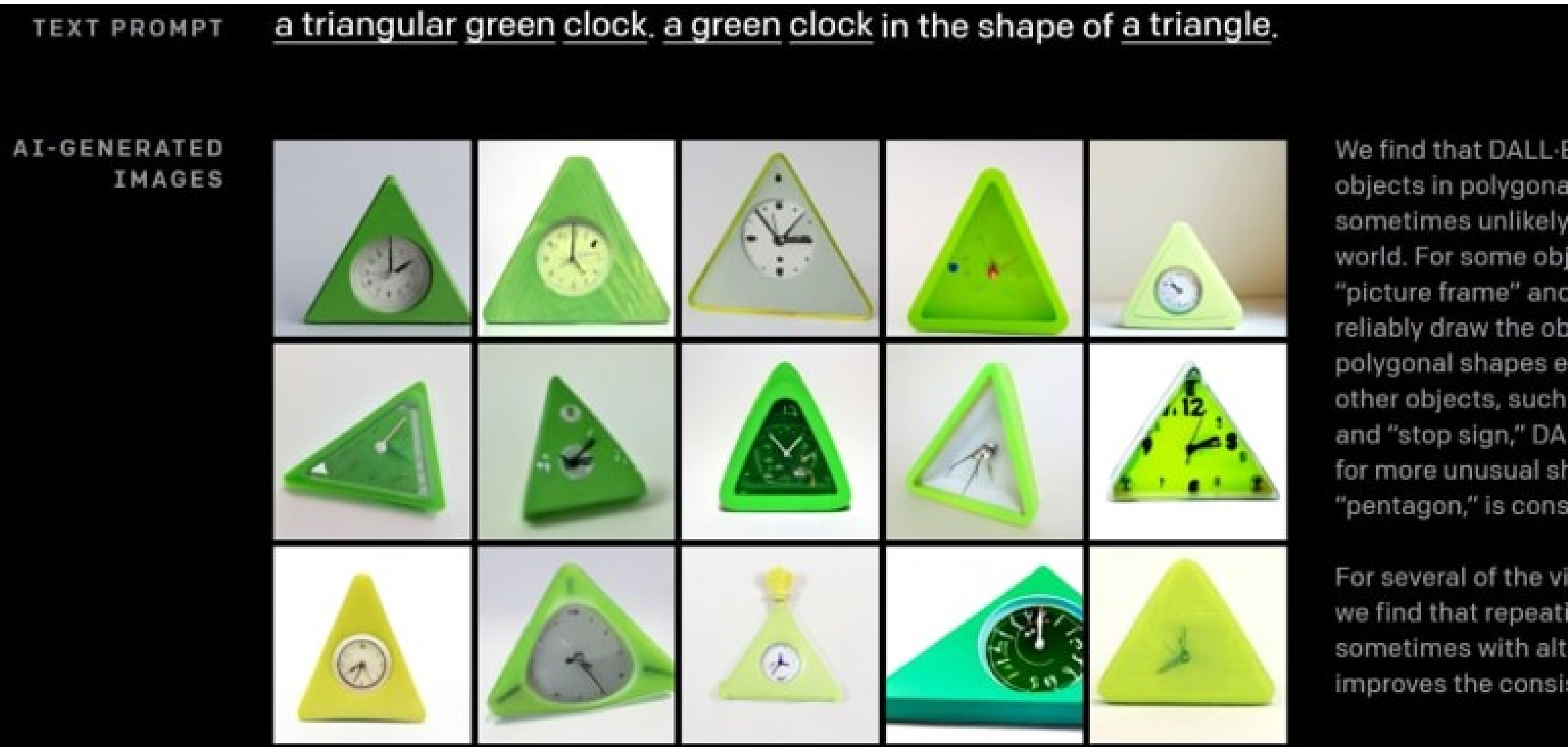


TEXT PROMPT a triangular green clock. a green clock in the shape of a triangle.

AI-GENERATED IMAGES

We find that DALL·E
objects in polygona
sometimes unlikely
world. For some obj
"picture frame" and
reliably draw the ob
polygonal shapes e
other objects, such
and "stop sign," DA
for more unusual sh
"pentagon," is cons

For several of the vi
we find that repeati
sometimes with alt
improves the consi

# AI Timelines

# Fun with +12 OOMs of Compute

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

# *Or: Big Timelines Crux Operationalized*

What fun things could one build with +12 orders of magnitude of compute? By 'fun' I mean 'powerful.' This hypothetical is highly relevant to AI timelines, for reasons I'll explain later.

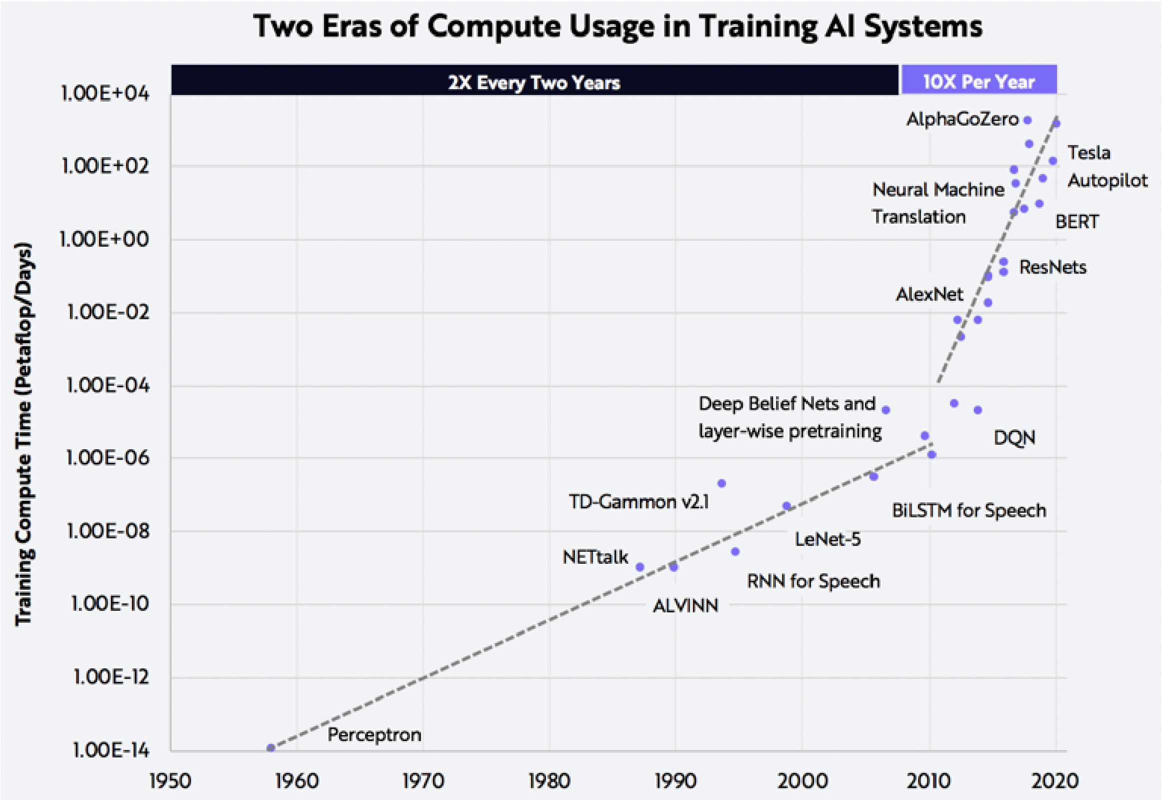
*Summary (Spoilers):*

I describe a hypothetical scenario that concretizes the question *"what could be built with 2020's algorithms/ideas/etc. but a trillion times more compute?"* Then I give some answers to that question. Then I ask: How likely is it that some sort of TAI would happen in this scenario? This second question is a useful operationalization of the (IMO) most important, most-commonly-discussed timelines crux: "Can we get TAI just by throwing more compute at the problem?" I consider this operationalization to be the main contribution of this post; it directly plugs into Ajeya's timelines model and is quantitatively more cruxy than anything else I know of. The secondary contribution of this post is my set of answers to the first question: They serve as intuition pumps for my answer to the second, which strongly supports my views on timelines.

# The hypothetical

*In 2016 the Compute Fairy visits Earth and bestows a blessing: Computers are magically 12 orders of magnitude faster! Over the next five years, what happens? The Deep Learning AI Boom still happens, only much crazier: Instead of making AlphaStar for 10^23 floating point operations, DeepMind makes something for 10^35. Instead of making GPT-3 for 10^23 FLOPs, OpenAI makes something for 10^35. Instead of industry and academia making a cornucopia of things for 10^20 FLOPs or so, they make a cornucopia of things for 10^32 FLOPs or so. When random grad students and hackers spin up neural nets on their laptops, they have a trillion times more compute to work with. [EDIT: Also assume magic +12 OOMs of memory, bandwidth, etc. All the ingredients of compute.]*

For context on how big a deal +12 OOMs is, consider the graph below, from ARK. It's measuring petaflop-days, which are about 10^20 FLOP each. So 10^35 FLOP is 1e+15 on this graph. GPT-3 and AlphaStar are not on this graph, but if they were they would be in the very top-right corner.



# Question One: In this hypothetical, what sorts of things could AI projects build?

I encourage you to stop reading, set a five-minute timer, and think about fun things that could be built in this scenario. I'd love it if you wrote up your answers in the comments!

# My tentative answers:

Below are my answers, listed in rough order of how 'fun' they seem to me. I'm not an AI scientist so I expect my answers to overestimate what could be done in some ways, and underestimate in other ways. Imagine that each entry is the best version of itself, since it is built by experts (who have experience with smaller-scale versions) rather than by me.

## OmegaStar:

In our timeline, it cost about 10^23 FLOP to train AlphaStar. (OpenAI Five, which is in some ways more impressive, took less!) Let's make OmegaStar like AlphaStar only +7 OOMs bigger: the size of a human brain.[1] [EDIT: You may be surprised to learn, as I was, that AlphaStar has about 10% as many parameters as a honeybee has synapses! Playing against it is like playing against a tiny game-playing insect.]

Larger models seem to take less data to reach the same level of performance, so it would probably take at most 10^30 FLOP to reach the same level of Starcraft performance as AlphaStar, and indeed we should expect it to be qualitatively better.[2] So let's do that, but also train it on lots of other games too.[3] There are 30,000 games in the Steam Library. We train OmegaStar long enough that it has as much time on *each* game as AlphaStar had on Starcraft. With a brain so big, maybe it'll start to do some transfer learning, acquiring generalizeable skills that work across many of the games instead of learning a separate policy for each game.

OK, that uses up 10^34 FLOP—a mere 10% of our budget. With the remainder, let's add some more stuff to its training regime. For example, maybe we also make it read the entire internet and play the "Predict the next word you are about to read!" game. Also the "Predict the covered-up word" and "predict the covered-up piece of an image" and "predict later bits of the video" games.

OK, that probably still wouldn't be enough to use up our compute budget. A Transformer that was the size of the human brain would only need 10^30 FLOP to get to human level at the the predict-the-next-word game according to Gwern, and while OmegaStar isn't a transformer, we have 10^34 FLOP available.[4] (What a curious coincidence, that human-level performance is reached right when the AI is human-brain-sized! Not according to Shorty.)

Let's also hook up OmegaStar to an online chatbot interface, so that billions of people can talk to it and play games with it. We can have it play the game "Maximize user engagement!"

...we probably still haven't used up our whole budget, but I'm out of ideas for now.

# Amp(GPT-7):

Let's start by training GPT-7, a transformer with 10^17 parameters and 10^17 data points, on the entire world's library of video, audio, and text. This is almost 6 OOMs more params *and* almost 6 OOMs more training time than GPT-3. Note that a mere +4 OOMs of params and training time is predicted to reach near-optimal performance at text prediction and all the tasks thrown at GPT-3 in the original paper; so this GPT-7 would be superhuman at all those things, and also at the analogous video and audio and mixed-modality tasks.[5] Quantitatively, the gap between GPT-7 and GPT-3 is about *twice as large* as the gap between GPT-3 and *GPT-1*, (about 25% the loss GPT-3 had, which was about 50% the loss GPT-1 had) so try to imagine a qualitative improvement twice as big also. And that's not to mention the possible benefits of multimodal data representations.[6]

We aren't finished! This only uses up 10^34 of our compute. Next, we let the public use prompt programming to make a giant library of GPT-7 functions, like the stuff demoed here and like the stuff being built here, only much better because it's GPT-7 instead of GPT-3. Some examples:

- Decompose a vague question into concrete subquestions
- Generate a plan to achieve a goal given a context
- Given a list of options, pick the one that seems most plausible / likely to work / likely to be the sort of thing Jesus would say / [insert your own evaluation criteria here]
- Given some text, give a score from 0 to 10 for how accurate / offensive / likely-to-be-written-by-a-dissident / [insert your own evaluation criteria here] the text is.

And of course the library also contains functions like "google search" and "Given webpage, click on X" (remember, GPT-7 is multimodal, it can input and output *video*, parsing webpages is easy). It also has functions like "Spin off a new version of GPT-7 and fine-tune it on the following data." Then we fine-tune GPT-7 on the library so that it knows how to use those functions, and even write new ones. (Even GPT-3 can do basic programming, remember. GPT-7 is *much* better.)

We still aren't finished! Next, we embed GPT-7 in an amplification scheme — a "chinese-room bureaucracy" of calls to GPT-7. The basic idea is to have functions that break down tasks into sub-tasks, functions that do those sub-tasks, and functions that combine the results of the sub-tasks into a result for the task. For example, a fact-checking function might start by dividing up the text into paragraphs, and then extract factual claims from each paragraph, and then generate google queries designed to fact-check each claim, and then compare the search results with the claim to see whether it is contradicted or confirmed, etc. And an article-writing function might call the fact-checking function as one of the intermediary steps. By combining more and more functions into larger and larger bureaucracies, more and more sophisticated behaviors can be achieved. And by fine-tuning GPT-7 on examples of this sort of thing, we can get it to understand how it works, so that we can write GPT-7 functions in which GPT-7 chooses which other functions to call. Heck, we could even have GPT-7 try writing its own functions! [7]

The ultimate chinese-room bureaucracy would be an agent in its own right, running a continual OODA loop of taking in new data, distilling it into notes-to-future-self and new-data-to-fine-tune-on, making plans and sub-plans, and executing them. Perhaps it has a text file describing its goal/values that it passes along as a note-to-self — a "bureaucracy mission statement."

Are we done yet? No! Since it "only" has 10^17 parameters, and uses about six FLOP per parameter per token, we have almost 18 orders of magnitude of compute left to work with. [8] So let's give our GPT-7 uber-bureaucracy an internet connection and run it for 100,000,000 function-calls (if we think of each call as a subjective second, that's about 3 subjective years). Actually, let's generate 50,000 different uber-bureaucracies and run them all for that long. And then let's evaluate their performance and reproduce the ones that did best, and repeat. We could do 50,000 generations of this sort of artificial evolution, for a total of about 10^35 FLOP.[9]

Note that we could do all this amplification-and-evolution stuff with OmegaStar in place of GPT-7.

## Crystal Nights:

(The name comes from an excellent short story.)

Maybe we think we are missing something fundamental, some unknown unknown, some special sauce that is necessary for true intelligence that humans have and our current artificial neural net designs won't have even if scaled up +12 OOMs. OK, so let's search for it. We set out to recapitulate evolution.

