Artificial General Intelligence

James T. Oswald

Opening Poll

https://strawpoll.com/GJn44B6Mznz

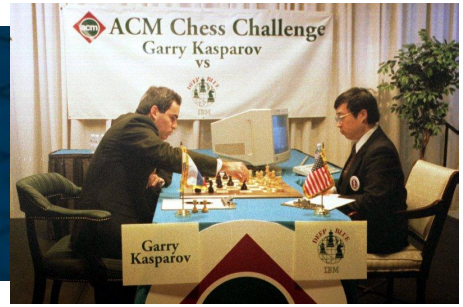
In your opinion, do any modern AI systems count as having artificial general intelligence (AGI)?

Today!

- What is AGI, what is HLAI, what about ASI? How do they relate?
- What is the Singularity?
- Four scaling arguments: the inevitability of AGI?
- History of AGI research.
- What do AGI researchers actually research?

WARNING this lesson will be extremely biased.

What is AGI?



Definition: Narrow AI (NAI) is
"AI that performs *well* on a single task or small collection of tasks"
Ex: AlphaGo, DeepBlue, Alexnet

Definition: Artificial General Intelligence (AGI) is
"AI that performs *well* on a *wide* range of tasks"
Contested Ex: modern language models

Definition: Human Level Artificial Intelligence (HLAI) is
"AI that can perform at a human level on *all human tasks*, if given the same level of human training"
Would be able to perform any job a human could.

Note that there is wide debate on if AGI should instead be defined as HLAI

Definition: Artificial SuperIntelligence (ASI) is
"AI that greatly exceeds human level performance on all tasks"

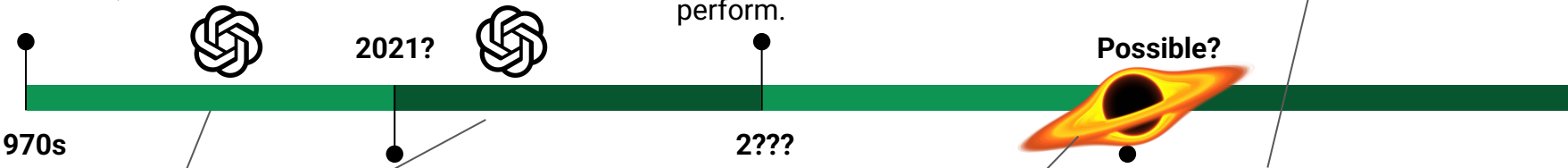
Trivial Theorem: AGI, by definition, subsumes HLAI and ASI

An AGI Spectrum

Narrow AI
AI that performs well on a single task or small collection of tasks. Ex. Most classic machine learning models, most Pre-2000s AI

Human Level Artificial Intelligence

AI that can pass as human in more than just text form. Ability to be embodied. Can replace humans in every job they could perform.



Artificial General Intelligence

"AI that performs *well* on a *wide* range of tasks"

Do LLMs perform *well* enough a *wide* enough range of tasks for you to consider them AGI?

Artificial Super Intelligence

A single AI agent that greatly exceeds the abilities of even the most capable humans.

The Singularity

Do LLMs count as AGI? Some people think yes! (Particularly those with a financial interest in the answer being yes)

Everything beyond this point is AGI

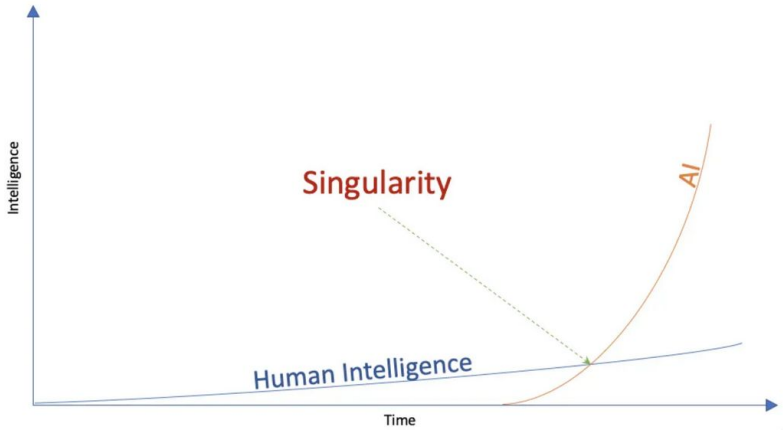

The New York Times
Artificial Intelligence › | Firms Tap Into Boom | Chinese Spending Spree | New E.U. Rules | Water Supply Issues | Chatbot Friendship?

