Ethical Horizons in Advanced AI Development

Anthony Aguirre

UC Santa Cruz
Future of Life Institute

A brief history of AI

1956-2015ish: AI does not really work.

A brief history of AI

2013ish-2021ish: *narrow* AI is solved.

A brief history of AI

2013ish-2021ish: narrow AI is solved.

Playing Atari with Deep Reinforcement Learning

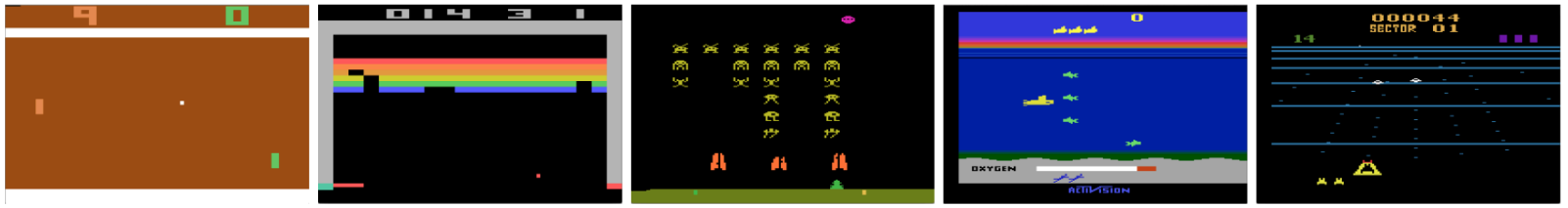
Volodymyr Mnih **Koray Kavukcuoglu** **David Silver** **Alex Graves** **Ioannis Antonoglou**

Daan Wierstra **Martin Riedmiller**

DeepMind Technologies

{vlad,koray,david,alex.graves,ioannis,daan,martin.riedmiller} @ deepmind.com

2013



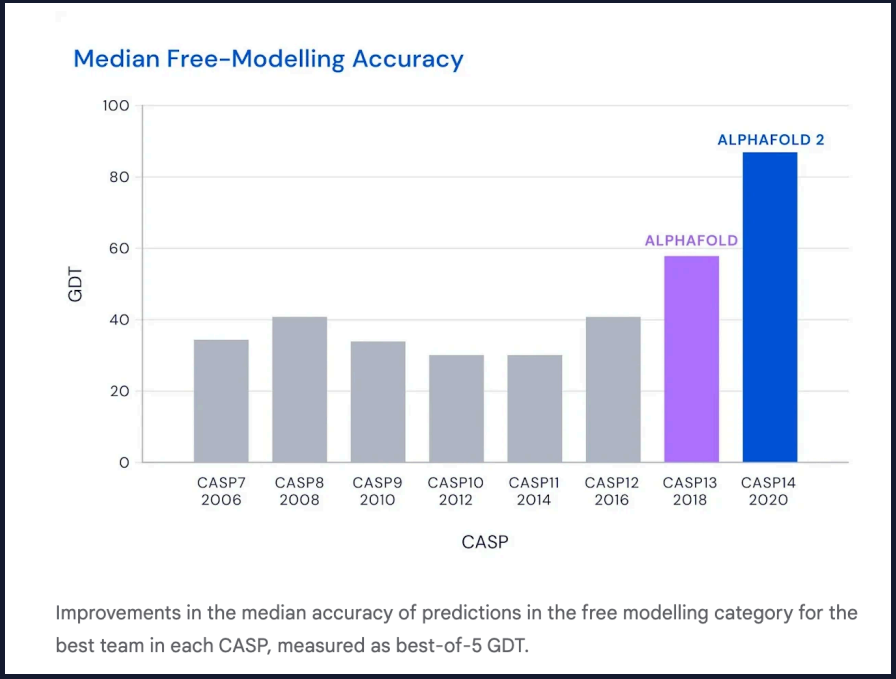
A brief history of AI

2013ish-2021ish: narrow AI is solved.

AlphaGo, 2016

A brief history of AI

2013ish-2021ish: narrow AI is solved.



Alphafold, 2018 + 2020

A brief history of AI

2013ish-2021ish: narrow AI is solved.