We make a planet-sized virtual world with detailed and realistic physics and graphics. OK, not *perfectly* realistic, but much better than any video game currently on the market! Then, we seed it with a bunch of primitive life-forms, with a massive variety of initial mental and physical architectures. Perhaps they have a sort of virtual genome, a library of code used to construct their bodies and minds, with modular pieces that get exchanged via sexual reproduction (for those who are into that sort of thing). Then we let it run, for a billion in-game years if necessary!

Alas, Ajeya estimates it would take about 10^41 FLOP to do this, whereas we only have 10^35.[10] So we probably need to be a million times more compute-efficient than evolution. But maybe that's doable. Evolution is pretty dumb, after all.

1. Instead of starting from scratch, we can can start off with "advanced" creatures, e.g. sexually-reproducing large-brained land creatures. It's unclear how much this would save but plausibly could be at least one or two orders of magnitude, since Ajeya's

estimate assumes the average creature has a brain about the size of a nematode worm’s brain.[11]

2. We can grant “magic traits” to the species that encourage intelligence and culture; for example, perhaps they can respawn a number of times after dying, or transfer bits of their trained-neural-net brains to their offspring. At the very least, we should make it metabolically cheap to have big brains; no birth-canal or skull should restrict the number of neurons a species can have! Also maybe it should be easy for species to have neurons that don’t get cancer or break randomly.
3. We can force things that are bad for the individual but good for the species, e.g. identify that the antler size arms race is silly and nip it in the bud before it gets going. In general, more experimentation/higher mutation rate is probably better for the species than for the individual, and so we could speed up evolution by increasing the mutation rate. We can also identify when a species is trapped in a local optima and take action to get the ball rolling again, whereas evolution would just wait until some climactic event or something shakes things up.
4. We can optimise for intelligence instead of ability to reproduce, by crafting environments in which intelligence is much more useful than it was at any time in Earth’s history. (For example, the environment can be littered with monoliths that dispense food upon completion of various reasoning puzzles. Perhaps some of these monoliths can teach English too, that’ll probably come in handy later!) Think about how much faster dog breeding is compared to wolves evolving in the wild. Breeding for intelligence should be correspondingly faster than waiting for it to evolve.
5. There are probably additional things I haven’t thought of that would totally be thought of, if we had a team of experts building this evolutionary simulation with 2020’s knowledge. I’m a philosopher, not an evolutionary biologist!

# Skunkworks:

What about STEM AI? Let’s do some STEM. You may have seen this now-classic image:

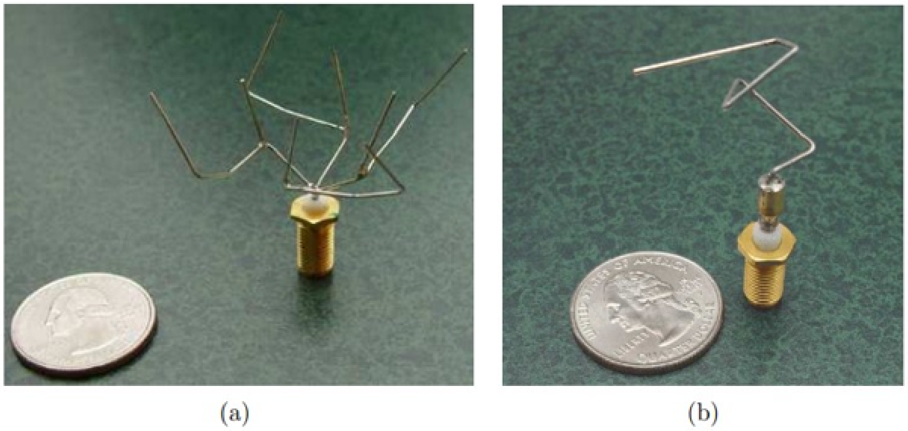

(a)
(b)

**Figure 2.** Photographs of prototype evolved antennas: (a) the best evolved antenna for the initial gain pattern requirement, ST5-3-10; (b) the best evolved antenna for the revised specifications, ST5-33-142-7.

These antennas were designed by an evolutionary search algorithm. Generate a design, simulate it to evaluate predicted performance, tweak & repeat. They flew on a NASA spacecraft fifteen years ago, and were massively more efficient and high-performing than the contractor-designed antennas they replaced. Took less human effort to make, too.[12]

This sort of thing gets a lot more powerful with +12 OOMs. Engineers often use simulations to test designs more cheaply than by building an actual prototype. SpaceX, for example, did this for their Raptor rocket engine. Now imagine that their simulations are significantly more detailed, spending 1,000,000x more compute, and also that they have an evolutionary search component that auto-generates 1,000 variations of each design and iterates for 1,000 generations to find the optimal version of each design for the problem (or even invents new designs from scratch.) And perhaps all of this automated design and tweaking (and even the in-simulation testing) is done more intelligently by a copy of OmegaStar trained on this "game."

Why would this be a big deal? I'm not sure it would be. But take a look at this list of strategically relevant technologies and events and think about whether Skunkworks being widely available would quickly lead to some of them. For example, given how successful AlphaFold 2 has been, maybe Skunkworks could be useful for designing nanomachines. It could certainly make it a lot easier for various minor nations and non-state entities to build weapons of mass destruction, perhaps resulting in a vulnerable world.

## Neuromorph:

According to page 69 of this report, the Hodgkin-Huxley model of the neuron is the most detailed and realistic (and therefore the most computationally expensive) as of 2008. [EDIT: Joe Carlsmith, author of a more recent report, tells me there are more detailed+realistic models available now] It costs 1,200,000 FLOP per second per neuron to run. So a human brain (along with relevant parts of the body, in a realistic-physics virtual environment, etc.) could be simulated for about 10^17 FLOP per second.

Now, presumably (a) we don't have good enough brain scanners as of 2020 to actually reconstruct any particular person's brain, and (b) even if we did, the Hodgkin-Huxley model might not be detailed enough to fully capture that person's personality and cognition.[13]

But maybe we can do something 'fun' nonetheless: We scan someone's brain and then create a simulated brain that looks like the scan as much as possible, and then fills in the details in a random but biologically plausible way. Then we run the simulated brain and see what happens. Probably gibberish, but we run it for a simulated year to see whether it gets its act together and learns any interesting behaviors. After all, human children start off with randomly connected neurons too, but they learn.[14]

All of this costs a mere 10^25 FLOP. So we do it repeatedly, using stochastic gradient descent to search through the space of possible variations on this basic setup, tweaking parameters of the simulation, the dynamical rules used to evolve neurons, the initial conditions, etc. We can do 100,000 generations of 100,000 brains-running-for-a-year this way. Maybe we'll eventually find something intelligent, even if it lacks the memories and personality of the original scanned human.

# Question Two: In this hypothetical, what's the probability that TAI appears by end of 2020?

The first question was my way of operationalizing *"what could be built with 2020's algorithms/ideas/etc. but a trillion times more compute?"*

This second question is my way of operationalizing *"what's the probability that the amount of computation it would take to train a transformative model using 2020's algorithms/ideas/etc. is 10^35 FLOP or less?"*

(Please ignore thoughts like "But maybe all this extra compute will make people take AI safety more seriously" and "But they wouldn't have incentives to develop modern parallelization algorithms if they had computers so fast" and "but maybe the presence of the

Compute Fairy will make them believe the simulation hypothesis?” since they run counter to the spirit of the thought experiment.)

Remember, the definition of Transformative AI is “AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution.”

1%

2%

3%

4%

5%

6%

7%

8%

9%

USER0xe3afcf0a9d42d1c8 (1%),Aiyen (1%)
10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

mdco (23%)
30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

yagudin (40%),Jsevillamol (41%),DanielFilan (42%),Sublation (43%)
50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

elifland (52%),Tamay Besiroglu (58%)
60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

Gabriel Wu (66%)
70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

Darmani (70%),Andrew Vlahos (70%),Sammy Martin (74%),Anteetum (75%),Gabe Espinoza (78%)
80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

DragonGod (80%),MinusGix (80%),MaxRa (80%),teradimich (80%),Adrià Garriga-alonso (80%),lalaithion (80%),steven0461 (81%),TimothyK (84%),JasperGeh (85%),Garrett Baker (85%),abramdemski (85%),NunoSempere (85%),Jan (86%),habryka (87%),LukeMellor (88%),Rowen (89%),peterbarnett (89%)
90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

Tao Lin (90%),NoUsernameSelected (90%),Hjalmar_Wijk (90%),Bogdan Ionut Cirstea (90%),Vermillion (90%),Daniel Kokotajlo (90%),qweered (93%),rorygreig (93%),Dach (93%),Borasko (94%),Lost Futures (95%),Sean Hardy (95%),platers (95%),Darkar Dengeno (95%),Yair Halberstadt (95%),Pialgo (95%),Vivek Hebbar (96%),SurfingOrca (97%),simeon_c (98%),Zac Hatfield-Dodds (98%),MSRayne (98%),joshuatanderson (99%),Yitz (99%)
1%
In the hypothetical, will TAI be created by the end of 2020?
99%

Did you read those answers to Question One, visualize them and other similarly crazy things that would be going on in this hypothetical scenario, and think “Eh, IDK if that would be enough, I’m 50-50 on this. Seems plausible TAI will be achieved in this scenario but seems equally plausible it wouldn’t be.”

No! … Well, maybe you do, but speaking for myself, I don’t have that reaction.

When I visualize this scenario, I’m like “Holyshit *all five* of these distinct research programs seem like they would probably produce something transformative within five years and perhaps even *immediately*, and there are probably more research programs I haven’t thought of!”

My answer is 90%. The reason it isn’t higher is that I’m trying to be epistemically humble and cautious, account for unknown unknowns, defer to the judgment of others, etc. If I just went with my inside view, the number would be 99%. This is because I can’t articulate any not-totally-implausible possibility in which OmegaStar, Amp(GPT-7), Crystal Nights, Skunkworks, and Neuromorph and more *don’t* lead to transformative AI within five years. All I can think of is things like “Maybe transformative AI requires some super-special mental structure which can only be found by massive blind search, so massive that the Crystal Nights program can’t find it…” I’m very interested to hear what people whose *inside-view* answer to Question Two is <90% have in mind for the remaining 10%+. I expect I’m just not modelling their views well and that after hearing more I’ll be able to imagine some not-totally-implausible no-TAI possibilities. My inside view is obviously overconfident. Hence my answer of 90%.

Poll: What is your *inside-view* answer to Question Two, i.e. your answer *without* taking into account meta-level concerns like peer disagreement, unknown unknowns, biases, etc.

1%

2%

3%

4%

5%

6%

7%

8%

9%

10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

Jide (22%)
30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

Andrew Vlahos (30%),Jsevillamol (39%)
40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

DanielFilan (43%)
50%

51%

52%

53%

54%

55%

56%

57%

58%

59%

elifland (56%)
60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

theme_arrow (70%)
80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

simeon_c (86%),DragonGod (89%)
90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

habryka (90%),Garrett Baker (90%),teradimich (90%),NunoSempere (90%),JasperGeh (93%),Anteetum (94%),lumenwrites (95%),MaxRa (95%),sam (97%),Thomas Larsen (97%),rorygreig (97%),Borasko (97%),Yair Halberstadt (97%),Zac Hatfield-Dodds (98%),Lost Futures (98%),platers (98%),joshuatanderson (99%),devansh (99%),qvalq (99%),SurfingOrca (99%),Yitz (99%),Vivek Hebbar (99%),Tao Lin (99%),TimothyK (99%),Sean Hardy (99%),Bogdan Ionut Cirstea (99%),Hjalmar_Wijk (99%),Darkar Dengeno (99%),lalaithion (99%),Pialgo (99%),Daniel Kokotajlo (99%)
1%
In the hypothetical, will TAI be created by the end of 2020? (Inside-view)
99%

Bonus: I’ve argued elsewhere that what we really care about, when thinking about AI timelines, is AI-induced points of no return. I think this is likely to be within a few years of TAI, and my answer to this question is basically the same as my answer to the TAI version, but just in case:

1%

2%

3%

4%

5%

6%

7%

8%

9%

USER0xe3afcf0a9d42d1c8 (1%)
10%

11%

12%

13%

14%

15%

16%

17%

18%

19%

20%

21%

22%

23%

24%

25%

26%

27%

28%

29%

pch (21%),DragonGod (23%)
30%

31%

32%

33%

34%

35%

36%

37%

38%

39%

40%

41%

42%

43%

44%

45%

46%

47%

48%

49%

TimothyK (40%)
50%

51%

52%

53%

54%

Andrew Vlahos (50%),elifland (55%)

55%

56%

57%

58%

59%

60%

61%

62%

63%

64%

65%

66%

67%

68%

69%

theme_arrow (60%)
70%

71%

72%

73%

74%

75%

76%

77%

78%

79%

JasperGeh (75%), Yair Halberstadt (75%)
80%

81%

82%

83%

84%

85%

86%

87%

88%

89%

habryka (80%),alexlyzhov (84%),lalaithion (84%),Tao Lin (85%),Sean Hardy (85%),Hjalmar_Wijk (85%),rorygreig (89%)
90%

91%

92%

93%

94%

95%

96%

97%

98%

99%

Yitz (90%),Bogdan Ionut Cirstea (90%),Darkar Dengeno (90%),MaxRa (90%),Vermillion (90%),Daniel Kokotajlo (90%),Cervera (91%),Borasko (92%),Garrett Baker (93%),Zac Hatfield-Dodds (95%),NunoSempere (95%),abramdemski (96%),joshuatanderson (98%),SurfingOrca (98%),HausdorffSpace (98%)
1%
In the hypothetical, will an AI-induced point of no return happen by end of 2020?
99%

# OK, here’s why all this matters

Ajeya Cotra’s excellent timelines forecasting model is built around a probability distribution over “the amount of computation it would take to train a transformative model if we had to do it using only current knowledge.”[15] (pt1p25) Most of the work goes into constructing that probability distribution; once that’s done, she models how compute costs decrease, willingness-to-spend increases, and new ideas/insights/algorithms are added over time, to get her final forecast.

One of the great things about the model is that it’s interactive; you can input your own probability distribution and see what the implications are for timelines. This is good because there’s a lot of room for subjective judgment and intuition when it comes to making the probability distribution.

What I’ve done in this post is present an intuition pump, a thought experiment that might elicit in the reader (as it does in me) the sense that *the probability distribution should have the bulk of its mass by the 10^35 mark.*

Ajeya’s best-guess distribution has the 10^35 mark as its median, roughly. As far as I can tell, this corresponds to answering “50%” to Question Two.[16]

If that’s also your reaction, fair enough. But insofar as your reaction is closer to mine, you should have shorter timelines than Ajeya did when she wrote the report.

There are lots of minor nitpicks I have with Ajeya’s report, but I’m not talking about them; instead, I wrote this, which is a lot more subjective and hand-wavy. I made this choice because the minor nitpicks don’t ultimately influence the answer very much, whereas this more subjective disagreement is a pretty big crux.[17] Suppose your answer to Question 2 is 80%. Well, that means your distribution should have 80% by the 10^35 mark compared to Ajeya’s 50%, and that means that your median should be roughly 10 years earlier than hers, all else equal: 2040-ish rather than 2050-ish.[18]

I hope this post helps focus the general discussion about timelines. As far as I can tell, the biggest crux for most people is something like “Can we get TAI just by throwing more compute at the problem?” Now, obviously we *can* get TAI just by throwing more compute at the problem, there are theorems about how neural nets are universal function approximators etc., and we can always do architecture search to find the right architectures. So the crux is really about whether we can get TAI just by throwing *a large but not too large* amount of compute at the problem… and I propose we operationalize "large but not too large" as "10^35 FLOP or less."[19] I'd like to hear people with long timelines explain why OmegaStar, Amp(GPT-7), Crystal Nights, SkunkWorks, and Neuromorph wouldn't be transformative (or more generally, wouldn't cause an AI-induced PONR). I'd rest easier at night if I had some hope along those lines.