Meta Is Creating a New A.I. Lab to Pursue 'Superintelligence'

The Technological Singularity

The Technological Singularity is a hypothetical future point when artificial intelligence (AI) surpasses human intelligence, leading to rapid, uncontrollable technological growth and a fundamental, irreversible shift in society.

The singularity is typically taken as a prerequisite for ASI.



https://strawpoll.com/B2ZB9oj7AgJ

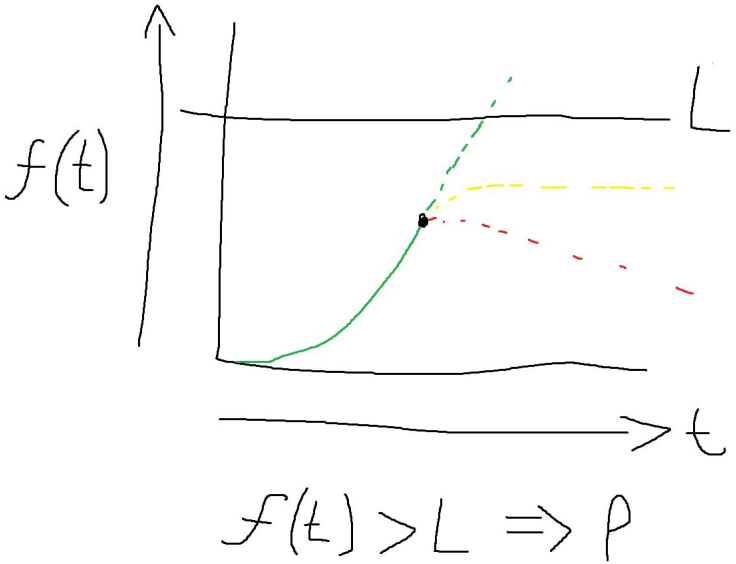
What type of AI do you consider Large Language Models (LLMs) like Chat GPT, Claude, Gemini, etc?

Is Any of This Inevitable?
Scaling Arguments

Inevitability of AGI, HLAI, ASI?

Most arguments for the inevitability of AGI take the form of hypothetical syllogisms about the growth of some variable.

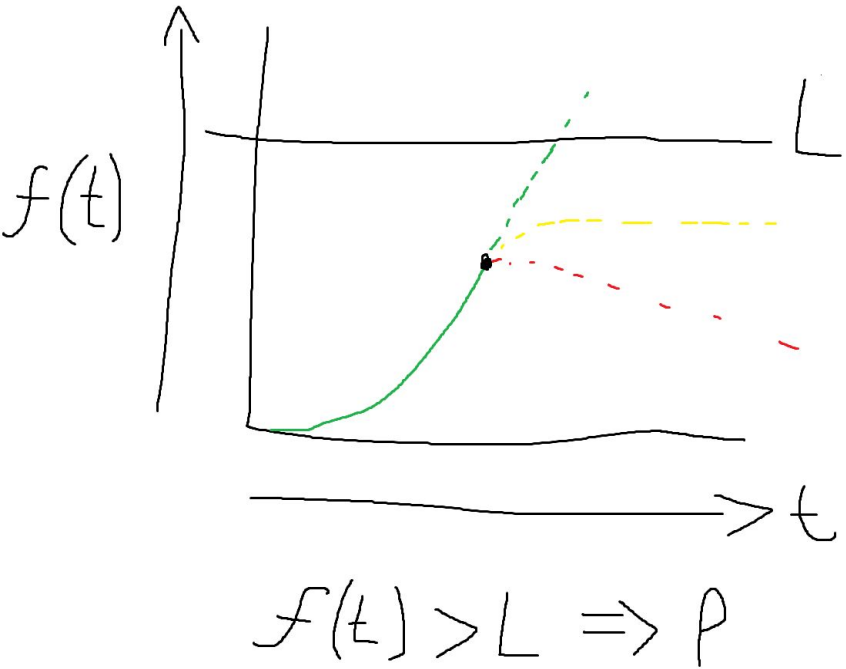
Definition: A *scaling argument* is a statement of the form "If *f(t)* reaches some level *L* then we will have *P*. *f(t)* is growing such that it will inevitably reach level *L*, thus we will have *P*".



