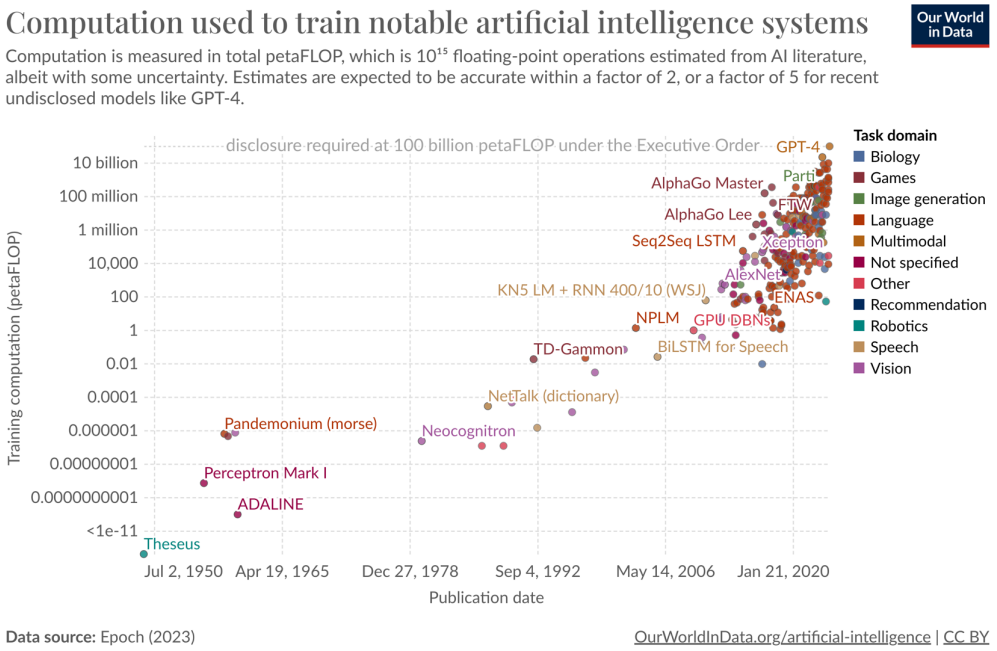
Ragnar Fjelland

Humanities and Social Sciences Communications **7**, Article number: 10 (2020) | Cite this article

186k Accesses | **109** Citations | **754** Altmetric | Metrics

AGI is coming soon:

- Deep learning explodes since mid-2010s.
- AI/ML needs:
  - Algorithms
  - Data
  - Compute