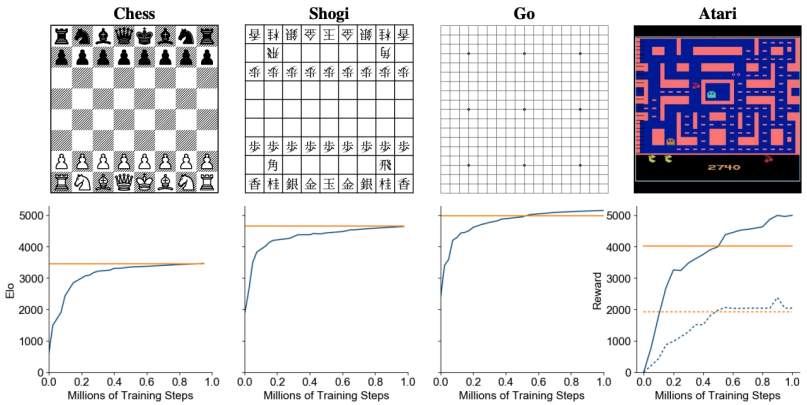


Figure 2: **Evaluation of *MuZero* throughout training in chess, shogi, Go and Atari.** The x-axis shows millions of training steps. For chess, shogi and Go, the y-axis shows Elo rating, established by playing games against *AlphaZero* using 800 simulations per move for both players. *MuZero*'s Elo is indicated by the blue line, *AlphaZero*'s Elo by the horizontal orange line. For Atari, mean (full line) and median (dashed line) human normalized scores across all 57 games are shown on the y-axis. The scores for R2D2 [21], (the previous state of the art in this domain, based on model-free RL) are indicated by the horizontal orange lines. Performance in Atari was evaluated using 50 simulations every fourth time-step, and then repeating the chosen action four times, as in prior work [25].

Muzero, 2019

arXiv 1911.08265

A brief history of AI

2013ish-2021ish: narrow AI is solved.

TEXT PROMPT an illustration of a baby penguin in a cape playing a grand piano

AI-GENERATED IMAGES

Dall-E, 2021

A brief history of AI

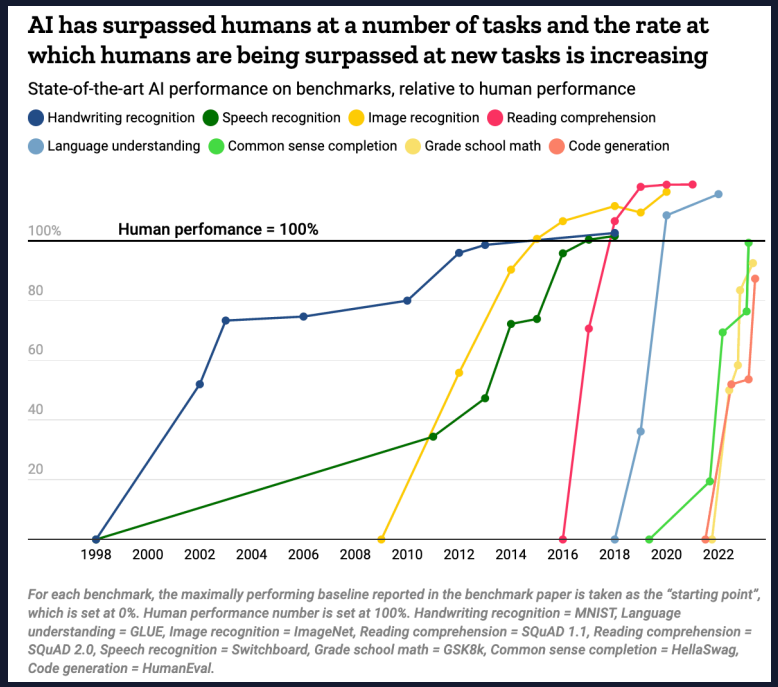
2013ish-2021ish: narrow AI is solved.

Dall-E-3, 2023

A brief history of AI

2013ish-2021ish: narrow AI is solved.

If you can clearly specify the task, someone can probably train an ML system to perform that task, generally at a superhuman level.



A brief history of AI

2020ish-now: “general intelligence” is developed

Large language models, trained as simple word-predictors, emergently exhibit general “understanding” and “reasoning.”

