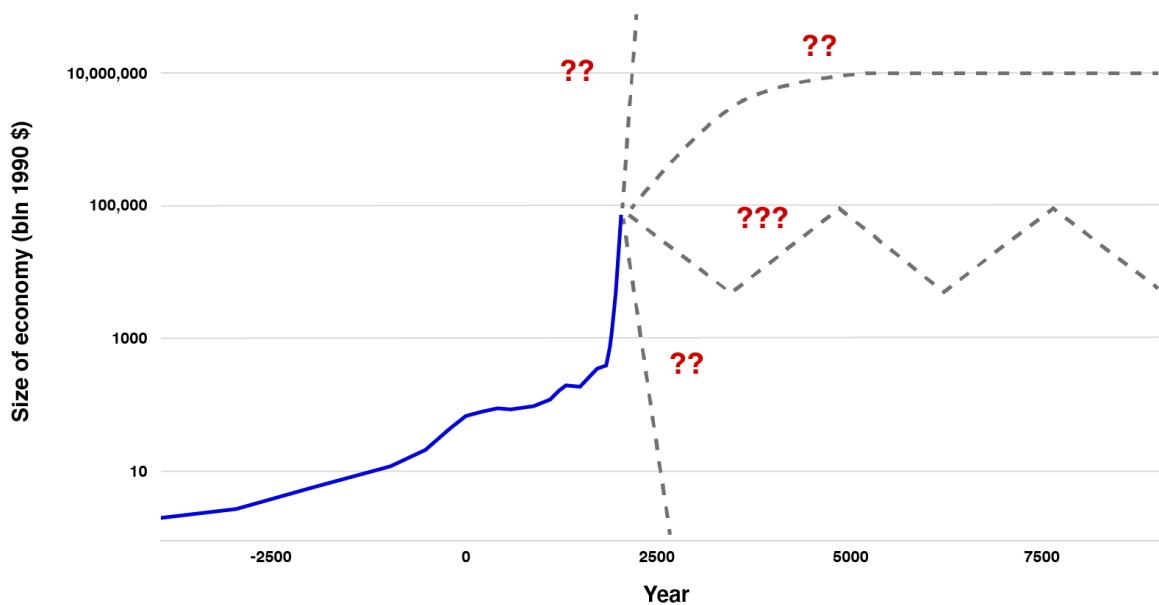


The Most Important Century



A blog post series from Cold Takes

(<https://www.cold-takes.com>)

For the web version, see

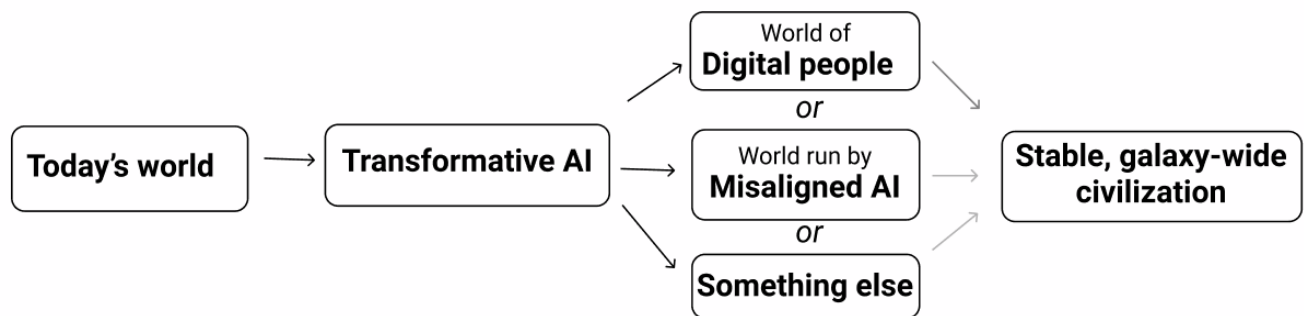
<https://www.cold-takes.com/most-important-century/>

Table of Contents

“Most Important Century” Series: Roadmap.....	1
All Possible Views About Humanity’s Future Are Wild.....	7
The Duplicator: Instant Cloning Would Make The World Economy Explode	17
Digital People Would Be An Even Bigger Deal (Intro).....	28
Digital People FAQ.....	34
Digital People Would Be An Even Bigger Deal (Final Section)	52
This Can’t Go On.....	61
Forecasting Transformative AI, Part 1: What Kind of AI?	73
Why AI Alignment Could Be Hard With Modern Deep Learning	83
Forecasting Transformative AI: What’s The Burden Of Proof?	100
Are We “Trending Toward” Transformative AI? (How Would We Know?)	116
Forecasting Transformative AI: The “Biological Anchors” Method In A Nutshell..	125
AI Timelines: Where The Arguments, And The “Experts,” Stand.....	142
How To Make The Best Of The Most Important Century?	153
Call To Vigilance.....	166
Appendices.....	170
Weak Point In “Most Important Century”: Full Automation.....	171
Weak Point In “Most Important Century”: Lock-In	174
“Biological Anchors” Is About Bounding, Not Pinpointing, AI Timelines	182
More On “Multiple World-Size Economies Per Atom”.....	192
A Note On Historical Economic Growth.....	198
Some Additional Detail On What I Mean By “Most Important Century”	202
Why Talk About 10,000 Years From Now?	205
Endnotes.....	208

“Most Important Century” Series: Roadmap

This is an outline of how each piece in the [“most important century”](#) series relates to the overall argument. I think it’s useful to read through this before reading through the series, to get a sense of where each piece fits in.



I think we have good reason to believe that the **21st century could be the most important century ever for humanity**. I think the most likely way this would happen would be via the development of advanced AI systems that lead to explosive growth and scientific advancement, getting us more quickly than most people imagine to a deeply unfamiliar future.

A bit more specifically,¹ I think there is a good chance that:

1. During the century we’re in right now, we will develop technologies that cause us to transition to a state in which humans as we know them are no longer the main force in world events. This is our last chance to shape how that transition happens.
2. Whatever the main force in world events is (perhaps digital people, misaligned AI, or something else) will create highly stable civilizations that populate our entire galaxy for billions of years to come. The transition taking place this century could shape all of that.

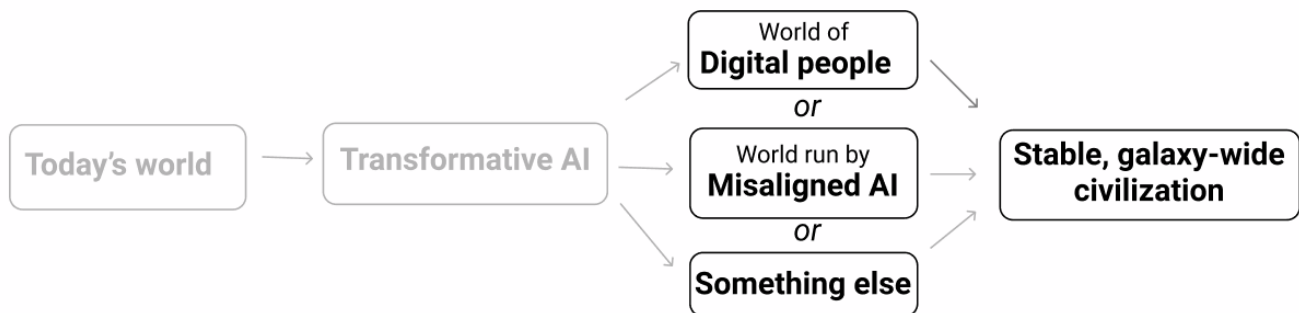
¹ For a more detailed elaboration of what I mean by “most important century,” see [here](#) (not likely to be of interest to most readers).

I think it's very unclear whether this would be a good or bad thing. What matters is that it could go a lot of different ways, and we have a chance to affect that.

I believe the above possibility doesn't get enough attention, discussion, or investment, particularly from people whose goal is to make the world better. By writing about it, I'd like to either help change that, or gain more opportunities to get criticized and change my mind.

This post serves as a summary/roadmap for an 11-post series arguing these points (and the posts themselves are often effectively summaries of longer analyses by others). I will add links as I put out posts in the series.

Our wildly important era

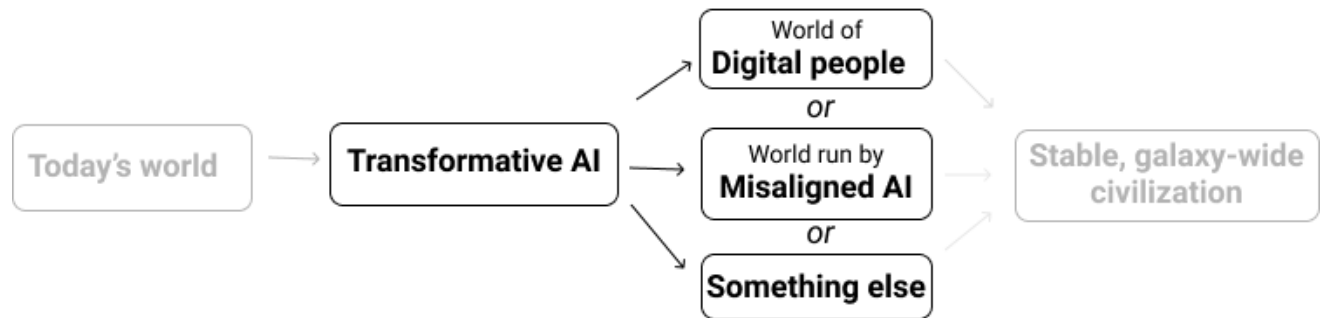


[**All Possible Views About Humanity's Long-Term Future Are Wild**](#) argues that two simple observations - (a) it appears likely that we will *eventually* be able to spread throughout the galaxy, and (b) it doesn't seem any other life form has done that yet - are sufficient to make the case that we live in an incredibly important time. I illustrate this with a timeline of the galaxy.

[**The Duplicator**](#) explains the basic mechanism by which “eventually” above could become “soon”: the ability to “copy human minds” could lead to a productivity explosion. This is background for the next few pieces.

[**Digital People Would Be An Even Bigger Deal**](#) discusses how achievable-seeming technology - in particular, [**mind uploading**](#) - could lead to unprecedented productivity, control of the environment, and more. The result could be a stable, galaxy-wide civilization that is deeply unfamiliar from today's vantage point.

Our century's potential for acceleration



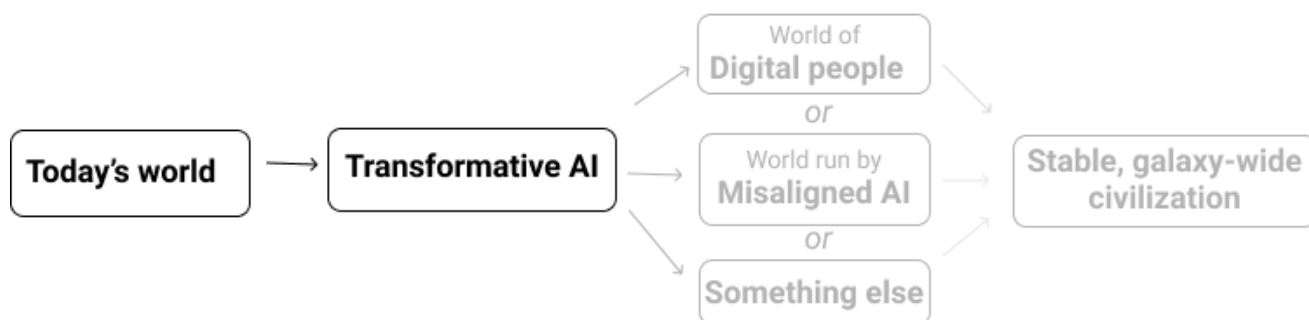
[**This Can't Go On**](#) looks at economic growth and scientific advancement over the course of human history. Over the last few generations, growth has been pretty steady. But zooming out to a longer time frame, it seems that growth has greatly accelerated recently; is near its historical high point; and is faster than it can be for all that much longer (there aren't enough atoms in the galaxy to sustain this rate of growth for even another 10,000 years).

The times we live in are unusual and unstable. Rather than planning on more of the same, we should anticipate stagnation (growth and scientific advancement slowing down), explosion (further acceleration) or collapse.

[**Forecasting Transformative AI, Part 1: What Kind Of AI?**](#) introduces the possibility of AI systems that automate scientific and technological advancement, which could cause explosive productivity. I argue that such systems would be “transformative” in the sense of bringing us into a new, qualitatively unfamiliar future.

[**Why AI Alignment Could Be Hard With Modern Deep Learning \(guest post\)**](#) goes into more detail on why advanced AI systems could be “misaligned,” with potentially catastrophic consequences.

Forecasting transformative AI this century



[Forecasting Transformative AI: What's The Burden Of Proof?](#) argues that we shouldn't have too high a "burden of proof" on believing that transformative AI could be developed this century, partly because our century is already special in many ways that you can see without detailed analysis of AI.

[Forecasting Transformative AI: Are We "Trending Toward" Transformative AI?](#) discusses the basic structure of forecasting transformative AI, the problems with trying to forecast it based on trends in "AI impressiveness," and the state of AI researcher opinion on transformative AI timelines.

[Forecasting Transformative AI: The "Biological Anchors" Method In A Nutshell](#) summarizes the **[biological anchors framework](#)** for forecasting AI. This framework is the main factor in my specific forecasts.

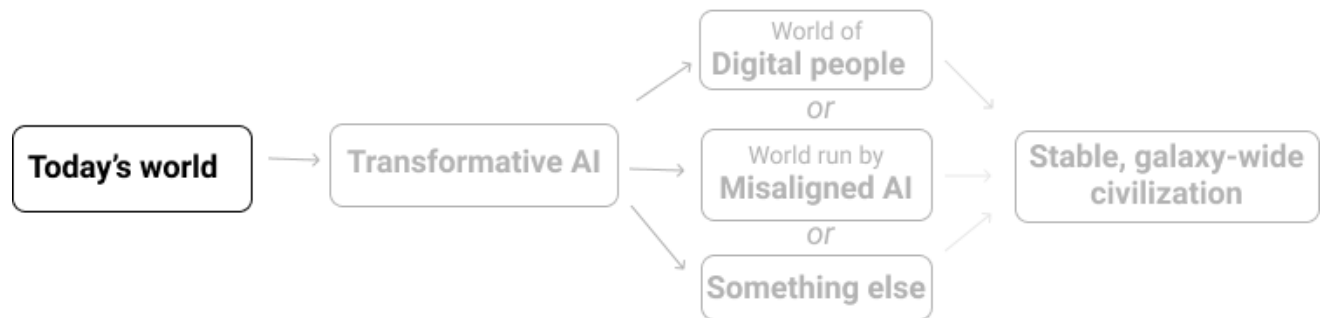
I am forecasting more than a 10% chance transformative AI will be developed within 15 years (by 2036); a ~50% chance it will be developed within 40 years (by 2060); and a ~2/3 chance it will be developed this century (by 2100).

[AI Timelines: Where The Arguments, And The "Experts," Stand](#) briefly summarizes the state of the arguments and addresses the question, "Where does expert opinion stand on all of this?"

- The claims I'm making neither *contradict* a particular expert consensus, nor are *supported* by one (though most of the key reports I cite have had external expert review). They are, rather, claims about topics that simply have no "field" of experts devoted to studying them.

- Some people might choose to ignore any claims that aren't actively supported by a robust expert consensus; but I don't think that is what we should be doing here.

Implications



[How To Make The Best Of The Most Important Century?](#) discusses different, contrasting views of how to help the most important century go as well as possible for humanity - and lists “robustly helpful actions” that seem worth taking regardless.

[Call to Vigilance](#) is in lieu of a “call to action” for the series. Given all the uncertainty we face, I don't think people should rush to “do something” and then move on. Instead, they should take whatever [robustly good actions](#) they can today, and otherwise put themselves in a better position to take important actions when the time comes.

