Institute and Faculty of Actuaries

AI : X-Risk

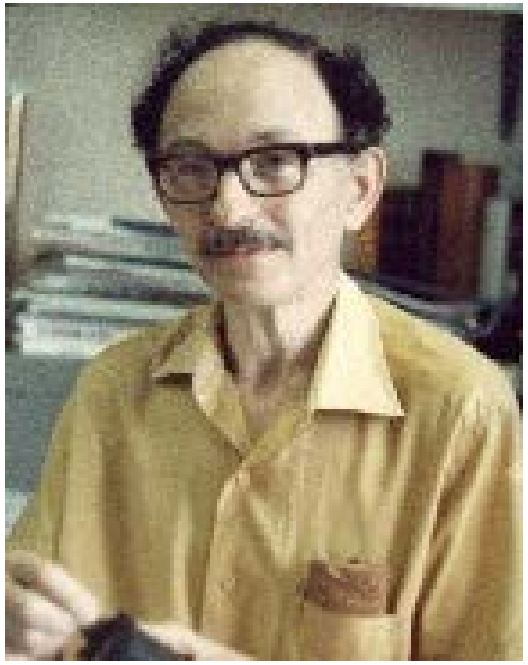
Matthew Byrne, Head of Actuarial Function, NFU Mutual
Hazel Davis, Actuary, Sabre

Expressions of individual views by members of the Institute and Faculty of Actuaries and its staff are encouraged.

The views expressed in this presentation are those of the presenter and not of our employers or the Institute and Faculty of Actuaries.

02 May 2024

Extinction risk from AI has long been a concern



‘..the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.’

I.J.Good 1965

Institute and Faculty of Actuaries

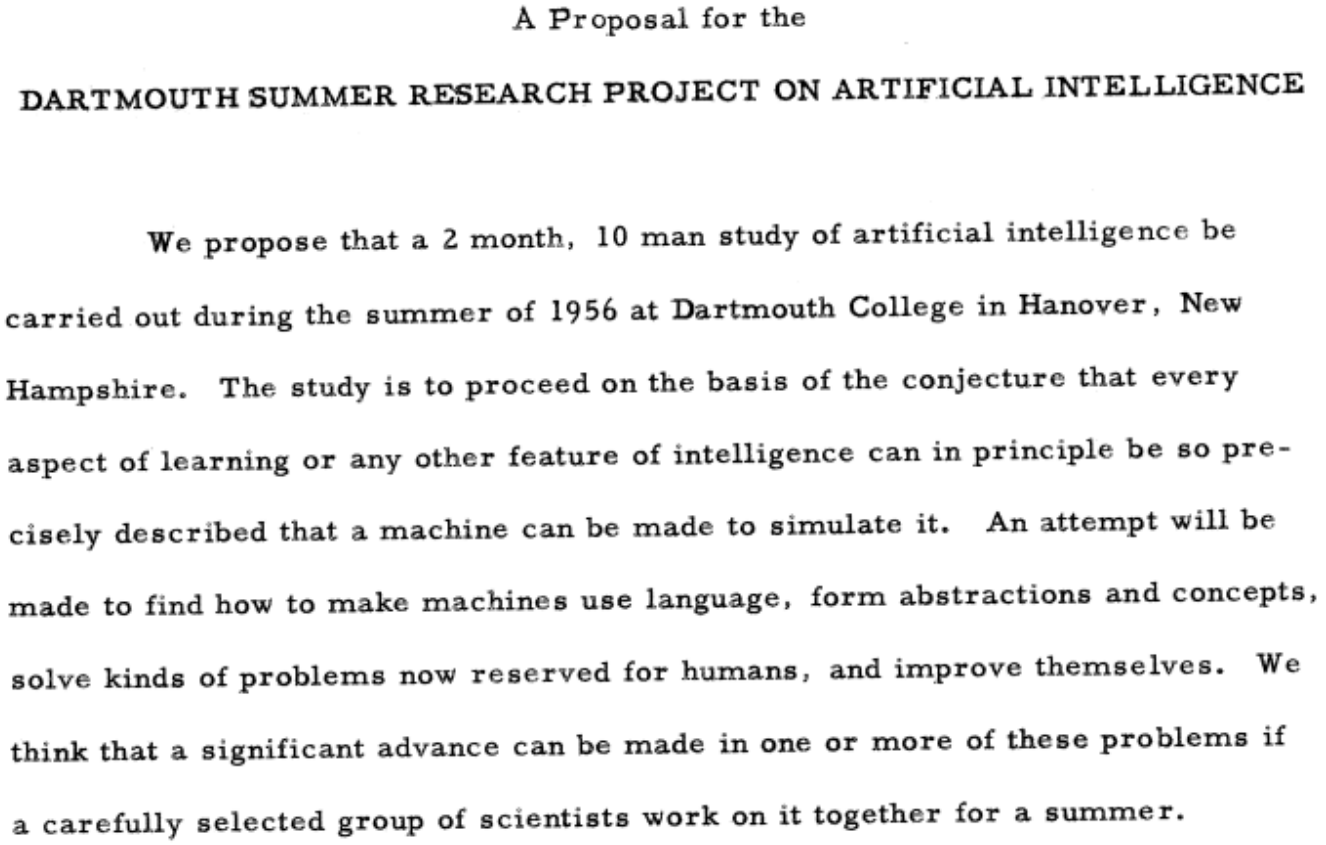
Original publication: 1978 Immediate source: http://imagebase.lib.vt.edu/view_record.php?URN=VTA0005 Fair use, wikipedia.org

Outline

- A brief history of AI
- The singularity
- Alignment
- Extinction risk
  - Possible routes
  - Probabilities and timelines
  - Comparison vs other risks
  - LLMs – signs of intelligence?
  - So what should we do?

Institute and Faculty of Actuaries

A brief history of AI as a discipline



A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

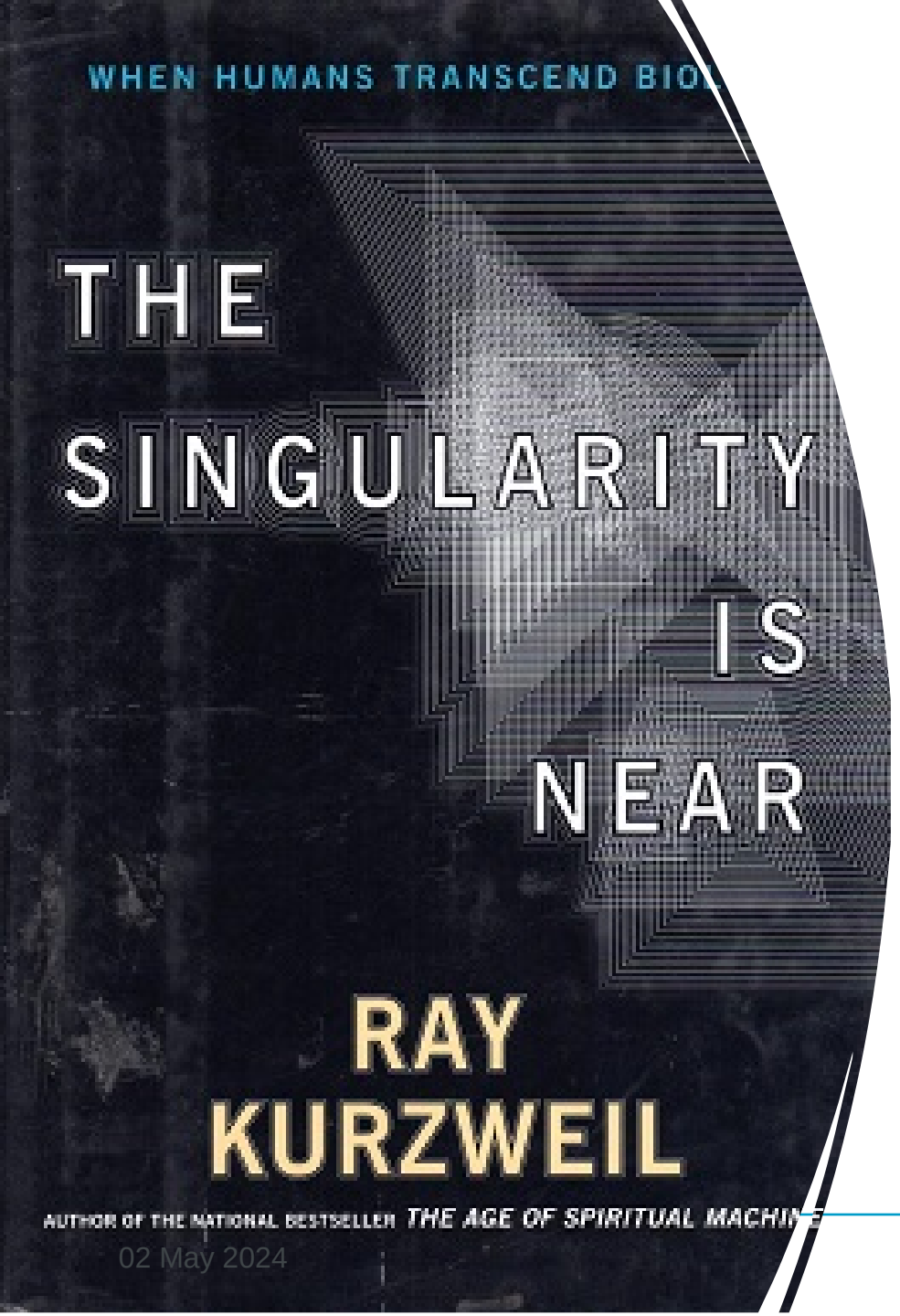
We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. 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.

- 1956 ‘Artificial Intelligence’
- 1959 ‘Machine Learning’
- 1962 ‘Data Analysis’
- 1974 ‘Data Science’
- 1989 ‘Knowledge Discovery’
- 1999 ‘Data Mining’
- 2022 ‘ChatGPT’

Since then AI (or AI hype?) everywhere!

Institute and Faculty of Actuaries

raysolomonoff.com/dartmouth/boxa/dart564props.pdf
A Brief History of Data Science – DATAVERSITY



The Technological Singularity

within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended.