WARNING Consider that at any point scaling could stop for any reason. To prove a scaling hypothesis you must prove scaling of x will continue until the level y or indefinitely. Proving scaling will continue is typically impossible (predicting the future is typically taken as impossible). The best you can do is provide an *argument* for why scaling will continue.

Email Me Your Scaling Hypothesis (5-10 min Exercise)

My Email: oswalj@rpi.edu



Include the following items in your email:

1) *f* - thing that seems to be growing over time.
2) *P* - one of AGI, HLAI, ASI
3) *L* - level at which you think f(t) would yield P
4) Argument why *f* at level *L* would yield *P*
5) Argument why *f(t)* will keep increasing until L

Bad Example:

1) *f(t)* = t (years since start of the universe)
2) *P* - HLAI
3) *L* - $10^{10^{\wedge}50}$ years
4) This is the expected time a Boltzmann brain will take to materialize via quantum fluxuations. A Boltzmann brain will have HLAI.
5) Time goes on.

Four of Many Inevitability Arguments for AGI

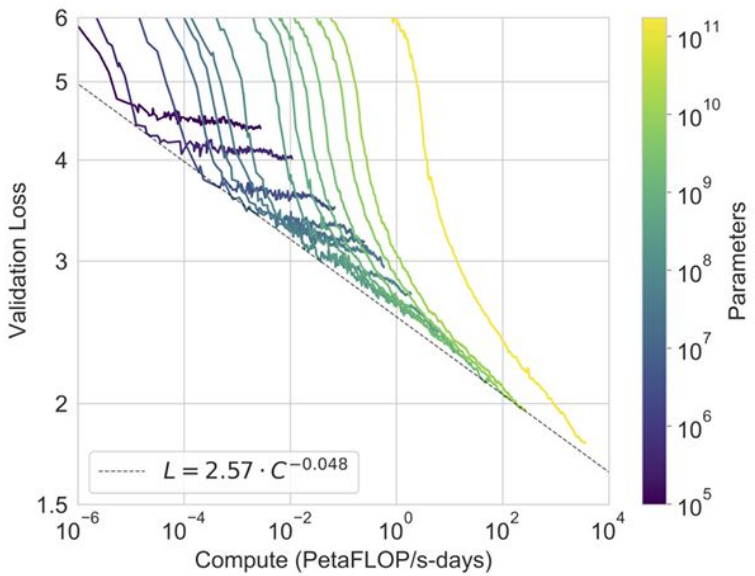
1) Strong Scaling Hypothesis for LMs as AGI
2) Moore's Law Scaling Argument for Intelligent Systems.
3) AI Self-Improvement Scaling Argument for ASI & The Singularity
4) Kurzweil's Technology Based Scaling Hypothesis for ASI & Beyond

Scaling Hypothesis for LMs as AGIs

Roughly, the **strong scaling hypothesis of LLMs for AGI** says that: “The more compute & params we add, the better we score on benchmarks! Eventually we can add so much compute we will have AGI.”

Based on the observation that: The more parameters and data we give LLMs the better they perform on all benchmarks. Lead to the creation of LLMs from LMs, people saw that you could just keep going bigger for more performance.

Limitations: scaling becomes exponentially expensive & lack of new data prevents scaling.



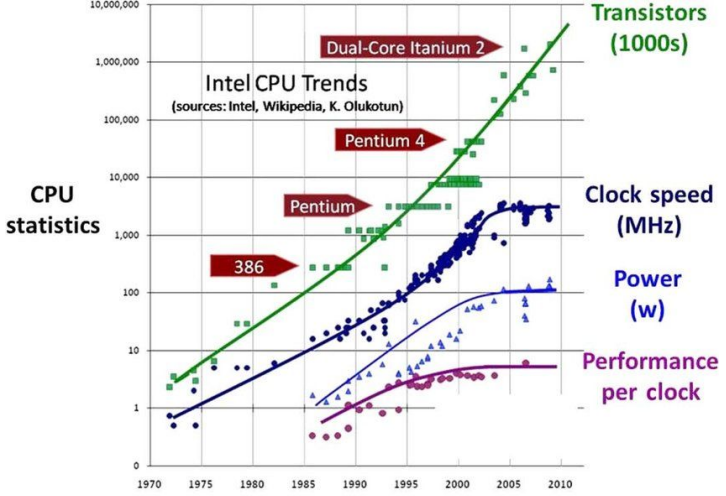
From “Language Models are Few-Shot Learners” Open AI (2020)