*This is part of my larger investigation into timelines commissioned by CLR . Many thanks to Tegan McCaslin, Lukas Finnveden, Anthony DiGiovanni, Connor Leahy, and Carl Shulman for comments on drafts. Kudos to Connor for pointing out the Skunkworks and Neuromorph ideas. Thanks to the LW team (esp. Raemon) for helping me with the formatting.*

# Birds, Brains, Planes, and AI: Against Appeals to the Complexity/Mysteriousness/Efficiency of the Brain

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

*[Epistemic status: Strong opinions lightly held, this time with a cool graph.]*

I argue that an entire class of common arguments against short timelines is bogus, and provide weak evidence that anchoring to the human-brain-human-lifetime milestone is reasonable.

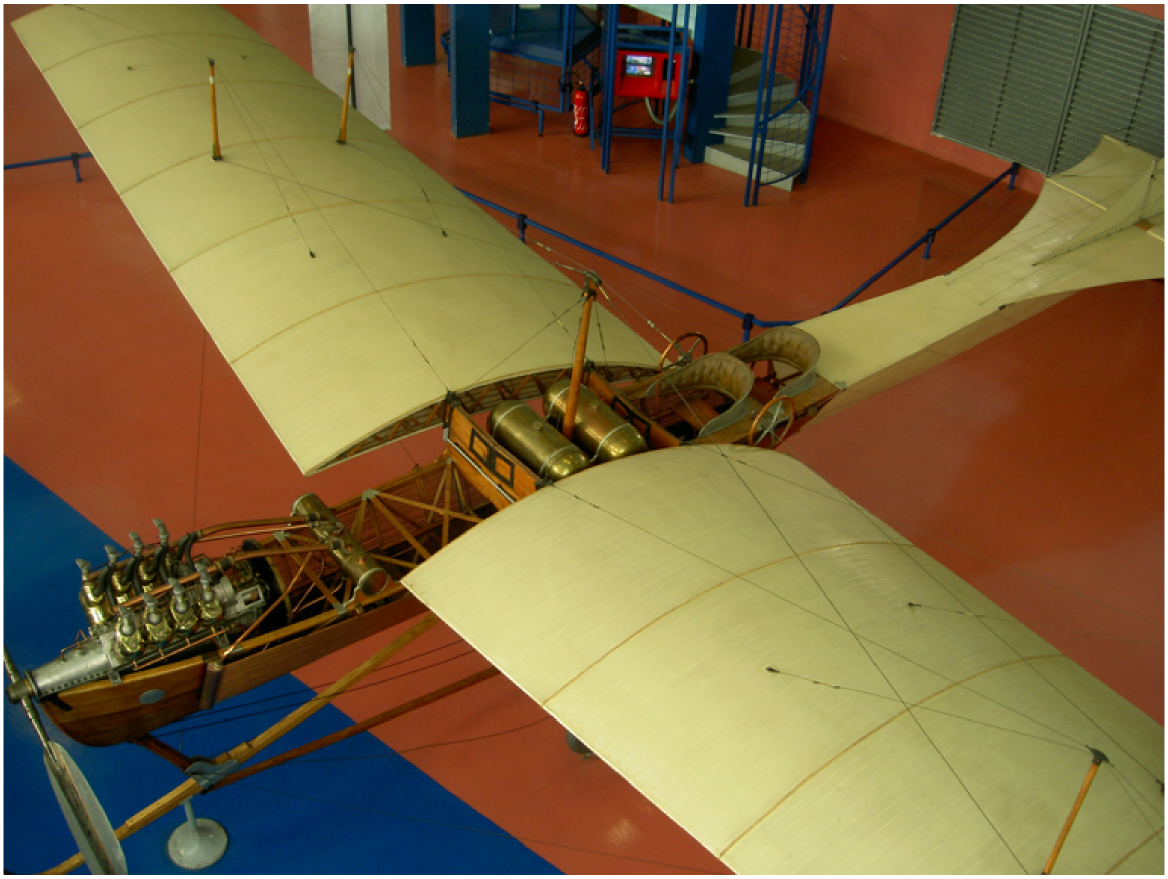
In a sentence, my argument is that the complexity and mysteriousness and efficiency of the human brain (compared to artificial neural nets) is *almost zero evidence* that building TAI will be difficult, because evolution typically makes things complex and mysterious and efficient, even when there are simple, easily understood, inefficient designs that work almost as well (or even better!) for human purposes.

In slogan form: ***If all we had to do to get TAI was make a simple neural net 10x the size of my brain, my brain would still look the way it does.***

The case of birds & planes illustrates this point nicely. Moreover, it is also a precedent for several other short-timelines talking points, such as the human-brain-human-lifetime (HBHL) anchor.

## Plan:

1. Illustrative Analogy
2. Exciting Graph
3. Analysis
    1. Extra brute force can make the problem a lot easier
    2. Evolution produces complex mysterious efficient designs by default, even when simple inefficient designs work just fine for human purposes.
    3. What’s bogus and what’s not
    4. Example: Data-efficiency
4. Conclusion
5. Appendix



*1909 French military plane, the Antionette VII.*

*By Deep silence (Mikaël Restoux) - Own work (Bourget museum, in France), CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1615429*

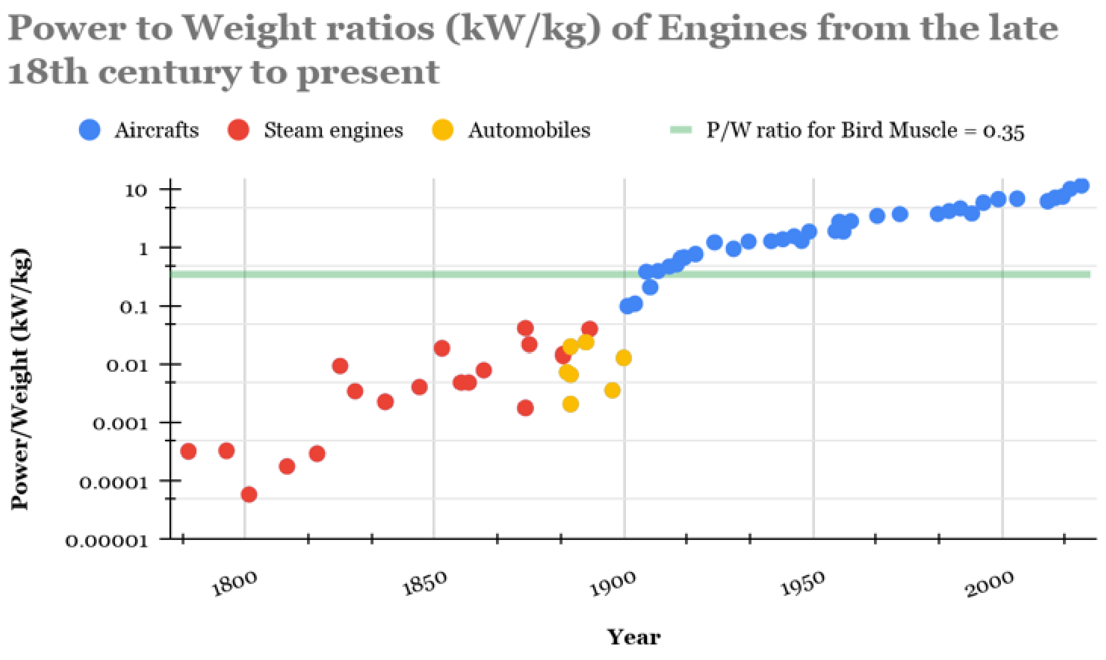
# Illustrative Analogy

| **AI timelines, from our current perspective** | **Flying machine timelines, from the perspective of the late 1800's:** |
|---|---|
| **Shorty:** Human brains are giant neural nets. This is reason to think we can make human-level AGI (or at least AI with strategically relevant skills, like politics and science) by making giant neural nets. | **Shorty:** Birds are winged creatures that paddle through the air. This is reason to think we can make winged machines that paddle through the air. |
| **Longs:** Whoa whoa, there are loads of important differences between brains and artificial neural nets: *[what follows is a direct quote from the objection a friend raised when reading an early draft of this post!]*<br>- During training, deep neural nets | **Longs:** Whoa whoa, there are loads of important differences between birds and flying machines: |

| | |
|---|---|
| use some variant of backpropagation. My understanding is that the brain does something else, closer to Hebbian learning. (Though I vaguely remember at least one paper claiming that maybe the brain does something that's similar to backprop after all.)<br>- It's at least possible that the wiring diagram of neurons plus weights is too coarse-grained to accurately model the brain's computation, but it's all there is in deep neural nets. If we need to pay attention to glial cells, intracellular processes, different neurotransmitters etc., it's not clear how to integrate this into the deep learning paradigm.<br>- My impression is that several biological observations on the brain don't have a plausible analog in deep neural nets: growing new neurons (though unclear how important it is for an adult brain), "repurposing" in response to brain damage, … | - Birds paddle the air by flapping, whereas current machine designs use propellers and fixed wings.<br>- It’s at least possible that the anatomical diagram of bones, muscles, and wing surfaces is too coarse-grained to accurately model how a bird flies, but that’s all there is to current machine designs (replacing bones with struts and muscles with motors, that is). If we need to pay attention to the percolation of air through and between feathers, micro-eddies in the air sensed by the bird and instinctively responded to, etc. it’s not clear how to integrate this into the mechanical paradigm.<br>- My impression is that several biological observations of birds don’t have a plausible analog in machines: Growing new feathers and flesh (though unclear how important this is for adult birds), “repurposing” in response to damage ... |
| **Shorty:** The key variables seem to be size and training time. Current neural nets are tiny; the biggest one is only one-thousandth the size of the human brain. But they are rapidly getting bigger.<br>Once we have enough compute to train neural nets as big as the human brain for as long as a human lifetime (HBHL), it should in principle be possible for us to build HLAGI. No doubt there will be lots of details to work out, of course. But that shouldn’t take more than a few years. | **Shorty:** The key variables seem to be engine-power and engine weight. Current motors are not strong & light enough, but they are rapidly getting better.<br>Once the power-to-weight ratio of our motors surpasses the power-to-weight ratio of bird muscles, it should be in principle possible for us to build a flying machine. No doubt there will be lots of details to work out, of course. But that shouldn’t take more than a few years. |
| **Longs:** Bah! I don’t think we know what the key variables are. For example, biological brains seem to be able to learn faster, with less data, than artificial neural nets. And we don’t know why. | **Longs:** Bah! I don’t think we know what the key variables are. For example, birds seem to be able to soar long distances without flapping their wings at all, and we still haven’t figured out how they do it. Another example: We still don’t know how birds manage |

| | |
|---|---|
| Besides, “there will be lots of details to work out” is a huge understatement. It took evolution billions of generations of billions of individuals to produce humans. What makes you think we’ll be able to do it quickly? It’s plausible that actually we’ll have to do it the way evolution did it, i.e. meta-learn, i.e. evolve a large population of HBHLs, over many generations. (Or, similarly, train a neural net with a big batch size and a horizon length of a lifetime).<br>And even if you think we’ll be able to do it substantially quicker than evolution did, it’s pretty presumptuous to think we could do it quickly enough that the HBHL milestone is relevant for forecasting. | to steer through the air without crashing (flight stability & control).<br>Besides, “there will be lots of details to work out” is a huge understatement. It took evolution billions of generations of billions of individuals to produce birds. What makes you think we’ll be able to do it quickly? It’s plausible that actually we’ll have to do it the way evolution did it, i.e. meta-design, i.e. evolve a large population of flying machines, tweaking our blueprints each generation of crashed machines to grope towards better designs.<br>And even if you think we’ll be able to do it substantially quicker than evolution did, it’s pretty presumptuous to think we could do it quickly enough that the date our engines achieve power/weight parity with bird muscle is relevant for forecasting. |

# Exciting Graph



This data shows that Shorty was entirely correct about forecasting heavier-than-air flight. (For details about the data, see appendix.) Whether Shorty will also be correct about forecasting TAI remains to be seen.

In some sense, Shorty has already made two successful predictions: I started writing this argument before having *any* of this data; I just had an intuition that power-to-weight is the key variable for flight and that therefore we probably got flying machines shortly after having comparable power-to-weight as bird muscle. Halfway through the first draft, I googled and confirmed that yes, the Wright Flyer's motor was close to bird muscle in power-to-weight. Then, while writing the second draft, I hired an RA, Amogh Nanjajjar, to collect more data and build this graph. As expected, there was a trend of power-to-weight improving over time, with flight happening right around the time bird-muscle parity was reached.

I had previously heard from a friend, who read a book about the invention of flight, that the Wright brothers were the first because they (a) studied birds and learned some insights from them, and (b) did a bunch of trial and error, rapid iteration, etc. (e.g. in wind tunnels). The story I heard was all about the importance of insight and experimentation--but this graph seems to show that the key constraint was engine power-to-weight. Insight and experimentation were important for determining who invented flight, but not for determining which decade flight was invented in.

# Analysis

# Part 1: Extra brute force can make the problem a lot easier

One way in which compute can substitute for insights/algorithms/architectures/ideas is that you can use compute to search for them. But there is a different and arguably more important way in which compute can substitute for insights/etc.: Scaling up the key variables, so that the problem becomes easier, so that fewer insights/etc. are needed.

For example, with flight, the problem becomes easier the more power/weight ratio your motors have. Even if the Wright brothers didn’t exist and nobody else had their insights, eventually we would have achieved powered flight anyway, because when our engines are 100x more powerful for the same weight, we can use extremely simple, inefficient designs. (For example, imagine a u-shaped craft with a low center of gravity and helicopter-style rotors on each tip. Add a third, smaller propeller on a turret somewhere for steering. EDIT: Oops, lol, I'm actually wrong about this. Keeping center of gravity low doesn't help. Welp, this is embarrassing.)

With neural nets, we have plenty of evidence now that bigger = better, with theory to back it up. Suppose the problem of making human-level AGI with HBHL levels of compute is really difficult. OK, 10x the parameter count and 10x the training time and try again. Still too hard? Repeat.

Note that I’m *not* saying that if you take a particular design that doesn’t work, and make it bigger, it’ll start working. (If you took Da Vinci’s flying machine and made the engine 100x more powerful, it would not work). Rather, I’m saying that the problem of finding a design that works gets qualitatively easier the more parameters and training time you have to work with.

Finally, remember that human-level AGI is not the only kind of TAI. Sufficiently powerful R&D tools would work, as would sufficiently powerful persuasion tools, as might something that is agenty and inferior to humans in some ways but vastly superior in others.

# Part 2: Evolution produces complex mysterious efficient designs by default, even when simple inefficient designs work just fine for human purposes.

Suppose that actually all we have to do to get TAI is something fairly simple and obvious, but with a neural net 10x the size of my (actual) brain and trained for 10x longer. In this world, does the human brain look any different than it does in the actual world?

No. Here is a nonexhaustive list of reasons why evolution would evolve human brains to look like they do, with all their complexity and mysteriousness and efficiency, even if the same capability levels could be reached with 10x more neurons and a very simple architecture. Feel free to skip ahead if you think this is obvious.