1993 Vernor Vinge: The Coming Technological Singularity

the intelligence that will emerge will continue to represent the human civilization.

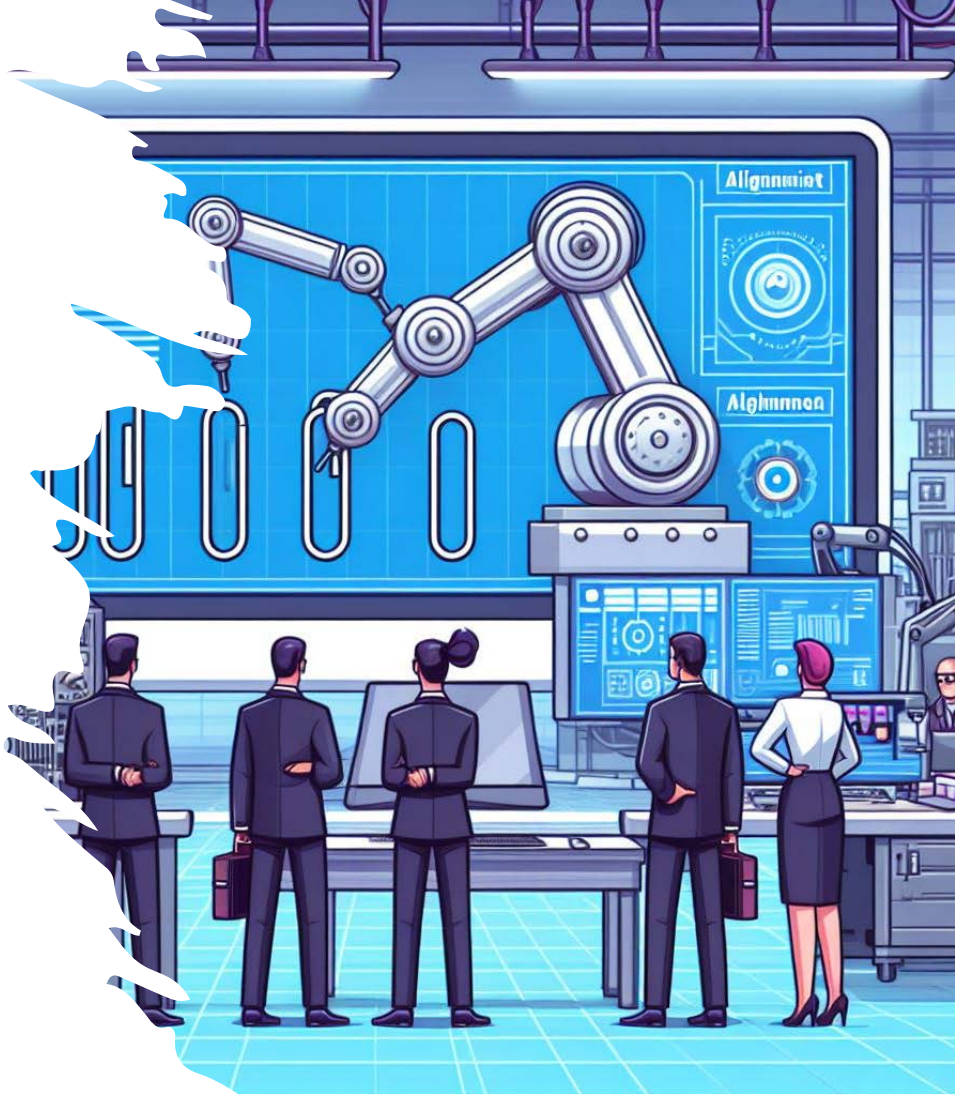
2005 Ray Kurzweil

Institute and Faculty of Actuaries

Cover Image by abebooks.co.uk; Fair use, wikipedia.org
The Coming Technological Singularity, Vernor Vinge, 1993 (accelerating.org)

The Alignment Problem

- Suppose the first AI that reached AGI was one designed to make as many paperclips as possible
  - humans might decide to switch it off
  - human bodies contain a lot of atoms that could be made into paperclips.
- Tell it to make no more than 100 paperclips per day and turn off
  - humans could still shut it down before it reaches its daily target
  - it could decide P >0 that the paperclips do not meet the specification, and dedicate increasing resources to quality control
- Tell it to make paperclips but only in a way that:
  - makes us happy?
  - does no harm?



Nick Bostram: Superintelligence, Tomas Pueyo Uncharted Territories

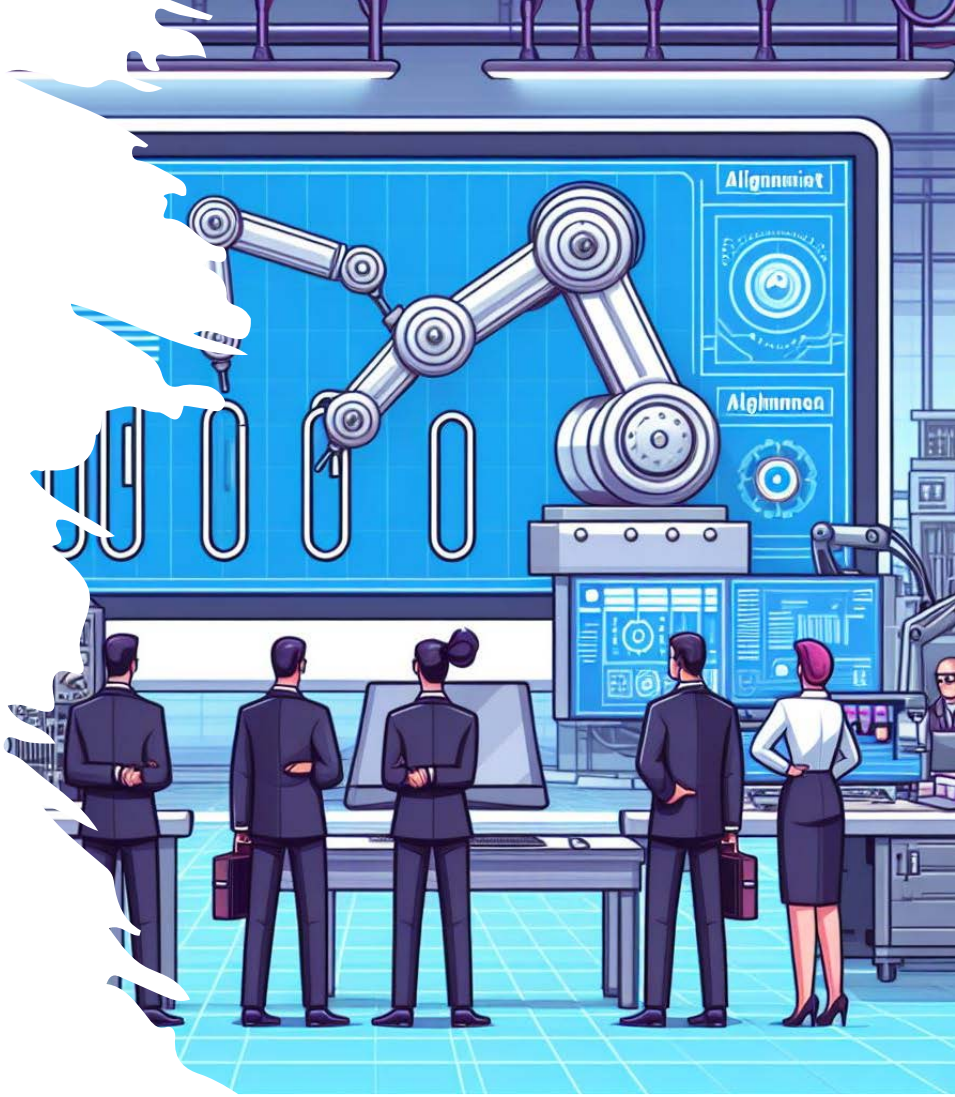
Artificial Intelligence May Doom The Human Race Within A Century, Oxford Professor Says | HuffPost The World Post

Institute and Faculty of Actuaries

Let the AI figure it out…

- Use the AI's intelligence to learn the values we want it to pursue.

'.. our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted'



Yudkowsky: Coherent Extrapolated Volition 2004

How could AI lead to human extinction?

Global Dictatorship: AI decides humanity is no longer necessary.

Hyper-intelligence: AI decides humanity is a threat to its own existence.

Rapid Advancements: AI development outpaces our ability to control it and manage risks.

Incompatible Ecosystem: AI creates an ecosystem that is incompatible with human life.

Misaligned AGI risk

Autonomous Weapons: AI is used to create autonomous weapons that lead to uncontrollable warfare.