Some supplemental posts that elaborate on points made in the series:

- [Some additional detail on what I mean by “most important century”](#)
- [A note on historical economic growth](#): How the “most important century” argument is affected if our picture of long-run economic history changes.
- [More on “multiple world-size economies per atom”](#): A follow up on “This Can't Go On” for the skeptical.
- [Weak point in “most important century”: full automation](#) (acknowledges that I could have done more to address the question of

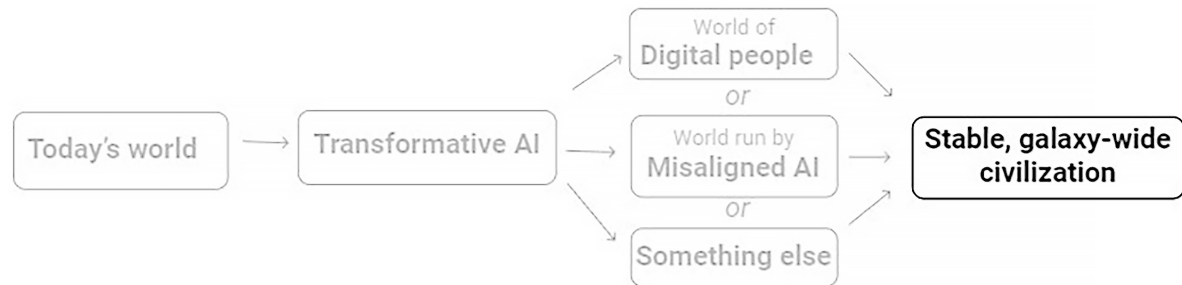
<https://www.cold-takes.com/most-important-century-series-roadmap/>

how *complete* AI automation has to be to bring about the consequences I discuss, and adds a bit more on this point)

- **Weak point in “most important century”: lock-in** (acknowledges that I could have done more to address how AI could lead to “lock-in” of the long-run future, and adds a bit more on this point)
- **“Biological anchors” is about bounding, not pinpointing, AI timelines**: more on how I’ve used the “biological anchors” framework, aimed at skeptical readers.

I’ve listed some key sources for this series in one place [here](#), for those interested in going much deeper.

All Possible Views About Humanity's Future Are Wild



Summary:

- In a series of posts starting with this one, I'm going to argue that the 21st century could see our civilization develop technologies allowing rapid expansion throughout our currently-empty galaxy. And thus, that **this century could determine the entire future of the galaxy for tens of billions of years, or more.**
- This view seems “wild”: we should be doing a double take at any view that we live in such a special time. I illustrate this with a timeline of the galaxy. (On a personal level, this “wildness” is probably the single biggest reason I was skeptical for many years of the arguments presented in this series. Such claims about the significance of the times we live in seem “wild” enough to be suspicious.)
- But I don't think it's really possible to hold a non-“wild” view on this topic. I discuss alternatives to my view: a “conservative” view that thinks the technologies I'm describing are possible, but will take much longer than I think, and a “skeptical” view that thinks galaxy-scale expansion will never happen. Each of these views seems “wild” in its own way.
- Ultimately, as hinted at by the [Fermi paradox](#), it seems that our species is simply in a wild situation.

Before I continue, I should say that I don't think humanity (or some digital descendant of humanity) expanding throughout the galaxy would necessarily be a good thing - especially if this prevents other life forms from ever emerging. I think it's quite hard to have a confident view on whether this would be good or bad. I'd like to keep the focus on the idea that our situation is "wild." I am not advocating excitement or glee at the prospect of expanding throughout the galaxy. I am advocating seriousness about the enormous potential stakes.

My view

This is the first in a series of pieces about the hypothesis that we live in the most important century for humanity.

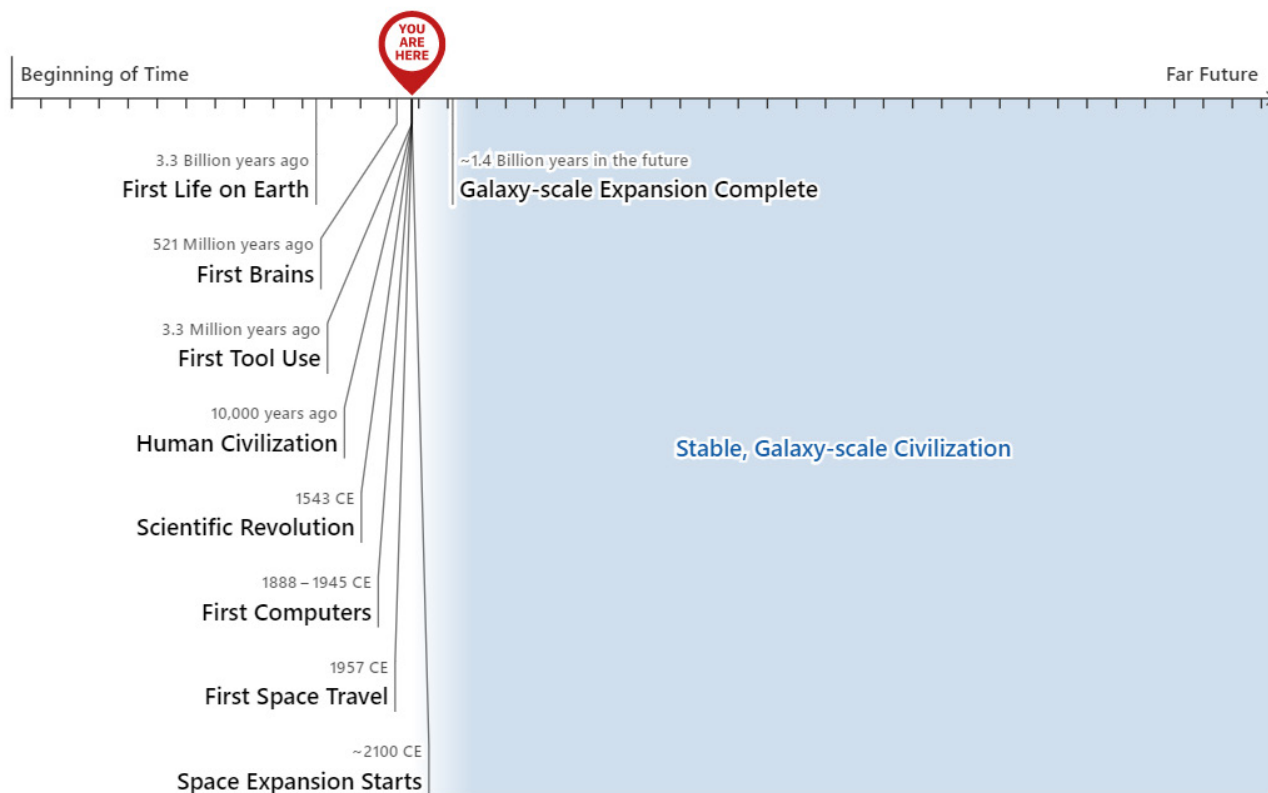
In this series, I'm going to argue that there's a good chance of a productivity explosion by 2100, which could quickly lead to what one might call a "technologically mature"² civilization. That would mean that:

- We'd be able to start sending spacecraft throughout the galaxy and beyond.
- These spacecraft could mine materials, build robots and computers, and construct very robust, long-lasting settlements on other planets, harnessing solar power from stars and supporting huge numbers of people (and/or our "[digital descendants](#)").
 - See [Eternity in Six Hours](#) for a fascinating and short, though technical, discussion of what this might require.
 - I'll also argue in future pieces (now available [here](#) and [here](#)) that there is a chance of "value lock-in": whoever is running the process of space expansion might be able to determine what sorts of people are in charge of the settlements and what sorts of societal values they have, in a way that is stable for many billions of years³.

² or [Kardashev Type III](#).

³ If we are able to create [mind uploads](#), or detailed computer simulations of people that are as conscious as we are, it could be possible to put them in virtual environments that automatically reset, or otherwise "correct" the environment, whenever the society would otherwise change in certain ways (for example, if a certain religion became dominant or lost dominance). This could give the designers

If that ends up happening, you might think of the story of our galaxy⁴ like this. I've marked major milestones along the way from “no life” to “intelligent life that builds its own computers and travels through space.”



Thanks to [Ludwig Schubert](#) for the visualization. Many dates are highly approximate and/or judgment-prone and/or just pulled from Wikipedia (sources [here](#)), but plausible changes wouldn't change the big picture. The ~1.4 billion years to complete space expansion is based on the distance to the outer edge of the Milky Way, divided by the speed of a fast existing human-made spaceship (details in spreadsheet just linked); IMO this is likely to be a massive overestimate of how long it takes to expand throughout the whole galaxy. See footnote for why I didn't use a logarithmic axis⁵

of these “virtual environments” the ability to “lock in” particular religions, rulers, etc. I'll discuss this more in future pieces (now available [here](#) and [here](#)).

⁴ I've focused on the “galaxy” somewhat arbitrarily. Spreading throughout all of the accessible universe would take a lot longer than spreading throughout the galaxy, and until we do it's still imaginable that some species from outside our galaxy will disrupt the “stable galaxy-scale civilization,” but I think accounting for this correctly would add a fair amount of complexity without changing the big picture. I may address that in some future piece, though.

⁵ A logarithmic version doesn't look any less weird, because the distances between the “middle” milestones are tiny compared to *both* the stretches of time before and after these milestones. More fundamentally, I'm talking about how remarkable it is to be in the most important [small number] of years out

??? That's crazy! According to me, there's a decent chance that we live at the very beginning of the tiny sliver of time during which the galaxy goes from nearly lifeless to largely populated. That out of a staggering number of persons who will ever exist, we're among the first. And that out of hundreds of billions of stars in our galaxy, ours will produce the beings that fill it.

I know what you're thinking: "The odds that we could live in such a significant time seem infinitesimal; the odds that Holden is having delusions of grandeur (on behalf of all of Earth, but still) seem far higher."⁶

But:

The "conservative" view

Let's say you agree with me about where humanity could *eventually* be headed - that we will eventually have the technology to create robust, stable settlements throughout our galaxy and beyond. But you think it will take far longer than I'm saying.

A key part of my view (which I'll write about more later) is that within this century, we could develop advanced enough AI to start a productivity explosion. Say you don't believe that.

- You think I'm underrating the fundamental limits of AI systems to date.
- You think we will need an enormous number of new scientific breakthroughs to build AIs that truly reason as effectively as humans.
- And even once we do, expanding throughout the galaxy will be a longer road still.

You don't think any of this is happening this century - you think, instead, that it will take something like **500 years**. That's 5-10x the time that has passed since we started building computers. It's more time than has passed since

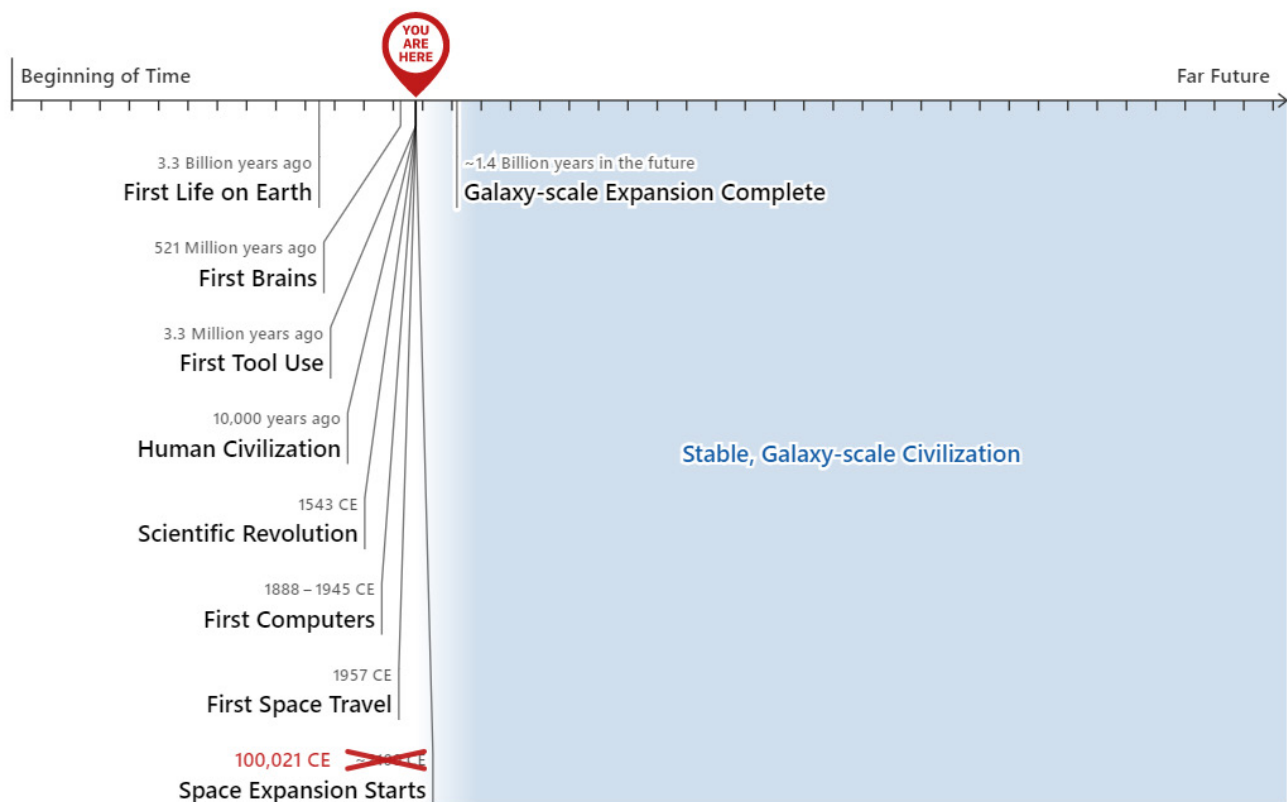
of [big number] of years - that's best displayed using a linear axis. It's often the case that weird-looking charts look more reasonable with logarithmic axes, but in this case I think the chart looks weird because the situation is weird. Probably the least weird-looking version of this chart would have the x-axis be something like the logged distance from the year 2100, but that would be a heck of a premise for a chart - it would basically bake in my argument that this appears to be a very special time period.

⁶ This is exactly the kind of thought that kept me skeptical for many years of the arguments I'll be laying out in the rest of this series about the potential impacts, and timing, of advanced technologies. Grappling directly with how "wild" our situation seems to ~undeniably be has been key for me.

Isaac Newton made the first credible attempt at laws of physics. It's about as much time has passed since the very start of the Scientific Revolution.

Actually, no, let's go even more conservative. You think our economic and scientific progress will stagnate. Today's civilizations will crumble, and many more civilizations will fall and rise. Sure, we'll *eventually* get the ability to expand throughout the galaxy. But it will take **100,000 years**. That's 10x the amount of time that has passed since human civilization began in the Levant.

Here's your version of the timeline:



The difference between your timeline and mine isn't even a pixel, so it doesn't show up on the chart. In the scheme of things, this "conservative" view and my view are the same.

It's true that the "conservative" view doesn't have the same urgency for our generation in particular. But it still places us among a tiny proportion of people in an incredibly significant time period. And it still raises questions of whether the things we do to make the world better - even if they only have a

tiny flow-through to the world 100,000 years from now - could be amplified to a galactic-historical-outlier degree.

The skeptical view

The “skeptical view” would essentially be that humanity (or some descendant of humanity, including a digital one) will *never* spread throughout the galaxy. There are many reasons it might not:

- Maybe something about space travel - and/or setting up mining robots, solar panels, etc. on other planets - is effectively impossible such that even another 100,000 years of human civilization won't reach that point.⁷
- Or perhaps for some reason, it will be technologically feasible, but it won't happen (because nobody wants to do it, because those who don't want to block those who do, etc.)
- Maybe it's possible to expand throughout the galaxy, but not possible to maintain a presence on many planets for billions of years, for some reason.
- Maybe humanity is destined to destroy itself before it reaches this stage.
 - But note that if the way we destroy ourselves is via misaligned AI,⁸ it would be possible for AI to build its own technology and spread throughout the galaxy, which still seems in line with the spirit of the above sections. In fact, it highlights that how we handle AI this century could have ramifications for many billions of years. So humanity would have to go extinct in some way that leaves no other intelligent life (or intelligent machines) behind.