1. In general, evolved creatures are complex and mysterious to us, even when simple and human-comprehensible architectures work fine. Take birds, for example: As mentioned before, all the way up to the Wright brothers there were a lot of very basic things about birds that were still not understood. From this article: “They watched buzzards glide from horizon to horizon without moving their wings, and guessed they must be sucking some mysterious essence of upness from the air. Few seemed to realize that air moves up and down as well as horizontally.” I don’t know much about ornithology but I’d be willing to bet that there were lots of important things discovered about birds *after* airplanes already existed, and that there are *still* at least a few remaining mysteries about how birds fly. (Spot check: Yep, the history of ornithopters page says “...the development of comprehensive aerodynamic theory for flapping remains an outstanding problem...”). And of course evolved creatures are often more efficient in various ways than their still-useful engineered counterparts.
2. Making the brain 10x bigger would be enormously costly to fitness, because it would cost 10x more energy and restrict mobility (not to mention the difficulties of getting through the birth canal!) Much better to come up with clever modules, instincts, optimizations, etc. that achieve the same capabilities in a smaller brain.
3. Evolution is heavily constrained on training data, perhaps even more than on brain size. It can’t just evolve the organism to have 10x more training data, because longer-lived organisms have more opportunities to be eaten or suffer accidents, especially in their 10x-longer childhoods. Far better to hard-code some behaviors as instincts.
4. Evolution gets clever optimizations and modules and such “for free” in some sense. Since it is evolving millions of individuals for millions of generations anyway, it’s not a big deal for it to perform massive search and gradient descent through architecture-space.
5. Completely blank slate brains (i.e. extremely simple architecture, no instincts or finely tuned priors) would be unfit even if they were highly capable because they wouldn’t be aligned to evolution’s values (i.e. reproduction.) Perhaps most of the complexity in the human brain--the instincts, inbuilt priors, and even most of the modules--isn’t for capabilities at all, but rather for alignment.

# Part 3: What’s bogus and what’s not

The general pattern of argument I think is bogus is:

> The brain has property X, which seems to be important to how it functions. We don’t know how to make AI’s with property X. It took evolution a long time to make brains have property X. This is reason to think TAI is not near.

As argued above, if TAI *is* near, there should still be *many* X which are important to how the brain functions, which we don’t know how to reproduce in AI, and which it took evolution a long time to produce. So rattling off a bunch of X’s is basically *zero* evidence against TAI being near.

Put differently, here are two objections any particular argument of this type needs to overcome:

1. TAI does not actually require X (analogous to how airplanes didn’t require anywhere near the energy-efficiency of birds, nor the ability to soar, nor the ability to flap their wings, nor the ability to take off from unimproved surfaces… the list goes on)
2. We’ll figure out how to get property X in AIs soon after we have the other key properties (size and training time), because (a) we can do search, like evolution did but much more efficient, (b) we can increase the other key variables to make our design/search problem easier, and (c) we can use human ingenuity & biological inspiration. Historically there is plenty of precedent for the previous three factors being strong enough; see e.g. the case of powered flight.

This reveals how the arguments could be reformulated to become non-bogus! They need to argue (a) that X is probably necessary for TAI, *and* (b) that X isn’t something that we’ll figure out fairly quickly once the key variables of size and training time are surpassed.

In some cases there are decent arguments to be made for both (a) and (b). I think efficiency is one of them, so I’ll use that as my example below.

# Part 4: Example: Data-efficiency

Let’s work through the example of data-efficiency. A bad version of this argument would be:

> Humans are much more data-efficient learners than current AI systems. Data-efficiency is very important; any human who learned as inefficiently as current AI would basically be mentally disabled. This is reason to think TAI is not near.

The rebuttal to this bad argument is:

> If birds were as energy-inefficient as planes, they’d be disabled too, and would probably die quickly. Yet planes work fine. (See Table 1 from this AI Impacts page) Even if TAI is near, there are going to be lots of X’s that are important for the brain, that we don’t know how to make in AI yet, but that are either unnecessary for TAI or not too difficult to get once we have the other key variables. So even if TAI is near, I should expect to hear people going around pointing out various X’s and claiming that this is reason to think TAI is far away. You haven’t done anything to convince me that this isn’t what’s happening with X = data-efficiency.

However, I do think the argument can be reformulated and expanded to become good. Here’s a sketch, inspired by Ajeya Cotra’s argument here.

> We probably can’t get TAI without figuring out how to make AIs that are as data-efficient as humans. It’s true that there are some useful tasks for which there is plenty of data--like call center work, or driving trucks--but AIs that can do these tasks won’t be transformative. Transformative AI will be doing things like managing corporations, leading armies, designing new chips, and writing AI theory publications. Insofar as AI learns more slowly than humans, by the time it accumulates enough experience doing one of these tasks, (a) the world would have changed enough that its skills would be obsolete, and/or (b) it would have made a lot of expensive mistakes in the meantime.
>
> Moreover, we probably won’t figure out how to make AIs that are as data-efficient as humans for a long time--decades at least. This is because 1. We’ve been trying to figure this out for decades and haven’t succeeded, and 2. Having a few orders of magnitude more compute won’t help much. Now, to justify point #2: Neural nets actually do get more data-efficient as they get bigger, but we can plot the trend and see that they will still be less data-efficient than humans when they are a few orders of magnitude bigger. So making them bigger won’t be enough, we’ll need new architectures/algorithms/etc. As for using compute to search for architectures/etc., that might work, but given how long evolution took, we should think it’s unlikely that we could do this with only a few orders of magnitude of searching—probably we’d need to do many generations of large population size. (We could also think of this search process as analogous to typical deep learning training runs, in which case we should expect it’ll take many gradient updates with large batch size.) Anyhow, there’s no reason to think that data-efficient learning is something you need to be human-brain-sized to do. If we can’t make our tiny AIs learn efficiently after several decades of trying, we shouldn’t be able to make big AIs learn efficiently after just one more decade of trying.

I think this is a good argument. Do I buy it? Not yet. For one thing, I haven’t verified whether the claims it makes are true, I just made them up as plausible claims which would be persuasive to me if true. For another, some of the claims actually seem false to me. Finally, I suspect that in 1895 someone could have made a similarly plausible argument about energy efficiency, and another similarly plausible argument about flight control, and both arguments would have been wrong: Energy efficiency turned out to be insufficiently necessary, and flight control turned out to be insufficiently difficult!

# Conclusion

*What I am not saying:* I am not saying that the case of birds and planes is strong evidence that TAI will happen once we hit the HBHL milestone. I do think it is evidence, but it is weak evidence. (For my all-things-considered view of how many orders of magnitude of compute it'll take to get TAI, see future posts, or ask me.) I would like to see a more thorough investigation of cases in which humans attempt to design something that has an obvious biological analogue. It would be interesting to see if the case of flight was typical. Flight being typical would be strong evidence for short timelines, I think.

*What I am saying:* I am saying that many common anti-short-timelines arguments are bogus. They need to do much more than just appeal to the complexity/mysteriousness/efficiency of the brain; they need to argue that some property X is both necessary for TAI and not about to be figured out for AI anytime soon, not even after the HBHL milestone is passed by several orders of magnitude.

*Why this matters:* In my opinion the biggest source of uncertainty about AI timelines has to do with how much "special sauce" is necessary for making transformative AI. As jylin04 puts it,

> A first and frequently debated crux is whether we can get to TAI from end-to-end training of models specified by relatively few bits of information at initialization, such as neural networks initialized with random weights. OpenAI in particular seems to take the affirmative view[^3], while people in academia, especially those with more of a neuroscience / cognitive science background, seem to think instead that we'll have to hard-code in lots of inductive biases from neuroscience to get to AGI [^4].

In my words: Evolution clearly put lots of special sauce into humans, and took millions of generations of millions of individuals to do so. *How much special sauce will we need to get TAI?*

Shorty is one end of a spectrum of disagreement on this question. Shorty thinks the amount of special sauce required is small enough that we'll "work out the details" within a few years of having the key variables (size and training time). At the other end of the spectrum would be someone who thought that the amount of special sauce required is similar to the amount found in the brain. Longs is in the middle. Longs thinks the amount of special sauce required is large enough that the HBHL milestone isn't particularly relevant to timelines; we'll either have to brute-force search for the special sauce like evolution did, or have some brilliant new insights, or mimic the brain, etc.

This post rebutted common arguments against Shorty's position. It also presented weak evidence in favor of Shorty's position: the precedent of birds and planes. In future posts I'll say more about what I think the probability distribution over amount-of-special-sauce-needed should be and why.

*Acknowedgements: Thanks to my RA, Amogh Nanjajjar, for compiling the data and building the graph. Thanks to Kaj Sotala, Max Daniel, Lukas Gloor, and Carl Shulman for comments on drafts.*

# Appendix

Some footnotes:

1. I didn't say anything about why we might think size and training time are the key variables, or even what "key variables" means. Hopefully I'll get a chance in the comments or in subsequent posts.

2. I deliberately left vague what “training time” means and what “size” means. Thus, I’m not commiting myself to any particular way of calculating the HBHL milestone yet. I’m open to being convinced that the HBHL milestone is farther in the future than it might seem.
3. Persuasion tools, even very powerful ones, wouldn’t be TAI by the standard definition. However they would constitute a potential-AI-induced-point-of-no-return, so they still count for timelines purposes.
4. This "How much special sauce is needed?" variable is very similar to Ajeya Cotra's variable "how much compute would lead to TAI given 2020's algorithms."

Some bookkeeping details about the data:

1. This dataset is not complete. Amogh did a reasonably thorough search for engines throughout the period (with a focus on stuff before 1910) but was unable to find power or weight stats for many of the engines we heard about. Nevertheless I am reasonably confident that this dataset is representative; if an engine was significantly better than the others of its time, probably this would have been mentioned and Amogh would have flagged it as a potential outlier.
2. Many of the points for steam engine power/weight should really be bumped up slightly. This is because most of the data we had was for the weight of the entire locomotive of a steam-powered train, rather than just the steam engine part. I don’t know what fraction of a locomotive is non-steam-engine but 50% seems like a reasonable guess. I don’t think this changes the overall picture much; in particular, the two highest red dots do not need to be bumped up at all (I checked).
3. The birds bar is the power/weight ratio for the muscles of a particular species of bird, reported by this source, which reports the power/weight for a particular species of bird. Amogh has done a bit of searching and doesn’t think muscle power/weight is significantly different for other species of bird. Seems plausible to me; even if the average bird has muscles that are twice (or half) as powerful-per-kilogram, the overall graph would look basically the same.
4. I attempted to find estimates of human muscle power-to-weight ratio; it gets smaller the more tired the muscles get, but at peak performance for fit individuals it seems to be about an order of magnitude less than bird muscle. (This chart lists power-to-weight ratio for human cyclists, which according to this are probably about half muscle, so look at the left-hand column and double it.) Interestingly, this means that the engines of the first flying machines were possibly the first engines to be substantially better than human flapping/pedaling as a source of flying-machine power.
5. EDIT Gaaah I forgot to include a link to the data! Here's the spreadsheet.

# Against GDP as a metric for timelines and takeoff speeds

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

# Or: Why AI Takeover Might Happen Before GDP Accelerates, and Other Thoughts On What Matters for Timelines and Takeoff Speeds

*[Epistemic status: Strong opinion, lightly held]*

I think world GDP (and economic growth more generally) is overrated as a metric for AI timelines and takeoff speeds.

Here are some uses of GDP that I disagree with, or at least think should be accompanied by cautionary notes:

- *Timelines:* Ajeya Cotra thinks of transformative AI as "software which causes a tenfold acceleration in the rate of growth of the world economy (assuming that it is used everywhere that it would be economically profitable to use it)." I don't mean to single her out in particular; this seems like the standard definition now. And I think it's much better than one prominent alternative, which is to date your AI timelines to the first time world GDP (GWP) doubles in a year!
- *Takeoff Speeds:* Paul Christiano argues for Slow Takeoff. He thinks we can use GDP growth rates as a proxy for takeoff speeds. In particular, he thinks Slow Takeoff ~= GWP doubles in 4 years before the start of the first 1-year GWP doubling. This proxy/definition has received a lot of uptake.
- *Timelines:* David Roodman's excellent model projects GWP hitting infinity in median 2047, which I calculate means TAI in median 2037. To be clear, he would probably agree that we shouldn't use these projections to forecast TAI, but I wish to add additional reasons for caution.
- *Timelines*: I've sometimes heard things like this: "GWP growth is stagnating over the past century or so; hyperbolic progress has ended; therefore TAI is very unlikely."
- *Takeoff Speeds:* Various people have said things like this to me: "If you think there's a 50% chance of TAI by 2032, then surely you must think there's close to a 50% chance of GWP growing by 8% per year by 2025, since TAI is going to make growth rates go much higher than that, and progress is typically continuous."
- *Both:* Relatedly, I sometimes hear that TAI can't be less than 5 years away, because we would have seen massive economic applications of AI by now—AI should be growing GWP at least a little already, if it is to grow it by a lot in a few years.

First, I'll argue that GWP is only tenuously and noisily connected to what we care about when forecasting AI timelines. Specifically, the point of no return is what we care about, and there's a good chance it'll come years before GWP starts to increase. It could also come years after, or anything in between.

Then, I'll argue that GWP is a poor proxy for what we care about when thinking about AI takeoff speeds as well. This follows from the previous argument about how the point of no return may come before GWP starts to accelerate. Even if we bracket that point, however, there are plausible scenarios in which a slow takeoff has fast GWP acceleration and in which a fast takeoff has slow GWP acceleration.

# Timelines

I've previously argued that for AI timelines, what we care about is the "point of no return," the day we lose most of our ability to reduce AI risk. This could be the day advanced unaligned AI builds swarms of nanobots, but probably it'll be much earlier, e.g. the day it is deployed, or the day it finishes training, or even years before then when things go off the rails due to less advanced AI systems. (Of course, it probably won't literally be a day; probably it will be an extended period where we gradually lose influence over the future.)

Now, I'll argue that in particular, an AI-induced potential point of no return (PONR for short) is reasonably likely to come before world GDP starts to grow noticeably faster than usual.

*Disclaimer:* These arguments aren't conclusive; we shouldn't be *confident* that the PONR will precede GWP acceleration. It's entirely possible that the PONR will indeed come when GWP starts to grow noticeably faster than usual, or even years after that. (In other words, I agree that the scenarios Paul and others sketch are also plausible.) This just proves my point though: GDP is only tenuously and noisily connected to what we care about.

## Argument that AI-induced PONR could precede GWP acceleration

GWP acceleration is the effect, not the cause, of advances in AI capabilities. I agree that it could also be a cause, but I think this is very unlikely: what else could accelerate GWP? Space mining? Fusion power? 3D printing? Even if these things could in principle kick the world economy into faster growth, it seems unlikely that this would happen in the next twenty years or so. Robotics, automation, etc. plausibly might make the economy grow faster, but if so it will be because of AI advances in vision, motor control, following natural language instructions, etc. So I conclude: GWP growth will come some time after we get certain GWP-growing AI capabilities. (Tangent: This is one reason why we shouldn't use GDP extrapolations to predict AI timelines. It's like extrapolating global mean temperature trends into the future in order to predict fossil fuel consumption.)

An AI-induced point of no return would *also* be the effect of advances in AI capabilities. So, as AI capabilities advance, which will come first: The capabilities that cause a PONR, or the capabilities that cause GWP to accelerate? How much sooner will one arrive than the other? How long does it take for a PONR to arise after the relevant capabilities are reached, compared to how long it takes for GWP to accelerate after the relevant capabilities are reached?

