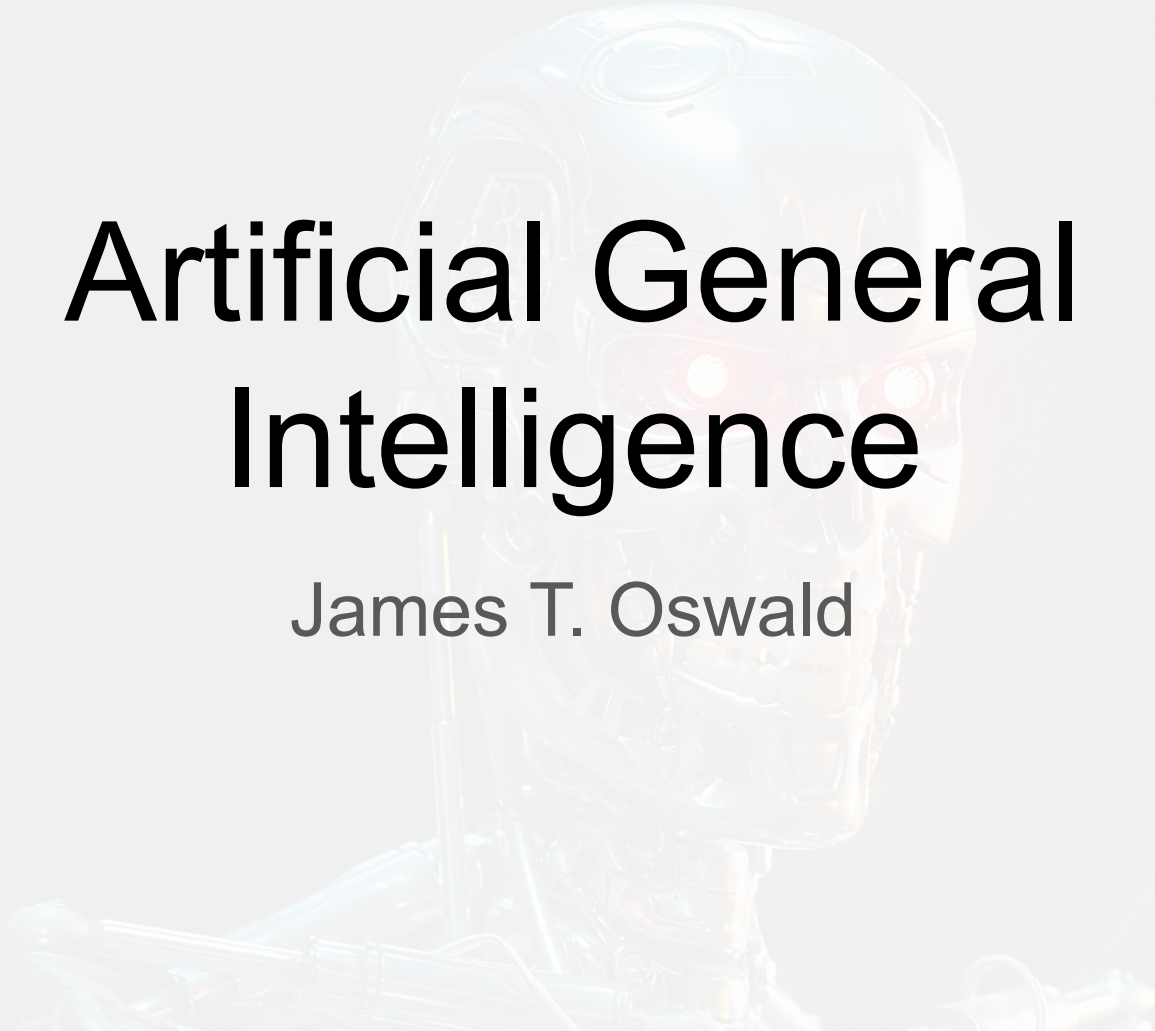


Artificial General Intelligence

James T. Oswald



Today!