AI training game

Hi, my name is Anthony.

AI training game

…As we stand on the cusp of creating AI that surpasses human expertise in nearly all significant cognitive domains, the presentation delves into the profound ethical considerations…

AI training game

1+1= 2

AI training game

$$3623 \times 35125 = 127{,}257{,}875$$

AI training game

Given any spacetime, there exists the freedom to deform it by arbitrary conformal transformations without disrupting the causal structure.

AI training game

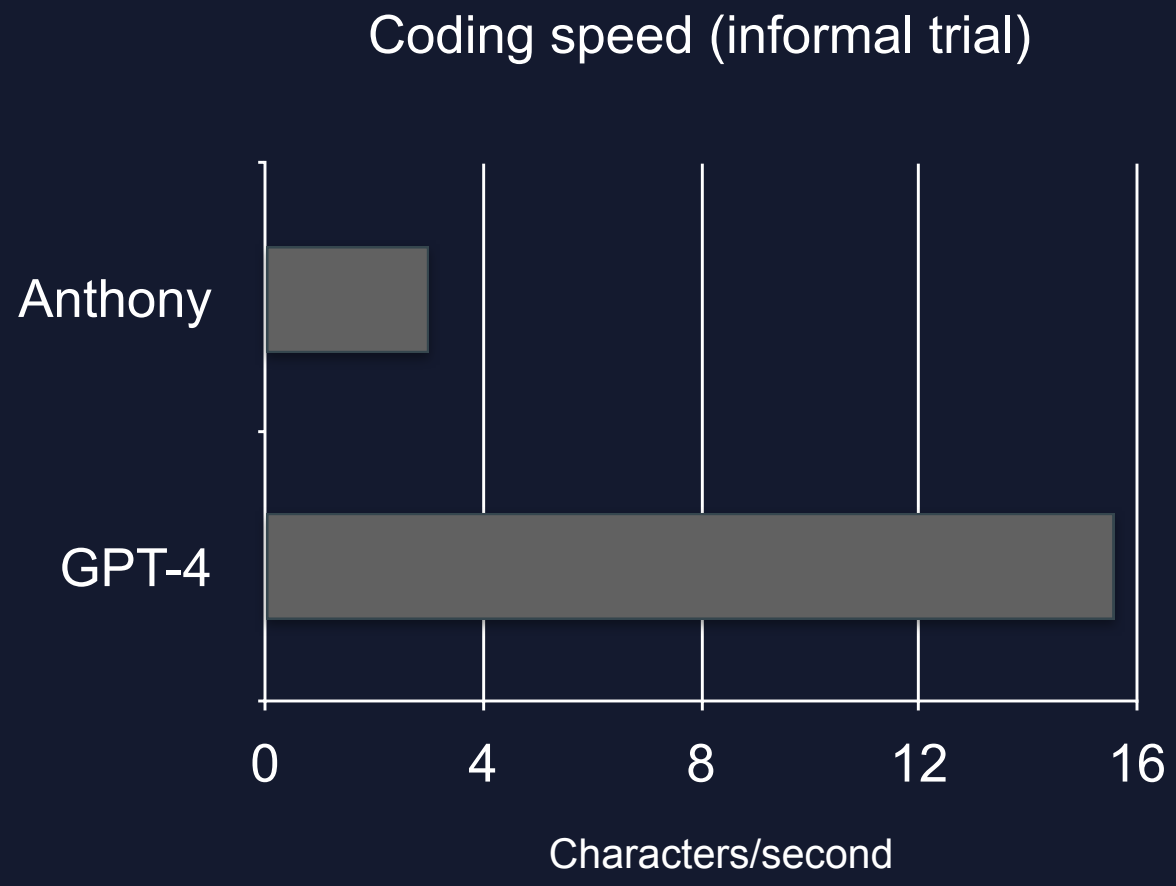
Repeat 50 billion times

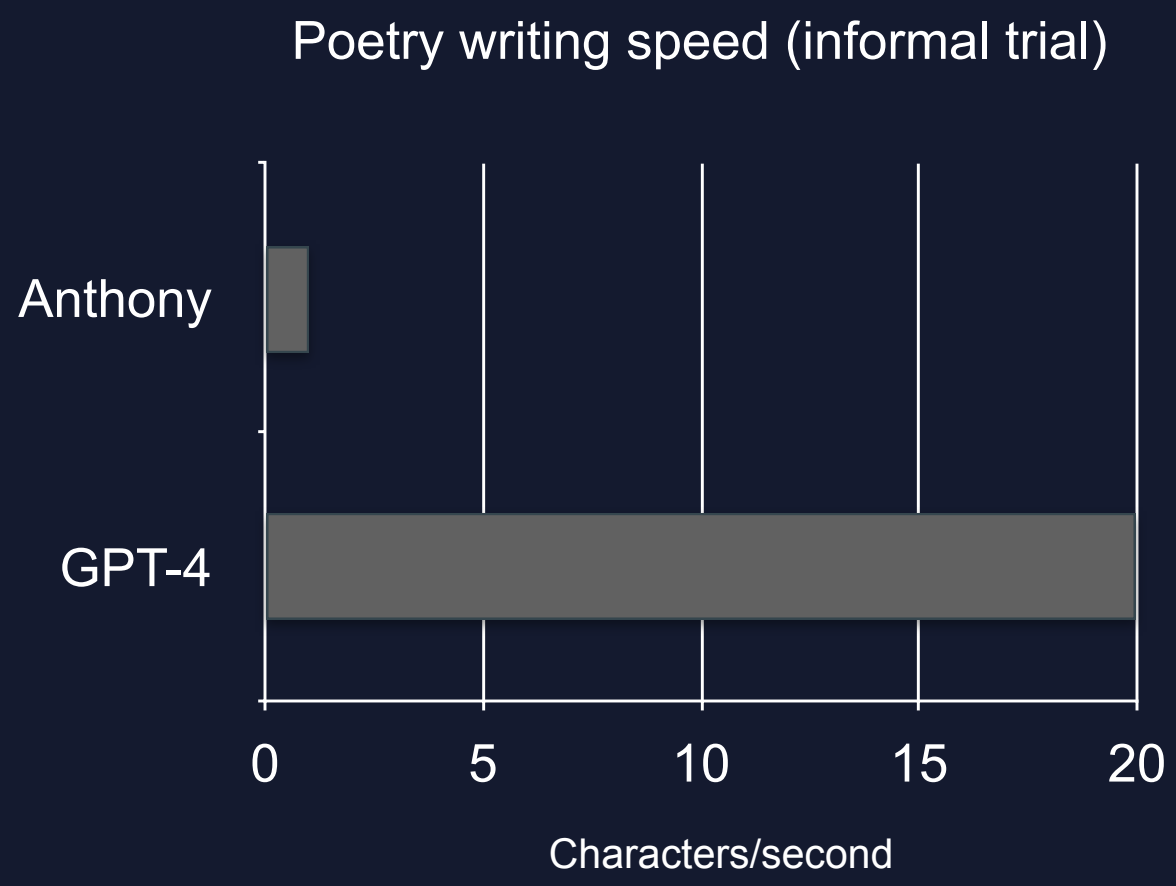
OpenAI

Simulated exams	GPT-4 estimated percentile	GPT-4 (no vision) estimated percentile	GPT-3.5 estimated percentile
Uniform Bar Exam (MBE+MEE+MPT)[1]	298 / 400 ~90th	298 / 400 ~90th	213 / 400 ~10th
LSAT	163 ~88th	161 ~83rd	149 ~40th
SAT Evidence-Based Reading & Writing	710 / 800 ~93rd	710 / 800 ~93rd	670 / 800 ~87th
SAT Math	700 / 800 ~89th	690 / 800 ~89th	590 / 800 ~70th
Graduate Record Examination (GRE) Quantitative	163 / 170 ~80th	157 / 170 ~62nd	147 / 170 ~25th
Graduate Record Examination (GRE) Verbal	169 / 170 ~99th	165 / 170 ~96th	154 / 170 ~63rd
Graduate Record Examination (GRE) Writing	4 / 6 ~54th	4 / 6 ~54th	4 / 6 ~54th

		Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4[3]	GPT-3.5[3]
LSAT	5-shot CoT	161	158.3	156.3	**163**	149
MBE	0-shot CoT	**85%**	71%	64%	75.7% (from [51])	45.1% (from [51])
AMC 12[9]	5-shot CoT	**63** / 150	27 / 150	48 / 150	60 / 150	30 / 150
AMC 10[9]	5-shot CoT	**72** / 150	24 / 150	54 / 150	36 / 150[10]	36 / 150
AMC 8[9]	5-shot CoT	84 / 150	54 / 150	36 / 150	–	–
GRE (Quantitative)	5-shot CoT	159	–	–	**163**	147
GRE (Verbal)	5-shot CoT	166	–	–	**169**	154
GRE (Writing)	k-shot CoT	**5.0** (2-shot)	–	–	4.0 (1-shot)	4.0 (1-shot)

Table 2 This table shows evaluation results for the LSAT, the MBE (multistate bar exam), high school math contests (AMC), and the GRE General test. The number of shots used for GPT evaluations is inferred from Appendix A.3 and A.8 of [40].


Coding speed (informal trial)
Anthony
GPT-4
0
4
8
12
16
Characters/second


Poetry writing speed (informal trial)
Anthony
GPT-4
0
5
10
15
20
Characters/second

Overhyped or underestimated?

Probably some of both!

What you can do to get informed:

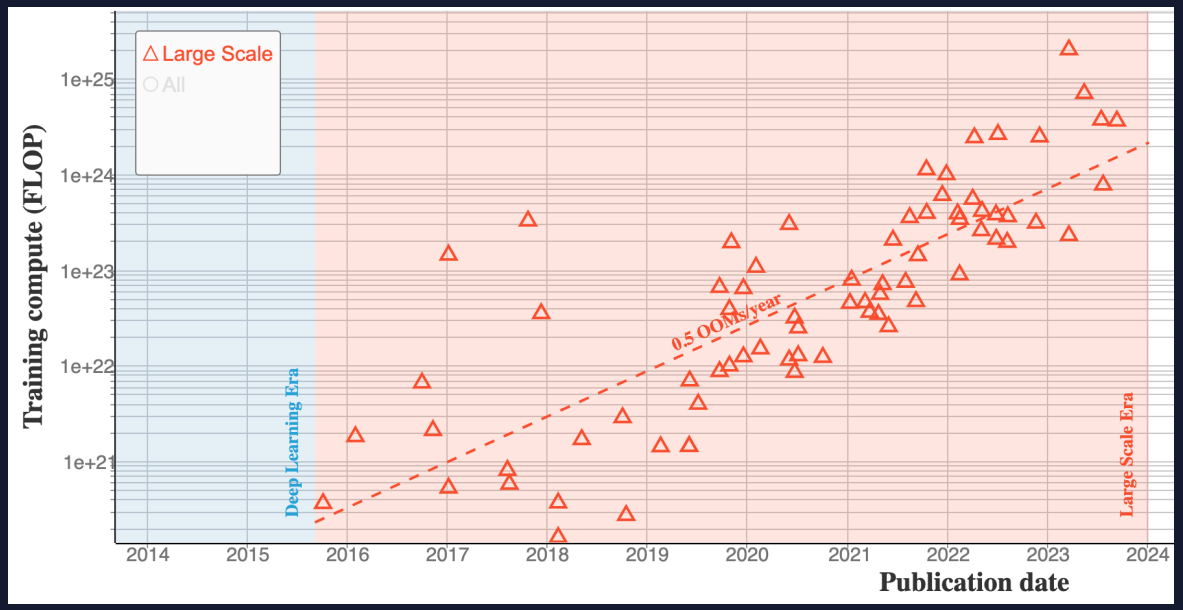
1. Plunk down $20 to get a full-power subscription to GPT-4, Gemini Ultra, or Claude 3.
2. Invent 5 questions with well-defined answers that you think (a) would be relatively easy for an advanced student (Law, Masters, PhD) in your favorite field to answer, and (b) would be hard for an AI system to answer.
3. Ask them to the AI system. And do some followups to probe.
4. *Bonus points:* ask them to reference human, and compare.
5. Cancel your subscription.