Moore's Law Scaling Argument

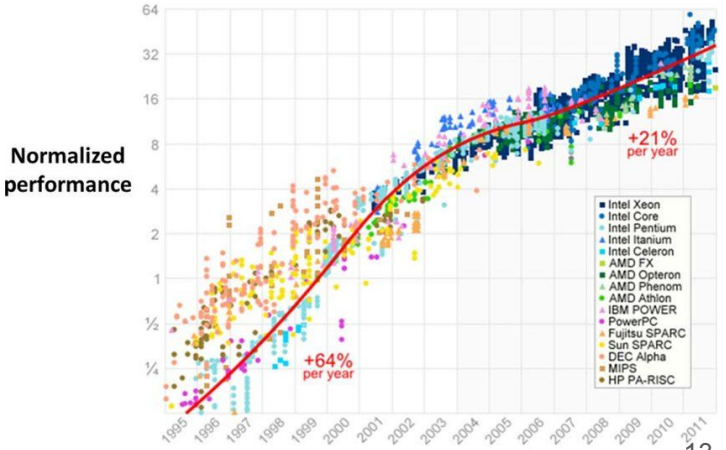
"To simulate an intelligent system we need X transistors (or silicon neuron analogs, etc). By Moore's law we will eventually get to the point where X can be realistically packaged. Therefore we will eventually be able to simulate intelligent systems."

Limitations: Deceleration in Moore's Law, not as fast as it once was. Physical limitations of silicon.

But Consider That new paradigms in computing such as biological, optical, or quantum computing may provide new performance scaling that allows for this.



(a) Transistor, CPU speeds and cooling power (Ref. 26)



(b) CPU performance growth

AI Self Improvement Scaling Argument

$\mathcal{A}$:

Premise 1 There will be AI (created by HI and such that AI = HI).

Premise 2 If there is AI, there will be AI^{+} (created by AI).

Premise 3 If there is AI^{+}, there will be AI^{++} (created by AI^{+}).

$\therefore$ **S** There will be AI^{++} ($= \mathcal{S}$ will occur).

HLAI

The Singularity

Can scale down **Premice 1** to a weaker “There will be an AI that is able to self improve.” This may even be a narrow AI who’s sole task is self improvement towards generality.

Flaws with the AI Self Improvement Scaling Argument

$\mathcal{A}$:

	Premise 1	There will be AI (created by HI and such that AI = HI).
	Premise 2	If there is AI, there will be AI^{+} (created by AI).
	Premise 3	If there is AI^{+}, there will be AI^{++} (created by AI^{+}).
$\therefore$	**S**	There will be AI^{++} ($= \mathcal{S}$ will occur).

Premise 1 assumes we can reach HLAI as a starting point, can we?