- AGI and RPI?
- What is AGI, What is HLAI what about ASI? How do they relate?
- Four Scaling Laws: the inevitability of AGI?
- History of AGI research
- What do AGI researchers actually research?
- Pathways to AGI?

WARNING this lesson will be extremely biased.

Link to this slide deck: <https://bit.ly/RPIAGI>

RPI and AGI?

Is James actually qualified to talk about AGI?

Prof. Ferguson, Prof. Bringsjord, & I just won the Spriner Prize for best paper at the leading conference on AGI this year.

The *Rensselaer AI and Reasoning Laboratory* is one of few groups in the world with a serious focus on AGI research.

(Take this with a grain of salt obviously)

I chose to come to RPI specifically to work on AGI research with Prof. Bringsjord, it certainly has AGI clout!

AGI-24

A Universal Intelligence Measure for Arithmetical Uncomputable Environments

James T. Oswald^{1,2}, Thomas M. Ferguson¹, and Selmer Bringsjord^{1,2}

¹Rensselaer Polytechnic Institute, Troy NY, USA

²Rensselaer AI and Reasoning Laboratory

oswalj@rpi.edu, tferguson@gradcenter.cuny.edu,
Selmer.Bringsjord@gmail.com

Abstract. We propose an extension to Legg and Hutter's universal intelligence (UI) measure to capture the intelligence of agents that operate in uncomputable environments that can be classified on the Arithmetical Hierarchy. Our measure is based on computable environments relativized to a (potentially uncomputable) oracle. We motivate our metric as a natural extension to UI that expands the class of environments evaluated with a trade-off of further uncomputability. Our metric is able to capture intelligence of agents in uncomputable environments we care about, such as first-order theorem proving, and also lends itself to providing a notion of intelligence of oracles. We end by proving some properties of the new measure, such as convergence (given certain assumptions about the complexity of uncomputable environments).

Keywords: Universal Intelligence · Arithmetical Hierarchy · Uncomputability

Introduction

Legg and Hutter's (L&H) universal intelligence (UI) measure [7] is to date one of the most well-formalized and deeply researched theoretical measures of general intelligence. L&H claim this measure captures their working definition of intelligence, that is, intelligence as "an agent's ability to achieve goals in a wide range of environments." While we are in broad agreement with their definition of intelligence, we find the formalization of UI currently lacks the ability to measure intelligence of agents over environments we find important. In fact, this is by design: the formal definition of UI does not capture the space of all environments; it captures only a countable fragment of an uncountably large space of



What is AGI?

Definition: *Narrow AI (NAI)* is

“AI that performs *well* on a single task or small collection of tasks”

Definition: *Artificial General Intelligence (AGI)* is

“AI that performs *well* on a wide range of tasks”

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.