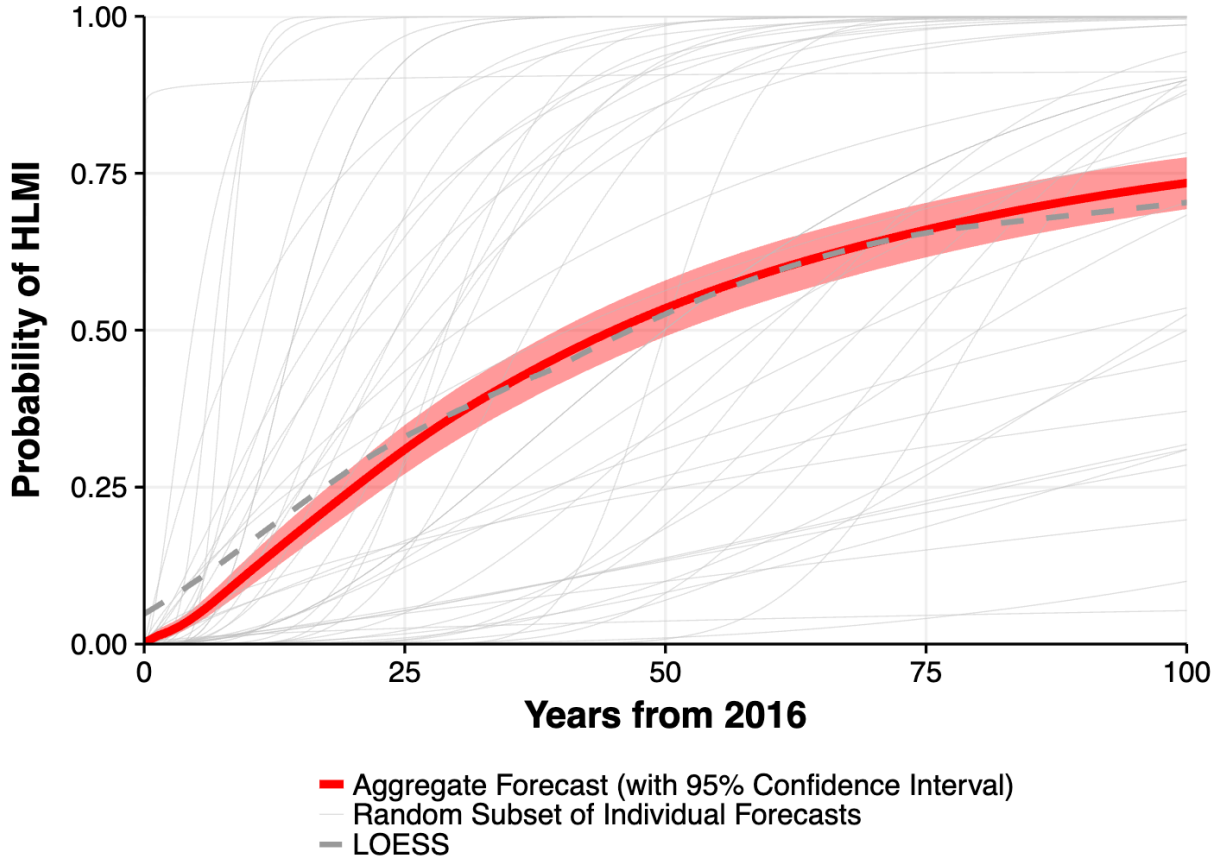
- [Scaling hypothesis:]
  - Computer vision: ImageNet
  - NLP: GPT-3

- AlphaStar, which can beat top professional players at StarCraft II (January 2019)
- MuZero, a single system that learned to win games of chess, shogi, and Go — without ever being told the rules (November 2019)
- GPT-3, a natural language model capable of producing high-quality text (May 2020)
- GPT-f, which can solve some Maths Olympiad problems (September 2020)
- AlphaFold 2, a huge step forward in solving the long-perplexing protein-folding problem (July 2021)
- Codex, which can produce code for programs from natural language instructions (August 2021)
- PaLM, a language model which has shown impressive capabilities to reason about things like cause and effect or explaining jokes (April 2022)
- DALL-E 2 (April 2022) and Imagen (May 2022), which are both capable of generating high-quality images from written descriptions
- SayCan, which takes natural language instructions and uses them to operate a robot (April 2022)
- Gato, a single ML model capable of doing a huge number of different things (including playing Atari, captioning images, chatting, and stacking blocks with a real robot arm), deciding based on its context what it should output (May 2022)
- Minerva can solve complex maths problems — fairly well at college level, and even better at high school maths competition level. (Minerva is far more successful than forecasters predicted in 2021.)

- GPT 3.5, 4…
- Bard
- Llama
- xAI?