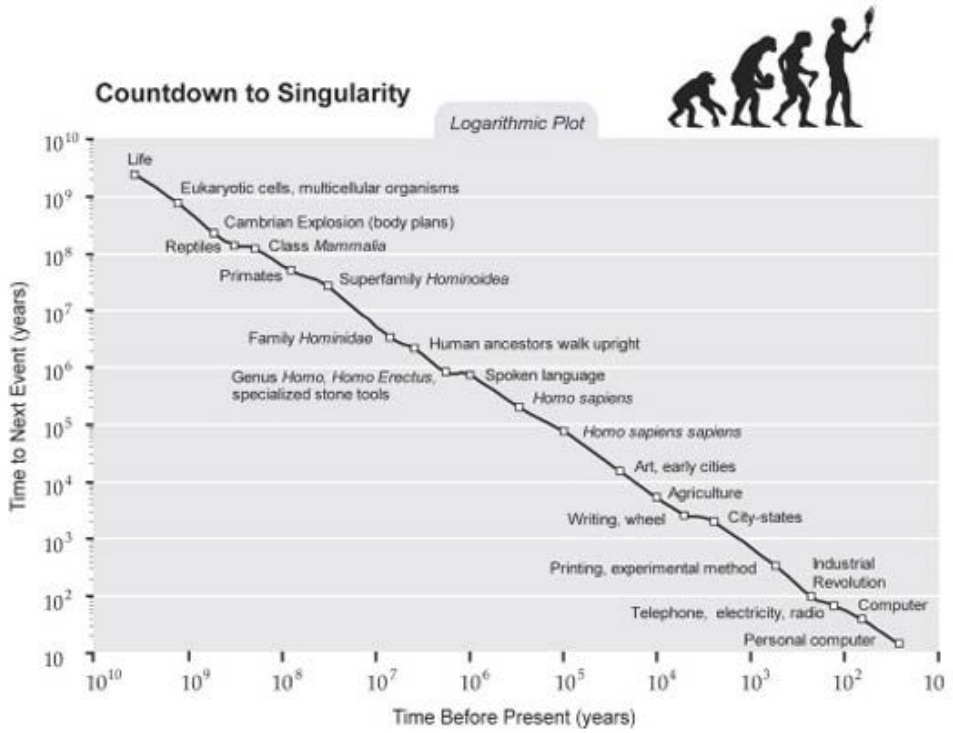
Premise 2 through n assumes AI can actually scale itself (f will increase until L)

Can we actually scale to S?

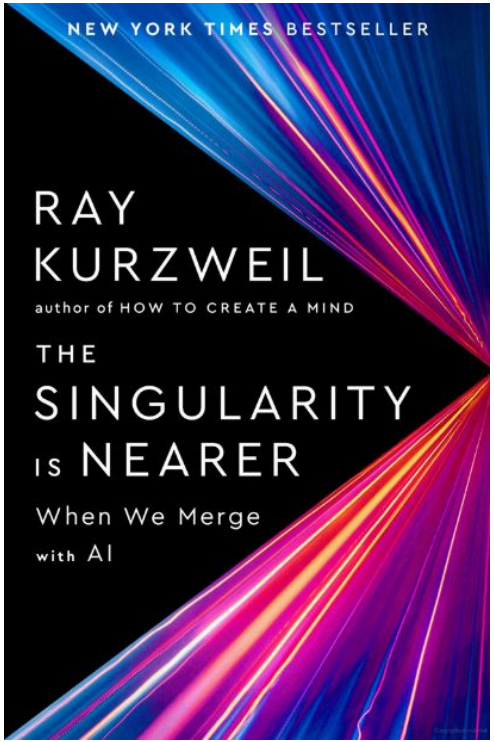
Kurzweil's Technology Scaling Argument

Given sufficient technology, we can do anything physically possible. Minimally HI is possible, we have it, AHI is possible.

Technology itself, including life itself, scales super-exponentially and has for billions of years. Thus we will eventually reach a point where AHI is technologically possible, and probably ASI and the Singularity.



Suggested Reading Material



Two Polls

How Many Years to AGI?

https://strawpoll.com/1MnwkbNBMn7

How Many Years to the Singularity?