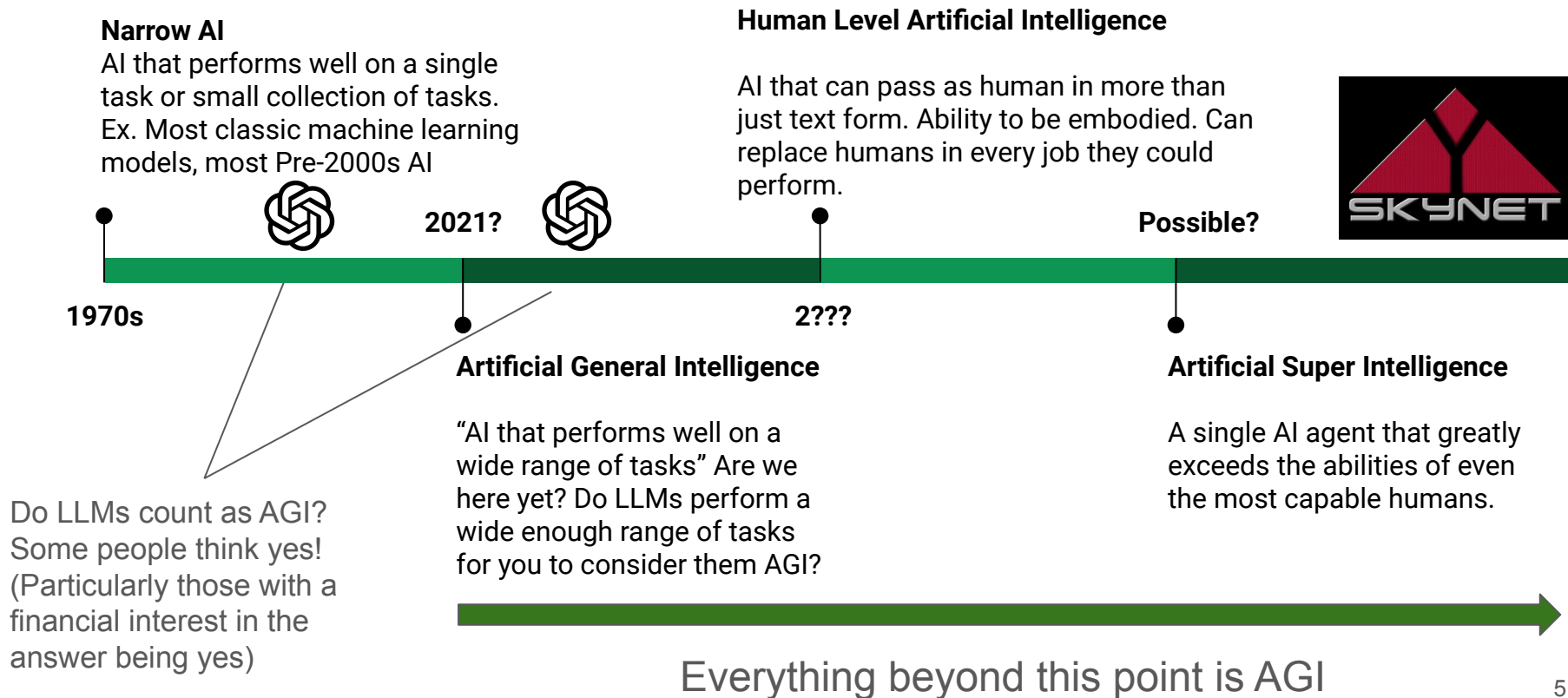
Definition: *Artificial SuperIntelligence (ASI)* is

“AI that greatly exceeds human level performance on all tasks”

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

*All of these definitions are hotly contested in the literature: these are my own working definitions based off a weak general consensus.

AGI Spectrum: My Conception



Inevitability of AGI : Scaling Hypotheses

Most Arguments For AGI take the form of *scaling hypotheses*.

Definition: A *scaling hypothesis* is a statement of the form “If x reaches some level y will have P . x is growing such that it will inevitably reach level y , therefore we will have P in the future”.