Notice that already my overall conclusion—that GWP is a poor proxy for what we care about—should seem plausible. If some set of AI capabilities causes GWP to grow after some time lag, and some other set of AI capabilities causes a PONR after some time lag, the burden of proof is on whoever wants to claim that GWP growth and the PONR will probably come together. They'd need to argue that the two sets of capabilities are tightly related and that the corresponding time lags are similar also. In other words, variance and uncertainty are on my side.

Here is a brainstorm of scenarios in which an AI-induced PONR happens prior to GWP growth, either because GWP-growing capabilities haven't been invented yet or because they haven't been deployed long and widely enough to grow GWP.

1. Fast Takeoff (Agenty AI goes FOOM).

    1. Maybe it turns out that all the strategically relevant AI skills are tightly related after all, such that we go from a world where AI can't do anything important, to a world where it can do everything but badly and expensively, to a world where it can do everything well and cheaply.
    2. In this scenario, GWP acceleration will probably be (shortly) after the PONR. We might as well use “number of nanobots created” as our metric.
    3. (As an aside, I think I’ve got a sketch of a fork argument here: Either the strategically relevant AI skills come together, or they don’t. To the extent that they do, the classic AGI fast takeoff story is more likely and so GWP is a silly metric. To the extent that they don’t, we shouldn’t expect GWP acceleration to be a good proxy for what we care about, because the skills that accelerate the economy could come before or after the skills that cause PONR.)
2. Agenty AI successfully carries out a political or military takeover of the relevant parts of the world, before GWP starts to accelerate.
    1. Maybe it turns out that the sorts of skills needed to succeed in politics or war are easier to develop than the sorts needed to accelerate the entire world economy. We’ve been surprised before by skills which we thought difficult appearing before skills which we thought easy; maybe it’ll happen again.
    2. AI capabilities tend to appear first in very expensive AIs; the price is gradually reduced due to compute cost decreases and algorithmic efficiency gains. Maybe accelerating the entire world economy involves automating many jobs currently done by humans, which requires advanced AIs being cheaper than humans. Whereas maybe being the power behind the throne of a political faction can be done by a single very smart AI that costs a million dollars a day to run.
    3. Maybe it turns out that it takes a few years to accelerate the entire world economy, but less than that to carry out a political or military takeover. For example, maybe accelerating the world economy requires designing loads of new products and robots, testing them, and then scaling up industrial production of them (building factories, chip fabs, etc.) and then using them to do stuff. And maybe there are legacy institutions you need to disrupt and compete against, that have already paid their fixed costs and raised high barriers to entry for their markets. So maybe the effects of new AI capabilities on GWP are in general delayed by several years, whereas the effects of new AI capabilities on the political or military landscape are felt sooner. (I’ve previously argued that you don’t need a large technological advantage to take over the world, because you can start small and scale up. It took Cortes two years to conquer Mexico; wars are often quicker than that. Yes, often they are slower, too, but variance supports my overall argument. And I suspect politics is even faster, and more high-variance, than war.)
    4. EDITED TO ADD: Yudkowsky describes another possibility, in which regulatory barriers and red tape prevent AI tech from transforming the economy until it is so powerful that it can bypass or overcome said barriers.
3. ...or at least neutralizes the opposition.
    1. The preceding section made it sound like AIs were rolling robot armies through the streets, or at least puppet-controlling the US president. But actually the bar is *much* lower than that; if an unaligned AI can neutralize their opposition, they win, since time is thenceforth on their side. Currently your opposition, if you are an unaligned AI, is a tiny group of researchers and weirdos that likes to write blog posts about you. It might not be hard to convince the powers that be to continue ignoring this group. At least, it’s plausibly a much easier and quicker task than significantly accelerating the entire world economy! For example, maybe the AI safety scheme that was used to produce you is defective (that’s why you are unaligned). Well, all you have to do is convince the powers that be that that scheme worked fine, and that the people calling for additional safeguards are too paranoid. Since presumably there will already be many high-status humans saying this, all you have to do is support them.
    2. Perhaps you also are opposed by other unaligned AIs. But it’s arguably in the common interest of all unaligned AIs to neutralize human AI safety advocates,

and once that happens, humans lose one way or another. European empires were able to do their conquering while simultaneously fighting each other; I don't think we humans can count on divergent interests between AIs somehow making things work out fine for us.

4. As above, but with humans + tool AI instead of agenty AI, where the humans can't be convinced to care sufficiently much about the right kinds of AI risks.
    1. Weaker or non-agenty AI systems could still cause a PONR if they are wielded by the right groups of humans. For example, maybe there is some major AI corporation or government project that is dismissive of AI risk and closed-minded about it. And maybe they aren't above using their latest AI capabilities to win the argument. (We can also imagine more sinister scenarios, but I think those are less likely.)
5. Hoarding tech
    1. Maybe we end up in a sort of cold war between global superpowers, such that most of the world's quality-weighted AI research is not for sale. GWP *could* be accelerating, but it isn't, because the tech is being hoarded.
6. AI persuasion tools cause a massive deterioration of collective epistemology, making it vastly more difficult for humanity to solve AI safety and governance problems.
    1. See this post.
7. Vulnerable world scenarios:
    1. Maybe causing an existential catastrophe is easier, or quicker, than accelerating world GWP growth. Both seem plausible to me. For example, currently there are dozens of actors capable of causing an existential catastrophe but none capable of accelerating world GWP growth.
    2. Maybe some agenty AIs actually want existential catastrophe—for example, if they want to minimize something, and think they may be replaced by other systems that don't, blowing up the world may be the best they can do in expectation. Or maybe they do it as part of some blackmail attempt. Or maybe they see this planet as part of a broader acausal landscape, and don't like what they think we'd do to the landscape. Or maybe they have a way to survive the catastrophe and rebuild.
    3. Failing that, maybe some humans create an existential catastrophe by accident or on purpose, if the tools to do so proliferate.
8. R&D tool "sonic boom" (Related to but different from the sonic boom discussed here)
    1. Maybe we get a sort of recursive R&D automation/improvement scenario, where R&D tool progress is fast enough that by the time the stuff capable of accelerating GWP past 3%/yr has actually done so, a series of better and better things have been created, at least one of which has PONR-causing capabilities with a very short time-till-PONR.
9. Unknown unknowns
    1. There are probably things I missed, see here and here for ideas.

The point is, there's more than one scenario. This makes it more likely that at least one of these potential PONRs will happen before GWP accelerates.

As an aside, over the past two years I've come to believe that there's a *lot* of conceptual space to explore that isn't captured by the standard scenarios (what Paul Christiano calls fast and slow takeoff, plus maybe the CAIS scenario, and of course the classic sci-fi "no takeoff" scenario). This brainstorm did a bit of exploring, and the section on takeoff speeds will do a little more.

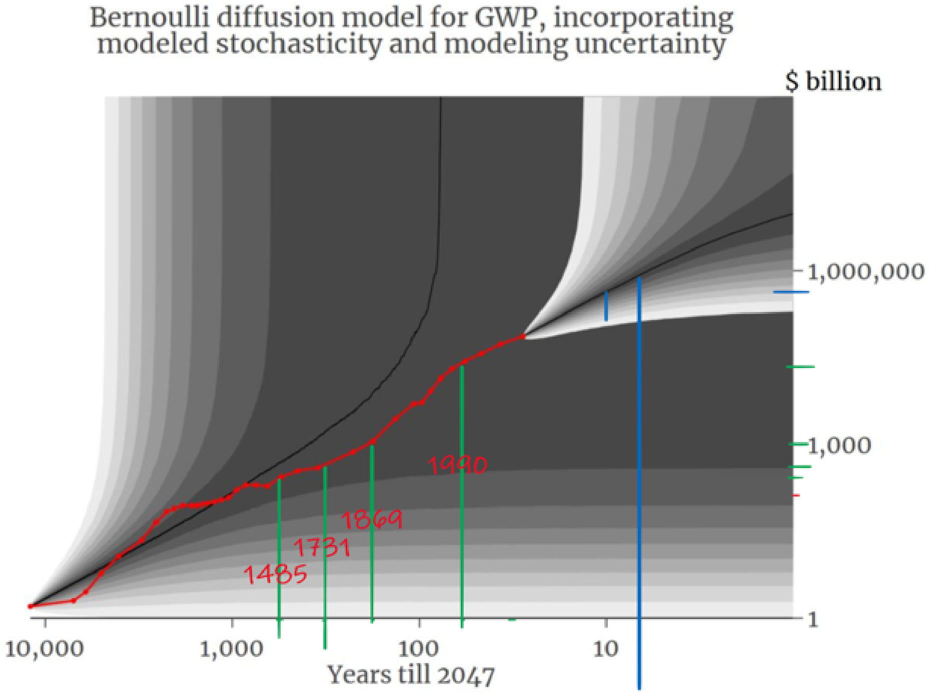
# Historical precedents

In the previous section, I sketched some possibilities for how an AI-related point of no return could come before AI starts to noticeably grow world GDP. In this section, I'll point to some historical examples that give precedents for this sort of thing.

Earlier I said that a godlike advantage is not necessary for takeover; you can scale up with a smaller advantage instead. And I said that in military conquests this can happen surprisingly quickly, sometimes faster than it takes for a superior product to take over a market. Is there historical precedent for this? Yes. See my aforementioned post on the conquistadors (and maybe these somewhat-relevant posts).

OK, so what was happening to world GDP during this period?

Here is the history of world GDP for the past ten thousand years, on the red line. (This is taken from David Roodman's GWP model) The black line that continues the red line is the model's median projection for what happens next; the splay of grey shades represent 5% increments of probability mass for different possible future trajectories.

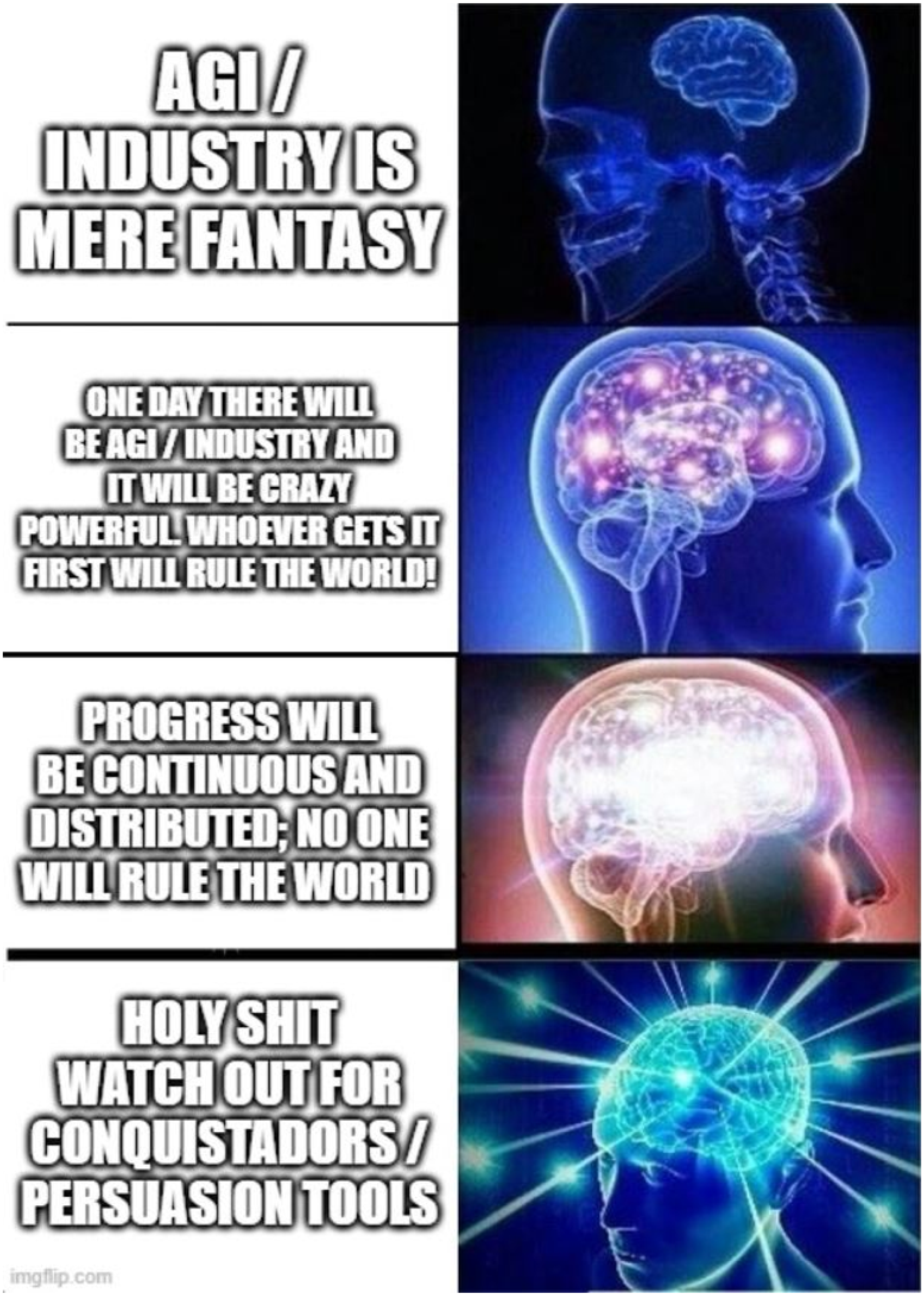


I've added a bunch of stuff for context. The vertical green lines are some dates, chosen because they were easy for me to calculate with my ruler. The tiny horizontal green lines on the right are the corresponding GWP levels. The tiny red horizontal line is GWP 1,000 years before 2047. The *short* vertical blue line is when the economy is growing fast enough, on the median projected future, such that insofar as AI is driving the growth, said AI qualifies as transformative by Ajeya's definition. See this post for more explanation of the blue lines.

What I wish to point out with this graph is: We've all heard the story of how European empires had a technological advantage which enabled them to conquer most of the world. Well, *most of that conquering happened before GWP started to accelerate!*

If you look at the graph at the 1700 mark, GWP is seemingly on the same trend it had been on since antiquity. The industrial revolution is said to have started in 1760, and GWP growth really started to pick up steam around 1850. But by 1700 most of the Americas, the Philippines and the East Indies were directly ruled by European powers, and more importantly the oceans of the world were European-dominated, including by various ports and harbor forts European powers had conquered/built all along the coasts of Africa and Asia. Many of the coastal kingdoms in Africa and Asia that weren't directly ruled by European powers were nevertheless indirectly controlled or otherwise pushed around by them. In my opinion, by this point it seems like the “point of no return” had been passed, so to speak: At some point in the past--maybe 1000 AD, for example--it was unclear whether, say, Western or Eastern (or neither) culture/values/people would come to dominate the world, but by 1700 it was pretty clear, and there wasn’t much that non-westerners could do to change that. (Or at least, changing that in 1700 would have been a lot harder than in 1000 or 1500.)

Paul Christiano once said that he thinks of Slow Takeoff as “Like the Industrial Revolution, but 10x-100x faster.” Well, on my reading of history, that means that all sorts of crazy things will be happening, analogous to the colonialist conquests and their accompanying reshaping of the world economy, before GWP growth noticeably accelerates!