⁷ Spreading throughout the galaxy would certainly be harder if nothing like [mind uploading](#) (which I discuss in a [separate piece](#), and which is part of why I think future space settlements could have “value lock-in” as discussed above) can ever be done. I would find a view that “mind uploading is impossible” to be “wild” in its own way, because it implies that human brains are so special that there is simply no way, ever, to digitally replicate what they're doing. (Thanks to David Roodman for this point.)

⁸ That is, advanced AI that pursues objectives of its own, which aren't compatible with human existence. I'll be writing more about this idea. Existing discussions of it include the books [Superintelligence](#), [Human Compatible](#), [life 3.0](#), and [The Alignment Problem](#). The shortest, most accessible presentation I know of is [The case for taking AI seriously as a threat to humanity](#) (Vox article by Kelsey Piper). This [report on existential risk from power-seeking AI](#), by Open Philanthropy's Joe Carlsmith, lays out a detailed set of premises that would collectively imply the problem is a serious one.

- Maybe an extraterrestrial species will spread throughout the galaxy before we do (or around the same time).
 - However, note that this doesn't seem to have happened in ~13.77 billion years so far since the universe began, and according to the above sections, there's only about 1.5 billion years left for it to happen before we spread throughout the galaxy.
- Maybe some extraterrestrial species already *has* spread throughout our galaxy, and for some reason we just don't see them. Maybe they are hiding their presence deliberately, for one reason or another, while being ready to stop us from spreading too far.
 - This would imply that they are choosing not to mine energy from any of the stars we can see, at least not in a way that we could see it. That would, in turn, imply that they're abstaining from mining a very large amount of energy that they could use to do whatever it is they want to do,⁹ including defend themselves against species like ours.
- Maybe this is all a dream. Or a **[simulation](#)**.
- Maybe something else I'm not thinking of.

That's a fair number of possibilities, though many seem quite "wild" in their own way. Collectively, I'd say they add up to more than 50% probability ... but I would feel very weird claiming they're collectively overwhelmingly likely.

Ultimately, it's very hard for me to see a case *against* thinking something like this is at least *reasonably* likely: "We will eventually create robust, stable settlements throughout our galaxy and beyond." It seems like saying "no way" to that statement would itself require "wild" confidence in something about the limits of technology, and/or long-run choices people will make, and/or the inevitability of human extinction, and/or something about aliens or simulations.

I imagine this claim will be intuitive to many readers, but not all. Defending it in depth is not on my agenda at the moment, but I'll rethink that if I get enough **[demand](#)**.

⁹ Thanks to Carl Shulman for this point.

Why all possible views are wild: the Fermi paradox

I'm claiming that it would be "wild" to think we're basically assured of *never* spreading throughout the galaxy, but also that it's "wild" to think that we have a decent chance of spreading throughout the galaxy.

In other words, I'm calling every possible belief on this topic "wild." That's because I think we're in a wild situation.

Here are some *alternative* situations we could have found ourselves in, that I wouldn't consider so wild:

- We could live in a mostly-populated galaxy, whether by our species or by a number of extraterrestrial species. We would be in some densely populated region of space, surrounded by populated planets. Perhaps we would read up on the history of our civilization. We would know (from history and from a lack of empty stars) that we weren't unusually early life-forms with unusual opportunities ahead.
- We could live in a world where the kind of technologies I've been discussing didn't seem like they'd ever be possible. We wouldn't have any hope of doing space travel, or successfully studying our own brains or building our own computers. Perhaps we could somehow detect life on other planets, but if we did, we'd see them having an equal lack of that sort of technology.

But space expansion seems feasible, *and* our galaxy is empty. These two things seem in tension. A similar tension - the question of why we see no signs of extraterrestrials, despite the galaxy having so many possible stars they could emerge from - is often discussed under the heading of the [Fermi Paradox](#).

Wikipedia has a list of [possible resolutions](#) of the Fermi paradox. Many correspond to the [skeptical view](#) possibilities I list above. Some seem less relevant to this piece. (For example, there are various reasons extraterrestrials might be present but not *detected*. But I think any world in which extraterrestrials don't *prevent* our species from galaxy-scale expansion ends up "wild," even if the extraterrestrials are there.)

My current sense is that the best analysis of the Fermi Paradox available today favors the explanation that **intelligent life is extremely rare**: something

about the appearance of life in the first place, or the evolution of brains, is so unlikely that it hasn't happened in many (or any) other parts of the galaxy.¹⁰

That would imply that **the hardest, most unlikely steps on the road to galaxy-scale expansion are the steps our species has *already taken***. And that, in turn, implies that we live in a strange time: extremely early in the history of an extremely unusual star.

If we started finding signs of intelligent life elsewhere in the galaxy, I'd consider that a big update away from my current "wild" view. It would imply that whatever has stopped other species from galaxy-wide expansion will also stop us.

This pale blue dot could be an awfully big deal

Describing Earth as a tiny dot in a [photo from space](#), Ann Druyan and Carl Sagan [wrote](#):

The Earth is a very small stage in a vast cosmic arena. Think of the rivers of blood spilled by all those generals and emperors so that, in glory and triumph, they could become the momentary masters of a [fraction of a dot](#) ... Our posturings, our imagined self-importance, the delusion that we have some privileged position in the Universe, are challenged by this point of pale light ... It has been said that astronomy is a humbling and character-building experience. There is perhaps no better demonstration of the folly of human conceits than this distant image of our tiny world.

This is a somewhat common sentiment - that when you pull back and think of our lives in the context of billions of years and billions of stars, you see how insignificant all the things we care about today really are.

But here I'm making the opposite point.

It looks for all the world as though our "tiny dot" has a real shot at being the origin of a galaxy-scale civilization. It seems absurd, even delusional to believe in this possibility. But given our observations, it seems equally strange to dismiss it.

¹⁰ See <https://arxiv.org/pdf/1806.02404.pdf>

And if that's right, the choices made in the next 100,000 years - or even this century - could determine whether that galaxy-scale civilization comes to exist, and what values it has, across billions of stars and billions of years to come.

So when I look up at the vast expanse of space, I don't think to myself, "Ah, in the end none of this matters." I think: "Well, *some* of what we do probably doesn't matter. But *some* of what we do might matter more than anything ever will again. ...It would be really good if we could keep our eye on the ball. ...[gulp]"

The Duplicator: Instant Cloning Would Make The World Economy Explode

This is the second post in a series explaining my view that we could be in the most important century of all time. [Here's the roadmap for this series.](#)

- The [first piece](#) in this series discusses our unusual era, which could be very close to the transition between an Earth-bound civilization and a stable galaxy-wide one.
- Future pieces will discuss how “digital people” - and/or advanced AI - could be key for this transition.
- This piece explores a particularly important dynamic that could make either digital people or advanced AI lead to explosive productivity.

I explore the simple question of how the world would change if people could be “copied.” I argue that this could lead to unprecedented economic growth and productivity. Later, I will describe how digital people or advanced AI could similarly cause a growth/productivity explosion.

When some people imagine the future, they picture the kind of thing you see in sci-fi films. But these sci-fi futures seem very tame, compared to the future I expect.

In sci-fi, the future is different mostly via:

- Shiny buildings, gadgets and holograms.
- Robots doing many of the things humans do today.
- Advanced medicine.
- Souped up transportation, from hoverboards to flying cars to space travel and teleportation.

But fundamentally, there are the same kinds of people we see today, with the same kinds of personalities, goals, relationships and concerns.

The future I picture is enormously bigger, faster, weirder, and either much much better or much much worse compared to today. It's also potentially a lot *sooner* than sci-fi futures:¹¹ I think particular, achievable-seeming technologies could get us there quickly.

Such technologies could include “digital people” or particular forms of advanced AI - each of which I'll discuss in a future piece.

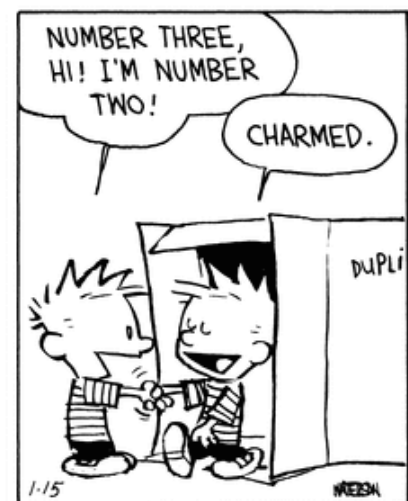
For now, I want to focus on just *one* aspect of what these sorts of technology would allow: the ability to make instant copies of people (or of entities with similar capabilities). Economic theory - and history - suggest that this ability, alone, could lead to unprecedented (in history or in sci-fi movies) levels of economic growth and productivity. This is via a self-reinforcing feedback loop in which innovation leads to more productivity, which leads to more “copies” of people, who in turn create more innovation and further increase productivity, which in turn ...

In this post, instead of directly discussing digital people or advanced AI, I'm going to keep things relatively simple and discuss a different hypothetical technology: the [Duplicator from Calvin & Hobbes](#), which simply copies people.

How the Duplicator works

The Duplicator is portrayed in [this series of comics](#). Its key feature is making an instant copy of a person: Calvin walks in, and two identical Calvins walk out

This is importantly different from the usual (and more realistic) version of “cloning,” in which a person's clone has the same DNA but has to start off as a baby and take years to become an adult.¹²



¹¹ For example, [Star Trek's](#) Captain Kirk first takes over the Enterprise in the mid-2200s. I think we could easily see a much more advanced, changed world than that of *Star Trek*, before 2100..

¹² [Example](#).

To flesh this out a bit, I'll assume that:

- The Duplicator allows any person to quickly make a copy of themselves, which starts from the same condition and mental state *or* from an earlier state (for example, I could make a replica of “Holden as of January 1, 2015”).¹³ Unlike in many sci-fi films, the copies function normally (they aren't evil or soulless or decaying or anything).
- It can be used to make an unlimited number of copies, though each has some noticeable cost of production (they aren't free).¹⁴

Productivity impacts

It seems that much of today's economy revolves around trying to make the most of “scarce human capital.” That is:

- Some people are “scarce” or “in demand.” Extreme examples include Barack Obama, Sundar Pichai, Beyonce Knowles and Jennifer Doudna.¹⁵ These people have some combination of skills, experience, knowledge, relationships, reputation, etc. that make it very hard for other people to do what they do. (Less extreme examples would be just about anyone who is playing a crucial role at an organization, hard to replace and often well paid.)
- These people end up overbooked, with far more demands on their time than they can fulfill. Armies of other people end up devoted to saving their time and working around their schedules.

The Duplicator would remove these bottlenecks. For example:

- Copies of Sundar Pichai could work at all levels of Google, armed with their ability to communicate easily with the CEO and make decisions as he would. They could also start new companies.
- Copies of the President of the U.S. could personally meet with any voter who wanted to interview the President, as well as with any Congresspeople or potential appointees or advisors the President didn't have time to

¹³ This isn't quite how it works in the comic, but it's how it'll work here.

¹⁴ The one in the comic burns out after a few copies, but that one's just a prototype.

¹⁵ Biologist who co-invented CRISPR and won a Nobel Prize in 2020.

meet with. They could deeply study key domestic and international issues and report back to the “original” President.

- Copies of Beyonce could make as many albums as the market could support. They could deeply study and specialize in different musical genres. They could even try living different lifestyles to gain different life experiences, all of which could inform different albums that still all shared Beyonce’s personal aesthetic and creativity. There would probably be at least one Beyonce copy whose music people considered better than the original’s; that one could further copy herself.
- Copies of Jennifer Doudna could investigate any of the ideas and experiments the original doesn’t have time to look into, as well as exploring the many fields she wasn’t able to specialize in. There could be Jennifer Doudna copies in physics, chemistry and computer science as well as biology, each collaborating with many other Jennifer Doudna copies.

(The ability to make copies for *temporary* purposes - and run them at different speeds - could further increase efficiency, as I’ll discuss in a future piece about digital people.)

Explosive growth

OK, the Duplicator would make the economy more productive - but *how much* more productive?

To answer, I’m going to briefly summarize what one might call the “**Population growth is the bottleneck to explosive economic growth**” viewpoint.

I would highly recommend reading more about this viewpoint at the following links, all of which I think are fascinating:

- [The Year The Singularity Was Cancelled](#) (Slate Star Codex - reasonably accessible if you have basic familiarity with **economic growth**)
- [Modeling the Human Trajectory](#) (Open Philanthropy’s David Roodman - reasonably accessible blog post, linking to dense technical report)
- [Could Advanced AI Drive Explosive Economic Growth?](#) (Open Philanthropy’s Tom Davidson - accessible blog post, linking to dense technical report) Here’s my rough summary.

In standard economic models, the total size of the economy (its total output, i.e., how much “stuff” it creates) is a function of:

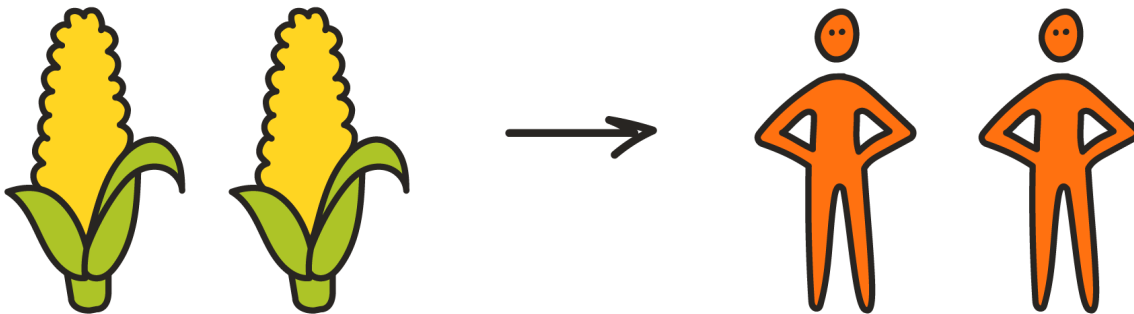
- How much total “labor” (people doing work) there is in the economy;
- How much “capital” (e.g., machines and energy sources - basically everything except labor) there is in the economy;
- How high productivity is, i.e., how much stuff is created for a given amount of labor and capital. (This is sometimes called “technology.”)

That is, the economy gets bigger when (a) there is more labor available, or (b) more capital (~everything other than labor) available, or when (c) productivity (“output per unit of labor/capital”) increases.

The total population (number of people) affects both labor and productivity, because people can have ideas that increase productivity.

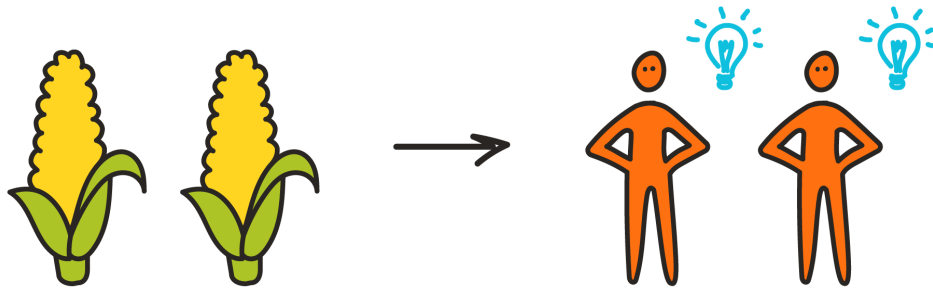
One way things *could* theoretically play out in an economy would be:

The economy starts with some set of resources (capital) supporting some set of people (population).

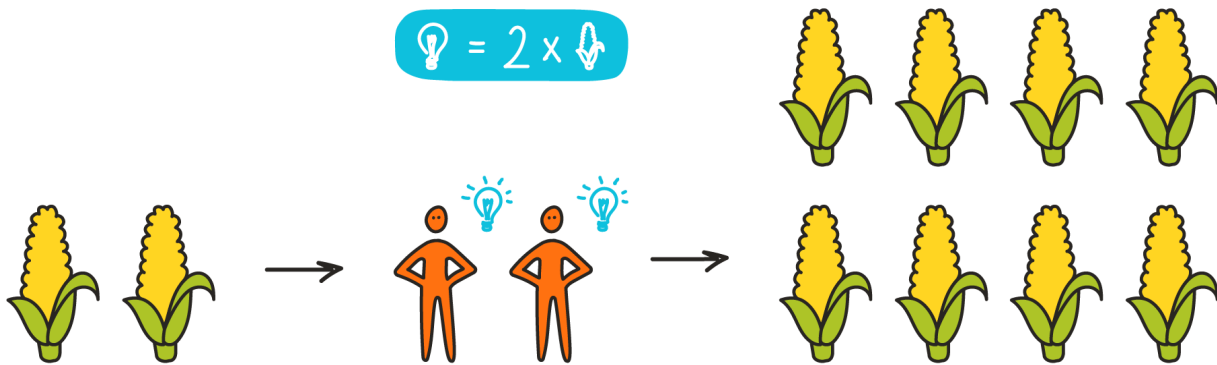


Thanks to María Gutiérrez Rojas for these graphics.