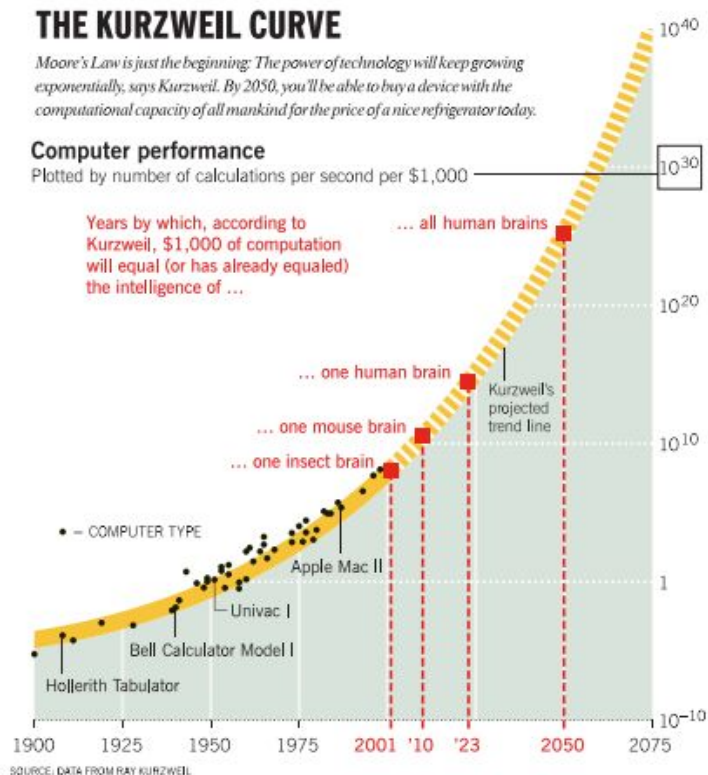
Example: Silly Scaling Hypothesis

If the water reaches the top of the 1L container, the container will be full. Water is filling the container at a rate of 0.1L/m, therefore we will inevitably have the container be full.

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.

Four of Many Scaling Hypotheses for AGI

- 1) Scaling Hypothesis for LMs as AGI
- 2) Moore's Law Scaling Hypothesis for Brain Simulation based AGI
- 3) AI Self-Improvement Scaling Hypothesis for ASI & The Singularity
- 4) Kurzweil's Technology Based Scaling Hypothesis for ASI & Beyond

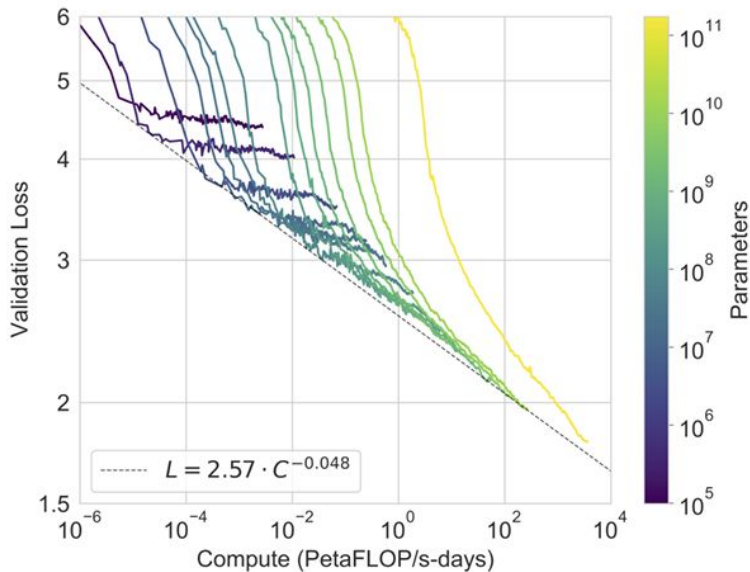


Scaling Hypothesis for LMs as AGIs

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.



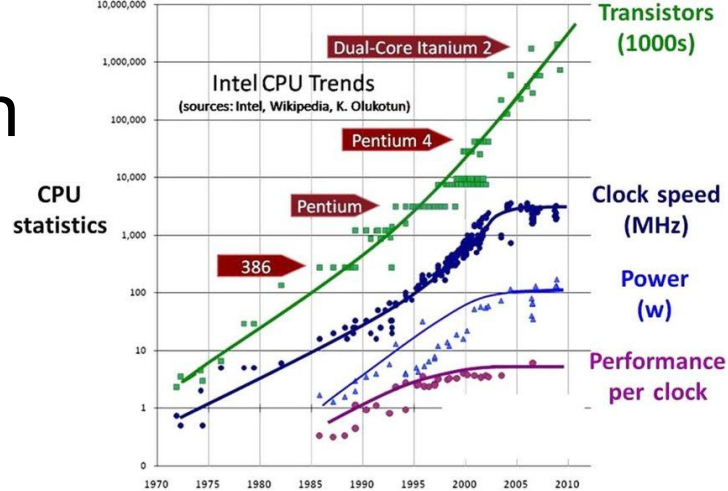
From “Language Models are Few-Shot Learners”