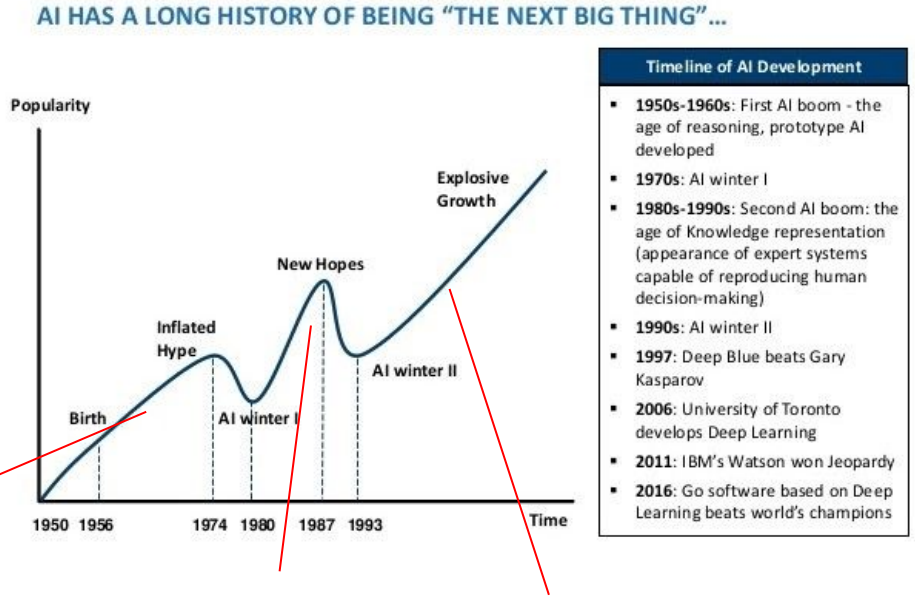
https://strawpoll.com/Q0Zp7pNWAgM

Modern AGI Research

A History of AGI Research Timeline

Almost every AI researcher before the second AI winter had AGI as a goal. “I’m working on AGI… but first I'll show how good my approach is by using it to solve a narrow problem.”

ML based approaches which work very well on narrow problems & largely took AGI off the table as a goal.



First wave of logic based AI, AGI seen to be “only about 20 years away”

Second wave of logic based AI via KRR approaches (Japanese 5th Generation Project, CYC)

ML approaches possible on new hardware offer never before seen performance on narrow tasks

Modern AGI Research

The modern AGI research community formed around 2005 to revive the original goal of AI, build agents that perform well on a wide range of tasks instead of just one.

Modern AGI Research Consists of:

- Defining AGI & Intelligence
- Theoretical analysis of AI Alignment and Safety Concerns with AGI.
- Investigating Pathways to AGI & Integrating & Generalizing narrow methods
- Creation and evaluation of AGI agents
- Proposing Tests of AGI

Core AGI Research Area: Formalizing Intelligence

Intelligence is ability to perform in all environments

$$\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_{\mu}^{\pi}.$$

Intelligence is skill acquisition efficiency

Intelligence of system IS over $scope$ (optimal case):

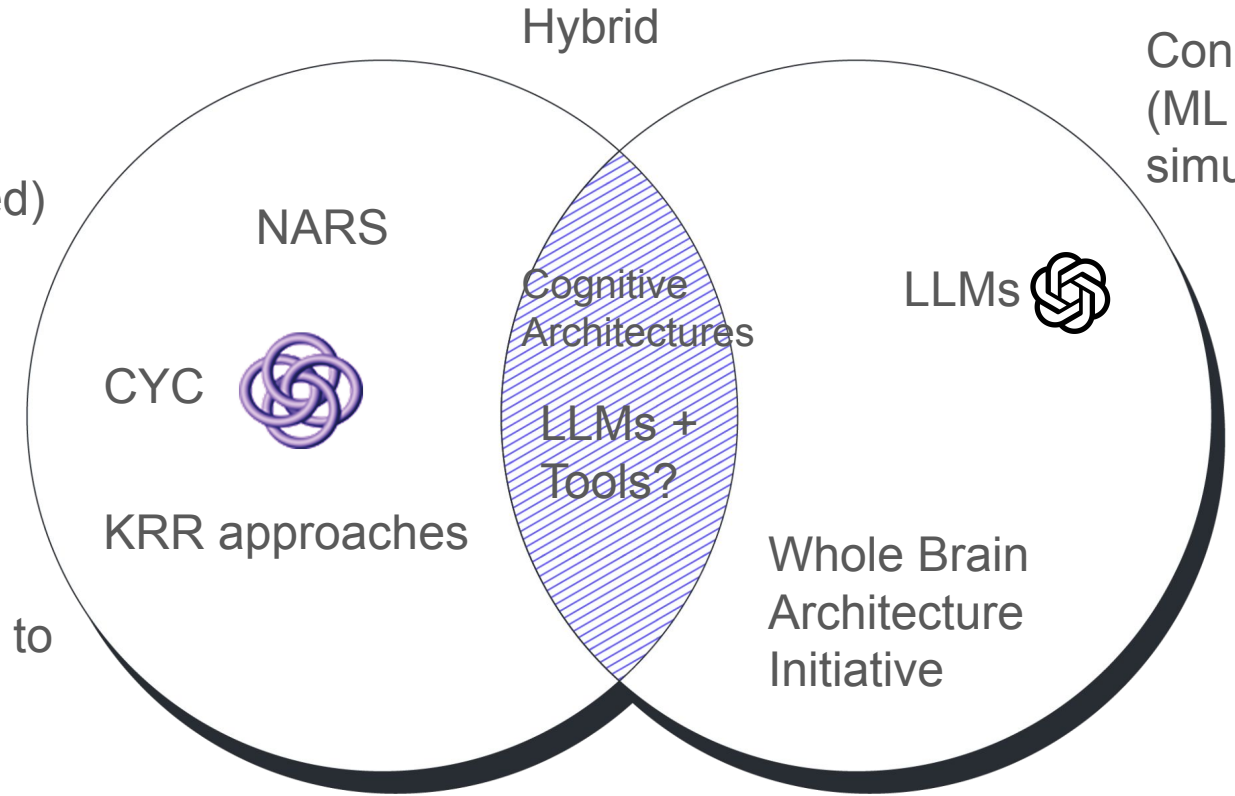
$$I^{opt}_{IS,scope} = \underset{T \in scope}{Avg} \left[\omega_{T,\Theta} \cdot \Theta \sum_{C \in Cur^{opt}_{T}} \left[P_C \cdot \frac{GD^{\Theta}_{IS,T,C}}{P^{\Theta}_{IS,T} + E^{\Theta}_{IS,T,C}} \right] \right]$$