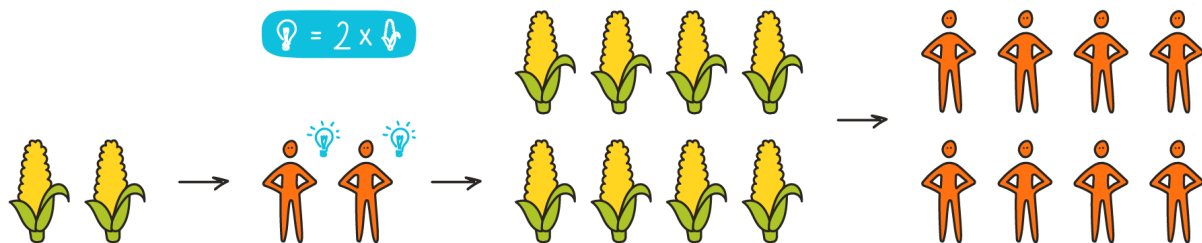
This set of people comes up with new ideas and innovations.



This leads to some amount of increased productivity, meaning there is more total economic output.¹⁶



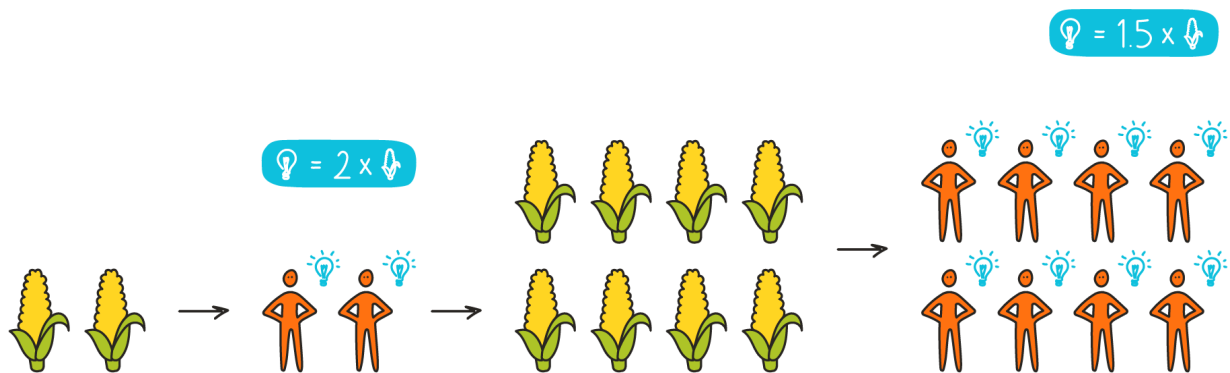
This means people can afford to have more children. They do, and the population grows more quickly.



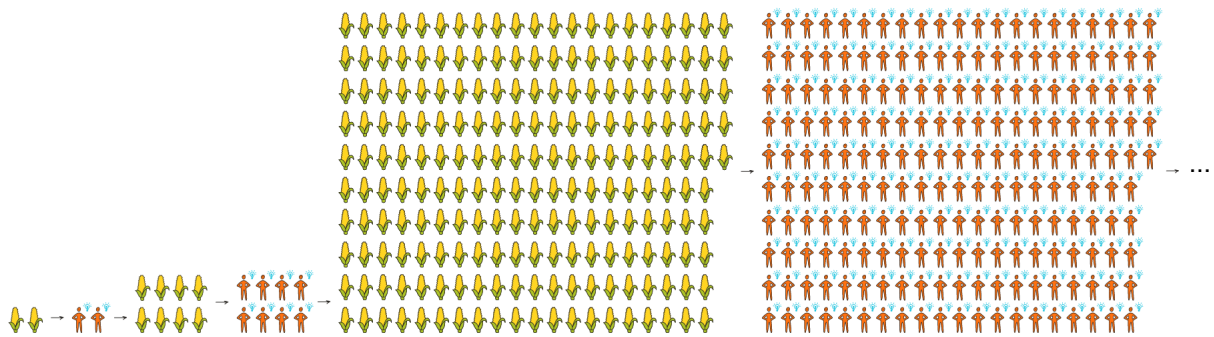
Because of that population growth, the economy comes up with new ideas and

¹⁶ Each idea doubled the amount of corn.

innovations *faster* than before (since more people means more new ideas).¹⁷



This leads to *even more* economic output and *even faster* population growth, in a self-reinforcing loop: *more ideas* → *more output* → *more people* → *more ideas* →



When you incorporate this full feedback loop into economic growth models,¹⁸ they predict that (under plausible assumptions) the world economy will

¹⁷ A faster-growing population doesn't *necessarily* mean faster technological advancement. There could be "diminishing returns": the first few ideas are easier to find than the next few, so even as the effort put into finding new ideas goes up, new ideas are found more slowly. ([Are Ideas Getting Harder To Find?](#) is a well-known paper on this topic.) More population = faster technological progress if the population is growing *faster* than the difficulty of finding new ideas is growing. This dynamic is portrayed in a simplified way in the graphic: initially people have ideas leading to doubling of corn output, but later the ideas only lead to a 1.5x'ing of corn output.

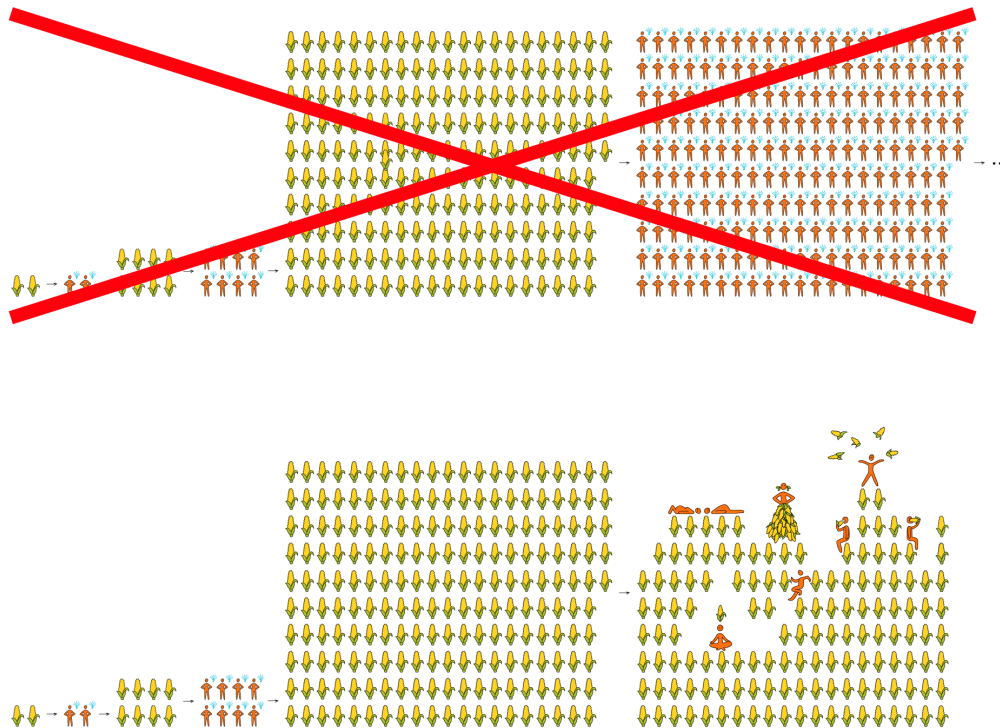
¹⁸ It's crucial to include the "more output -> more people" step, which is often not there by default, and doesn't describe today's world (but could describe a world with The Duplicator). It's standard for growth models to incorporate the other parts of the feedback loop: more people --> more ideas --> more

see **accelerating growth**.¹⁹ “Accelerating growth” is a fairly “explosive” dynamic in which the economy can go from small to extremely large with disorienting speed.

The pattern of growth predicted by these models seems like a reasonably good fit with the data on the world economy over the last 5,000 years (see [Modeling the Human Trajectory](#), though there is an open [debate](#) on this point; I discuss how the debate could change my conclusions [here](#)). **However, over the last few hundred years, growth has not accelerated; it has been “constant”** (a less explosive dynamic) **at around today’s level.**

Why did accelerating growth transition to constant growth?

This change coincided with the [demographic transition](#). In the demographic transition it **stopped being the case that having more output -> having more children**. Instead, more output just meant richer people, and people actually had *fewer* children as they became richer. This broke the self-reinforcing loop described above.



The demographic transition.

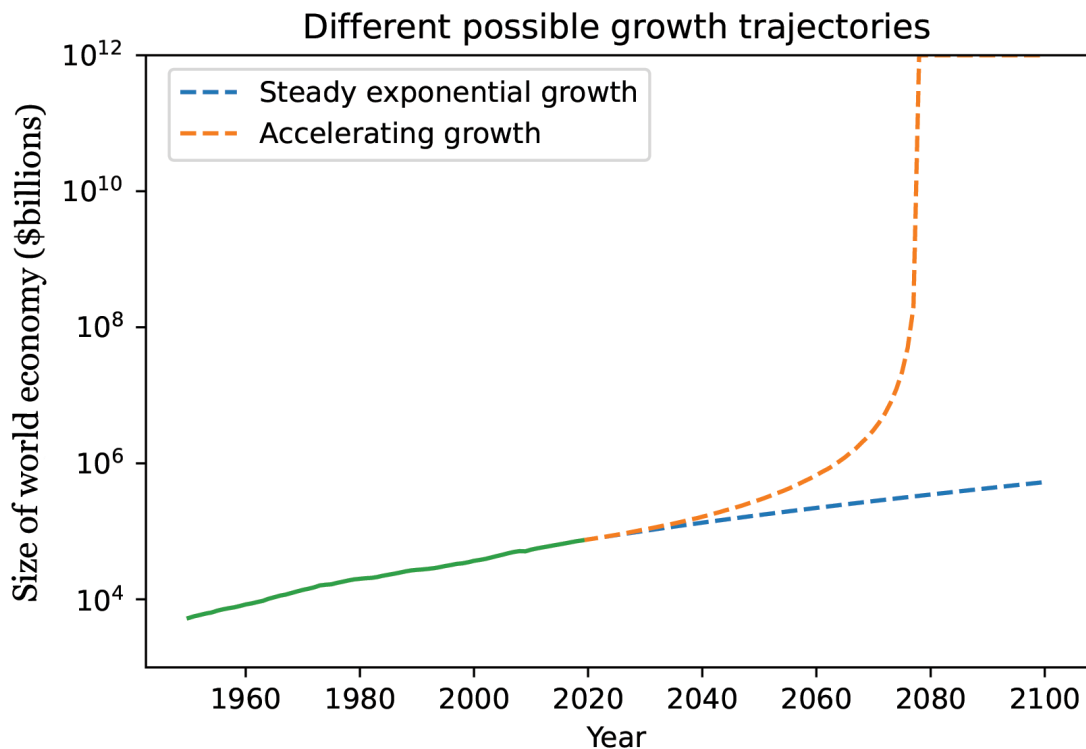
output.

¹⁹ This claim is defended in detail in [Could Advanced AI Drive Explosive Economic Growth?](#)

Raising children is a massive investment (of time and personal energy, not just “capital”), and children take a long time to mature. By changing what it takes to grow the population, the Duplicator could restore the accelerating feedback loop.

Period	Feedback loop?	Pattern of growth
Before the demographic transition	Yes: more ideas → more output → more people → more ideas→	Accelerating growth (economy can go from small to large disorientingly quickly)
Since the demographic transition	No: more ideas → more output → richer people	Constant growth (less explosive)
With the Duplicator	Yes: more ideas → more output → more people → more ideas→	Accelerating growth

This figure from [Could Advanced AI Drive Explosive Economic Growth?](#) illustrates how the next decades might look different with steady exponential growth vs. accelerating growth:



To see more detailed (but simplified) example numbers demonstrating the explosive growth, see footnote.²⁰

If we wanted to guess what a Duplicator might do in real life, we might imagine that it would get back to the kind of acceleration the world economy had historically, which loosely implies (based on [Modeling the Human Trajectory](#)) that **the economy would reach infinite size sometime in the next century.**²¹

Of course, that can't happen - at some point the size of the economy would be limited by fundamental natural resources, such as the number of atoms or amount of energy available in the galaxy. But in between here and running out of space/atoms/energy/something, we could easily see levels of economic growth that are massively faster than anything in history.

Over the last 100 years or so, the economy has doubled in size every few decades. With a Duplicator, it could double in size every year or month, on its way to hitting the limits.

Depending on how things played out, such productivity could result in an end to scarcity and material need, or in a dystopian race between different people making as many copies of themselves as possible in the hopes of taking over the population. (Or many in-between and other scenarios.)

Conclusion

I think the Duplicator would be a more powerful technology than warp drives, tricorders, laser guns²² or even teleporters. Minds are the source of innovation that can lead to all of those other things. So being cheaply able to duplicate them would be an extraordinary situation.

A harder-to-intuit, but even more powerful, technology would be **digital people**, e.g., the ability to run detailed simulations of people²³ on a comput-

²⁰ See endnotes ([1](#))

²¹ As noted above, there is an open [debate](#) on whether past economic growth actually follows the pattern described in [Modeling the Human Trajectory](#). I discuss how the debate could change my conclusions [here](#); I think there is a case either way for explosive growth this century.

²² TBH, I've never been able to figure out why these are better than regular guns.

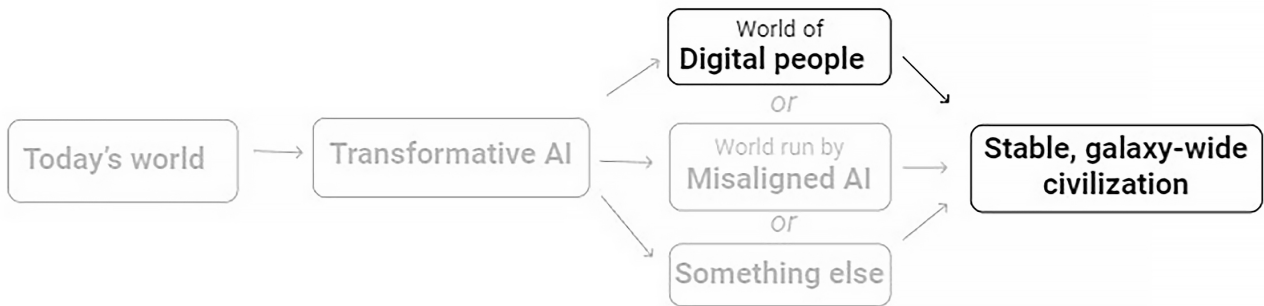
²³ Or of some sort of entity that's properly described as a "descendant" of people, as I'll discuss in the piece on digital people.

<https://www.cold-takes.com/the-duplicator>

er. Such simulated people could be copied Duplicator-style, and could also be sped up, slowed down, and reset, with virtual environments that were fully controlled.

I think that sort of technology is probably possible, and I expect a world with it to be even wilder than a world with the Duplicator. I'll elaborate on this in the next piece.