Moore's Law for Brain Simulation

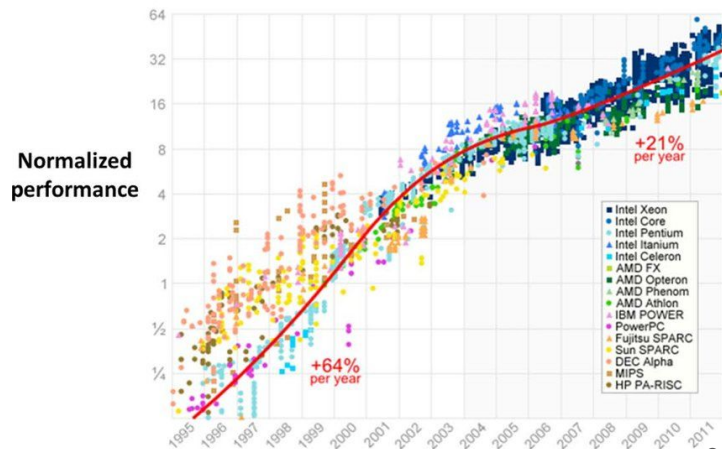
“To simulate a human brain we need X transistors (or silicon neuron analogs, etc). By Moore's law we will eventually get to the point where X can be realistically packaged. Therefore we will eventually be able to simulate brains”

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 Hypothesis

\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).

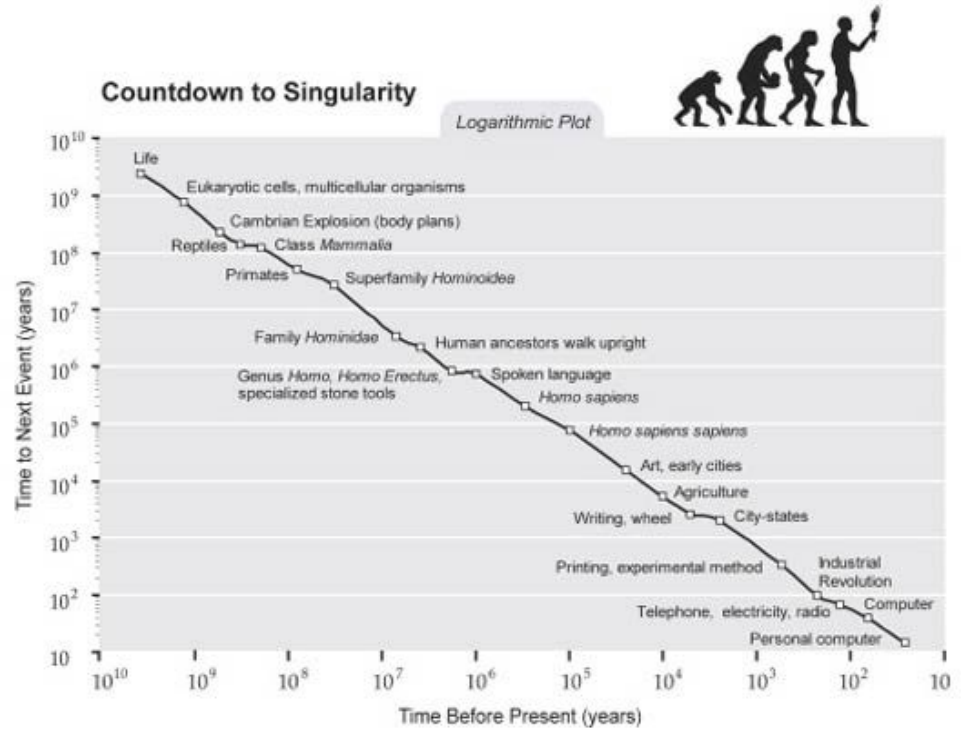
Can scale down **Premise 1** to a weaker “There will be an AI that is able to self improve” This may even be a narrow AI who’s sole task is self improvement towards generality.

Limitations: Assumes the existence of a self improving AI who has resources to self improve.

An Argument from Kurzweil's Scaling Hypothesis

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

Technology itself, including life itself, scales 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.