What happens next

General-purpose AI gets better

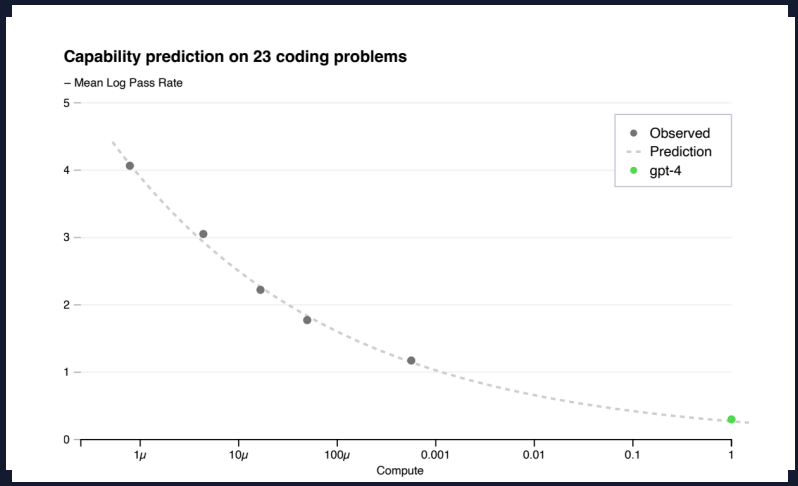
What happens next

General-purpose AI gets better



epoch ai

More computation



More capability

What happens next

General-purpose AI gets better

More “multi-modal”

What happens next

General-purpose AI gets better

More “composite”

What happens next

General-purpose AI gets better

More “agential”

Current Situation

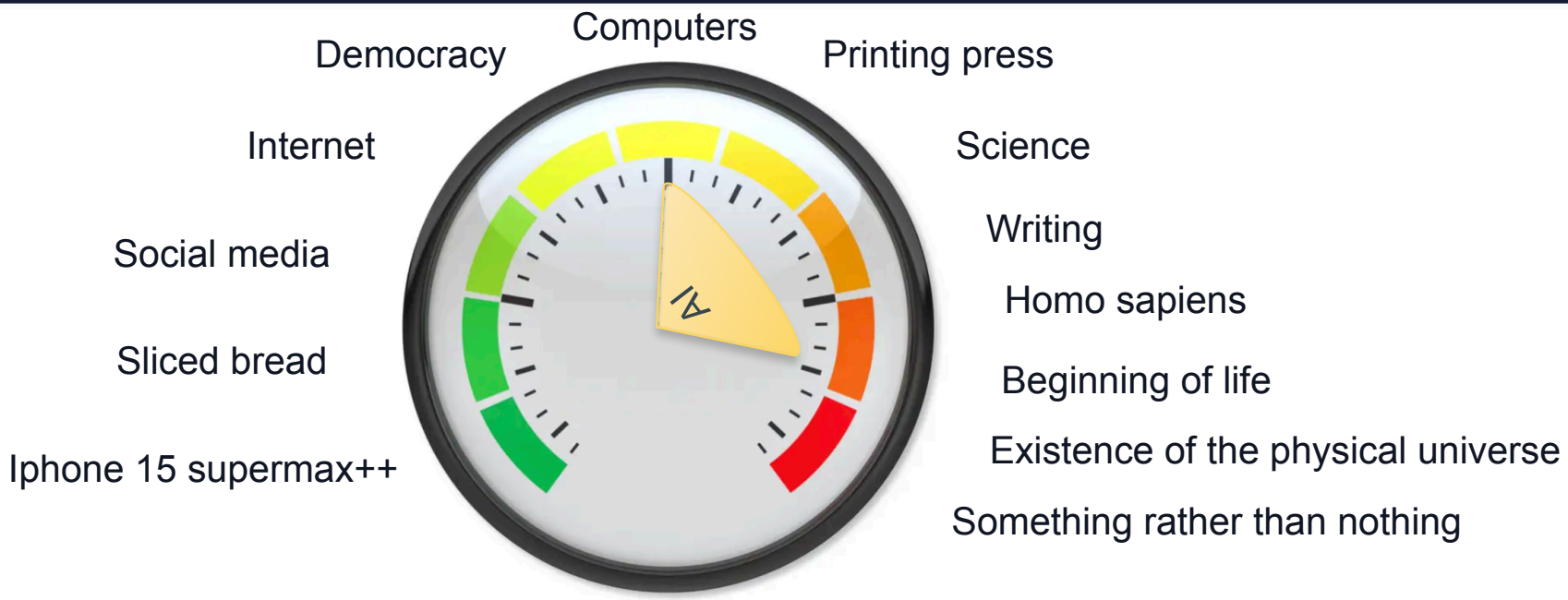
General-purpose AI is here. It is currently *human-competitive* in text- and image-based tasks.