Digital People Would Be An Even Bigger Deal (Intro)



This is the third post in a series explaining my view that we could be in the most important century of all time. ([Here's the roadmap for this series.](#))

- The [first piece](#) in this series discusses our unusual era, which could be very close to the transition between an Earth-bound civilization and a stable galaxy-wide civilization.
- This piece discusses “digital people,” a category of technology that could be key for this transition (and would have even bigger impacts than the hypothetical [Duplicator](#) discussed previously).
- Many of the ideas here appear somewhere in sci-fi or speculative non-fiction, but I’m not aware of another piece laying out (compactly) the basic idea of digital people and the key reasons that a world of digital people would be so different from today’s.
- The idea of digital people provides a concrete way of imagining how the right kind of technology (which I believe to be almost certainly feasible) could change the world **radically**, such that “humans as we know them” would no longer be the main force.
- It will be important to have this picture, because I’m going to argue that AI advances this century could quickly lead to digital people

or similarly significant technology. The transformative potential of something like digital people, combined with how quickly AI could lead to it, form the case that we could be in the most important century.

Intro

[Previously](#), I wrote:

When some people imagine the future, they picture the kind of thing you see in sci-fi films. But these sci-fi futures seem very tame, compared to the future I expect ...

The future I picture is enormously bigger, faster, weirder, and either much much better or much much worse compared to today. It's also potentially a lot sooner than sci-fi futures: I think particular, achievable-seeming technologies could get us there quickly.

This piece is about **digital people**, one example²⁴ of a technology that could lead to an extremely big, fast, weird future.

To get the idea of digital people, imagine a computer simulation of a specific person, in a virtual environment. For example, a simulation of you that reacts to all “virtual events” - virtual hunger, virtual weather, a virtual computer with an inbox - just as you would. (Like [The Matrix](#)? See footnote.²⁵) I explain in more depth in the [FAQ companion piece](#).

The central case I'll focus on is that of digital people just like us, perhaps created via [mind uploading](#) (simulating human brains). However, one could also imagine entities unlike us in many ways, but still properly thought of as “descendants” of humanity; those would be digital people as well. (More on my choice of term [in the FAQ](#).)

²⁴ The best example I can think of, but surely not the only one.

²⁵ The movie *The Matrix* gives a decent intuition for the idea with its fully-immersive virtual reality, but unlike the heroes of *The Matrix*, a digital person need not be connected to any physical person - they could exist as pure software.

The agents (“bad guys”) are more like digital people than the heroes are. In fact, one extensively [copies himself](#).

Popular culture on this sort of topic tends to focus on the prospect of **digital immortality**: people avoiding death by taking on a digital form, which can be backed up just like you back up your data. But I consider this to be small potatoes compared to other potential impacts of digital people, in particular:

- **Productivity.** Digital people could be copied, just as we can easily make copies of ~any software today. They could also be run much faster than humans. Because of this, digital people could have effects comparable to those of the **Duplicator**, but more so: unprecedented (in history or in sci-fi movies) levels of economic growth and productivity.
- **Social science.** Today, we see a lot of progress on understanding scientific laws and developing cool new technologies, but not so much progress on understanding human nature and human behavior. Digital people would fundamentally change this dynamic: people could make copies of themselves (including sped-up, temporary copies) to explore how different choices, lifestyles and environments affected them. Comparing copies would be informative in a way that current social science rarely is.
- **Control of the environment.** Digital people would experience whatever world they (or the controller of their virtual environment) wanted. Assuming digital people had true conscious experience (an assumption discussed **in the FAQ**), this could be a good thing (it should be possible to eliminate disease, material poverty and non-consensual violence for digital people) or a bad thing (if human rights are not protected, digital people could be subject to scary levels of control).
- **Space expansion.** The population of digital people might become staggeringly large, and the computers running them could end up distributed throughout our galaxy and beyond. Digital people could exist anywhere that computers could be run - so space settlements could be more straightforward for digital people than for biological humans.
- **Lock-in.** In today's world, we're used to the idea that the future is unpredictable and uncontrollable. Political regimes, ideologies, and cultures all come and go (and evolve). But a community, city or nation of digital people could be much more stable.
 - Digital people need not die or age.
 - Whoever sets up a "virtual environment" containing a community of

digital people could have quite a bit of long-lasting control over what that community is like. For example, they might build in software to reset the community (both the virtual environment and the people in it) to an earlier state if particular things change - such as who's in power, or what religion is dominant.

- I consider this a disturbing thought, as it could enable long-lasting authoritarianism, though it could also enable things like permanent protection of particular human rights.

I think these effects (elaborated below) could be a very good or a very bad thing. How the early years with digital people go could irreversibly determine which.

I think similar consequences would arise from any technology that allowed (a) extreme control over our experiences and environment; (b) duplicating human minds. This means there are potentially **many ways for the future to become as wacky as what I sketch out here**. I discuss digital people because doing so provides a particularly easy way to imagine the consequences of (a) and (b): it is essentially about transferring the most important building block of our world (human minds) to a domain (software) where we are used to the idea of having a huge amount of control to program whatever behaviors we want.





Much of this piece is inspired by [Age of Em](#), an unusual and fascinating book. It tries to describe a hypothetical world of digital people (specifically mind uploads) in a lot of detail, but (unlike science fiction) it also aims for predictive accuracy rather than entertainment. In many places I find it overly specific, and overall, I don't expect that the world it describes will end up having much in common with a real digital-people-filled world. However, it has a number of sections that I think illustrate how powerful and radical a technology digital people could be.

Below, I will:

- Describe the basic idea of digital people, and link to a [FAQ](#) on the idea.
- Go through the potential implications of digital people, listed above.

This is a piece that different people may want to read in different orders. Here's an overall guide to the piece and FAQ:

	Normal humans	Digital people
Possible today (More)	✓	✗
Probably possible someday (More)	✓	✓
Can interact with the real world , do most jobs (More)	✓	✓
Conscious , should have human rights (More)	✓	✓
Easily duplicated , ala The Duplicator (More)	✗	✓
Can be run sped-up (More)	✗	✓
Can make " temporary copies " that run fast, then retire at slow speed (More)	✗	✓
Productivity and social science : could cause unprecedented economic growth, productivity, and knowledge of human nature and behavior (More)	✗	✓
Control of the environment : can have their experiences altered in any way (More)	✗	✓

Lock-in: could live in highly stable civilizations with no aging or death, and "digital resets" stopping certain changes (More)		
Space expansion: can live comfortably anywhere computers can run, thus highly suitable for galaxy-wide expansion (More)		
Good or bad? (More)	Outside the scope of this piece	Could be very good or bad

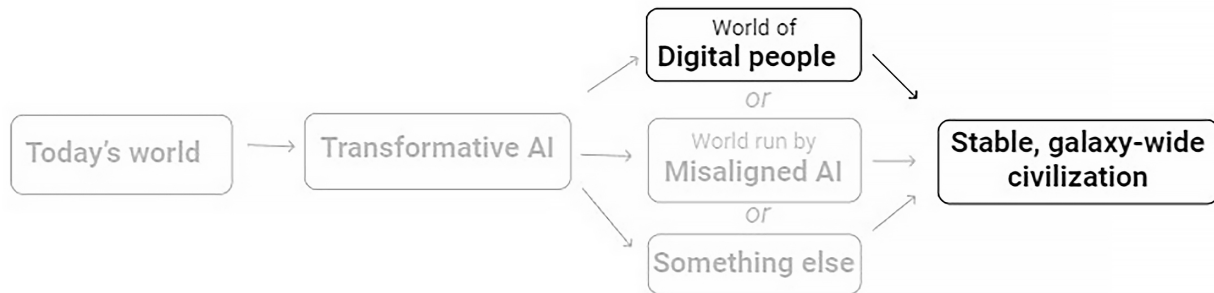
Premises

This piece focuses on how digital people could change the world. I will mostly assume that **digital people are just like us, except that they can be easily copied, run at different speeds, and embedded in virtual environments.** In particular, I will assume that digital people are conscious, have human rights, and can do most of the things humans can, including interacting with the real world.

I expect **many readers will have trouble engaging with this until they see answers to some more basic questions about digital people.** Therefore, I encourage readers to click on any questions that sound helpful from the [companion FAQ](#), or just read the FAQ straight through. If you are reading the “eBook” or “consolidated PDF” version of this series, the FAQ will be next, followed by the rest of this piece. It probably makes sense to skim the FAQ’s table of contents and then move on, depending on whether any of the questions seem interesting or important.

Digital People FAQ

Companion piece to “*Digital People Would Be An Even Bigger Deal.*”



This is a companion piece to [Digital People Would Be An Even Bigger Deal](#), which is the third in a series of posts about the possibility that we are in the [most important century for humanity](#)

This piece discusses basic questions about “digital people,” e.g., extremely detailed, realistic computer simulations of specific people. This is a hypothetical (but, I believe, realistic) technology that could be key for a transition to a [stable, galaxy-wide civilization](#). (The [other piece](#) describes the *consequences* of such a technology; this piece focuses on basic questions about how it might work.)

It will be important to have this picture, because I’m going to argue that AI advances this century could quickly lead to digital people or similarly significant technology. The transformative potential of something like digital people, combined with how quickly AI could lead to it, form the case that we could be in the most important century.

This table (also in the other piece) serves as a summary of the two pieces together:

	Normal humans	Digital people
Possible today (More)	✓	✗
Probably possible someday (More)	✓	✓
Can interact with the real world, do most jobs (More)	✓	✓
Conscious, should have human rights (More)	✓	✓
Easily duplicated, ala The Duplicator (More)	✗	✓
Can be run sped-up (More)	✗	✓
Can make " temporary copies " that run fast, then retire at slow speed (More)	✗	✓
Productivity and social science: could cause unprecedented economic growth, productivity, and knowledge of human nature and behavior (More)	✗	✓
Control of the environment: can have their experiences altered in any way (More)	✗	✓
Lock-in: could live in highly stable civilizations with no aging or death, and "digital resets" stopping certain changes (More)	✗	✓



Space expansion: can live comfortably anywhere computers can run, thus highly suitable for galaxy-wide expansion (More)		
Good or bad? (More)	Outside the scope of this piece	Could be very good or bad

Table of contents for this FAQ

- **Basics**
 - Basics of digital people
 - I'm finding this hard to imagine. Can you use an analogy?
 - Could digital people interact with the real world? For example, could a real-world company hire a digital person to work for it?
- **Humans and digital people**
 - Could digital people be conscious? Could they deserve human rights?
 - Let's say you're wrong, and digital people couldn't be conscious. How would that affect your views about how they could change the world?
- **Feasibility**
 - Are digital people possible?
 - How soon could digital people be possible?
- **Other questions**
 - I'm having trouble picturing a world of digital people - how the technology could be introduced, how they would interact with us, etc. Can you lay out a detailed scenario of what the transition from today's world to a world full of digital people might look like?
 - Are digital people different from mind uploads?
 - Would a digital copy of me be me?
 - What other questions can I ask?
 - Why does all of this matter?

Basics

Basics of digital people

To get the idea of digital people, imagine a computer simulation of a specific person, in a virtual environment. For example, a simulation of you that reacts to all “virtual events” (virtual hunger, virtual weather, a virtual computer with an inbox) just as you would.

The movie *The Matrix* gives a decent intuition for the idea with its fully-immersive virtual reality. But unlike the heroes of *The Matrix*, a digital person need not be connected to any physical person - they could exist as pure software.²⁶

Like other software, digital people could be copied (ala [The Duplicator](#)) and run at different speeds. And their virtual environments wouldn't have to obey the rules of the real world - they could work however the environment designers wanted. These properties drive most of the [consequences](#) I talk about in the main piece.

I'm finding this hard to imagine. Can you use an analogy?

There isn't anything today that's much like a digital person, but to start approaching the idea, consider this simulated person:



²⁶ The agents (“bad guys”) are more like digital people. In fact, one extensively [copies himself](#).

That's legendary football player Jerry Rice, as portrayed in the video game [Madden NFL 98](#). He probably represents the best anyone at that time (1997) could do to simulate the real Jerry Rice, in the context of a football game.

The idea is that this video game character runs, jumps, makes catches, drops the ball, and responds to tackles as closely as possible to how the real Jerry Rice would, in analogous situations. (At least, this is what he does when the video game player isn't explicitly controlling him.) The simulation is a very crude, simplified, limited-to-football-games version of real life.

Over the years, video games have advanced, and their simulations of Jerry Rice - as well as the rest of the players, the football field, etc. - have become more and more realistic:²⁷



OK, the last one is a photo of the real Jerry Rice. But imagine that the video game designers kept making their Jerry Rice simulations more and more realistic and the game's universe more and more expansive,²⁸ to the point where their simulated Jerry Rice would give interviews to virtual reporters, joke around with his virtual children, file his virtual taxes, and do *everything* else *exactly* how the real Jerry Rice would.

In this case, the simulated Jerry Rice would have a mind that works just like the real Jerry Rice's. It would be a "digital person" version of Jerry Rice.

Now imagine that one could do the same for ~everyone, and you're imagining a world of digital people.

²⁷ These are all taken from [this video](#), except for the last one.

²⁸ Football video games have already expanded to simulate [offseason tradings, signings and setting ticket prices](#)..

Could digital people interact with the real world? For example, could a real-world company hire a digital person to work for it?

Yes and yes.

- A digital person could be connected to a robot body. Cameras could feed in light signals to the digital person's mind, and microphones could feed in sound signals; the digital person could send out signals to e.g. move their hand, which would go to the robot. Humans can generally learn to control implants this way, so it seems very likely that digital people could learn to pilot robots.
- Digital people might inhabit a virtual "office" with a virtual monitor displaying their web browser, a virtual keyboard they could type on, etc. They could use this setup to send information over the internet just as biological humans do (and as today's bots do). So they could answer emails, write and send memos, tweet, and do other "remote work" pretty normally, without needing any real-world "body."
 - The virtual office need not be like the real world in all its detail - a pretty simple virtual environment with a basic "virtual computer" could be enough for a digital person to do most "remote work."
- They could also do phone and video calls with biological humans, by transmitting their "virtual face/voice" back to the biological human on the other end.

Overall, it seems you could have the same relationship to a digital person that you can have to any person whom you never meet in the flesh.

Humans and digital people

Could digital people be conscious? Could they deserve human rights?

Say there is a detailed digital copy of you, sending/receiving signals to/from a virtual body in a virtual world. The digital person sends signals telling the virtual body to put their hand on a virtual stove. As a consequence, the digital person receives signals that correspond to their hand burning. The digital

person processes these signals and sends further signals to their mouth to cry out “Ow!” and to their hand to jerk away from the virtual stove.

Does this digital person feel pain? Are they really “conscious” or “sentient” or “alive?” Relatedly, should we consider their experience of burning to be an unfortunate event, one we wish had been prevented so they wouldn’t have to go through this?

This is a question not about physics or biology, but about philosophy. And a full answer is outside the scope of this piece.

I believe **sufficiently detailed and accurate simulations of humans would be conscious, to the same degree and for the same reasons that humans are conscious.**²⁹

It’s hard to put a probability on this when it’s not totally clear what the statement even means, but I believe it is the best available conclusion given the state of academic philosophy of mind. I expect this view to be fairly common, though not universal, among philosophers of mind.³⁰

I will give an abbreviated explanation for why, via a couple of thought experiments.

Thought experiment 1. Imagine one could somehow replace a neuron in my brain with a “digital neuron”: an electrical device, made out of the same sorts of things today’s computers are made out of instead of what my neurons are made out of, that recorded input from other neurons (perhaps using a camera to monitor the various signals they were sending) and sent output to them in exactly the same pattern as the old neuron.

²⁹ It’s also possible there could be conscious “digital people” who did not resemble today’s humans, but I won’t go into that here - I’ll just focus on the concrete example of “digital people” that are virtual versions of humans.

³⁰ According to the [PhilPapers Surveys](#), 56.5% of philosophers endorse [physicalism](#), vs. 27.1% who endorse non-physicalism and 16.4% “other.” I expect the vast majority of philosophers who endorse [physicalism](#) to agree that a sufficiently detailed simulation of a human would be conscious. (My understanding is that [biological naturalism](#) is a fringe/unpopular position, and that physicalism + rejecting biological naturalism would imply believing that sufficiently detailed simulations of humans would be conscious.) I also expect that *some* philosophers who don’t endorse physicalism would still believe that such simulations would be conscious (David Chalmers is an example - see [The Conscious Mind](#)). These expectations are just based on my impressions of the field.