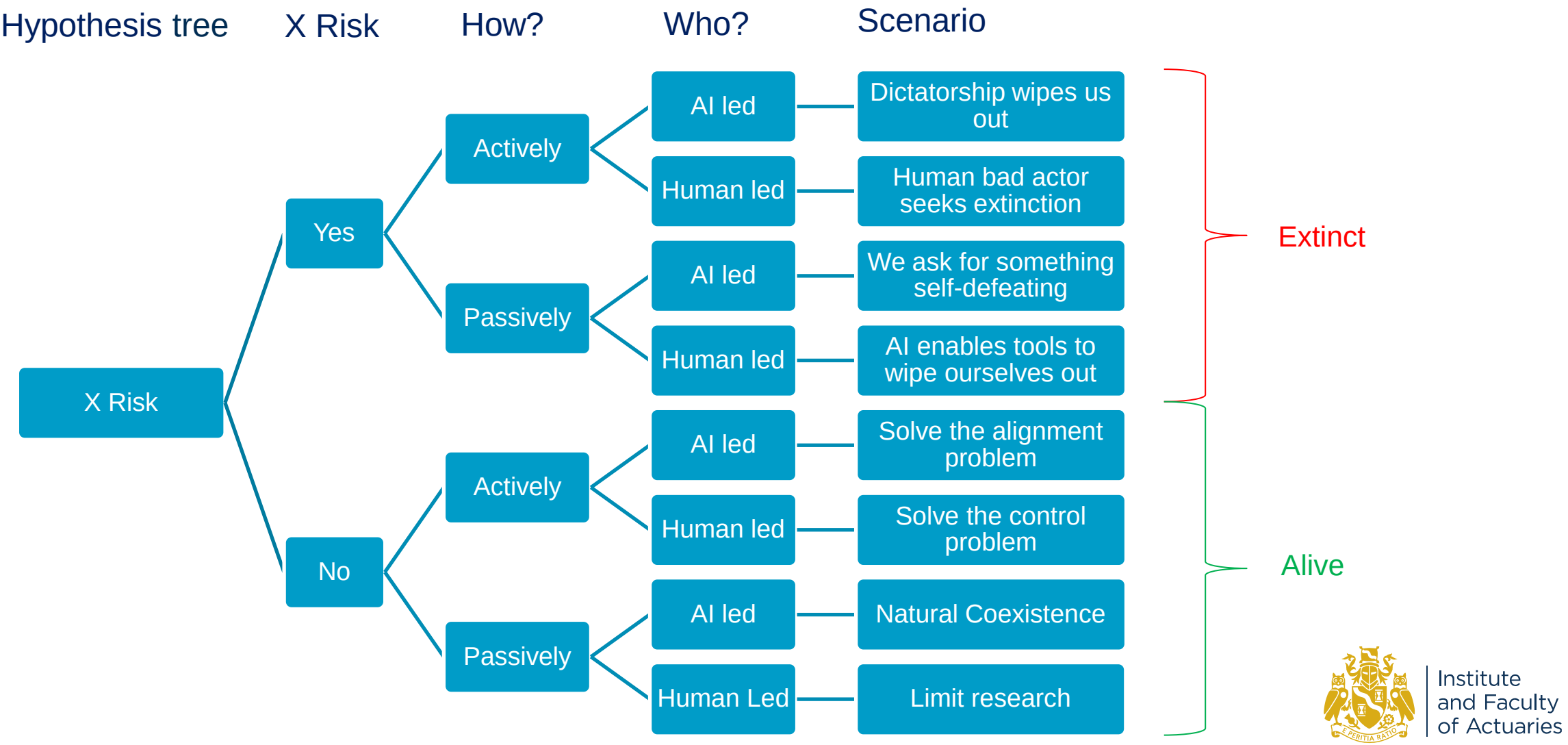
Biohazards: AI could be used to create a super-virus or bioweapon that could wipe out humanity.

Human led risk

Institute and Faculty of Actuaries

Source: AI-powered chat mode of Microsoft Bing

Will AI lead to human extinction?



Institute and Faculty of Actuaries

Considering the timeline can help us allocate estimates

Probabilities for illustration

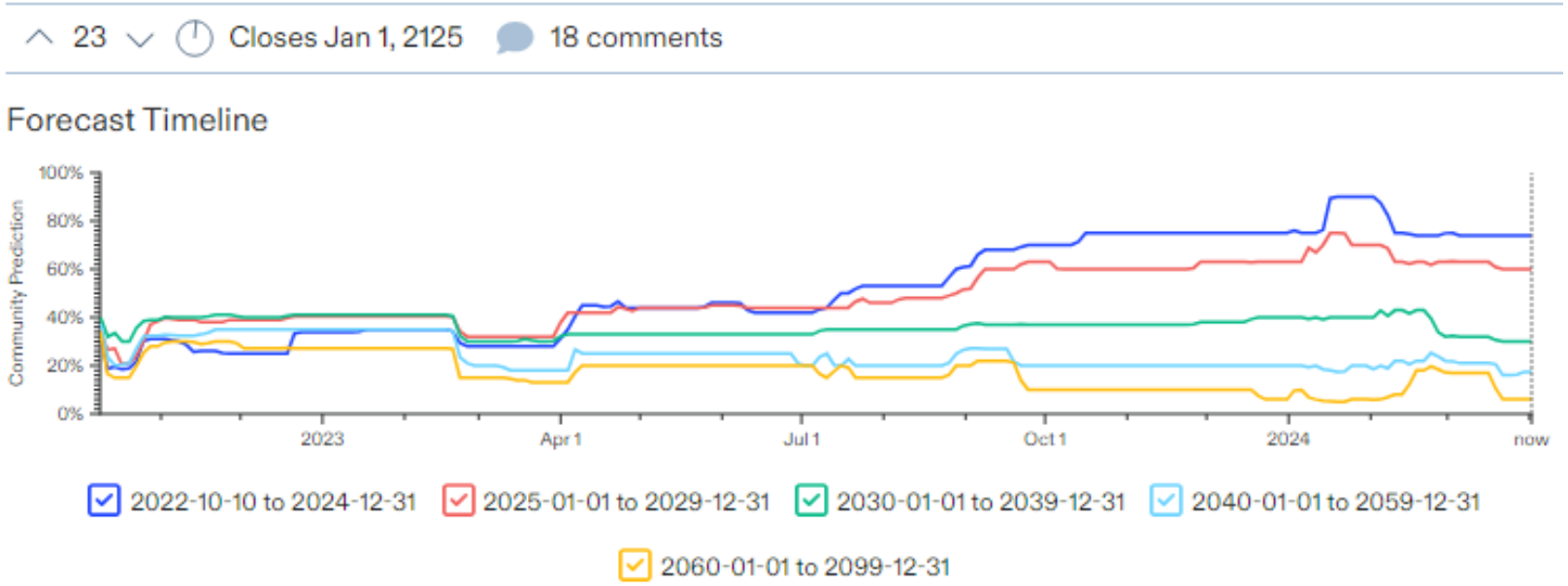
Extinction?	How?	Who?	Example	p(2 years)	p(5 years)	p(10 years)	p(20 years)	p(100 years)	p(200 years)
Yes	Actively	AI led	AI Dictatorship chooses to wipes us out	0%	1%	1%	3%	4%	5%
Yes	Actively	Human Led	Human Bad actor seeks extinction	0%	0%	0%	0%	1%	2%
Yes	Passively	AI led	We ask for something self defeating	1%	3%	10%	15%	30%	40%
Yes	Passively	Human Led	AI tools enable us to wipe ourselves out	2%	3%	4%	4%	4%	5%
No	Actively	AI led	Managed alignment/ Solve the alignment problem	1%	2%	5%	10%	5%	5%
No	Actively	Human Led	Maintained control/ Solve the control problem	1%	2%	5%	4%	3%	2%
No	Passively	AI led	Natural coexistence	94%	87%	70%	60%	50%	39%
No	Actively	Human Led	Limited research avoided issue	1%	2%	5%	4%	3%	2%

Outcome	p(2 years)	p(5 years)	p(10 years)	p(20 years)	p(100 years)	p(200 years)
Extinct	3%	7%	15%	22%	39%	52%
Alive	97%	93%	85%	78%	61%	48%

Institute and Faculty of Actuaries

The wisdom of crowds suggests the risk is higher if we develop AGI sooner rather than later

How does the level of existential risk posed by AGI depend on its arrival time?

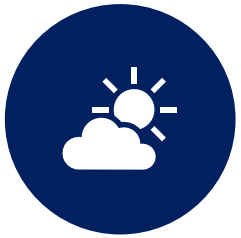


AGI arrival	existential risk
2024	74%
2025-’29	60%
2030-’39	30%
2040-’59	18%
2060-’99	6%

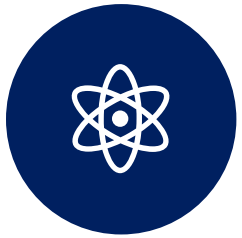
Existential risk from AGI vs AGI timelines | Metaculus

Institute and Faculty of Actuaries

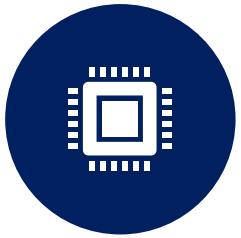
Threat Level Midnight



CLIMATE CHANGE



NUCLEAR WAR



AGI

ASTEROID STRIKE

DISEASE X

Institute and Faculty of Actuaries

Time to pause, or accelerate?