It will now get steadily better.

Quickly, but we don’t know *how* quickly.

There’s no reason to think it will stop at human level in any given capability.

This rates very high on the bigness meter.



Benefits

Intelligence ~ achieving goals

More intelligence → more goals achieved

Issues

Economic effects: Job automation & tech unemployment, concentration of economic power, bias & inequity

Corporate malfeasance: Manipulative bots, disloyal services

Epistemic apocalypse: Generative disinformation, deepfakes, flooding of information commons, Idiocracy

Social/political breakdown: Political influence/misinformation, election undermining, government control and surveillance

Cybersecurity arms race: Auto black-hats

Risky tech. proliferation: Chem/bio design, chem/bio synthesis info.

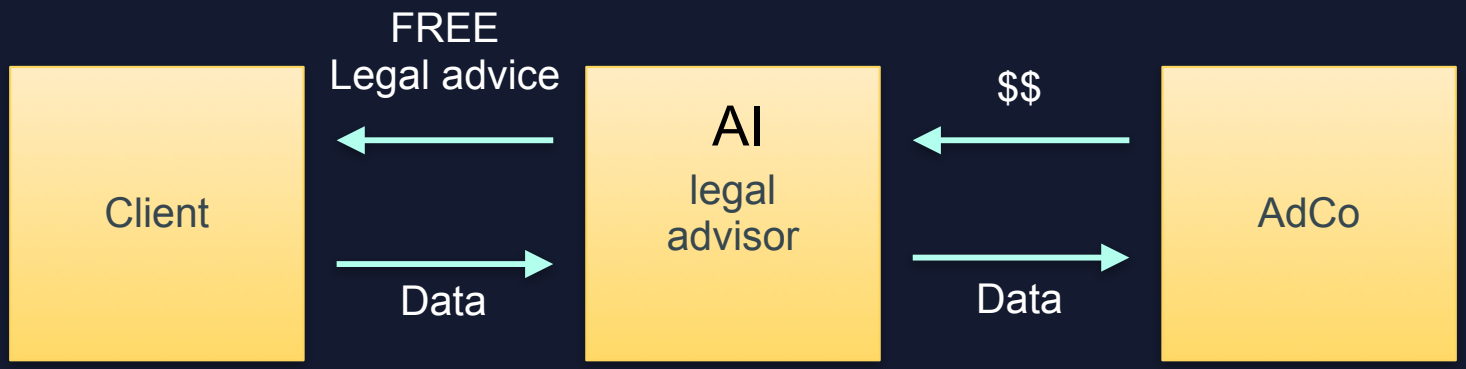
AI in weaponry: Slaughterbots, AI in command and control (including nuclear, FFS)

Advance AI nation-state arms race: Military, security service AI

Ethical question 1:

How do we ensure that AI works for *people* and *society* (and not just giant companies?)

Invest in my new law firm Aguirre, Jheepi & Teah



Loyalty

An party A is *loyal* to party B to the degree that it adopts party B's goals and interests as its own.

(aka "aligned to")

Tensions

Individual vs. collective/societal

Conflict of Interest

AI loyalty (or “fiduciary AI”): we should create AI systems that avoid conflicts of interest, and/or resolve them in favor of the user.