If we did this, I wouldn't behave differently in any way, or have any way of "noticing" the difference.

Now imagine that one did the same to every other neuron in my brain, one by one - such that my brain ultimately contained only "digital neurons" connected to each other, receiving input signals from my eyes/ears/etc. and sending output signals to my arms/feet/etc. I would still not behave differently in any way, or have any way of "noticing."

As you swapped out all the neurons, I would not notice the vividness of my thoughts dimming. Reasoning: if I did notice the vividness of my thoughts dimming, the "noticing" would affect me in ways that could ultimately change my behavior. For example, I might remark on the vividness of my thoughts dimming. But we've already specified that nothing about the inputs and outputs of my brain change, which means nothing about my behavior could change.

Now imagine that one could remove the set of interconnected "digital neurons" from my head, and feed in similar input signals and output signals directly (instead of via my eyes/ears/etc.). This would be a digital version of me: a simulation of my brain, running on a computer. And at no point would I have noticed anything changing - no diminished consciousness, no muted feelings, etc.

Thought experiment 2. Imagine that I was talking with a digital copy of myself - an extremely detailed simulation of me that reacted to every situation just as I would.

If I asked my digital copy whether he's conscious, he would insist that he is (just as I would in response to the same question). If I explained and demonstrated his situation (e.g., that he's "virtual") and asked whether he still thinks he's conscious, he would continue to insist that he is (just as I would, if I went through the experience of being shown that I was being simulated on some computer - something my current observations can't rule out).

I doubt there's any argument that could ever convince my digital counterpart that he's not conscious. If a reasoning process that works just like mine, with access to all the same facts I have access to, is convinced of "digital-Holden is conscious," what rational basis could I have for thinking this is wrong?

General points:

- I imagine that whatever else consciousness is, it is the cause of things like “I say that that I am conscious,” and the source of my observations about my own conscious experience. The fact that my brain is made out of neurons (as opposed to computer chips or something else) isn’t something that plays any role in my propensity to say I’m conscious, or in the observations I make about my own conscious experience: if my brain were a computer instead of a set of neurons, sending the same output signals, I would express all of the same beliefs and observations about my own conscious experience.
- The cause of my statements about consciousness and the source of my observations about my own consciousness is not something about *the material my brain is made of*; rather, it is something about *the patterns of information processing my brain performs*. A computer performing the same patterns of information processing would therefore have as much reason to think itself conscious as I do.
- Finally, my understanding from talking to physicists is that many of them believe there is some important sense in which “the universe can only be fundamentally understood as patterns of information processing,” and that the distinction between e.g. neurons and computer processors seems unlikely to have anything “deep” to it.³¹

³¹ From an email from a physicist friend: “I think a lot of people have the intuition that real neural activity, produced by real chemical reactions from real neurotransmitters, and real electrical activity that you can feel with your hand, somehow has some property that mere computer code can’t have. But one of the overwhelming messages of modern physics has been that everything that exists -- particles, fields, atoms, etc, is best thought of in terms of information, and may simply **be** information. The universe may perhaps be best described as a mathematical abstraction. Chemical reactions don’t come from some essential property of atoms but instead from subtle interactions between their valence electron shells. Electrons and protons aren’t well-defined particles, but abstract clouds of probability mass. Even the concept of “particles” is misleading; what seems to actually exist is quantum fields which are the solutions of abstract mathematical equations, and some of whose states are labeled by humans as “1 particle” or “2 particles”. To be a bit metaphorical, we are like tiny ripples on vast abstract mathematical waves, ripples whose patterns and dynamics happen to execute the information processing corresponding to what we call sentience. If you ask me our existence and the substrate we live on is already much weirder and more ephemeral than anything we might upload humans onto.”

For longer takes on this topic, see:

- Section 9 of [The Singularity: A Philosophical Analysis](#) by David Chalmers. Similar reasoning appears in part III of Chalmers's book [The Conscious Mind](#).
- [Zombies Redacted](#) by Eliezer Yudkowsky. This is more informal and less academic, and its arguments are more similar to the one I make above.

Let's say you're wrong, and digital people couldn't be conscious. How would that affect your views about how they could change the world?

Say we could make digital duplicates of today's humans, but they weren't conscious. In that case:

- They could still be enormously productive compared to biological humans. And studying them could still shed light on human nature and behavior. So the [Productivity](#) and [Social Science](#) sections would be pretty unchanged.
- They would still believe themselves to be conscious (since we do, and they'd be simulations of us). They could still seek to expand throughout space and establish stable/"locked-in" communities to preserve the values they care about.
- Due to their productivity and huge numbers, I'd expect the population of digital people to determine what the long-run future of the galaxy looks like - including for biological humans.
- The overall stakes would be lower, if the massive numbers of digital people throughout the galaxy and the virtual experiences they had "didn't matter." But the stakes would still be quite high, since how digital people set up the galaxy would determine what life was like for biological humans.

Feasibility

Are digital people possible?

They certainly aren't possible today. We have no idea how to create a piece of software that would "respond" to video and audio data (e.g., sending the same signals to talk, move, etc.) the way a particular human would.

We can't simply copy and simulate human brains, because relatively little is known about what the human brain does. Neuroscientists have very limited ability to make observations about it.³² (We can do a pretty good job simulating some of the key *inputs* to the brain - cameras seem to capture images about as well as human eyes, and microphones seem to capture sound about as well as human ears.³³)

Digital people are a hypothetical technology, and we may one day discover that they are impossible. But to my knowledge, there isn't any current reason to believe they're impossible.

I personally would bet that they will eventually be possible - at least via mind uploading (scanning and simulating human brains).³⁴ I think it is a matter of (a) neuroscience advancing to the point where we can thoroughly observe and characterize the key details of what human brains are doing - potentially a very long road, but not an endless one; (b) writing software that simulates those key details; (c) running the software simulation on a computer; (d) providing a "good enough" virtual body and virtual environment, which could be quite simple (enabling e.g. talking, reading, and typing would go a long way). I'd guess that (a) is the hard part, and would guess that (c) could be done even on today's computer hardware.³⁵

³² For an illustration of this, see this report: [How much computational power does it take to match the human brain?](#) (Particularly the [Uncertainty in neuroscience](#) section.) Even estimating *how many* meaningful operations the human brain performs is, today, very difficult and fraught - let alone characterizing what those operations are.

³³ This statement is based on my understanding of conventional wisdom plus the fact that recorded video and audio often seems quite realistic, implying that the camera/microphone didn't fail to record much important information about its source.

³⁴ This is assuming technology continues to advance, the species doesn't go extinct, etc.

³⁵ [This report concludes](#) that a computer costing ~\$10,000 today has enough computational power (10^{14} FLOP/s, a measure of computational power) to be within 1/10 of the author's best guess at

I won't elaborate on this in this piece, but might do so in the future if there's [interest](#).

How soon could digital people be possible?

I don't think we have a good way of forecasting when neuroscientists will understand the brain well enough to get started on mind uploading - other than to say that we don't seem anywhere near this today.

The reason I think digital people could come in the next few decades is different: I think we could invent *something else* (mainly, advanced artificial intelligence) that dramatically speeds up scientific research. If that happens, we could see all sorts of new world-changing technologies emerge quickly - including digital people.

I also think that thinking about digital people helps form intuitions about just how productive and powerful advanced AI could be (I'll discuss this in a future piece).

Other questions

I'm having trouble picturing a world of digital people - how the technology could be introduced, how they would interact with us, etc. Can you lay out a detailed scenario of what the transition from today's world to a world full of digital people might look like?

I'll give one example of how things could go. It's skewed somewhat to the optimistic side so it doesn't immediately become dystopia. And it's skewed toward the "familiar" side: I don't explore all the potential radical consequences of digital people.

what it would take to replicate the input-output behavior of a human brain (10^{15} FLOP/s). If we take the author's [high-end estimate](#) rather than best guess, it is about 10 million times as much computation (10^{22} FLOP/s), which would presumably cost \$1 trillion today - probably too high to be worth it, but computing is still getting cheaper. It's possible that replicating the input-output behavior alone wouldn't be enough detail to attain "consciousness," though I'd guess it would be, and either way it would be sufficient for the [productivity](#)" and [social science](#)" consequences.

Nothing else in the piece depends on this story being accurate; the only goal is to make it a bit easier to picture this world and think about the motivations of the people in it.

So imagine that:

One day, a working mind uploading technology becomes available. For simplicity, let's assume that it is modestly priced from the beginning.³⁶ What this means: anyone who wants can have their brain scanned, creating a “digital copy” of themselves.

A few tens of thousands of people create “digital copies” of themselves. So there are now tens of thousands of digital people living in a simple virtual environment, consisting of simple office buildings, apartments and parks.

Initially, each digital person thinks just like some non-digital person they were copied from, although as time goes on, their life experiences and thinking styles diverge.

Each digital person gets to design their own “virtual body” that represents them in the environment. (This is a bit like choosing an avatar - the bodies need to be in a normal range of height, weight, strength, etc. but are pretty customizable.)

The computer server running all of the digital people, and the virtual environment they inhabit, is privately owned. However, thanks to prescient regulation, the digital people themselves are considered to be people with full legal rights (not property of their creators or of the server company). They make their own choices, subject to the law, and they have some basic initial protections, such as:

- In order for them to continue existing, the owner of the server they're on must choose to run them. However, each digital person initially must have a pre-paid long-term contract with whatever server company is running them at first, so they can be assured of existing for a long time - say, at least 100 years from their biological copy's date of birth - if they want to.

³⁶ I actually expect it would start off very expensive, but become cheaper very quickly due to a productivity explosion, discussed below.

- They must be fully informed of their situation as a digital person and be given other information about what's going on, how to contact key people, etc. (Relatedly, initially only people 18 years and older can be digitally copied, although later digital people can have their own “digital children” - see below.)
- Their initial virtual environment has to initially meet certain criteria (e.g., no violence or suffering inflicted on them, ample virtual food and water). They have their own bank account that starts with some money in it, and they can make more just like biological people do (e.g., by doing work for some company).
- The server owner cannot make any significant changes to their virtual environment without their consent (other than ceasing to run them at all, which they can do after the contract runs out after some number of decades). Digital people may request, and offer money for, changes to their virtual environment (though any other affected digital people would need to give their consent too).
- The server owner must cease running any digital people who requests to stop existing.

Digital people form professional and personal relationships with each other. They also form personal and professional relationships with biological humans, whom they communicate with via email, video chat, etc.

- They might work for the first company offering digital copying of humans, doing research on how to make future digital people cheaper to run.
- They might stay in touch with the biological person they were copied from, exchanging emails about their personal lives.
- They would almost certainly be interested in ensuring that no biological humans interfered with their server in unwelcome ways (such as by shutting it off).

Some digital people fall in love and get married. A couple is able to “have children” by creating a new digital person whose mind is a hybrid of their two minds. Initially (subject to child abuse protections) they can decide how their child appears in the virtual environment, and even make some tweaks such as “When the child’s brain sends a signal to poop, a rainbow comes out instead.” The child gains rights as they age, as biological humans do.

Digital people are also allowed to copy themselves, as long as they are able to meet the requirements for new digital people (guarantee of being able to live for a reasonably long time, etc.) Copies have their own rights and don't owe anything to their creators.

The population of digital people grows, via people copying themselves and having children. Eventually (perhaps quickly, as discussed below), there are far more digital people than biological humans. Still, some digital people work for, employ or have personal relationships (via email, video chat, etc.) with biological humans.

- Many digital people work on making further population growth possible - by making it cheaper to run digital people, by building more computers (in the "real" world), by finding new sources of raw materials and energy for computers (also in the "real" world), etc.
- Many other digital people work on designing ever-more-creative virtual environments, some based on real-world locations, some more exotic (altered physics, etc.) Some virtual environments are designed to be lived in, while others are designed to be visited for recreation. Access is sold to digital people who want to be transferred to these environments.

So digital people are doing work, entertaining themselves, meeting each other, reproducing, etc. In these respects their lives have a fair amount in common with ours.

- Like us, they have some incentive to work for money - they need to pay for server costs if they want to keep existing for more than their initial contract says, or if they want to copy themselves or have children (they need to buy long server contracts for any such new digital people), or if they want to participate in various recreational environments and activities.
- Unlike us, they can do things like copying themselves, running at different speeds, changing their virtual bodies, entering exotic virtual environments (e.g., zero gravity), etc.

The prescient regulators have carved out ways for large groups of digital people to form their own virtual states and civilizations, which can set and change their own regulations.

Dystopian alternatives. A world of digital people could very quickly get dystopian if there were worse regulation, or no regulation. For example, imag-

ine if the rule were “Whoever owns the server can run whatever they want on it.” Then people might make digital copies of themselves that they ran experiments on, forced to do work, and even open-sourced, so that anyone running a server could make and abuse copies. [This very short story](#) (recommended, but chilling) gives a flavor for what that might be like.

There are other (more gradual) ways for a world of digital people to become dystopian, as outlined [here](#) (unassailable authoritarianism) and in [The Duplicator](#) (people racing to make copies of each other and dominate the population).

And what are the biological humans up to? Throughout this section, I’ve talked about how the world would be *for digital people*, not for normal biological humans. I’m more focused on that, because I expect that digital people would quickly become most of the population, and I think we should [care about them as much as we care about biological humans](#). But if you’re wondering what things would be like for biological humans, I’d expect that:

- Digital people, due to their numbers and running speeds, would become the dominant political and military players in the world. They would probably be the people determining what biological humans’ lives would be like.
- There would be very rapid scientific and technological advancement (as discussed below). So assuming digital people and biological humans stayed on good terms, I’d expect biological humans to have access to technology far beyond today’s. At a minimum, I expect this would mean pretty much unlimited medical technologies (including e.g. “curing” aging and having indefinitely long lifespans).

Are digital people different from mind uploads?

[Mind uploading](#) refers to simulating a human brain on a computer. (It is usually implied that this would not literally be an isolated brain, i.e., it would include some sort of virtual environment and body for the person being simulated, or perhaps they would be piloting a robot)

A mind upload would be one form of digital person, and most of this piece could have been written about mind uploads. Mind uploads are the most easy-to-imagine version of digital people, and I focus on them when I talk

about [why I think digital people will someday be possible](#) and [why they would be conscious like we are](#).

But I could also imagine a future of “digital people” that are not derived from copying human brains, or even all that similar to today’s humans. I think it’s reasonably likely that by the time digital people are possible (or pretty soon afterward), they will be quite different from today’s humans.³⁷

Most of this piece would apply to roughly any digital entities that (a) had moral value and human rights, like non-digital people; (b) could interact with their environments with equal (or greater) skill and ingenuity to today’s people. With enough understanding of how (a) and (b) work, it could be possible to design digital people without imitating human brains.

I’ll be referring to digital people a lot throughout [this series](#) to indicate how radically different the future could be. I don’t want to be read as saying that this would necessarily involve copying actual human brains.

Would a digital copy of me be me?

Say that someone scanned my brain and created a simulation of it on a computer: a digital copy of me. Would this count as “me”? Should I hope that this digital person has a good life, as much as I hope that for myself?

This is another philosophy question. My basic answer is “Sort of, but it doesn’t really matter much.” This piece is about how radically digital people could change the world; this doesn’t depend on whether we identify with our own digital copies.

It *does* depend (somewhat) on whether digital people should be considered “full persons” in the sense that we care about them, want them to avoid bad experiences, etc. The section on consciousness is more relevant to this question.

³⁷ I could also imagine a future in which the two key properties I list in the next paragraph - (a) moral value and human rights (b) human-level-or-above capabilities - were totally separated. That is, there could be a world full of (a) AIs with human-level-or-above capabilities, but no consciousness or moral value; (b) digital entities with moral value and conscious experience, but very few skills compared to AIs and even compared to today’s people. Most of what I say in this piece about a world of “digital people” would apply to such a world; in this case you could sort of think of a “digital people” as “teams” of AIs and morally-valuable-but-low-skill entities.

What other questions can I ask?

So many more!

E.g.: <https://tvtropes.org/pmwiki/pmwiki.php/Analysis/BrainUploading>

Why does all of this matter?

The piece that this is a companion for, [digital people would be an even bigger deal](#), spells out a number of ways in which digital people could lead to a radically unfamiliar future.