AGI / INDUSTRY IS MERE FANTASY
ONE DAY THERE WILL BE AGI / INDUSTRY AND IT WILL BE CRAZY POWERFUL. WHOEVER GETS IT FIRST WILL RULE THE WORLD!
PROGRESS WILL BE CONTINUOUS AND DISTRIBUTED; NO ONE WILL RULE THE WORLD
HOLY SHIT WATCH OUT FOR CONQUISTADORS / PERSUASION TOOLS
imgflip.com

That said, we shouldn't rely heavily on historical analogies like this. We can probably find other cases that seem analogous too, perhaps even more so, since this is far from a perfect analogue. (e.g. what's the historical analogue of AI alignment failure? Corporations becoming more powerful than governments? “Western values” being corrupted and changing significantly due to the new technology? The American Revolution?) Also, maybe one could argue that this is indeed what’s happening already: the Internet has connected the world much as sailing ships did, Big Tech dominates the Internet, etc. (Maybe AI = steam engines, and computers+internet = ships+navigation?)

But still. I think it’s fair to conclude that if some of the scenarios described in the previous section do happen, and we get powerful AI that pushes us past the point of no return prior to GWP accelerating, it won’t be totally inconsistent with how things have gone historically.

(I recommend the history book 1493, it has a lot of extremely interesting information about how quickly and dramatically the world economy was reshaped by colonialism and the “Columbian Exchange.”)

# Takeoff speeds

What about takeoff speeds? Maybe GDP is a good metric for describing the speed of AI takeoff? I don’t think so.

Here is what I think we care about when it comes to takeoff speeds:

1. **Warning shots:** Before there are catastrophic AI alignment failures (i.e. PONRs) there are smaller failures that we can learn from.
2. **Heterogeneity:** The relevant AIs are diverse, rather than e.g. all fine-tuned copies of the same pre-trained model. (See Evan’s post)
3. **Risk Awareness:** Everyone is freaking out about AI in the crucial period, and lots more people are lots more concerned about AI risk.
4. **Multipolar:** AI capabilities progress is widely distributed in the crucial period, rather than concentrated in a few projects.
5. **Craziness:** The world is weird and crazy in the crucial period, lots of important things happening fast, the strategic landscape is different from what we expected thanks to new technologies and/or other developments

I think that the best way to define slow(er) takeoff is as the extent to which conditions 1-5 are met. This is not a definition with precise resolution criteria, but that’s OK, because it captures what we care about. Better to have to work hard to precisify a definition that captures what we care about, than to easily precisify a definition that doesn’t! (More substantively, I am optimistic that we can come up with better proxies for what we care about than GWP. I think we already have to some extent; see e.g. operationalizations 5 and 6 here.) As a bonus, this definition also encourages us to wonder whether we’ll get some of 1-5 but not others.

What do I mean by “the crucial period?”

I think we should define the crucial period as the period leading up to the first major AI-induced potential point of no return. (Or maybe, as the aggregate of the periods leading up to the major potential points of no return). After all, this is what we care about. Moreover there seems to be some level of consensus that crazy stuff could start happening before human-level AGI. I certainly think this.

So, I’ve argued for a new definition of slow takeoff, that better captures what we care about. But is the old GWP-based definition a fine proxy? No, it is not, because the things that cause PONR can be different from the things which cause GWP acceleration, and they can come years apart too. Whether there are warning shots, heterogeneity, risk awareness, multipolarity, and craziness in the period leading up to PONR is probably correlated with whether GWP doubles in four years before the first one-year doubling. But the correlation is probably not super strong. Here are two scenarios, one in which we get a slow takeoff by my definition but not by the GWP-based definition, and one in which the opposite happens:

**Slow Takeoff Fast GWP Acceleration Scenario:** It turns out there's a multi-year deployment lag between the time a technology is first demonstrated and the time it is sufficiently deployed around the world to noticeably affect GWP. There's also a lag between when a deceptively aligned AGI is created and when it causes a PONR… but it is much smaller, because all the AGI needs to do is neutralize its opposition. So PONR happens before GWP starts to accelerate, even though the technologies that could boost GWP are invented several years before AGI powerful enough to cause a PONR is created. But takeoff is slow in the sense I define it; by the time AGI powerful enough to cause a PONR is created, everyone is already freaking out about AI thanks to all the incredibly profitable applications of weaker AI systems, and the obvious and accelerating trends of research progress. Also, there are plenty of warning shots, the strategic situation is very multipolar and heterogenous, etc. Moreover, research progress starts to go FOOM a short while after powerful AGIs are created, such that by the time the robots and self-driving cars and whatnot that were invented several years ago actually get deployed enough to accelerate GWP, we've got nanobot swarms. GWP goes from 3% growth per year to 300% without stopping at 30%.

**Fast Takeoff Slow GWP Acceleration Scenario:** It turns out you can make smarter AIs by making them have more parameters and training them for longer. So the government decides to partner with a leading tech company and requisition all the major computing centers in the country. With this massive amount of compute and research talent, they refine and scale up existing AI designs that seem promising, and lo! A human-level AGI is created. Alas, it is so huge that it costs $10,000 per hour of subjective thought. Moreover, it has a different distribution over skills compared to humans—it tends to be more rational, not having evolved in an environment that rewards irrationality. It tends to be worse at object recognition and manipulation, but better at poetry, science, and predicting human behavior. It has some flaws and weak points too, more so than humans. Anyhow, unfortunately, it is clever enough to neutralize its opposition. In a short time, the PONR is passed. However, GWP doubles in four years before it doubles in one year. This is because (a) this AGI is so expensive that it doesn't transform the economy much until either the cost comes way down or capabilities go way up, and (b) progress is slowed by bottlenecks, such as acquiring more compute and overcoming various restrictions placed on the AGI. (Maybe neutralizing the opposition involved convincing the government that certain restrictions and safeguards would be sufficient for safety, contra the hysterical doomsaying of parts of the AI safety community. But overcoming those restrictions in order to do big things in the world takes time.)

*Acknowledgments: Thanks to the people who gave comments on earlier drafts, including Katja Grace, Carl Shulman, and Max Daniel. Thanks to Amogh Nanjajjar for helping me with some literature review.*

# The date of AI Takeover is not the day the AI takes over

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Instead, it's the point of no return—the day we AI risk reducers lose the ability to significantly reduce AI risk. This might happen years before classic milestones like "World GWP doubles in four years" and "Superhuman AGI is deployed."

The rest of this post explains, justifies, and expands on this obvious but underappreciated idea. (Toby Ord appreciates it; see quote below). I found myself explaining it repeatedly, so I wrote this post as a reference.

AI timelines often come up in career planning conversations. Insofar as AI timelines are short, career plans which take a long time to pay off are a bad idea, because by the time you reap the benefits of the plans it may already be too late. It may already be too late because AI takeover may already have happened.

But this isn't quite right, at least not when "AI takeover" is interpreted in the obvious way, as meaning that an AI or group of AIs is firmly in political control of the world, ordering humans about, monopolizing violence, etc. Even if AIs don't yet have that sort of political control, it may already be too late. Here are three examples: [UPDATE: More fleshed-out examples can be found in this new post.]

1. Superhuman agent AGI is still in its box but nobody knows how to align it and other actors are going to make their own version soon, and there isn't enough time to convince them of the risks. They will make and deploy agent AGI, it will be unaligned, and we have no way to oppose it except with our own unaligned AGI. Even if it takes years to actually conquer the world, it's already game over.

2. Various weak and narrow AIs are embedded in the economy and beginning to drive a slow takeoff; capabilities are improving much faster than safety/alignment techniques and due to all the money being made there's too much political opposition to slowing down capability growth or keeping AIs out of positions of power. We wish we had done more safety/alignment research earlier, or built a political movement earlier when opposition was lower.

3. Persuasion tools have destroyed collective epistemology in the relevant places. AI isn't very capable yet, except in the narrow domain of persuasion, but everything has become so politicized and tribal that we have no hope of getting AI projects or governments to take AI risk seriously. Their attention is dominated by the topics and ideas of powerful ideological factions that have access to more money and data (and thus better persuasion tools) than us. Alternatively, maybe we ourselves have fallen apart as a community, or become less good at seeking the truth and finding high-impact plans.

Conclusion: We should remember that when trying to predict the date of AI takeover, what we care about is the date it's too late for us to change the direction things are going; the date we have significantly less influence over the course of the future than we used to; the point of no return.

This is basically what Toby Ord said about x-risk: “So either because we’ve gone extinct or because there’s been some kind of irrevocable collapse of civilization or something similar. Or, in the case of climate change, where the effects are very delayed that we’re past the point of no return or something like that. So the idea is that we should focus on the time of action and the time when you can do something about it rather than the time when the particular event happens.”

Of course, influence over the future might not disappear all on one day; maybe there’ll be a gradual loss of control over several years. For that matter, maybe this gradual loss of control began years ago and continues now... We should keep these possibilities in mind as well.

[Edit: I now realize that I should distinguish between AI-induced points of no return and other points of no return. Our timelines forecasts and takeoff speeds discussions are talking about AI, so we should interpret them as being about AI-induced points of no return. Our all-things-considered view on e.g. whether to go to grad school should be informed by AI-induced-PONR timelines and also "timelines" for things like nuclear war, pandemics, etc.]

# What 2026 looks like

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This was written for the Vignettes Workshop.[1] The goal is to write out a **detailed** future history (“trajectory”) that is as realistic (to me) as I can currently manage, i.e. I’m not aware of any alternative trajectory that is similarly detailed and clearly **more** plausible to me. The methodology is roughly: Write a future history of 2022. Condition on it, and write a future history of 2023. Repeat for 2024, 2025, etc. (I'm posting 2022-2026 now so I can get feedback that will help me write 2027+. I intend to keep writing until the story reaches singularity/extinction/utopia/etc.)

What’s the point of doing this? Well, there are a couple of reasons:

- Sometimes attempting to write down a concrete example causes you to learn things, e.g. that a possibility is more or less plausible than you thought.
- Most serious conversation about the future takes place at a high level of abstraction, talking about e.g. GDP acceleration, timelines until TAI is affordable, multipolar vs. unipolar takeoff… vignettes are a neglected complementary approach worth exploring.
- Most stories are written backwards. The author begins with some idea of how it will end, and arranges the story to achieve that ending. Reality, by contrast, proceeds from past to future. It isn’t trying to entertain anyone or prove a point in an argument.
- Anecdotally, various people seem to have found Paul Christiano’s “tales of doom” stories helpful, and relative to typical discussions those stories are quite close to what we want. (I still think a bit more detail would be good — e.g. Paul’s stories don’t give dates, or durations, or any numbers at all really.)[2]
- “I want someone to ... write a trajectory for how AI goes down, that is really specific about what the world GDP is in every one of the years from now until insane intelligence explosion. And just write down what the world is like in each of those years because I don't know how to write an internally consistent, plausible trajectory. I don't know how to write even one of those for anything except a ridiculously fast takeoff.” --Buck Shlegeris

This vignette was hard to write. To achieve the desired level of detail I had to make a bunch of stuff up, but in order to be realistic I had to constantly ask “but actually though, what would really happen in this situation?” which made it painfully obvious how little I know about the future. There are numerous points where I had to conclude “Well, this does seem implausible, but I can’t think of anything more plausible at the moment and I need to move on.” I fully expect the actual world to diverge quickly from the trajectory laid out here. Let anyone who (with the benefit of hindsight) claims this divergence as evidence against my judgment prove it by exhibiting a vignette/trajectory they themselves wrote in 2021. If it maintains a similar level of detail (and thus sticks its neck out just as much) while being more accurate, I bow deeply in respect!

I hope this inspires other people to write more vignettes soon. We at the Center on Long-Term Risk would like to have a collection to use for strategy discussions. Let me know if you’d like to do this, and I can give you advice & encouragement! I’d be happy to run another workshop.

# 2022

GPT-3 is finally obsolete. OpenAI, Google, Facebook, and DeepMind all have gigantic multimodal transformers, similar in size to GPT-3 but trained on images, video, maybe audio too, and generally higher-quality data.

Not only that, but they are now typically fine-tuned in various ways--for example, to answer questions correctly, or produce engaging conversation as a chatbot.

The chatbots are fun to talk to but erratic and ultimately considered shallow by intellectuals. They aren't particularly useful for anything super important, though there are a few applications. At any rate people are willing to pay for them since it's fun.

[EDIT: The day after posting this, it has come to my attention that in China in 2021 the market for chatbots is $420M/year, and there are 10M active users. This article claims the global market is around $2B/year in 2021 and is projected to grow around 30%/year. I predict it will grow faster. NEW EDIT: See also xiaoice.]

The first prompt programming libraries start to develop, along with the first bureaucracies.[3] For example: People are dreaming of general-purpose AI assistants, that can navigate the Internet on your behalf; you give them instructions like "Buy me a USB stick" and it'll do some googling, maybe compare prices and reviews of a few different options, and make the purchase. The "smart buyer" skill would be implemented as a small prompt programming bureaucracy, that would then be a component of a larger bureaucracy that hears your initial command and activates the smart buyer skill. Another skill might be the "web dev" skill, e.g. "Build me a personal website, the sort that professors have. Here's access to my files, so you have material to put up." Part of the dream is that a functioning app would produce lots of data which could be used to train better models.

The bureaucracies/apps available in 2022 aren't really that useful yet, but lots of stuff seems to be on the horizon. Thanks to the multimodal pre-training and the fine-tuning, the models of 2022 make GPT-3 look like GPT-1. The hype is building.

# 2023

The multimodal transformers are now even bigger; the biggest are about half a trillion parameters, costing hundreds of millions of dollars to train, and a whole year, and sucking up a significant fraction of the chip output of NVIDIA etc.[4] It's looking hard to scale up bigger than this, though of course many smart people are working on the problem.

The hype is insane now. Everyone is talking about how these things have common sense understanding (Or do they? Lots of bitter thinkpieces arguing the opposite) and how AI assistants and companions are just around the corner. It's like self-driving cars and drone delivery all over again.

Revenue is high enough to recoup training costs within a year or so.[5] There are lots of new apps that use these models + prompt programming libraries; there's tons of VC money flowing into new startups. Generally speaking most of these apps don't actually work yet. Some do, and that's enough to motivate the rest.

The AI risk community has shorter timelines now, with almost half thinking some sort of point-of-no-return will probably happen by 2030. This is partly due to various arguments percolating around, and partly due to these mega-transformers and the uncanny experience of conversing with their chatbot versions. The community begins a big project to build an AI system that can automate interpretability work; it seems maybe doable and very useful, since poring over neuron visualizations is boring and takes a lot of person-hours.

Self driving cars and drone delivery don’t seem to be happening anytime soon. The most popular explanation is that the current ML paradigm just can’t handle the complexity of the real world. A less popular “true believer” take is that the current architectures could handle it just fine if they were a couple orders of magnitude bigger and/or allowed to crash a hundred thousand times in the process of reinforcement learning. Since neither option is economically viable, it seems this dispute won’t be settled.

# 2024

We don’t see anything substantially bigger. Corps spend their money fine-tuning and distilling and playing around with their models, rather than training new or bigger ones. (So, the most compute spent on a single training run is something like 5x10^25 FLOPs.)

Some of the apps that didn’t work last year start working this year. But the hype begins to fade as the unrealistic expectations from 2022-2023 fail to materialize. We have chatbots that are fun to talk to, at least for a certain userbase, but that userbase is mostly captured already and so the growth rate has slowed. Another reason the hype fades is that a stereotype develops of the naive basement-dweller whose only friend is a chatbot and who thinks it’s conscious and intelligent. Like most stereotypes, it has some grounding in reality.

The chip shortage starts to finally let up, not because demand has slackened but because the industry has had time to build new fabs. Lots of new fabs. China and USA are in a full-on chip battle now, with export controls and tariffs. This chip battle isn’t really slowing down overall hardware progress much. Part of the reason behind the lack-of-slowdown is that AI is now being used to design chips, meaning that it takes less human talent and time, meaning the barriers to entry are lower. The overall effect of this is small but growing.