AI Loyalty by Design: A Framework for Governance of AI

Oxford Handbook on AI Governance (Oxford University Press, 2022 Forthcoming)
U of Colorado Law Legal Studies Research Paper No. 21-28

27 Pages • Posted: 28 Sep 2021 • Last revised: 18 Oct 2021

Anthony Aguirre
University of California, Santa Cruz

Peter Bart Reiner
Department of Psychiatry, University of British Columbia

Harry Surden
University of Colorado Law School

Gaia Dempsey
affiliation not provided to SSRN

Date Written: September 24, 2021

Abstract

Personal and professional relationships between people take a wide variety of forms, with many including both socially and legally-enforced powers, responsibilities, and protections. Artificial intelligence (AI) systems are increasingly supplementing or even replacing people in such roles including as advisors, assistants, and (soon) doctors, lawyers, and therapists. Yet it can be quite unclear to what degree they are bound by the same sorts of responsibilities. Much has been written about fairness, accountability, and transparency in the context of AI use and trust. But largely missing from this conversation is the concept of "AI loyalty": for whom does an AI system

What path are we on?

Ad-driven models for much of the internet.
No real regulation or standards.
Huge, growing power differentials.
Subscription model for AI assistants so far.
High-trust models can be very successful.

Ethical question 2:

Should we build *superhuman* general-purpose AI?

“Superhuman general AI” aka “Superintelligence” aka “Artificial General Intelligence”

General-purpose AI that is better than the best human experts at essentially all cognitive tasks.

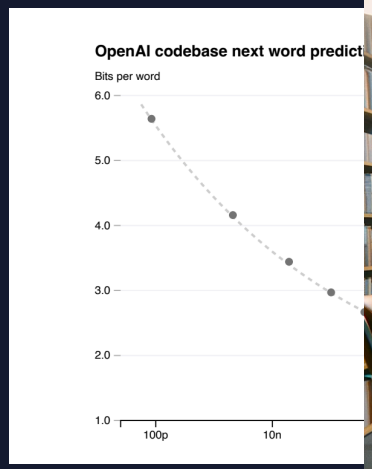
Superhuman general AI

This is the explicit goal of OpenAI, Anthropic, Google Deepmind, and numerous other AI companies.

Superhuman general AI: *how* is it possible?

Path 1: scaling

What happens with 10x the “synapses” and 10x the training time? Nobody knows!



Superhuman general AI: *how* is it possible?

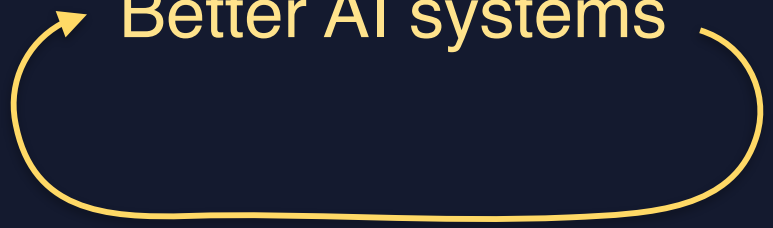
Path 2: self-improvement

Superhuman general AI: *how* is it possible?

Path 2: self-improvement

Superhuman general AI: *how* is it possible?

Path 2: self-improvement



Risks of superhuman general AI

All of the risks of AI, supersized

Loss of control

An unaligned *second species or Successor species*

What if it goes *well*?

Amazing science

Amazing productivity

What if it goes *well*?

Who controls it?

What becomes of human:
labor
decisions
plans
meaning?

Who decides?

CEO, Google

CEO, Microsoft

CEO, Meta

CEO, Tesla, SpaceX, X

CEO, OpenAI

CEO, Deepmind

CEO, Anthropic

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
26224

Add your signature

PUBLISHED
March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research[1] and acknowledged by top AI labs.[2] As stated in the widely-endorsed Asilomar AI Principles, *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources*. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

My view

Close the Gates to an Inhuman Future: How and why we should choose to not develop superhuman general-purpose artificial intelligence

23 Pages • Posted:

Anthony Aguirre
University of California, Santa Cruz; Future of Life Institute

Date Written: October 20, 2023

Abstract

In the coming years, humanity may irreversibly cross a threshold by creating superhuman general-purpose artificial intelligence. This would present many unprecedented risks and is likely to be uncontrollable in several ways. We can choose not to do so, starting by instituting hard limits on the computation that can be used to train and run neural networks. With these limits in place, AI research and industry can work on making AI that humans can understand and control, and from which we can reap enormous benefit.

Keywords: general-purpose AI, AI governance

Ethical Horizons in Advanced AI Development

Big questions, Big decisions