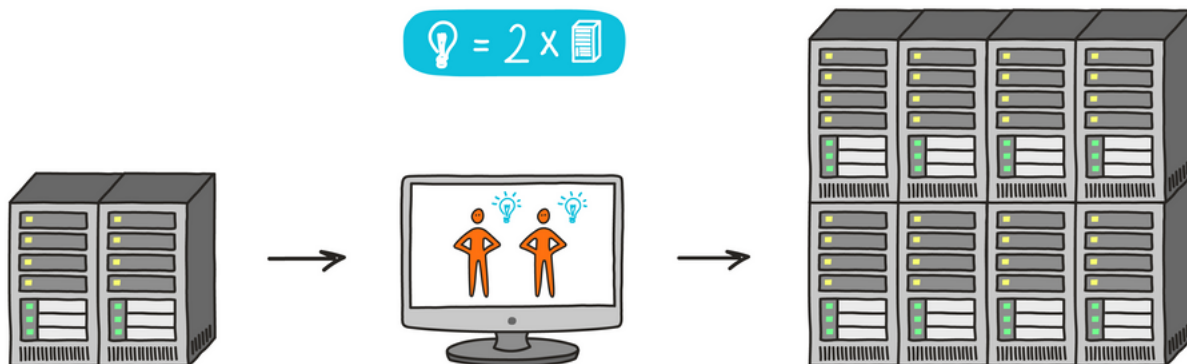
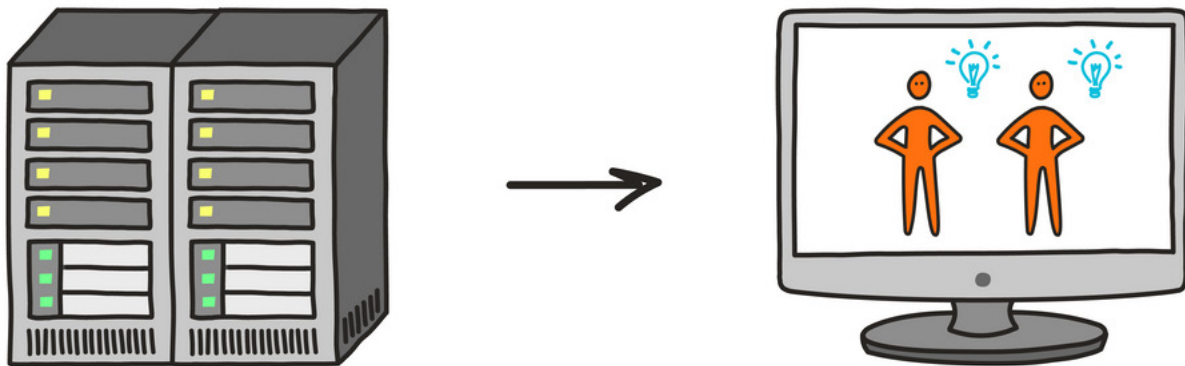
Elsewhere in [this series](#), I'm going to argue that AI advances this century could quickly lead to digital people or similarly significant technology. The transformative potential of something like digital people, combined with how quickly AI could lead to it, form the case that we could be in the most important century.

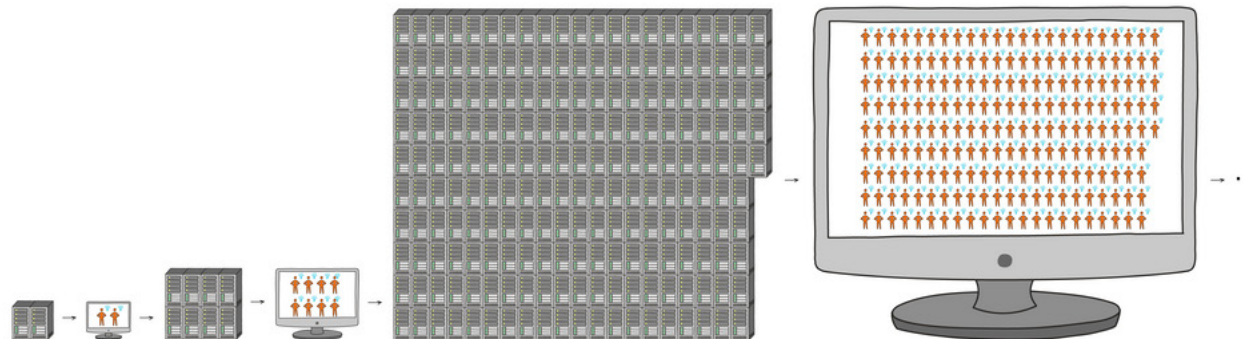
Digital People Would Be An Even Bigger Deal (Final Section)

How could digital people change the world?

Productivity

Like any software, digital people could be instantly and accurately copied. [The Duplicator](#) argues that the ability to “copy people” could lead to rapidly accelerating economic growth: “Over the last 100 years or so, the economy has doubled in size every few decades. With a Duplicator, it could double in size every year or month, on its way to hitting the limits.”





Thanks to María Gutiérrez Rojas for these graphics, a variation on a similar set of graphics from [The Duplicator](#) illustrating how duplicating people could cause explosive growth.

Digital people could create a more dramatic effect than this, because of their ability to be sped up (perhaps by thousands or millions of times)³⁸ as well as slowed down (to save on costs). This could further increase both speed and coordinating ability.³⁹

Another factor that could increase productivity: “Temporary” digital people could complete a task and then retire to a nice virtual life, while running very slowly (and cheaply).⁴⁰ This could make some digital people comfortable copying themselves for temporary purposes. Digital people could, for example, copy themselves hundreds of times to try different approaches to figuring out a problem or gaining a skill, then keep only the most successful version and make many copies of that version.

It's possible that digital people could be *less* of an economic force than [The Duplicator](#) since digital people would lack human bodies. But this seems likely to be only a minor consideration (details in footnote).⁴¹

³⁸ See *Age of Em* Chapter 6, starting with “Regarding the computation ...”

³⁹ For example, when multiple teams of digital people need to coordinate on a project, they might speed up (or slow down) particular steps and teams in order to make sure that each piece of the project is completed just on time. This would allow more complex, “fragile” plans to work out. (This point is from *Age of Em* Chapter 17, “Preparation” section.)

⁴⁰ See *Age of Em* Chapter 11, “Retirement” section.

⁴¹ See endnotes (2).

Social science

Today, we see a lot of impressive innovation and progress in some areas, and relatively little in other areas.

For example, we're constantly able to buy cheaper, faster computers and more realistic video games, but we don't seem to be constantly getting better at making friends, falling in love, or finding happiness.⁴² We also aren't clearly getting better at things like fighting addiction, and getting ourselves to behave as we (on reflection) want to.

One way of thinking about it is that *natural sciences* (e.g. physics, chemistry, biology) are advancing much more impressively than *social sciences* (e.g. economics, psychology, sociology). Or: "We're making great strides in understanding natural laws, not so much in understanding ourselves."

Digital people could change this. It could address what I see as perhaps the **fundamental reason social science is so hard to learn from: it's too hard to run true experiments and make clean comparisons.**

Today, if we we want to know whether meditation is helpful to people:

- We can compare people who meditate to people who don't, but there will be lots of differences between those people, and we can't isolate the effect of meditation itself. (Researchers try to do so with various statistical techniques, but these raise their own issues.)
- We could also try to run an experiment in which people are randomly assigned to meditate or not. But we need a lot of people to participate, all at the same time and under the same conditions, in the hopes that the differences between meditators and non-meditators will statistically "wash out" and we can pick up the effects of meditation. Today, these kinds of experiments - known as "randomized controlled trials" - are expensive, logistically challenging, time-consuming, and almost always end up with ambiguous and difficult-to-interpret results.

But in a world with digital people:

⁴² It is debatable whether the world is getting somewhat better at these things, somewhat worse, or neither. But it seems pretty clear that the progress isn't as impressive as in computing.

- Anyone could make a copy of themselves to try out meditation, perhaps even dedicating themselves to it for several years (possibly sped-up).⁴³ If they liked the results, they could then meditate for several years themselves, and ensure that all future copies were made from someone who had reaped the benefits of meditation.
- Social scientists could study people who had tried things like this and look for patterns, which would be much more informative than social science research tends to be now. (They could also run deliberate experiments, recruiting/paying people to make copies of themselves to try different lifestyles, cities, schools, etc. - these could be much smaller, cheaper, *and* more definitive than today's social science experiments.⁴⁴)

The ability to run experiments could be good or bad, depending on the robustness and enforcement of scientific ethics. If informed consent weren't sufficiently protected, digital people could open up the potential for an enormous amount of abuse; if it were, it could hopefully primarily enable learning.

Digital people could also enable:

- **Overcoming bias.** Digital people could make copies of themselves (including temporary, sped-up copies) to consider arguments delivered in different ways, by different people, including with different apparent race and gender, and see whether the copies came to different conclusions. In this way they could explore which cognitive biases - from sexism and rac-

⁴³ Why would the copy cooperate in the experiment? Perhaps because they simply were on board with the goal (I certainly would cooperate with a copy of myself trying to learn about meditation!). Perhaps because they were paid (in the form of a nice retirement after the experiment). Perhaps because they saw themselves and their copies (and/or original) as [the same person](#) (or at least cared a lot about these very similar people). A couple of factors that would facilitate this kind of experimentation: (a) digital people could examine their own state of mind to get a sense of the odds of cooperation (since the copy would have the same state of mind); (b) if only a small number of digital people experimented, large numbers of people could still learn from the results.

⁴⁴ I'd also expect them to be able to try more radical things. For example, in today's world, it's unlikely that you could run a randomized experiment on what happens if people currently living in New York just decide to move to Chicago. It would be too hard to find people willing to be randomly assigned to stay in New York or move to Chicago. But in a world of digital people, experimenters could pay New Yorkers to make copies of themselves who move to Chicago. And after the experiment, each Chicago copy that wished it had stayed in New York could choose to replace itself with another copy of the New York version. (The latter brings up questions about [philosophy of personal identity](#), but for social science purposes, all that matters is that *some* people would be happy to participate in experiments due to this option, and everyone could learn from the experiments.)

ism to wishful thinking and ego - affected their judgments, and work on improving and adapting to these biases. (Even if people weren't excited to do this, they might have to, as others would be able to ask for information on how biased they are and expect to get clear data.)

- **Bonanzas of reflection and discussion.** Digital people could make copies of themselves (including sped-up, temporary copies) to study and discuss particular philosophy questions, psychology questions, etc. in depth, and then summarize their findings to the original.⁴⁵ By seeing how different copies with different expertises and life experiences formed different opinions, they could have much more thoughtful, informed answers than I do to questions like “What do I want in life?”, “Why do I want it?”, “How can I be a person I’m proud of being?”, etc.

Virtual reality and control of the environment

As stated above, digital people could live in “virtual environments.” In order to design a virtual environment, programmers would systematically generate the right sort of light signals, sound signals, etc. to send to a digital person as if they were “really there.”

One could say the historical role of science and technology is to give people more control over their environment. And one could think of digital people almost as the logical endpoint of this: digital people would experience whatever world they (or the controller of their virtual environment) wanted.

This could be a very bad or good thing:

Bad thing. Someone who controlled a digital person’s virtual environment could have almost unlimited control over them.

- For this reason, it would be important for a world of digital people to in-



⁴⁵ See footnote from the first bullet point on why people’s copies might cooperate with them.

clude effective enforcement of basic human rights for all digital people. (More on this idea [in the FAQ.](#))

- A world of digital people could very quickly get dystopian if digital people didn't have human rights protections. For example, imagine if the rule were “Whoever owns a server can run whatever they want on it, including digital copies of anyone.” Then people might make “digital copies” of themselves that they ran experiments on, forced to do work, and even open-sourced, so that anyone running a server could make and abuse copies. [This very short story](#) (recommended, but chilling) gives a flavor for what that might be like.

Good thing. On the other hand, if a digital person were in control of their own environment (or someone else was and looked out for them), they could be free from any experiences they wanted to be free from, including hunger, violence, disease, other forms of ill health, and debilitating pain of any kind. Broadly, they could be “free from material need” - other than the need for computing resources to be run at all.



- This is a big change from today's world. Today, if you get cancer, you're going to suffer pain and debilitation even if everyone in the world would prefer that you didn't. Digital people need not experience having cancer if they and others don't want this to happen.
- In particular, physical coercion within a virtual environment could be made impossible (it could simply be impossible to transmit signals to another digital person corresponding to e.g. being punched or shot).
- Digital people might also have the ability to experience a lot of things we can't experience now - inhabiting another person's body, going to outer space, being in a “dangerous” situation without actually being in danger, eating without worrying about health consequences, changing from one apparent race or gender to another, etc.

Space expansion

If digital people underwent an explosion of economic growth as discussed above, this could come with an explosion in the *population* of digital people (for reasons discussed in [The Duplicator](#)).

It might reach the point where they needed to build spaceships and leave the solar system in order to get enough energy, metal, etc. to build more computers and enable more lives to exist.

Settling space could be much easier for digital people than for biological humans. They could exist anywhere one could run computers, and the basic ingredients needed to do that - raw materials, energy, and “real estate”⁴⁶ - are all super-abundant throughout our galaxy, not just on Earth. Because of this, the population of digital people could end up becoming staggeringly large.⁴⁷

Lock-in

In today’s world, we’re used to the idea that the future is unpredictable and uncontrollable. Political regimes, ideologies, and cultures all come and go (and evolve). Some are good, and some are bad, but it generally doesn’t seem as though anything will last forever. But communities, cities, and nations of digital people could be much more stable.

First, because digital people need not die or physically age, and their environment need not deteriorate or run out of anything. As long as they could keep their server running, everything in their virtual environment would be physically capable of staying as it is.

Second, because an environment could be designed to *enforce* stability. For example, imagine that:

- A community of digital people forms its own government (this would require either overpowering or getting consent from their original government).

⁴⁶ And air for cooling.

⁴⁷ See the estimates in [Astronomical Waste](#) for a rough sense of how big the numbers can get here (although these estimates are extremely speculative).

- The government turns authoritarian and repeals the basic human rights protections discussed [in the FAQ](#).
- The head wants to make sure that they - or perhaps their ideology of choice - stays in power forever.
- They could overhaul the virtual environment that they and all of the other citizens are in (by gaining access to the source code and reprogramming it, or operating robots that physically alter the server), so that certain things about the environment can never be changed - such as who's in power. If such a thing were about to change, the virtual environment could simply prohibit the action or reset to an earlier state.
- It would still be possible to change the virtual environment from outside - e.g., to physically destroy, hack or otherwise alter the server running it. But if this were taking place after a long period of population growth and space colonization, then the server might be way out in outer space, light-years from anyone who'd be interested in doing such a thing.

Alternatively, “digital correction” could be a force for good if used wisely enough. It could be used to ensure that no dictator ever gains power, or that certain basic human rights are always protected. If a civilization became “mature” enough - e.g., fair, equitable and prosperous, with a commitment to freedom and self-determination and a universally thriving population - it could keep these properties for a very long time.

I'm not aware of many in-depth analyses of the “lock-in” idea, but I elaborate further on this idea [here](#). (Additionally, [here are some informal notes](#) from physicist [Jess Riedel](#).)

Would these impacts be a good or bad thing?

Throughout this piece, I imagine many readers have been thinking “That sounds terrible! Does the author think it would be good?” Or “That sounds great! Does the author disagree?”

My take on a future with digital people is that it **could be very good or very bad, and how it gets set up in the first place could irreversibly determine which.**

- Hasty use of lock-in (discussed [above](#)) and/or overly quick spreading out through the galaxy (discussed [above](#)) could result in a huge world full of digital people (as conscious as we are) that is heavily dysfunctional, dystopian or at least falling short of its potential.
- But acceptably good initial conditions (protecting basic human rights for digital people, at a minimum), plus a lot of patience and accumulation of wisdom and self-awareness we don't have today (perhaps facilitated by [better social science](#)), could lead to a large, stable, much better world. It should be possible to eliminate disease, material poverty and non-consensual violence, and create a society much better than today's.