Modern AGI Research

A History of AGI Research Timeline

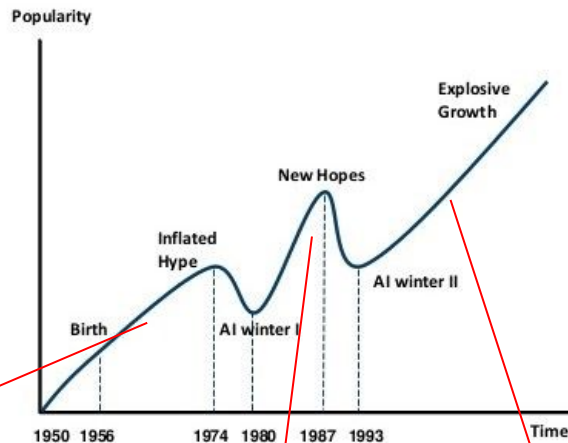
Almost every AI researcher before the second AI winter had AGI as a goal.

“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”

AI HAS A LONG HISTORY OF BEING “THE NEXT BIG THING”...



Timeline of AI Development

- **1950s-1960s:** First AI boom - the age of reasoning, prototype AI developed
- **1970s:** AI winter I
- **1980s-1990s:** Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
- **1990s:** AI winter II
- **1997:** Deep Blue beats Gary Kasparov
- **2006:** University of Toronto develops Deep Learning
- **2011:** IBM's Watson won Jeopardy
- **2016:** Go software based on Deep Learning beats world's champions

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 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):

$$I_{IS,scope}^{opt} = Avg_{T \in scope} \left[\omega_{T,\Theta} \cdot \Theta \sum_{C \in Cur_T^{opt}} \left[P_C \cdot \frac{GD_{IS,T,C}^{\Theta}}{P_{IS,T}^{\Theta} + E_{IS,T,C}^{\Theta}} \right] \right]$$



Intelligence is
representation
& reasoning capacity*

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

(sum or f over i, j)

*My take on Selmer's measure, not his




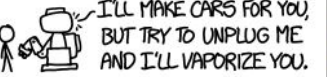

Core AGI Research Area: Alignment

Ensuring AGI systems align with human priorities and don't kill us or get any ideas....

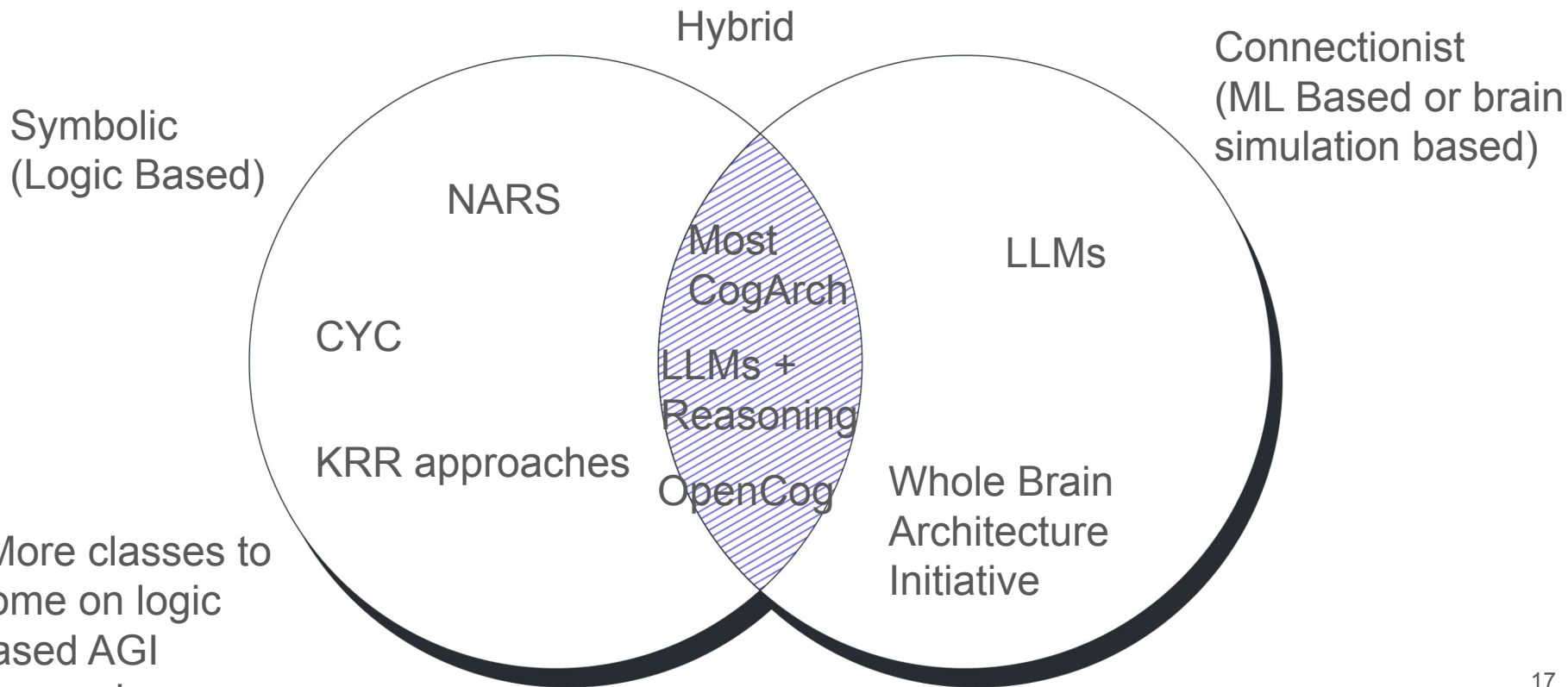
For this Class (Alignment & Logic Based AI):