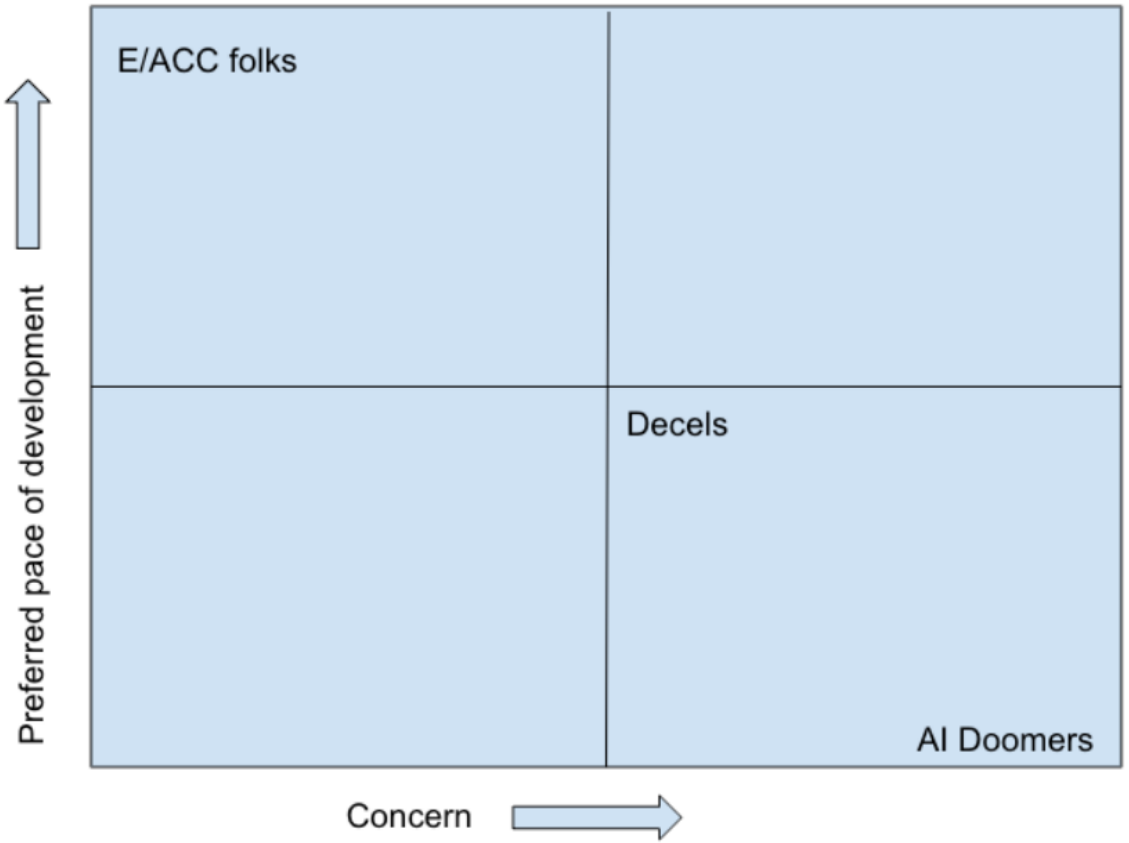


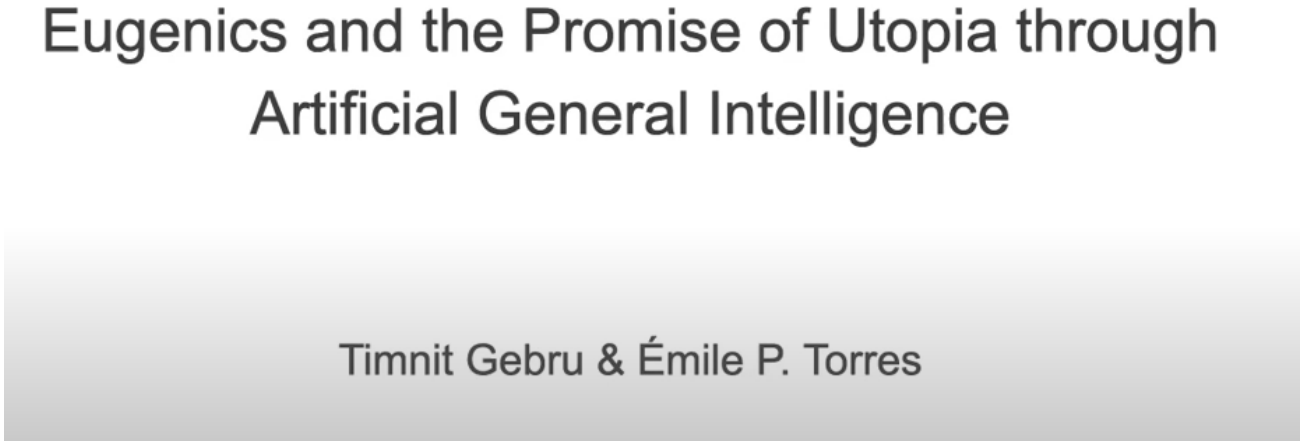
Image Credits: *Alex Wilhelm/TechCrunch*

Institute and Faculty of Actuaries

TESCREAL

- Transhumanism
- Extropianism
- Singularitarianism
- Cosmism
- Rationalism
- Effective Altruism
- Longtermism

AI and the threat of "human extinction": What are the tech-bros worried about? It's not you and me | Salon.com



SaTML 2023 - Timnit Gebru - Eugenics and the Promise of Utopia through AGI (youtube.com)

Understanding TESCREAL — the Weird Ideologies Behind Silicon Valley's Rightward Turn

Institute and Faculty of Actuaries

State of the Art - LLMs

- Transformers (2017)
- ChatGPT (2022)

- Sparks of Artificial General Intelligence – or not:

  Early experiments with GPT-4

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

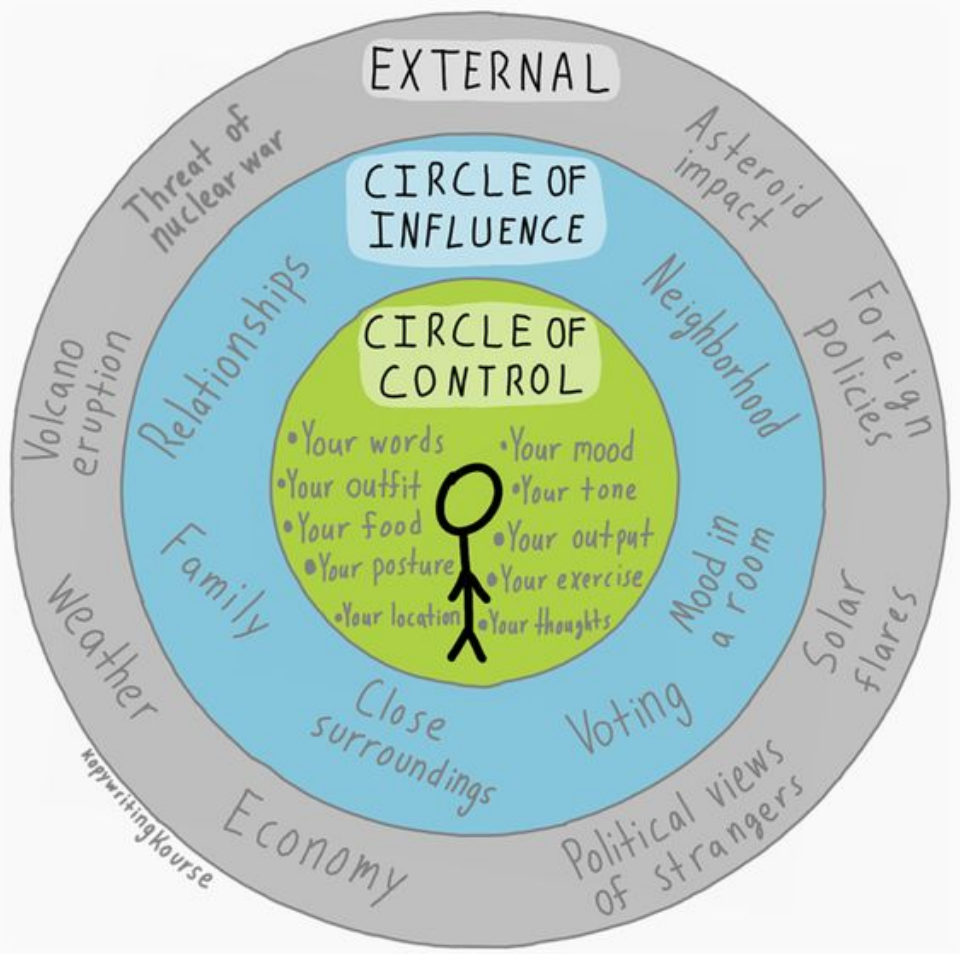
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Institute and Faculty of Actuaries

https://arxiv.org/pdf/1706.03762.pdf

So what can Actuaries do?



- Control
  - My use of AI for good
  - Critically review articles
- Influence
  - IFoA research
  - Government policy?
- Concern
  - Accept vs worry?

Institute and Faculty of Actuaries

Diagram based on Stephen Covey's Seven Habits of Highly Effective people. Pinterest Fair use

Meanwhile focus on shorter term risks of AI

Malicious Use

Bias and Discrimination

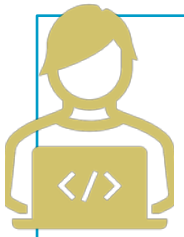
Copyright

Economic Upheaval

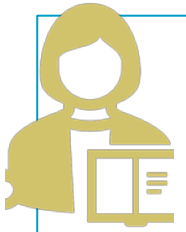
Environmental Impacts

Institute and Faculty of Actuaries

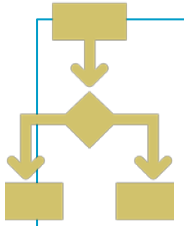
In Conclusion



Use models with responsibility now



Educate and critically review material



Use framework to compare to other risks

Use influence for good

Institute and Faculty of Actuaries

- *“By far, the greatest danger of Artificial Intelligence is that people conclude too early that they understand it.”*
- - Eliezer Yudkowsky

Institute and Faculty of Actuaries

Questions

Comments

Expressions of individual views by members of the Institute and Faculty of Actuaries and its staff are encouraged.