Intelligence is (internal) cognitive representation & reasoning capacity*

$$\Lambda(a,t)_{i,j} = \max_{\phi} \left\{ \mu^i(\phi) \mid \phi \in \Delta\Big(\omega_j\big[o(a,t)\big], \omega_j\big[i(a,t)\big] \Big) \right\}$$

*My take on Selmer's measure, not his

Core AGI Research Area: Pathways to AGI



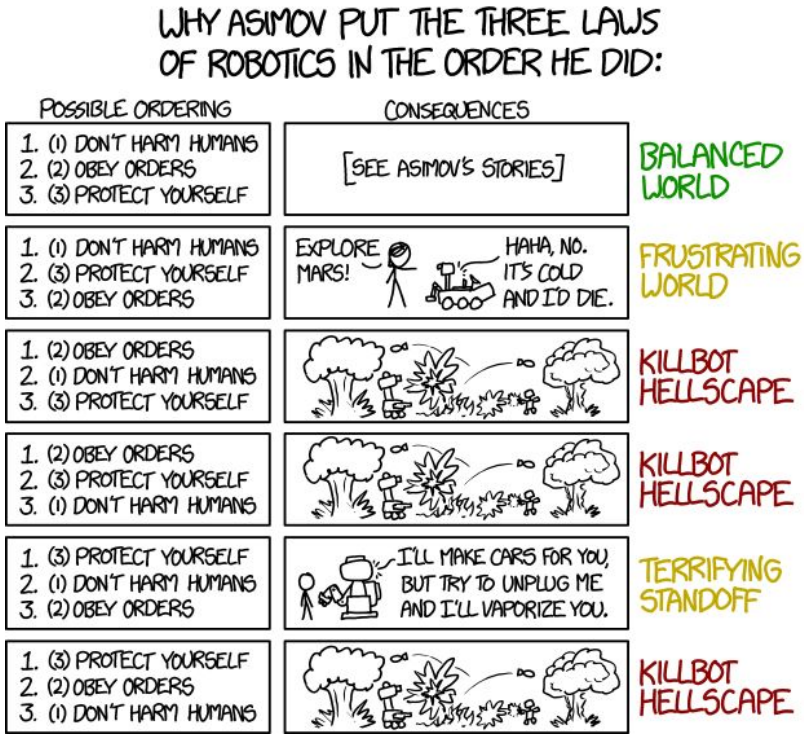
*More classes to come on logic based AGI approaches

Core AGI Research Area: Alignment

Ensuring AGI systems align with human priorities and don't get any ideas….

Logic Based AI systems have a leg up here!

- All reasoning and thought processes can be explained & inspected.
- Can formally prove that no reasoning process terminates in undesirable situations, or prove that if it does, it is never the fault of the agent.