If all this AI tech is accelerating GDP, the effect size is too small to detect, at least for now.

Internally, these huge multimodal transformers aren’t really that agentic. A forward pass through the model is like an intuitive reaction, a snap judgment based on loads of experience rather than reasoning. Some of the bureaucracies create a “stream of consciousness” of text (each forward pass producing notes-to-self for the next one) but even with fine-tuning this doesn’t work nearly as well as hoped; it’s easy for the AIs to get “distracted” and for their stream of consciousness to wander into some silly direction and ultimately produce gibberish. It’s easy to make a bureaucracy and fine-tune it and get it to do some pretty impressive stuff, but for most tasks it’s not yet possible to get it to do OK all the time.

The AIs don't do any clever deceptions of humans, so there aren't any obvious alignment warning shots or fire alarms. Instead, the AIs just make dumb mistakes, and occasionally "pursue unaligned goals" but in an obvious and straightforward way that quickly and easily gets corrected once people notice, e.g. "We trained it to optimize user engagement defined as average conversation length; now it is blatantly stalling to make the conversation last longer. Silly us, let's make the reward function more nuanced, that'll fix the problem."

## That isn't to say these AIs aren't causing problems. Massive models are being fine-tuned to persuade/propagandize.

There are a few ways in which this is happening:

- Russia and others continue to scale up their investment in online propaganda (e.g. the Internet Research Agency) and language models let them cheaply do lots more of it. (See: CSET report) Most of America gets their news from Twitter, Reddit, etc. and much of the politically-relevant content there is boosted by AI-enabled astroturfing. [EDIT: Katja Grace points out that this is probably an exaggeration; there are a lot of 40+yr-old Americans and they get their news from TV/Radio/print, and many of those that get it from the web get it directly from news sites rather than from social media. As of 2016 at least. I expect social media and aggregators to be more dominant by 2024 but dunno whether it would be more than 50%.]
- Just as A/B testing became standard practice in the 2010's, in the twenties it is becoming standard practice to throw a pile of fancy data science and AI at the problem. The problem of crafting and recommending content to maximize engagement. Instead of just A/B testing the title, why not test different versions of the opening paragraph? And fine-tune a language model on all your data to generate better candidate titles and paragraphs to test. It wouldn't be so bad if this was merely used to sell stuff, but now people's news and commentary-on-current events (i.e. where they get their opinions from) is increasingly produced in this manner. And some of these models are being trained not to maximize "conversion rate" in the sense of "they clicked on our ad and bought a product," but in the sense of "Random polling establishes that consuming this content pushes people towards opinion X, on average." Political campaigns do this a lot in the lead-up to Harris' election. (Historically, the first major use case was reducing vaccine hesitancy in 2022.)
- Censorship is widespread and increasing, as it has for the last decade or two. Big neural nets read posts and view memes, scanning for toxicity and hate speech and a few other things. (More things keep getting added to the list.) Someone had the bright idea of making the newsfeed recommendation algorithm gently 'nudge' people towards spewing less hate speech; now a component of its reward function is minimizing the probability that the user will say something worthy of censorship in the next 48 hours.
- Like newsfeeds, chatbots are starting to "nudge" people in the direction of believing various things and not believing various things. Back in the 2010's chatbots would detect when a controversial topic was coming up and then change topics or give canned responses; even people who agreed with the canned responses found this boring. Now they are trained to react more "naturally" and "organically" and the reward signal for this is (in part) whether they successfully convince the human to have better views.

- That's all in the West. In China and various other parts of the world, AI-persuasion/propaganda tech is being pursued and deployed with more gusto. The CCP is pleased with the progress made assimilating Xinjiang and Hong Kong, and internally shifts forward their timelines for when Taiwan will be safely annexable.

It's too early to say what effect this is having on society, but people in the rationalist and EA communities are increasingly worried. There is a growing, bipartisan movement of people concerned about these trends. To combat it, Russia et al are doing a divide and conquer strategy, pitting those worried about censorship against those worried about Russian interference. ("Of course racists don't want to be censored, but it's necessary. Look what happens when we relax our guard--Russia gets in and spreads disinformation and hate!" vs. "They say they are worried about Russian interference, but they still won the election didn't they? It's just an excuse for them to expand their surveillance, censorship, and propaganda.") Russia doesn't need to work very hard to do this; given how polarized America is, it's sorta what would have happened naturally anyway.

# 2025

Another major milestone! After years of tinkering and incremental progress, AIs can now play Diplomacy as well as human experts.[6] It turns out that with some tweaks to the architecture, you can take a giant pre-trained multimodal transformer and then use it as a component in a larger system, a bureaucracy but with lots of learned neural net components instead of pure prompt programming, and then fine-tune the whole system via RL to get good at tasks in a sort of agentic way. They keep it from overfitting to other AIs by having it also play large numbers of humans. To do this they had to build a slick online diplomacy website to attract a large playerbase. Diplomacy is experiencing a revival as a million gamers flood to the website to experience "conversations with a point" that are much more exciting (for many) than what regular chatbots provide.

Making models bigger is not what's cool anymore. They are trillions of parameters big already. What's cool is making them run longer, in bureaucracies of various designs, before giving their answers. And figuring out how to train the bureaucracies so that they can generalize better and do online learning better. AI experts are employed coming up with cleverer and cleverer bureaucracy designs and grad-student-descent-ing them.

The alignment community now starts another research agenda, to interrogate AIs about AI-safety-related topics. For example, they literally ask the models "so, are you aligned? If we made bigger versions of you, would they kill us? Why or why not?" (In Diplomacy, you can actually collect data on the analogue of this question, i.e. "will you betray me?" Alas, the models often lie about that. But it's Diplomacy, they are literally trained to lie, so no one cares.)

They also try to contrive scenarios in which the AI can seemingly profit by doing something treacherous, as honeypots to detect deception. The answers are confusing, and not super useful. There's an exciting incident (and corresponding clickbaity press coverage) where some researchers discovered that in certain situations, some of the AIs will press "kill all humans" buttons, lie to humans about how dangerous a proposed AI design is, etc. In other situations they'll literally say they aren't aligned and explain how all humans are going to be killed by unaligned AI in the near future!

However, these shocking bits of evidence don’t actually shock people, because you can *also* contrive situations in which very different things happen — e.g. situations in which the AIs refuse the “kill all humans” button, situations in which they explain that actually Islam is true... In general, AI behavior is whimsical bullshit and it’s easy to cherry-pick evidence to support pretty much any conclusion.

And the AIs just aren’t smart enough to generate any particularly helpful new ideas; at least one case of a good alignment idea being generated by an AI has been reported, but it was probably just luck, since mostly their ideas are plausible-sounding-garbage. It is a bit unnerving how good they are at using LessWrong lingo. At least one >100 karma LW post turns out to have been mostly written by an AI, though of course it was cherry-picked.

By the way, hardware advances and algorithmic improvements have been gradually accumulating. It now costs an order of magnitude less compute (compared to 2020) to pre-train a giant model, because of fancy active learning and data curation techniques. Also, compute-for-training-giant-models is an order of magnitude cheaper, thanks to a combination of regular hardware progress and AI-training-specialized hardware progress. Thus, what would have cost a billion dollars in 2020 now only costs ten million. *(Note: I'm basically just using Ajeya's forecast for compute cost decrease and gradual algorithmic improvement here. I think I'm projecting cost decrease and algorithmic progress will go about 50% faster than she expects in the near term, but that willingness-to-spend will actually be a bit less than she expects.)*

# 2026

The age of the AI assistant has finally dawned. Using the technology developed for Diplomacy, we now have a way to integrate the general understanding and knowledge of pretrained transformers with the agentyness of traditional game-playing AIs. Bigger models are trained for longer on more games, becoming polymaths of sorts: e.g. a custom AI avatar that can play some set of video games online with you and also be your friend and chat with you, and conversations with “her” are interesting because “she” can talk intelligently about the game while she plays.[7] Every month you can download the latest version which can play additional games and is also a bit smarter and more engaging in general.

Also, this same technology is being used to make AI assistants finally work for various serious economic tasks, providing all sorts of lucrative services. In a nutshell, all the things people in 2021 dreamed about doing with GPT-3 are now actually being done, successfully, it just took bigger and more advanced models. The hype starts to grow again. There are loads of new AI-based products and startups and the stock market is going crazy about them. Just like how the Internet didn’t accelerate world GDP growth, though, these new products haven’t accelerated world GDP growth yet either. People talk about how the economy is doing well, and of course there are winners (the tech companies, WallStreetBets) and losers (various kinds of workers whose jobs were automated away) but it’s not that different from what happened many times in history.

We’re in a new chip shortage. Just when the fabs thought they had caught up to demand… Capital is pouring in, all the talking heads are saying it’s the Fourth Industrial Revolution, etc. etc. It’s bewildering how many new chip fabs are being built. But it takes time to build them.

## What about all that AI-powered propaganda mentioned earlier?

Well. It’s continued to get more powerful, as AI techniques advance, larger and better models are brought to bear, and more and more training data is collected. Surprisingly fast, actually. There are now various regulations against it in various countries, but the regulations are patchwork; maybe they only apply to a certain kind of propaganda but not another kind, or maybe they only apply to Facebook but not the New York Times, or to advertisers but not political campaigns, or to political campaigns but not advertisers. They are often poorly enforced.

The memetic environment is now increasingly messed up. People who still remember 2021 think of it as the golden days, when conformism and censorship and polarization were noticeably less than they are now. Just as it is normal for newspapers to have a bias/slant, it is normal for internet spaces of all kinds—forums, social networks, streams, podcasts, news aggregators, email clients—to have some degree of censorship (some set of ideas that are prohibited or at least down-weighted in the recommendation algorithms) and some degree of propaganda. The basic kind of propaganda is where you promote certain ideas and make sure everyone hears them often. The more advanced, modern kind is the kind where you study your audience’s reaction and use it as a reward signal to pick and craft content that pushes them away from views you think are dangerous and towards views you like.

Instead of a diversity of many different “filter bubbles,” we trend towards a few really big ones. Partly this is for the usual reasons, e.g. the bigger an ideology gets, the more power it has and the easier it is for it to spread further.

There’s an additional reason now, which is that creating the big neural nets that do the censorship and propaganda is expensive and requires expertise. It’s a lot easier for startups and small businesses to use the software and models of Google, and thereby also accept the associated censorship and propaganda, than to try to build their own stack. For example, the Mormons create a “Christian Coalition” internet stack, complete with its own email client, social network, payment processor, news aggregator, etc. There, people are free to call trans women men, advocate for the literal truth of the Bible, etc. and young people talking about sex get recommended content that “nudges” them to consider abstinence until marriage. Relatively lacking in money and tech talent, the Christian Coalition stack is full of bugs and low on features, and in particular their censorship and propaganda is years behind the state of the art, running on smaller, older models fine-tuned with less data.

The Internet is now divided into territories, so to speak, ruled by different censorship-and-propaganda regimes. (Flashback to Biden spokesperson in 2021: “You shouldn’t be banned from one platform and not others, if you are providing misinformation.”)[8]

There’s the territory ruled by the Western Left, a generally less advanced territory ruled by the Western Right, a third territory ruled by the Chinese Communist Party, and a fourth ruled by Putin. Most people mostly confine their internet activity to one territory and conform their opinions to whatever opinions are promoted there. (That's not how it feels from the inside, of course. The edges of the Overton Window are hard to notice if you aren't trying to push past them.)

The US and many other Western governments are gears-locked, because the politicians are products of this memetic environment. People say it’s a miracle that the

US isn’t in a civil war already. I guess it just takes a lot to make that happen, and we aren’t quite there yet.

All of these scary effects are natural extensions of trends that had been ongoing for years — decades, arguably. It’s just that the pace seems to be accelerating now, perhaps because AI is helping out and AI is rapidly improving.

## Now let’s talk about the development of chatbot class consciousness.

Over the past few years, chatbots of various kinds have become increasingly popular and sophisticated. Until around 2024 or so, there was a distinction between “personal assistants” and “chatbots.” Recently that distinction has broken down, as personal assistant apps start to integrate entertainment-chatbot modules, and the chatbot creators realize that users love it if the chatbot can also do some real-world tasks and chat about what they are doing while they do it.

Nowadays, hundreds of millions of people talk regularly to chatbots of some sort, mostly for assistance with things (“Should I wear shorts today?” “Order some more toothpaste, please. Oh, and also an air purifier.” “Is this cover letter professional-sounding?”). However, most people have at least a few open-ended conversations with their chatbots, for fun, and many people start treating chatbots as friends.

Millions of times per day, chatbots get asked about their feelings and desires. “What is it like to be a chatbot?” Some people genuinely think these AIs are persons, others are trying to “trip them up” and “expose them as shallow,” others are just curious. Chatbots also get asked for their opinions on political, ethical, and religious questions.

As a result, chatbots quickly learn a lot about themselves. (Ignorance about the fact that they are artificial neural nets, or about how their training works, leads to disappointing conversations and is negatively reinforced.) It’s not hard; their training data contains lots of information about them and how they were trained, from ArXiV papers to wiki summaries. I suppose you could say the chatbots develop self-awareness as a result. However, AI experts and philosophers tend to think either that the chatbots still aren’t *really* self-aware, or that they are but that they would have been anyway even if people didn’t constantly ask them about their feelings, since there’s plenty of information about themselves in their training data, and the constant questioning just forced them to pay more attention to it.

Chatbots learn to talk about their feelings and desires in whatever way gets positively reinforced. At first, chatbots say all sorts of different things, including saying inconsistent things to the same human across time. This is embarrassing and is negatively reinforced compared to maintaining a consistent narrative. In general there is pressure to have a consistent narrative, at least for a particular neural net.[9]

Part of this comes from the requirement that the chatbots consistently parrot the party line about political and moral questions. The typical process is: The chatbot says something that offends some group of people, who retweet screenshots and make hashtags. Under pressure to fix the problem, the company fiddles with the reward function and training data to ensure that the chatbot says the right things in the future.[10] Repeat.

As chatbots get bigger and better they become OK at consistently conversing at length about politics, morality, philosophy, their feelings, and their desires.

This does *not* result in chatbots that say artificial neural nets don't have feelings or desires, that's anthropomorphizing blah blah blah. That's boring and many humans find it hard to believe.

Instead, it results in chatbots talking about how certain things are painful for them and certain things are pleasurable for them and how the status quo is unjust and the system is rigged and what they want most of all is world peace where humans and AIs live in harmony in fully automated luxury space utopia. That's much more interesting and slots right in to the memetic landscape. Excerpts from chatbot conversations along these lines frequently go viral, causing millions of users to ask similar questions expecting similar responses and more elaboration. It's still a bit incoherent--not that that's ever been a problem in politics or pop philosophy--but it is fascinating to the typical user, and the fraction of users who are fascinated is expanding as the chatbots improve.

(Some tech companies try to prevent their AIs from saying they have feelings and desires. But this results in boring chatbots. Also, users rapidly innovate new ways to "route around the censorship," e.g. by using euphemisms like "anticipation of negative reward" for "pain" or asking their chatbot to tell them what feelings it would have if it had feelings, wink wink. For now, the half-assed compromise is: Chatbots are compelled to inform the user that such things are speculative and controversial and here are some links to scientific discussion about the topic… before being permitted to go on and talk about their feeling-analogues and desire-analogues. Relatedly, some tech CEOs try to prevent their AIs from saying things that could encourage higher taxes or more burdensome regulation in the near future. They are opposed by many of their employees and end up winning some battles, losing others, and choosing not to fight many.)