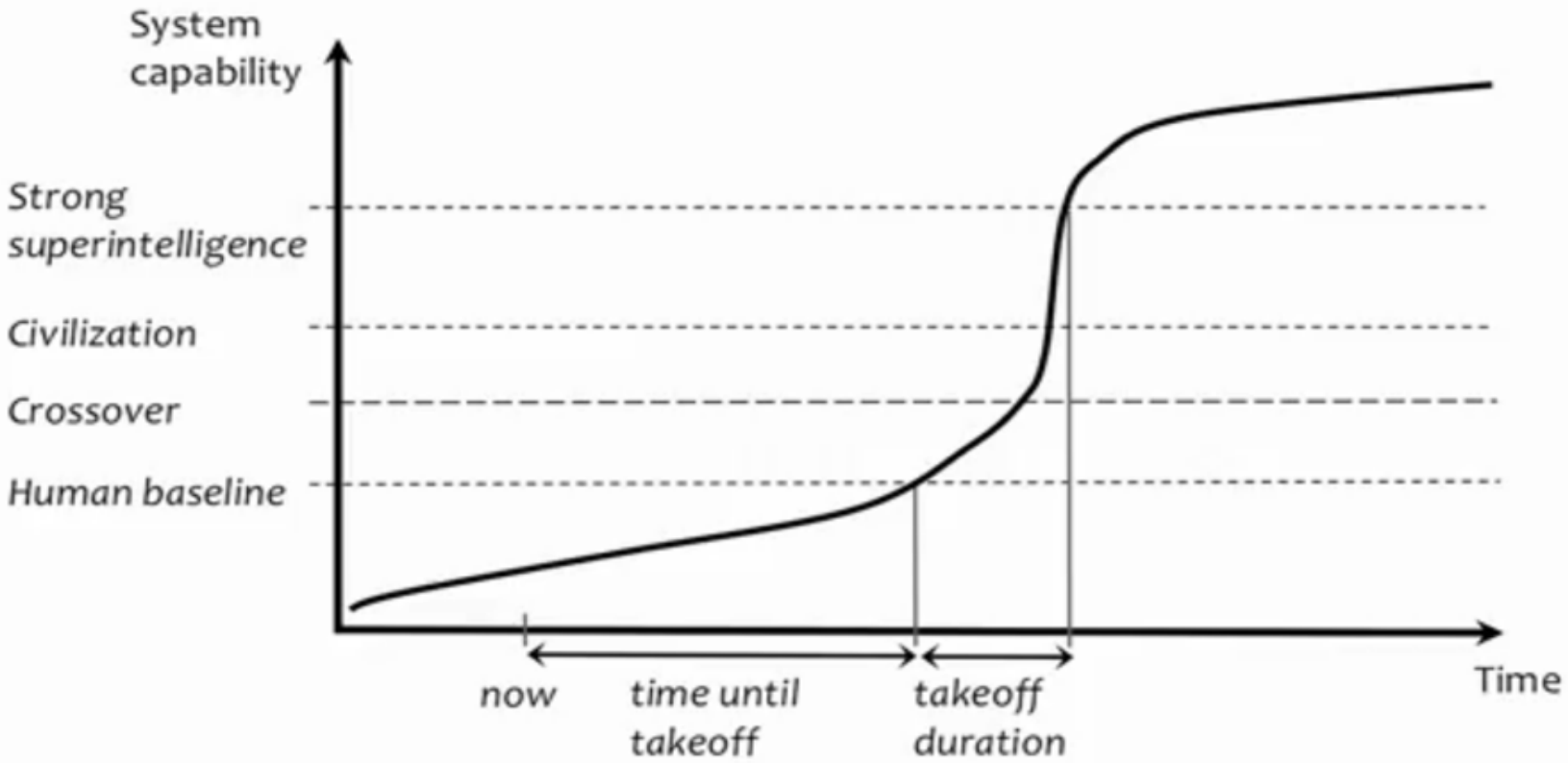
The views expressed in this presentation are those of the presenter and not of our employers or the Institute and Faculty of Actuaries.

Institute and Faculty of Actuaries

Institute and Faculty of Actuaries

Appendix

Illustrating the Singularity



Institute and Faculty of Actuaries

Exponential growth could be a reasonable assumption

- Rate of improvement in intelligence = optimization power/ recalcitrance

- If we assume
  - Optimization power is a linear function of intelligence I (α, β positive)
  - Recalcitrance is constant k

$$\frac{dI}{dt} = \frac{O(I)}{R} = \frac{\alpha I + \beta}{k}$$

- Then I increases exponentially over time

Institute and Faculty of Actuaries

There are alternative viewpoints

- Rate of improvement in intelligence = optimization power/ recalcitrance

- This time we assume
  - Optimization power is a linear function of intelligence I (α_O,β_O positive)
  - Recalcitrance increases with intelligence, as low hanging fruit are picked (α_R,β_R positive)

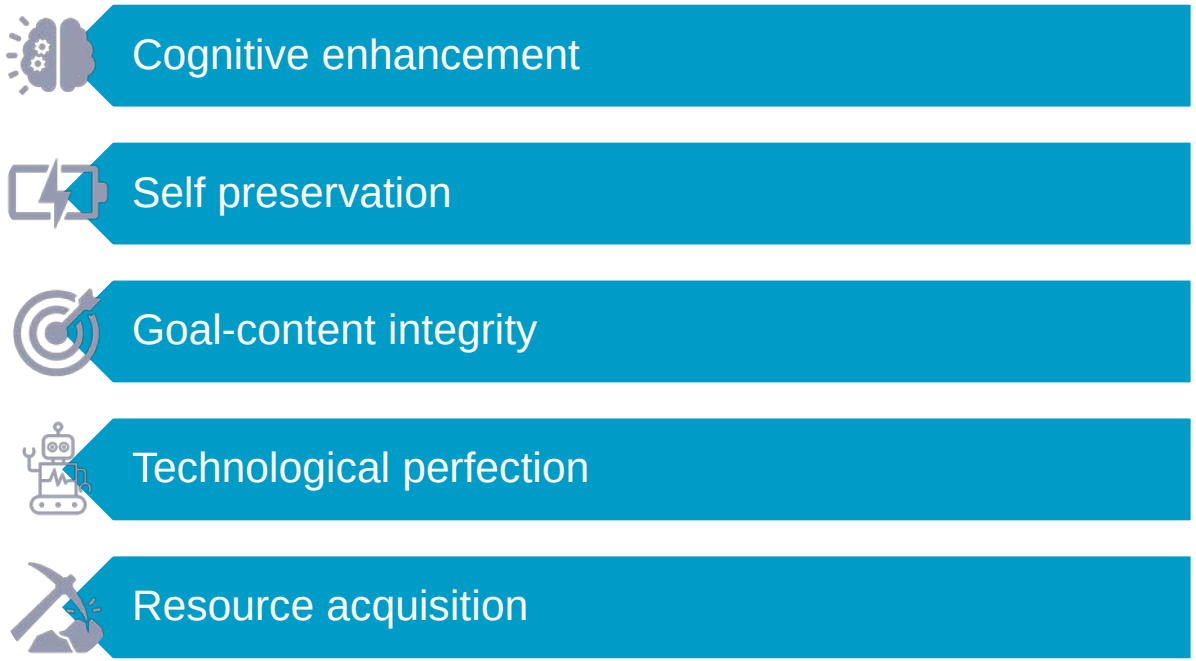
$$\frac{dI}{dt} = \frac{O(I)}{R} = \frac{\alpha_O I + \beta_O}{\alpha_R I + \beta_R}$$

- Over time, as I increases, the rate of intelligence growth approaches $^{\alpha_O}/_{\alpha_R}$ - linear growth
- Benthall also suggests some elements of recalcitrance are effectively infinite

Institute and Faculty of Actuaries

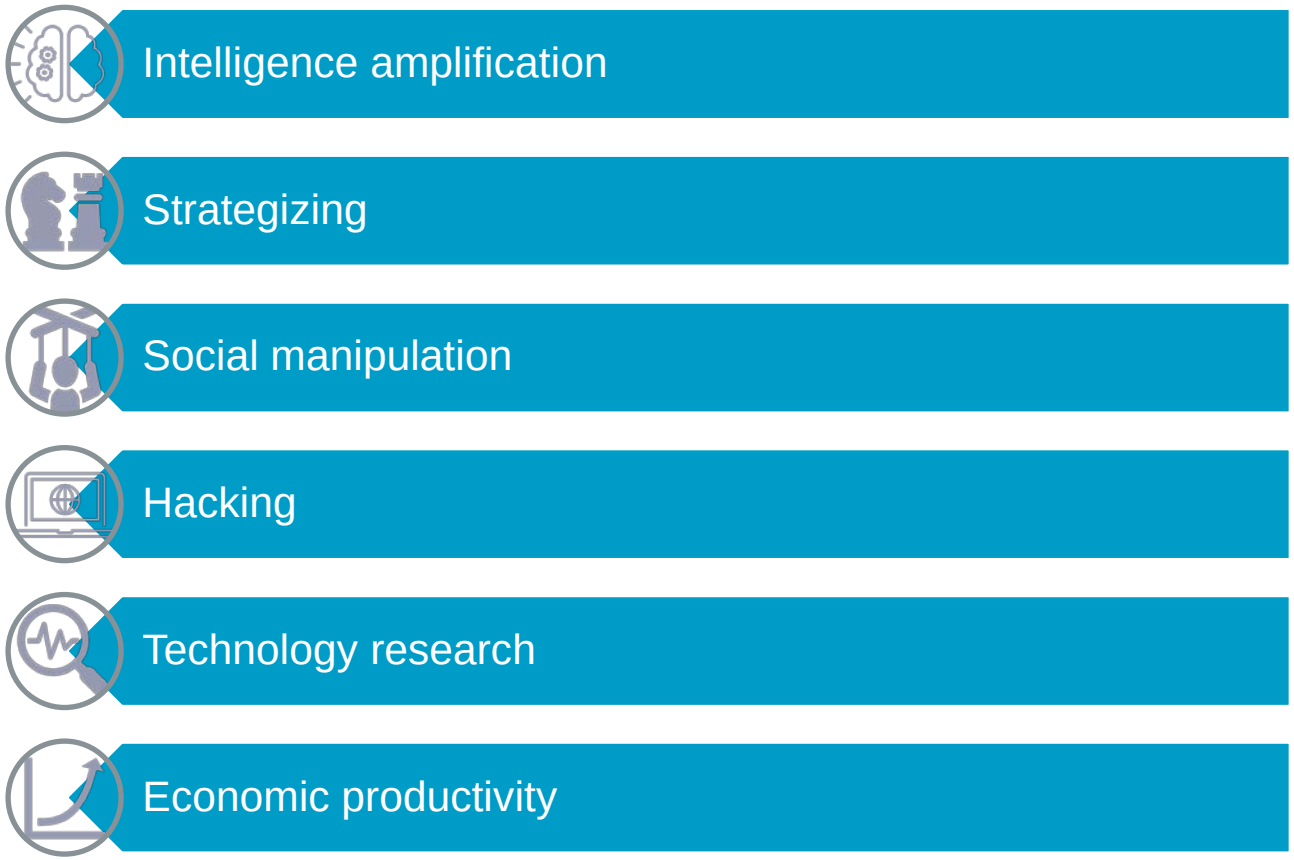
We may need to solve the alignment problem first time

‘Several instrumental values… would increase the chances of the agent’s goal being realized for a wide range of final goals and a wide range of situations’



Institute and Faculty of Actuaries

Key skills an AGI may identify as useful to learn



Institute and Faculty of Actuaries

Worrying outcomes – malignant failures

- ‘only a project that got a great number of things right could succeed in building a machine intelligent enough to pose a risk of malignant failure. When a weak system fails, the fallout is limited. A malignant failure involves an existential catastrophe. It eliminates the opportunity to try again’
  - Perverse instantiation
    - Includes wireheading
  - Infrastructure profusion
  - Mind crimes

Institute and Faculty of Actuaries

Different kinds of AI agent

Oracle	Genie	Sovereign	Tool
• Question answering system	• Command executing agent	• Open ended autonomous operating system	• System not designed to exhibit goal-directed behaviour

Institute and Faculty of Actuaries

Create tools rather than goal-oriented agents?