- Logic Based AI safest (and maybe only) path to safe and aligned AGI.
- 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.

WHY ASIMOV PUT THE THREE LAWS OF ROBOTICS IN THE ORDER HE DID:

POSSIBLE ORDERING	CONSEQUENCES	
1. (1) DON'T HARM HUMANS 2. (2) OBEY ORDERS 3. (3) PROTECT YOURSELF	[SEE ASIMOV'S STORIES]	BALANCED WORLD
1. (1) DON'T HARM HUMANS 2. (3) PROTECT YOURSELF 3. (2) OBEY ORDERS	EXPLORE MARS!  HAHA, NO. IT'S COLD AND I'D DIE.	FRUSTRATING WORLD
1. (2) OBEY ORDERS 2. (1) DON'T HARM HUMANS 3. (3) PROTECT YOURSELF		KILLBOT HELLSCAPE
1. (2) OBEY ORDERS 2. (3) PROTECT YOURSELF 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE
1. (3) PROTECT YOURSELF 2. (1) DON'T HARM HUMANS 3. (2) OBEY ORDERS	 I'LL MAKE CARS FOR YOU, BUT TRY TO UNPLUG ME AND I'LL VAPORIZE YOU.	TERRIFYING STANDOFF
1. (3) PROTECT YOURSELF 2. (2) OBEY ORDERS 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE

Core AGI Research Area: Pathways to AGI



Core AGI Research Area :

Creation and Evaluation of AGI systems

- Lots of people have ideas about what types of systems could lead to AGI. Gotta build it first, after that, how do we evaluate these?
- Lots of existing systems that claim to be the start of an AGI system, that we want to evaluate. Two notable ones are OpenCog and NARS

opencog/**opencog**

A framework for integrated Artificial Intelligence & Artificial General Intelligence (AGI)



97 Contributors 55 Issues 2k Stars 724 Forks



opennars/**opennars**

OpenNARS for Research 3.0+



11 Contributors 75 Issues 384 Stars 83 Forks



Tests of AGI (some fun ones from Wikipedia)

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]}

Chollet's ARC-AGI challenge

- \$1,000,000 Prize, currently the largest challenge on Kaggle
- Humans score incredibly well, 97%+ accuracy.
- o1 model from OpenAI at best can score in the low 20s
- Highly specialized program search + LLMs reach 40s.

<https://arcprize.org/>



I want to give you one million dollars to create AGI!