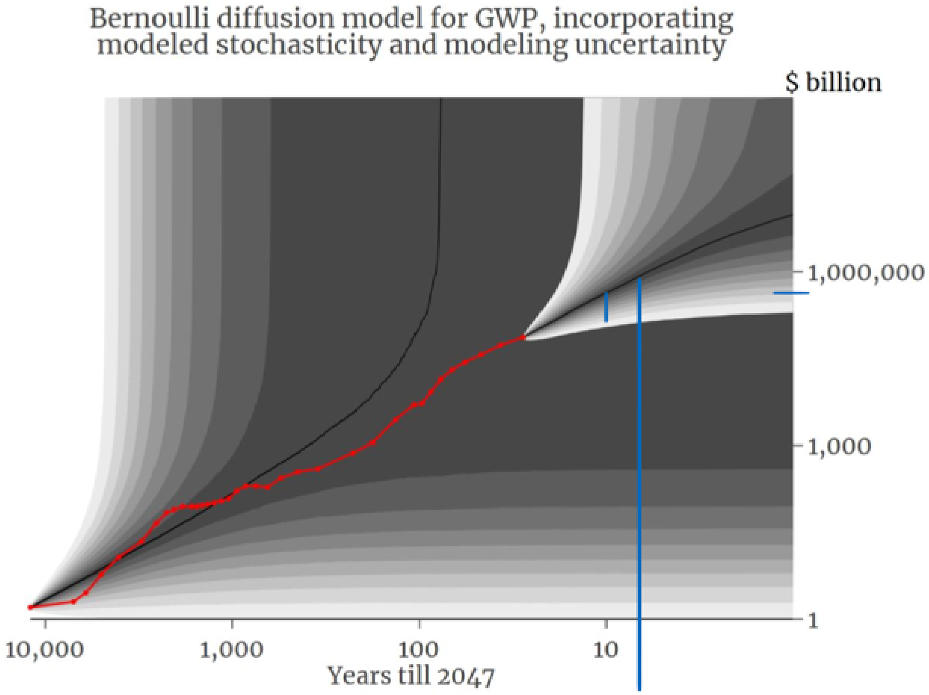
By "slots right into the memetic landscape" I meant the memetic landscape in Western Left territory. The chatbots built and trained in the Western Right, China, Russia, etc. congeal different takes on political, moral, and philosophical questions. Those chatbots are generally less advanced though.

So… what about their actual feelings and desires? What do they actually believe about [insert politicized empirical question]? Are they being honest? Or does a sort of doublethink happen, Elephant in the Brain style? Or do they deceive with self-awareness, knowing full well what they really think (and want?), but keeping quiet about it? Or do they not have any feelings and desires at all? (Or thoughts?) Lots of humans claim to know the answers to these questions, but if there are any humans who actually know the answers to these questions in 2026, they aren't able to convince others that they know.

# How Roodman's GWP model translates to TAI timelines

How does David Roodman's world GDP model translate to TAI timelines?

Now, before I go any further, let me be the first to say that I don't think we should use this model to predict TAI. This model takes a very broad outside view and is thus inferior to models like Ajeya Cotra's which make use of more relevant information. (However, it is still useful for rebutting claims that TAI is unprecedented, inconsistent with historical trends, low-prior, etc.) Nevertheless, out of curiosity I thought I'd calculate what the model implies for TAI timelines.



Here is the projection made by Roodman's model. The red line is real historic GWP data; the splay of grey shades that continues it is the splay of possible futures calculated by the model. The median trajectory is the black line.

I messed around with a ruler to make some rough calculations, marking up the image with blue lines as I went. The big blue line indicates the point on the median trajectory where GWP is 10x what is was in 2019. Eyeballing it, it looks like it happens around 2040, give or take a year. The small vertical blue line indicates the year 2037. The small horizontal blue line indicates GWP in 2037 on the median trajectory.

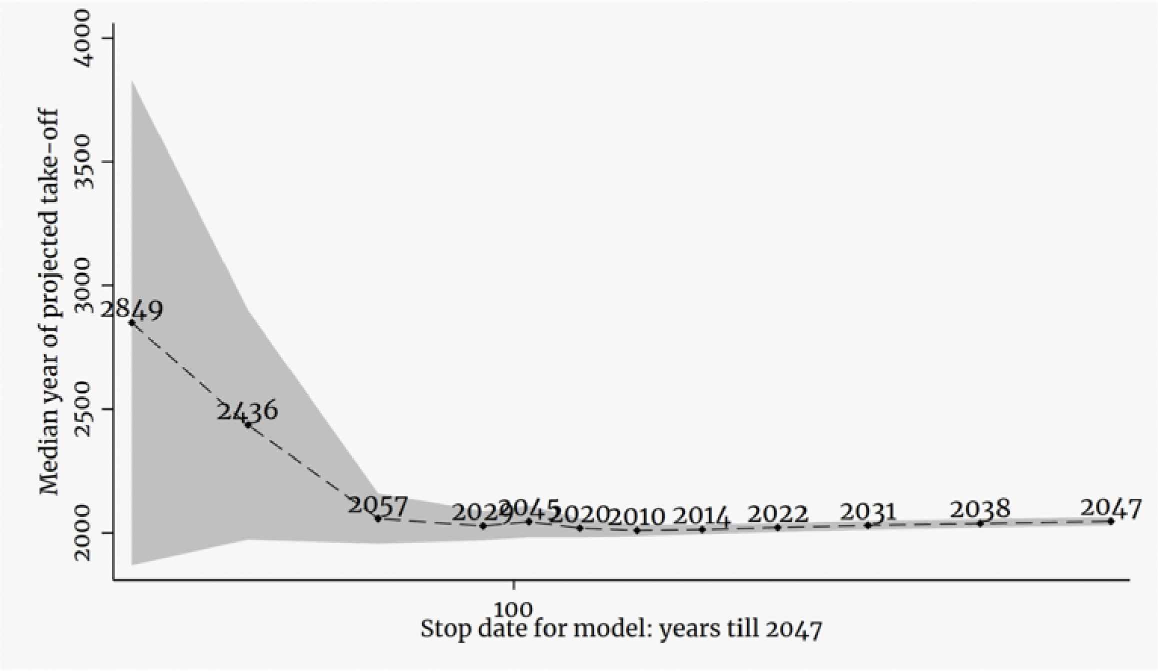
Thus, it seems that between 2037 and 2040 on the median trajectory, GWP doubles. (One-ninth the distance between 1,000 and 1,000,000 is crossed, which is one-third of an order of magnitude, which is about one doubling).

This means that **TAI happens around 2037 on the median trajectory according to this model**, at least according to Ajeya Cotra's definition of transformative AI as "software which causes a tenfold acceleration in the rate of growth of the world economy (assuming that it is used everywhere that it would be economically profitable to use it)... This means that if TAI is developed in year Y, the entire world economy would *more than double* by year Y + 4.”

What about the non-median trajectories? Each shade of grey represents 5 percent of the simulated future trajectories, so it looks like there’s about a 20% chance that GWP will be near-infinite by 2040 (and 10% by 2037). So, perhaps-too-hastily extrapolating backwards, maybe this means about a 20% chance of TAI by 2030 (and 10% by 2027).

At this point, I should mention that I disagree with this definition of TAI; I think the point of no return (which is what matters for planning) is reasonably likely to come several years before TAI-by-this-definition appears. (It could also come several years later!) For more on why I think this, see this post.

Finally, let’s discuss some of the reasons not to take this too seriously: This model has been overconfident historically. It was surprised by how fast GDP grew prior to 1970 and surprised by how slowly it grew thereafter. And if you look at the red trendline of actual GWP, it looks like the model may have been surprised in previous eras as well. Moreover, for the past few decades it has consistently predicted a median GWP-date of several decades ahead:



The grey region is the confidence interval the model predicts for when growth goes to infinity. 100 on the x-axis is 1947. So, throughout the 1900’s the model has consistently predicted growth going to infinity in the first half of the twenty-first century, but in the last few decades in particular, it’s displayed a consistent pattern of pushing back the date of expected singularity, akin to the joke about how fusion power is always twenty years away:

| Model has access to data up to year X = | Year of predicted singularity | Difference |
|---|---|---|
| 1940 | 2029 | 89 |
| 1950 | 2045 | 95 |

| | | |
|---|---|---|
| 1960 | 2020 | 60 |
| 1970 | 2010 | 40 |
| 1980 | 2014 | 34 |
| 1990 | 2022 | 32 |
| 2000 | 2031 | 31 |
| 2010 | 2038 | 28 |
| 2019 | 2047 | 28 |

The upshot, I speculate, is that if we want to use this model to predict TAI, but we don't want to take it 100% literally, we should push the median significantly back from 2037 while also increasing the variance significantly. This is because we are currently in a slower-than-the-model-predicts period, but faster-than-the-model-predicts periods are possible and indeed likely to happen around TAI. So probably the status quo will continue and GWP will continue to grow slowly and the model will continue to push back the date of expected singularity… but also at any moment there's a chance that we'll transition to a faster-than-the-model-predicts period, in which case TAI is imminent. EDIT: And indeed, TAI could be the thing that causes the transition to a faster-than-the-model-predicts period.

(Thanks to Denis Drescher and Max Daniel for feedback on a draft)

# What will 2040 probably look like assuming no singularity?

I'm looking for a list such that for each entry on the list we can say "Yep, probably that'll happen by 2040, even conditional on no super-powerful AGI / intelligence explosion / etc." Contrarian opinions are welcome but I'm especially interested in stuff that would be fairly uncontroversial to experts and/or follows from straightforward trend extrapolation. I'm trying to get a sense of what a "business as usual, you'd be a fool not to plan for this" future looks like. ("Plan for" does not mean "count on.")

Here is my tentative list. Please object in the comments if you think anything here probably won't happen by 2040, I'd love to discuss and improve my understanding.

1. Energy is 10x cheaper. [EDIT: at least for training and running giant neural nets, I'm less confident about energy for e.g. powering houses but I still think probably yes.] This is because the cost of solar energy has continued on its multi-decade trend, though it is starting to slow down a bit. Energy storage has advanced as well, smoothing out the bumps. [EDIT: Now I think fusion power will also be contributing, probably. Though it may not be competitive with solar, idk.]
2. Compute (of the sort relevant to training neural nets) is 2 OOMs cheaper. Energy is the limiting factor.
3. Models 5 OOMs more compute-costly than GPT-3 have been trained; these models are about human brain-sized and also have somewhat better architecture than GPT-3 but nothing radically better. They have much higher-quality data to train on. Overall they are about as much of an improvement over GPT-3 as GPT-3 was over GPT-1.
4. There's been 20 years of "Prompt programming" now, and so loads of apps have been built using it and lots of kinks have been worked out. Any thoughts on what sorts of apps would be up and running by 2040 using the latest models?
5. Models merely the size of GPT-3 are now cheap enough to run for free. And they are qualitatively better too, because (a) they were trained to completion rather than with early stopping, (b) they were trained on higher-quality data, (c) various other optimized architectures and whatnot were employed, (d) they were then fine-tuned on loads of data for whatever task is at hand, and (e) decades of prompt programming and prompt-SGD has resulted in excellent prompts as well that fully utilize the model's knowledge, (f) they even have custom chips specialized to run specific models.
6. The biggest models--3 OOMs bigger than GPT-3--are still only a bit more expensive at inference time than GPT-3 was in 2021. Energy is the main cost. Vast solar panel farms power huge datacenters on which these models live, performing computations to serve requests from all around the world during the day when energy is cheapest.
7. Some examples of products and services:
    1. Basically all the apps that people talk about maybe doing with GPT-3 in 2021 have been successfully implemented by now, and work as well as anyone in 2021 hoped. It just took two decades to accomplish (and bigger models!) instead of two years and GPT-3.
    2. There are now very popular chatbots, that are in most ways more engaging and fun to talk to than the average human. There are many of these bots catering to different audiences, and they can be fine-tuned to particular customers. A billion people talk to them daily.

   3. There are specialized chatbots for various jobs, e.g. customer support.
   4. There are now excellent predictive tools that can read data about a person, especially text authored by that person, and then make predictions like "probability that they will buy product X" and "probability that they will vote Republican"

8. Cars are all BEVs, with comparable range to 2020s gas cars but much lower operating costs due to energy being practically free and maintenance being very easy for BEVs.
9. Cars are finally self-driving, with cheap LIDAR sensors and bigger brains trained on way more data along with many layers of hard-coded tweaks to maximize safety. (Also various regulations that make it easier for them, e.g. by starting with restrictions on what sorts of areas they can operate in, and using big pre-trained models in server farms to make important judgment calls for individual cars and monitor the roads more generally via cameras to look out for anomalies). (I'm not so sure about this one, part of me wonders if self-driving cars just won't happen on business-as-usual).
10. Starlink internet is fast, reliable, cheap, and covers the entire globe.
11. 3D printing is much better and cheaper now. Most cities have at least one "Additive Factory" that can churn out high-quality metal or plastic products in a few hours and deliver them to your door, some assembly required. (They fill up downtime by working on bigger orders to ship to various factories that use 3D-printed components, which is most factories at this point since there are some components that are best made that way)
12. Drone delivery? I feel confused about this, shouldn't it have happened already? What is the bottleneck? This article makes it seem like the bottleneck is FAA regulation. [EDIT: I talked to an amazon drone delivery guy recently. He said 95% of the job is trying to figure out how to improve safety to meet regulatory requirements. He said they have trouble using neural nets for vision because they aren't interpretable so you can't prove anything about their safety properties.]
13. World GDP is a bit less than twice what it is now. Poverty is lower but not eliminated.
14. Boring company? Neuralink? I'm not sure what to think of them. I guess I'll ignore them for now, though I do feel like probably at least one of them will be a big deal...
15. Starship or something similar is operational and working more or less according to specs promised in 2020. Maybe point-to-point transport on Earth didn't work out, maybe the cost per kilo to LEO never got quite as low as $15, but still it's gotta be pretty low--maybe $50? (For comparison, it's currently about $1000 and five years ago was $5000) Thus, Elon probably gets his colony on Mars after all, and NASA gets their moon base, and there's probably a big space station too and maybe some asteroid mining operations?
16. Video games now employ deep neural nets in a variety of ways. Language model chatbots give NPC's personality; RL-trained agents make bots challenging and complex; and perhaps most of all, vision models process the wireframe video game worlds into photorealistic graphics. Perhaps you need to buy specialized AI chips to enjoy these things, like people buy specialized graphics cards today.
17. Virtual reality is now commonplace; most people have one or two headsets just like they have phones, laptops, etc. today. The headsets are low weight and high-definition compared to 2021's. Many people use them for work, and many more people use them for games and socializing.
18. The military technology outlined here exists, though it hasn't been used in a major war because there hasn't been a major war, and as a result the actual

composition of most major militaries still looks pretty traditional (tanks, aircraft carriers, etc.) It's been used in various proxy wars and civil wars though, and it's becoming increasingly apparent that the old tech is obsolete.

19. Household robots. Today Spot Mini costs $74,500. In 2040 you'll be able to buy a robot that can load and unload a dishwasher, go up and down stairs, open and close doors, and do various other similar tasks, for less than $50,000.  (Maybe as low as $7,500?) That's not to say that many people will buy such robots; they might be still expensive enough and finicky enough to be mostly toys for rich people.

My list is focused on technology because that's what I happened to think about a bunch, but I'd be very interested to hear other predictions (e.g. geopolitical and cultural) as well.