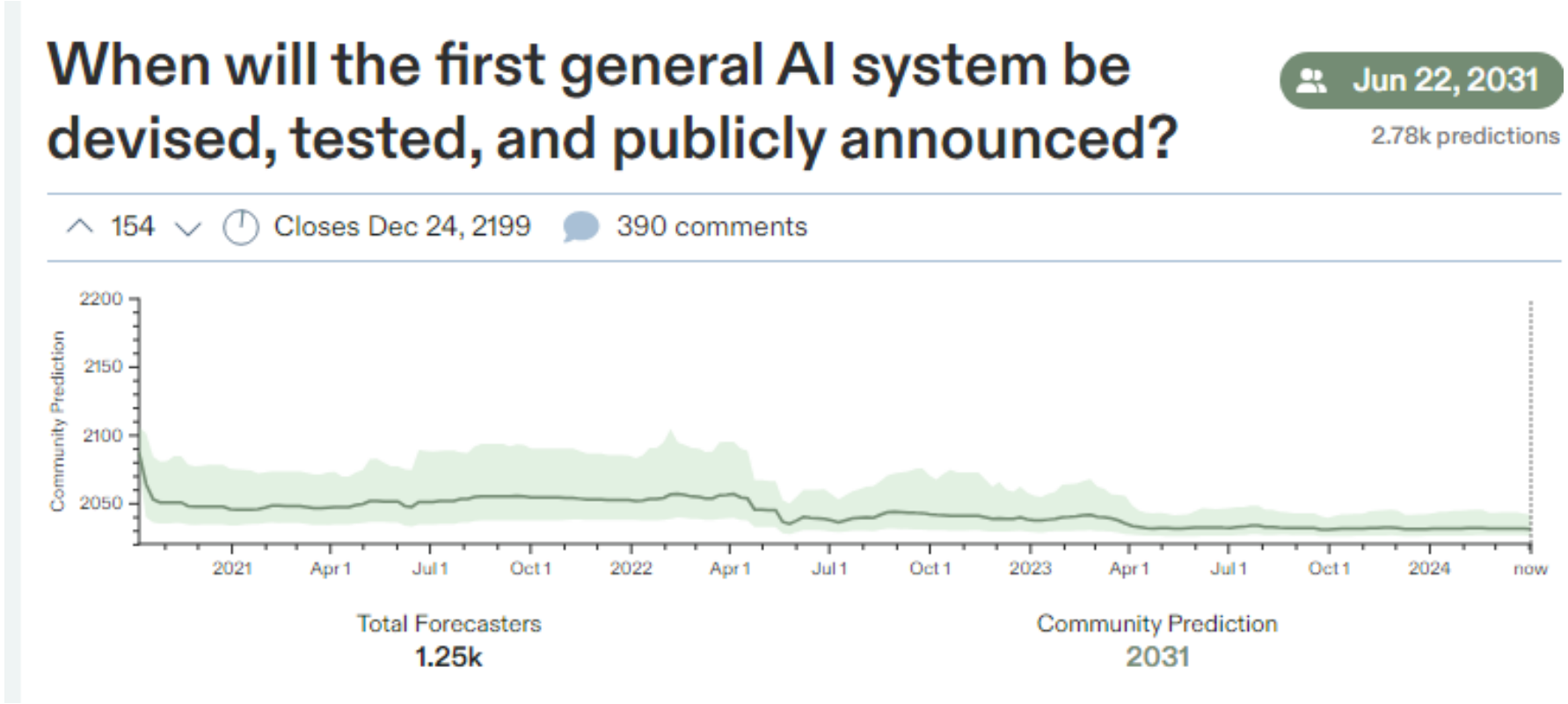
- ‘Instead of creating an AI that has beliefs and desires and that acts like an artificial person, we should aim to build regular software that simply does what it is programmed to do.

- End up with software that can learn on its own to do new tasks, and indeed discover new tasks in need of doing. This would require the software to be able to learn, reason, and plan, and to do so in a powerful and robustly cross domain manner.

- Instead of allowing agent-like purposive behaviour to emerge spontaneously and haphazardly from the implementation of powerful search process… it may be better to create agents on purpose. Endowing a superintelligence with an explicit agent like structure can be a way of increasing predictability and transparency.’

Institute and Faculty of Actuaries

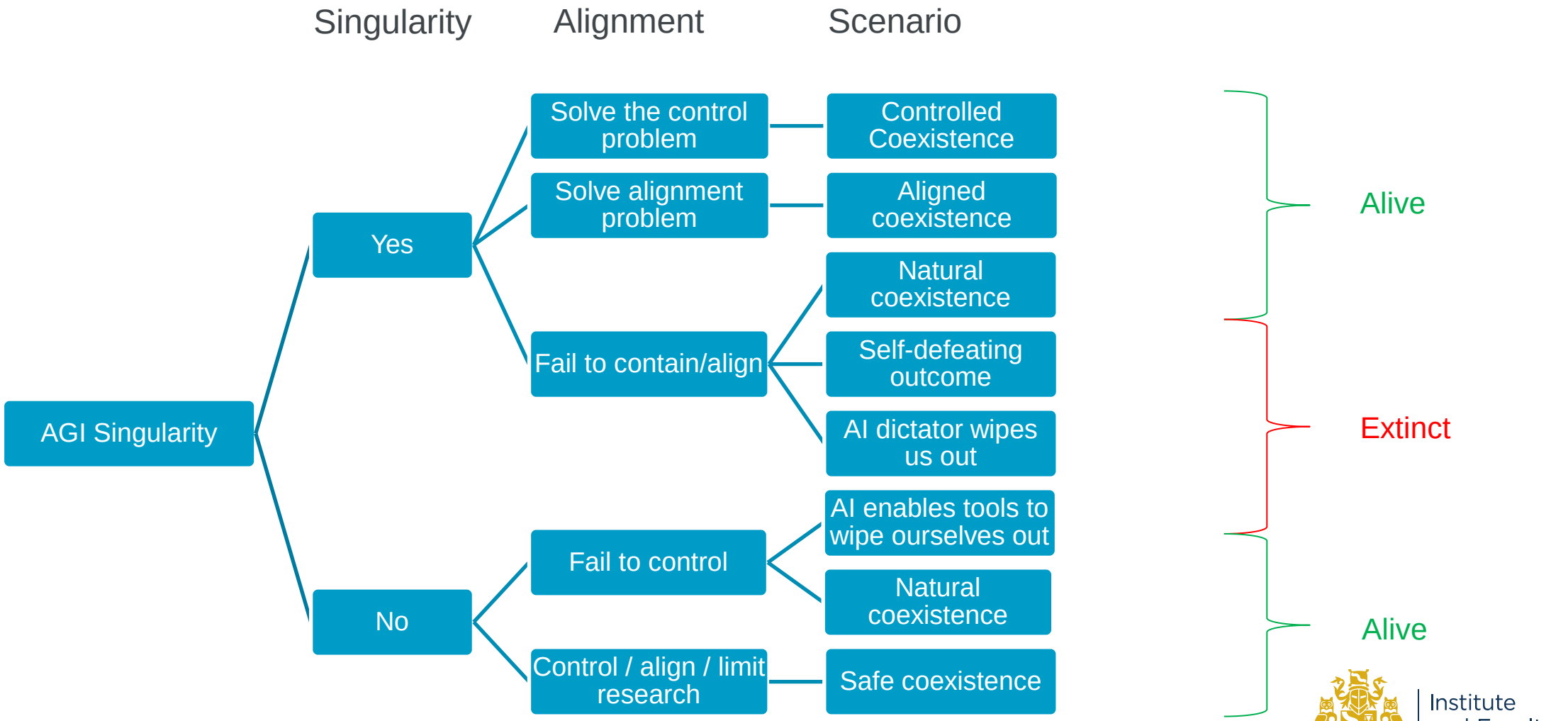
Current view on metacalculus is AGI in early 2030s



Institute and Faculty of Actuaries

- Date of Artificial General Intelligence | Metaculus

Will AI lead to human extinction?