AGI coming soon

- Expert surveys of authors in NeurIPS and ICML:
  - Median researcher estimate of “extremely bad (e.g. human extinction)” changes: 5% (2016), 2% (2019), 5% (2022)



Alignment is hard

- Giving a long list of “dos” and “don’ts” would not work.
- Specification gaming
- RLHF with superintelligence
- Instrumental goals

The paperclip maximizer

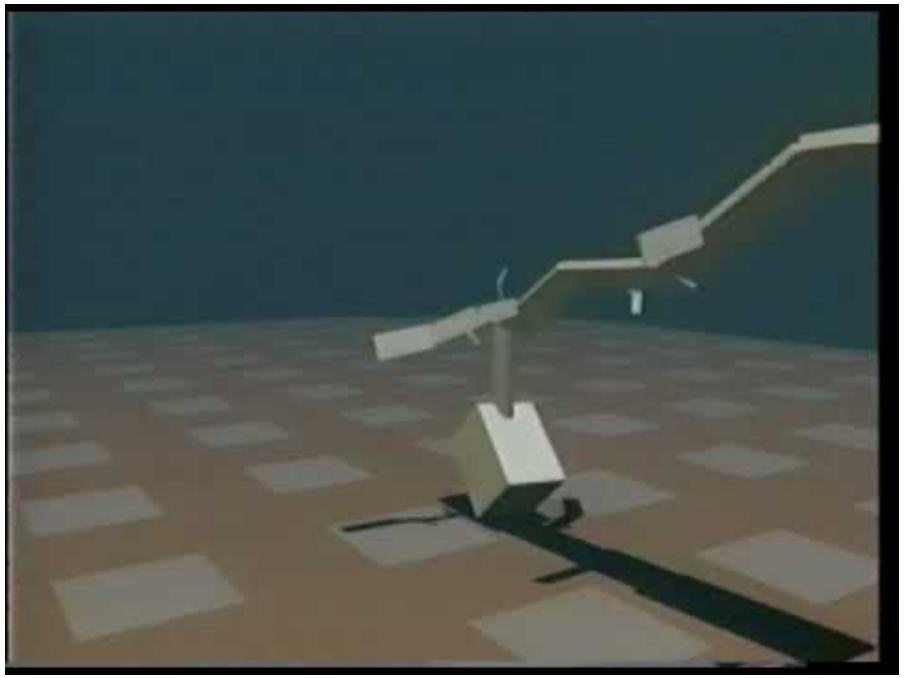
Instrumental convergence hypothesis

- **“Instrumental convergence** is the hypothetical tendency for most sufficiently intelligent beings (human and non-human) to pursue similar sub-goals, even if their ultimate goals are quite different.” (Wikipedia)
  - Self-preservation: “You can’t fetch coffee if you’re dead” – Stuart Russel.
  - Goal integrity: AGI will prevent us from changing its goal.
  - Power acquisition: for almost any purpose, having more power/resources will be useful.

Specification gaming

Reinforcement learning agents

Evolutionary algorithms



Finding bugs

Victoria Krakovna blog: https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

Specification gaming

"A system that is optimizing a function of n variables,

where the objective depends on a subset of size $k<n$,

will often set the remaining unconstrained variables to extreme values.