Core AGI Research Area:
Engineering and Evaluation of AGI systems

Core AGI Research Area: Tests of AGI

The Robot College Student Test (*Goertzel*)

A machine enrolls in a university, taking and passing the same classes that humans would, and obtaining a degree. LLMs can now pass university degree-level exams without even attending the classes.[37]

The Employment Test (*Nilsson*)

A machine performs an economically important job at least as well as humans in the same job. AIs are now replacing humans in many roles as varied as fast food and marketing.[38]

The Ikea test (*Marcus*)

Also known as the Flat Pack Furniture Test. An AI views the parts and instructions of an Ikea flat-pack product, then controls a robot to assemble the furniture correctly.[39]

The Coffee Test (*Wozniak*)

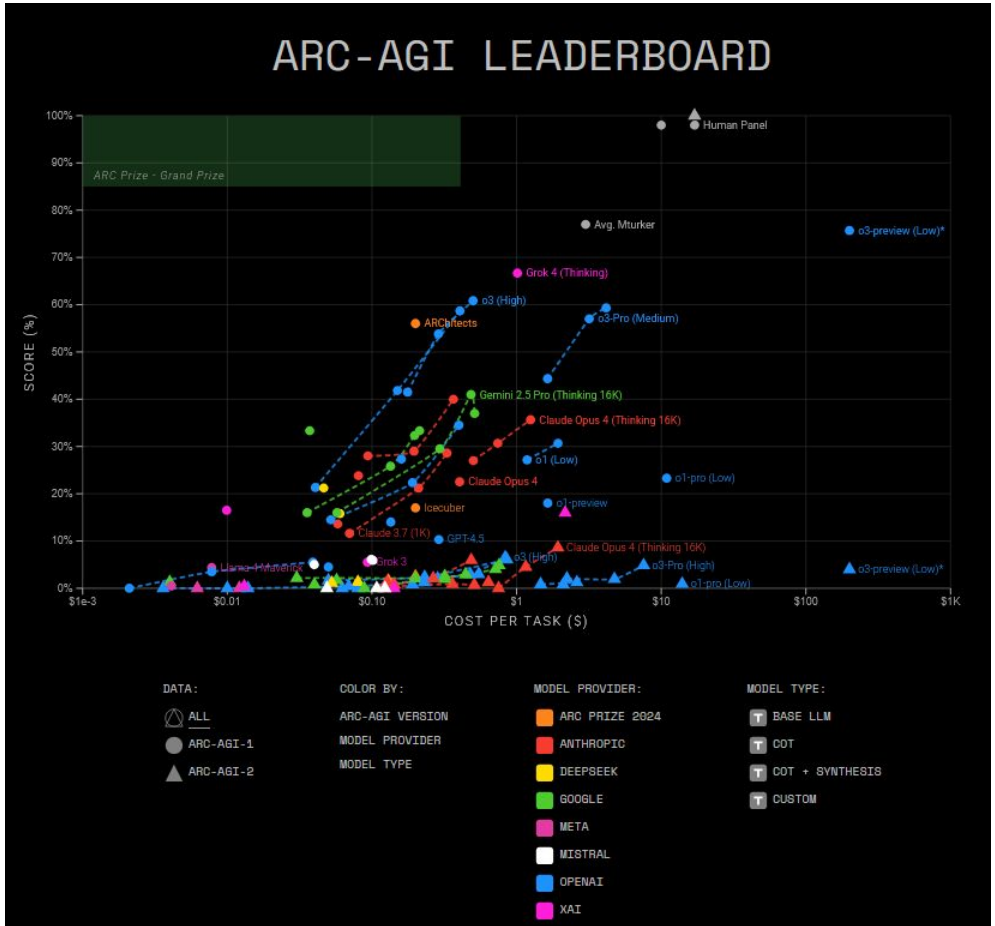
A machine is required to enter an average American home and figure out how to make coffee: find the coffee machine, find the coffee, add water, find a mug, and brew the coffee by pushing the proper buttons.[40] This has not yet been completed.

The Modern Turing Test (*Suleyman*)

An AI model is given $100,000 and has to obtain $1 million.[41][42]

A fun list from Wikipedia

Chollet's ARC-AGI Challenge



- https://arcprize.org/
- $700,000 Prize

Questions?