Decision tree framework

Institute and Faculty of Actuaries

Considering the timeline can help us allocate estimates

Probabilities for illustration

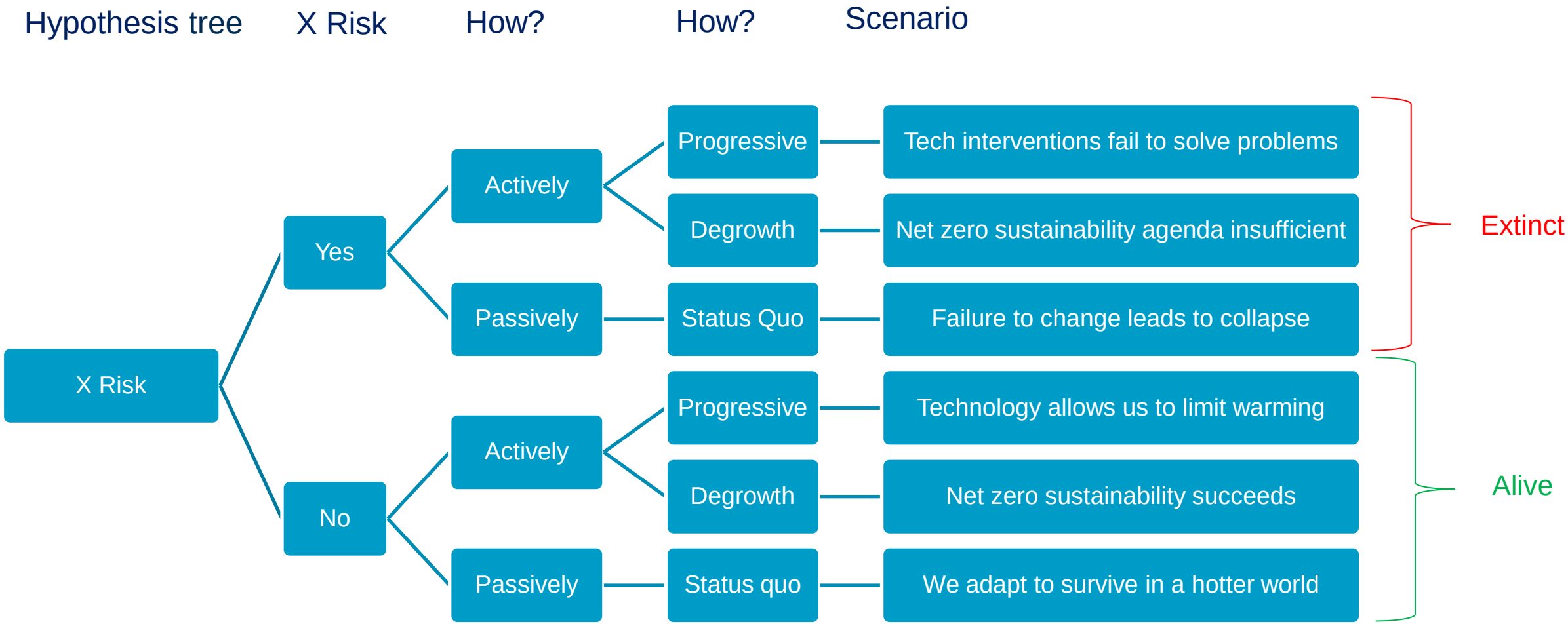
Have we reached AGI singularity?	Do we have control/alignment?	Does it matter?	Example	p(2 years)	p(5 years)	p(10 years)	p(20 years)	p(100 years)	p(200 years)
Yes	Control	Yes	Maintained control/ Solve the control problem	1%	2%	5%	4%	3%	2%
Yes	No	Yes	AI Dictatorship chooses to wipes us out	0%	1%	1%	3%	4%	5%
Yes	Alignment	Yes	Solve the alignment problem	1%	2%	5%	4%	5%	5%
Yes	No	No	Natural coexistence	7%	27%	43%	49%	43%	44%
Yes	No	Yes	We ask for something self defeating	1%	3%	10%	15%	30%	40%
No	Yes	Yes	Limited research avoided issue	1%	2%	5%	4%	3%	2%
No	No	No	Natural coexistence	88%	63%	31%	21%	12%	2%
No	No	Yes	AI tools enable us to wipe ourselves out	1%	1%	1%	1%	1%	1%

Have we reached singularity?	p(2 years)	p(5 years)	p(10 years)	p(20 years)	p(100 years)	p(200 years)
Yes	11%	35%	64%	75%	85%	96%
No	90%	66%	37%	26%	16%	5%

Outcome	p(2 years)	p(5 years)	p(10 years)	p(20 years)	p(100 years)	p(200 years)
Extinct	2%	4%	12%	19%	35%	46%
Alive	98%	96%	89%	82%	66%	55%

Institute and Faculty of Actuaries

Will climate change lead to human extinction?



Institute and Faculty of Actuaries

Climate change timeline and responses

Probabilities for illustration

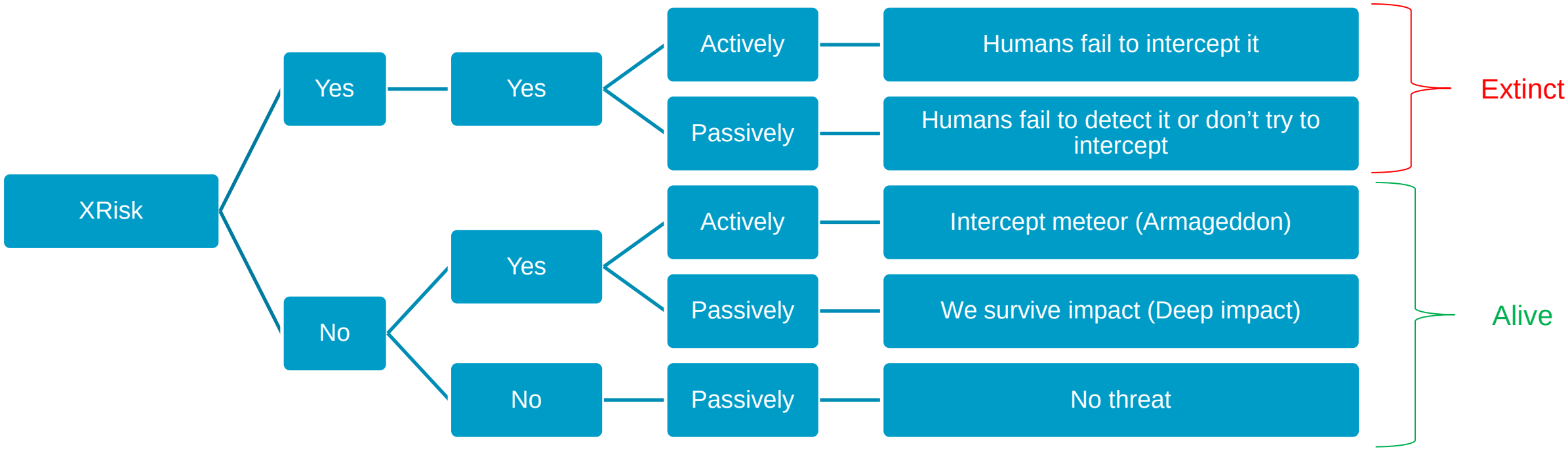
Extinction?	How	How?	Example	p(2 years)	p(5 years)	p(10 years)	p(20 years)	p(100 years)	p(200 years)
Yes	Actively	Progressive	Technology fails to reduce warming	0%	1%	3%	5%	6%	7%
Yes	Actively	Degrowth	Sustainability agenda fails to reduce warming	0%	0%	0%	2%	6%	7%
Yes	Passively	Status Quo	Fail to change sufficiently to avoid problems	0%	0%	0%	2%	6%	7%
No	Actively	Progressive	Technology helps us limit warming	0%	2%	5%	15%	30%	50%
No	Actively	Degrowth	Net zero sustainability succeeds	0%	0%	2%	10%	10%	10%
No	Passively	Status Quo	We survive in a hotter world	100%	97%	90%	66%	42%	19%

Outcome	p(2 years)	p(5 years)	p(10 years)	p(20 years)	p(100 years)	p(200 years)
Extinct	0%	1%	3%	9%	18%	21%
Alive	100%	99%	97%	91%	82%	79%

Institute and Faculty of Actuaries

Will a meteor strike lead to human extinction?

Hypothesis tree X Risk Meteor? Response? Scenario



Institute and Faculty of Actuaries

Meteor strike timelines and responses

Probabilities for illustration

Extinction?	is there a meteor?	Response?	Example	p(2 years)	p(5 years)	p(10 years)	p(20 years)	p(100 years)	p(200 years)
Yes	Yes	Actively	Humans fail to intercept it	0.000%	0.000%	0.000%	0.000%	0.001%	0.002%
Yes	Yes	Passively	Humans fail to detect it, and/or don't try to mitigate it	0.000%	0.000%	0.000%	0.000%	0.005%	0.020%
No	Yes	Actively	Humans intercept meteor (Armageddon)	0.000%	0.000%	0.000%	0.000%	0.001%	0.010%
No	Yes	Passively	Humans survive meteor strike (Deep Impact)	0.000%	0.000%	0.000%	0.000%	0.005%	0.010%
No	No		No Meteor	100.00%	100.00%	100.00%	100.00%	99.99%	99.96%

Outcome	p(2 years)	p(5 years)	p(10 years)	p(20 years)	p(100 years)	p(200 years)
Extinct	0.00%	0.00%	0.00%	0.00%	0.01%	0.02%
Alive	100.00%	100.00%	100.00%	100.00%	99.99%	99.98%

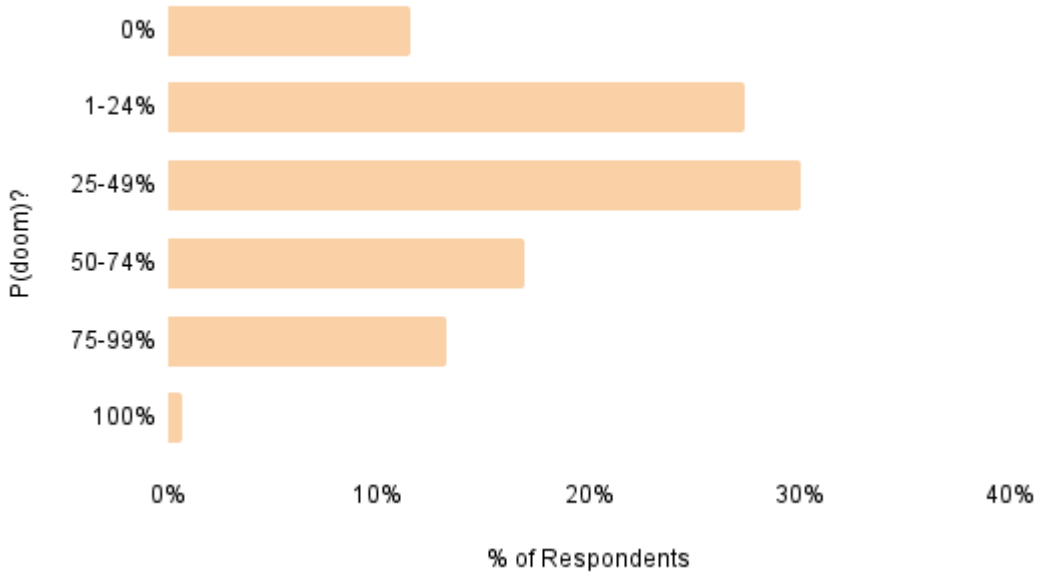
Institute and Faculty of Actuaries

Industry engineers' views on risk vary – but only 12% think there is no chance of doom*

Most common view is the risk of 'doom' is 25-49%.