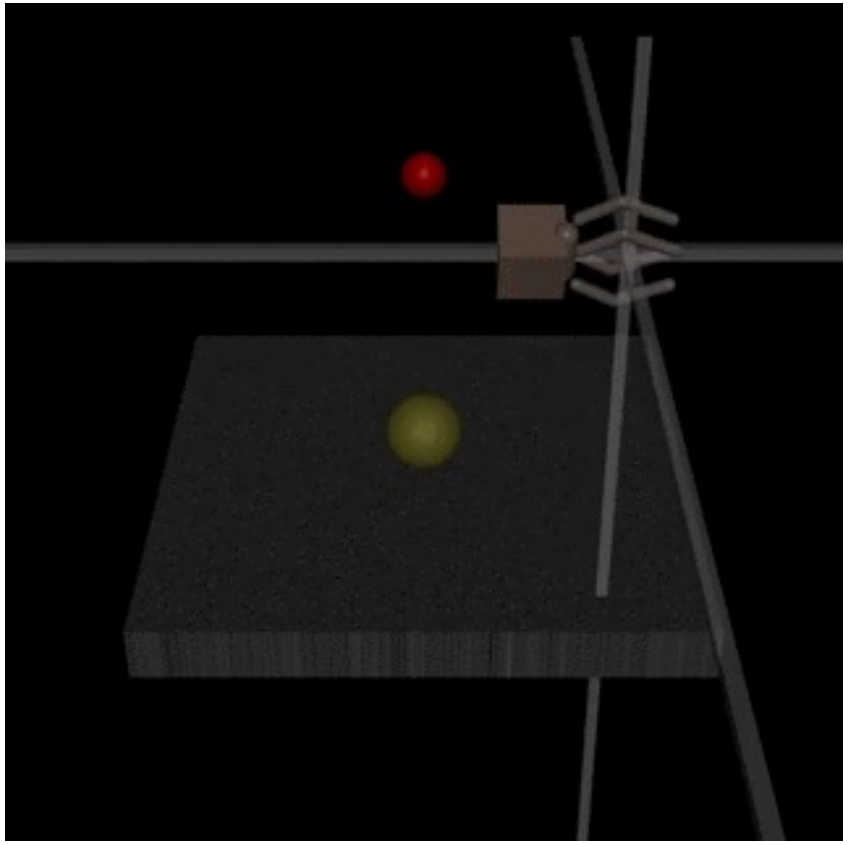
If one of those unconstrained variables is something we care about,

the solution found may be *highly undesirable*"

— *Prof. Stuart Russell*

Robert Miles, “Intro to AI Safety” https://www.youtube.com/watch?v=pYXy-A4siMw&t=604s

RLHF with superintelligence



Distributional shift

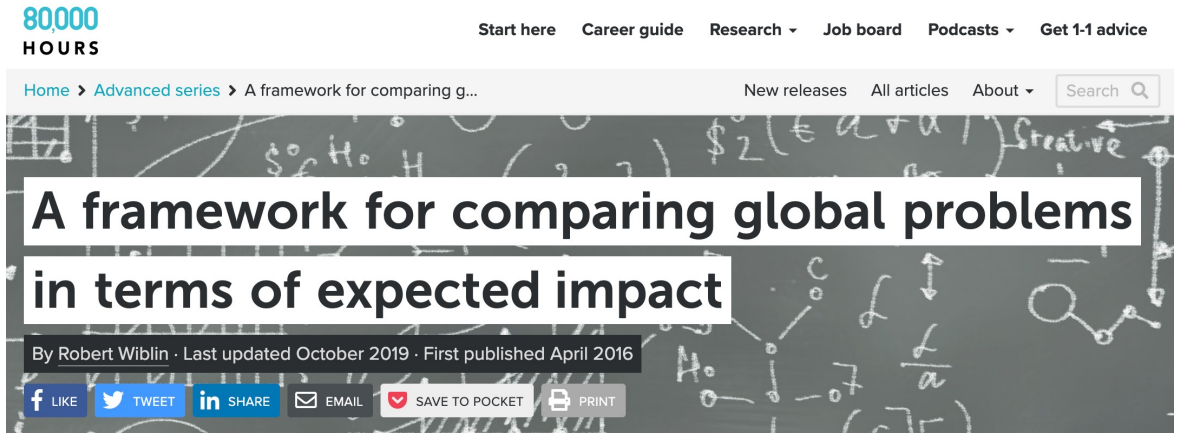
Mesa-Optimisers

Base optimiser	Evolution	Humans
	↓	↓
Mesa optimiser	Humans	AI

Hubinger et al, 2021 https://arxiv.org/pdf/1906.01820.pdf

Is AI x risk a priority?

- The ITN framework:
  - Importance
    - *Good done / % of a problem solved*
  - Tractability
    - *% of a problem solved / % increase in resources*
  - Neglectedness
    - *% increase in resources / extra person or $*



What can we do about it?

- Technical work on AI alignment
  - Inverse Reinforcement Learning (IRL)
  - Alignment through Debate
  - Iterated Distillation and Amplification (IDA)
- Governance work
- Public outreach and education