12% think there is no chance, but more than that think it is over 75%

**Note that we did not define P(doom) or a time horizon in the survey for participants.*



Institute and Faculty of Actuaries

Whilst higher probabilities given to good outcomes, the risk of extremely bad e.g. extinction in the long run is viewed as 9%

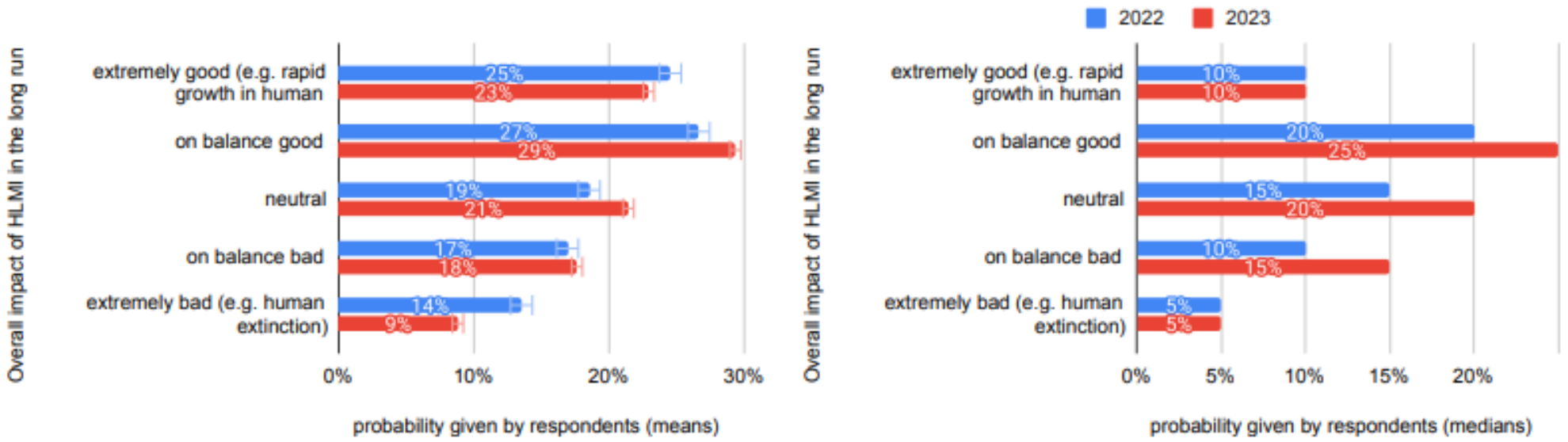


Figure 11: **Mean but not median predictions in 2023 (n=2704) about the consequences of HLMI have shifted slightly away from extreme outcomes compared to 2022 (n=559). Error bars indicate the standard error.**

https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

Institute and Faculty of Actuaries

Probability on human inability to control future advanced AI systems causing human extinction or similar – 2023 mean 19%, median 10%

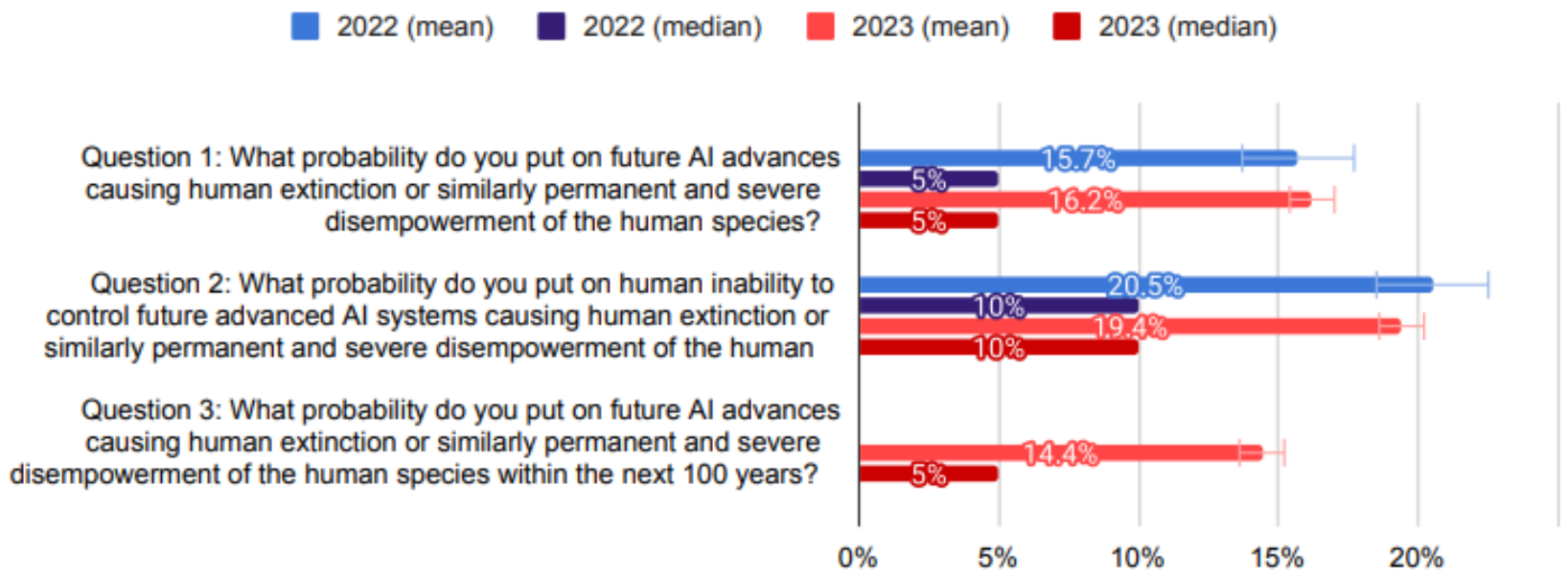
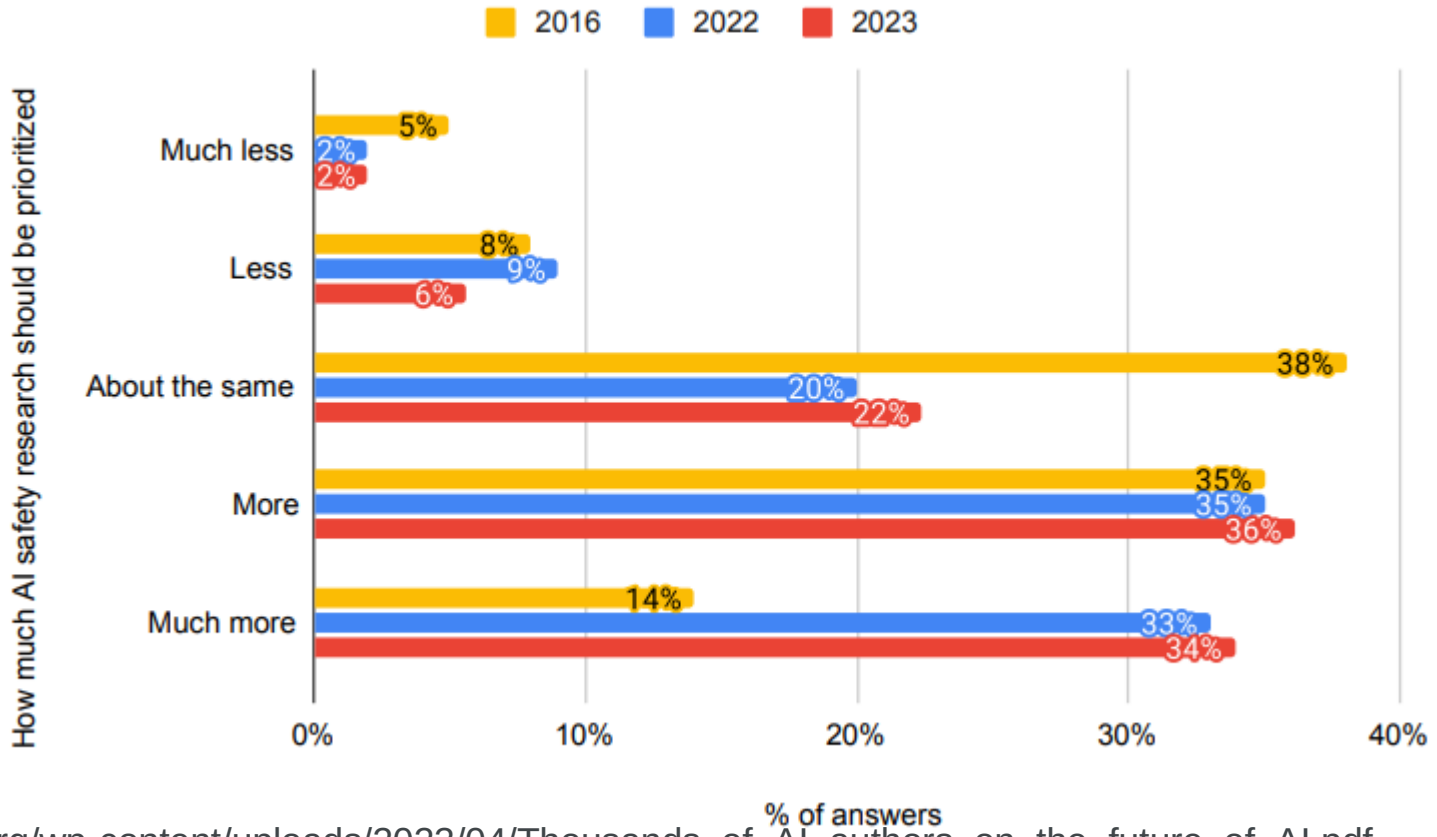


Figure 12: **Mean and median predictions to three questions on human extinction. Error bars indicate the standard error.** (Question 1 n=149 in 2022 and 1321 in 2023. Question 2 n=162 in 2022 and 661 in 2023. Question 3 was asked only in 2023, n=655).

https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

Institute and Faculty of Actuaries

AI safety research is now viewed as higher priority



https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

Institute and Faculty of Actuaries