This Can't Go On

This piece starts to make the case that we live in a remarkable century, not just a remarkable era. Previous pieces in this [series](#) talked about the strange future that could be ahead of us eventually (maybe 100 years, maybe 100,000).

Summary of this piece:

- We're used to the world economy growing a few percent per year. This has been the case for many generations.
- However, this is a very unusual situation. Zooming out to all of history, we see that growth has been accelerating; that it's near its historical high point; and that it's faster than it can be for all that much longer (there aren't enough atoms in the galaxy to sustain this rate of growth for even another 10,000 years).
- The world can't just keep growing at this rate indefinitely. We should be ready for other possibilities: stagnation (growth slows or ends), explosion (growth accelerates even more, before hitting its limits), and collapse (some disaster levels the economy).

The times we live in are unusual and unstable. We shouldn't be surprised if something wacky happens, like an explosion in economic and scientific progress, leading to [technological maturity](#). In fact, such an explosion would arguably be right on trend.

For as long as any of us can remember, the world economy has grown⁴⁸ a few percent per year, on average. Some years see more or less growth than other years, but growth is pretty steady overall.⁴⁹ I'll call this the **Business As Usual** world.

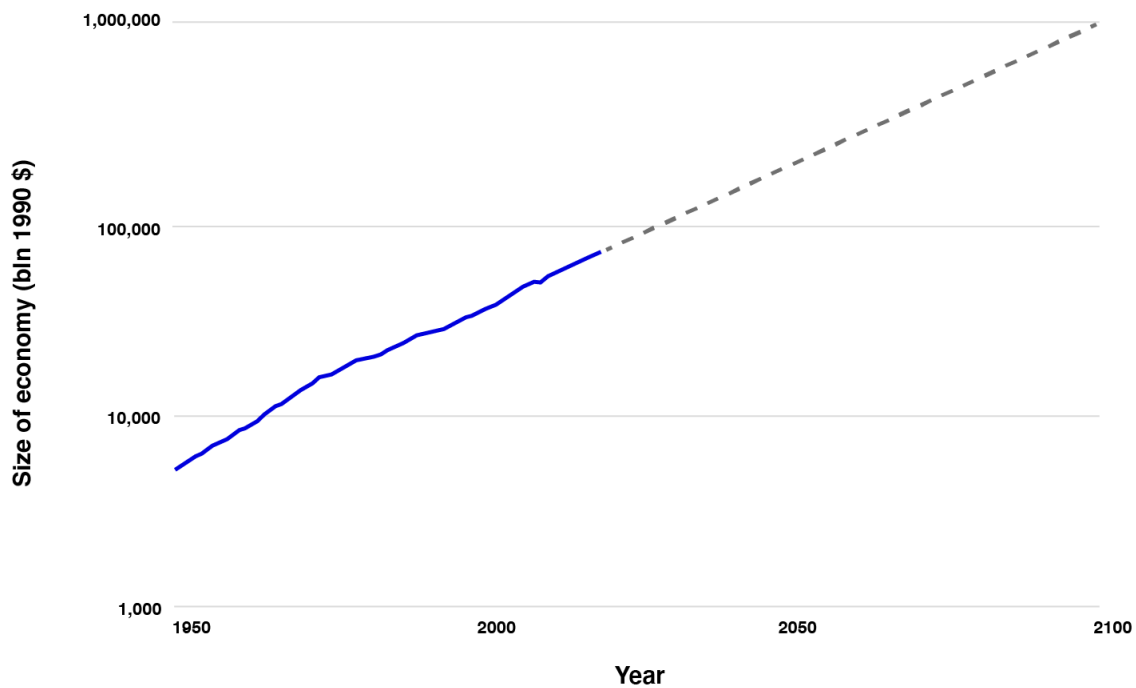
In Business As Usual, the world is constantly changing, and the change is noticeable, but it's not overwhelming or impossible to keep up with. There is a constant stream of new opportunities and new challenges, but if you want

⁴⁸ If you have no idea what that means, try my short [economic growth explainer](#).

⁴⁹ Global real growth has generally ranged from slightly negative to ~7% per year.

to take a few extra years to adapt to them while you mostly do things the way you were doing them before, you can usually (personally) get away with that. In terms of day-to-day life, 2019 was pretty similar to 2018, noticeably but not hugely different from 2010, and hugely but not crazily different from 1980.⁵⁰

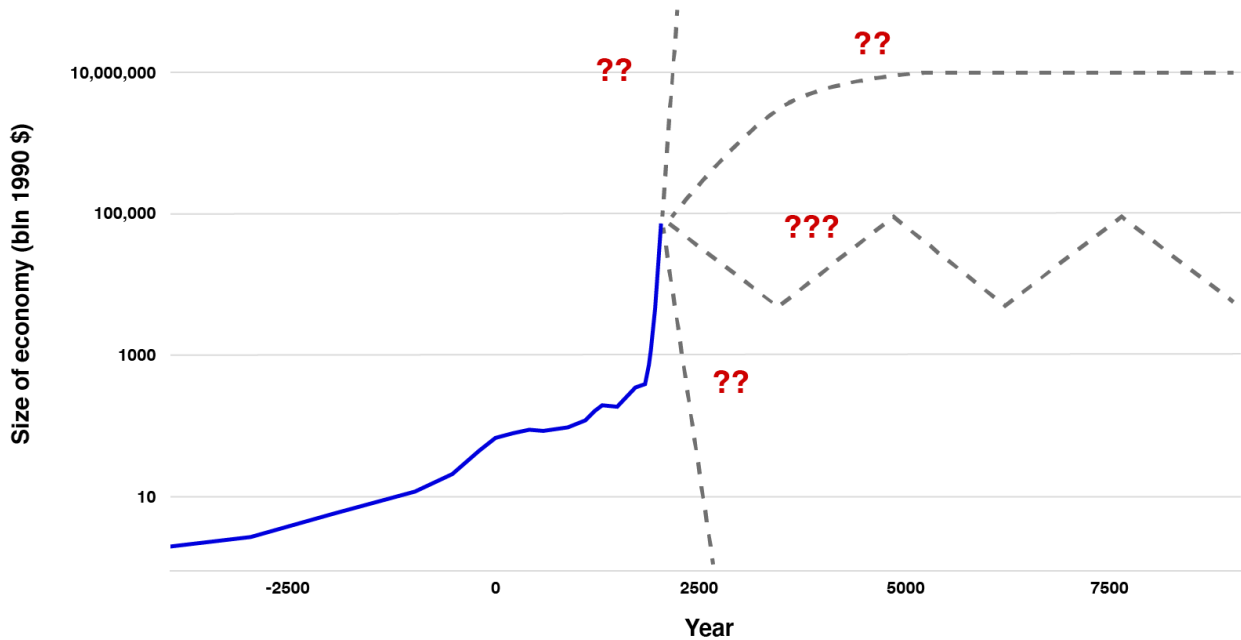
If this sounds right to you, and you're used to it, and you picture the future being like this as well, then you live in the Business As Usual headspace. When you think about the past and the future, you're probably thinking about something kind of like this:



Business As Usual

I live in a different headspace, one with a more turbulent past and a more uncertain future. I'll call it the **This Can't Go On** headspace. Here's my version of the chart:

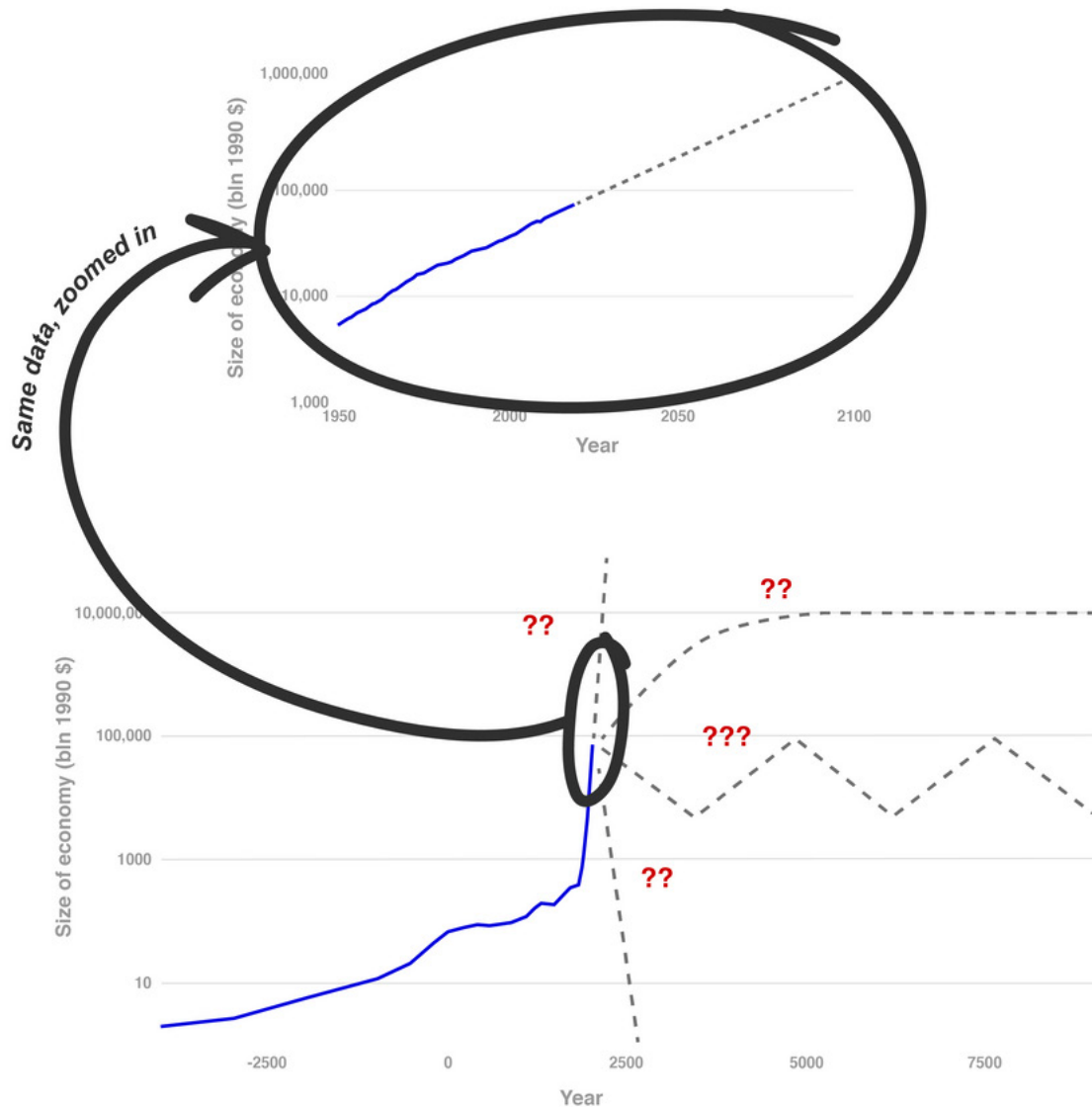
⁵⁰ I'm skipping over 2020 here since it was unusually different from past years, due to the global pandemic and other things.



This Can't Go On⁵¹.

Which chart is the right one? Well, they're using exactly the same historical data - it's just that the Business As Usual chart starts in 1950, whereas This Can't Go On starts all the way back in 5000 BC. **“This Can't Go On” is the whole story; “Business As Usual” is a tiny slice of it.**

⁵¹ For the historical data, see [Modeling the Human Trajectory](#). The projections are rough and meant to be visually suggestive rather than using the best modeling approaches..



Growing at a few percent a year is what we're all used to. But in full historical context, growing at a few percent a year is crazy. (It's the part where the blue line goes near-vertical.)

This growth has gone on for longer than any of us can remember, but that isn't very long in the scheme of things - just a couple hundred years, out of thousands of years of human civilization. It's a huge acceleration, and it can't go on all that much longer. (I'll flesh out "it can't go on all that much longer" [below](#).)

The first chart suggests regularity and predictability. The second suggests volatility and dramatically different possible futures.

One possible future is **stagnation**: we'll reach the economy's "maximum size" and growth will essentially stop. We'll all be concerned with how to divide up the resources we have, and the days of a growing pie and a dynamic economy will be over forever.

Another is **explosion**: growth will accelerate further, to the point where the world economy is doubling every year, or week, or hour. A **Duplicator**-like technology (such as **digital people** or, as I'll discuss in future pieces, advanced AI) could drive growth like this. If this happens, everything will be changing far faster than humans can process it.

Another is **collapse**: a global catastrophe will bring civilization to its knees, or wipe out humanity entirely, and we'll never reach today's level of growth again.

Or maybe something else will happen.

Why can't this go on?

A good starting point would be [this analysis from Overcoming Bias](#), which I'll give my own version of here:

- Let's say the world economy is currently getting 2% bigger each year.⁵² This implies that the economy would be doubling in size about every 35 years.⁵³
- If this holds up, then 8200 years from now, the economy would be about 3×10^{70} times its current size.

⁵² This refers to real GDP growth (adjusted for inflation). 2% is lower than the current world growth figure, and using the world growth figure would make my point stronger. But I think that 2% is a decent guess for "frontier growth" - growth occurring in the already-most-developed economies - as opposed to total world growth, which includes "catchup growth" (previously poor countries growing rapidly, such as China today).

To check my 2% guess, I downloaded [this US data](#) and looked at the annualized growth rate between 2000-2020, 2010-2020, and 2015-2020 (all using July since July was the latest 2020 point). These were 2.5%, 2.2% and 2.05% respectively.

⁵³ 2% growth over 35 years is $(1 + 2\%)^{35} = 2x$ growth

- There are likely fewer than 10^{70} atoms in our galaxy,⁵⁴ which we would not be able to travel beyond within the 8200-year time frame.⁵⁵
- So if the economy were $3 \cdot 10^{70}$ times as big as today's, and could only make use of 10^{70} (or fewer) atoms, we'd need to be sustaining **multiple economies as big as today's entire world economy *per atom***

8200 years might sound like a while, but it's far less time than humans have been around. In fact, it's less time than human (agriculture-based) civilization has been around.

Is it *imaginable* that we could develop the technology to support multiple equivalents of today's entire civilization, per atom available? Sure - but this would require a radical degree of transformation of our lives and societies, far beyond how much change we've seen over the course of human history to date. And I wouldn't exactly *bet* that this is how things are going to go over the next several thousand years. (**Update:** for people who aren't convinced yet, I've [expanded on this argument in another post.](#))

It seems much more likely that we will "run out" of new scientific insights, technological innovations, and resources, and the regime of "getting richer by a few percent a year" will come to an end. After all, this regime is only a couple hundred years old.

([This post](#) does a similar analysis looking at energy rather than economics. It projects that the limits come even sooner. It assumes 2.3% annual growth in energy consumption (less than the historical rate for the USA since the 1600s), and estimates this would use up as much energy as is produced by all the stars in our galaxy within 2500 years.⁵⁶)

⁵⁴ [Wikipedia](#)'s highest listed estimate for the Milky Way's mass is $4.5 \cdot 10^{12}$ solar masses, each of which [is](#) about $2 \cdot 10^{30}$ kg. The mass of a (hydrogen) atom [is](#) estimated as the equivalent of about $1.67 \cdot 10^{-27}$ kg. (Hydrogen atoms have the lowest mass, so assuming each atom is hydrogen will overestimate the total number of atoms.) So a high-end estimate of the total number of atoms in the Milky Way would be $(4.5 \cdot 10^{12} \cdot 2 \cdot 10^{30}) / (1.67 \cdot 10^{-27}) \approx 5.4 \cdot 10^{69}$.

⁵⁵ [Wikipedia](#): "In March 2019, astronomers reported that the mass of the Milky Way galaxy is 1.5 trillion solar masses within a radius of about 129,000 light-years." I'm assuming we can't travel more than 129,000 light-years in the next 8200 years, because this would require far-faster-than-light travel.

⁵⁶ This calculation isn't presented straightforwardly in the post. The key lines are "No matter what the technology, a sustained 2.3% energy growth rate would require us to produce as much energy as the entire sun within 1400 years" and "The Milky Way galaxy hosts about 100 billion stars. Lots of energy just spewing into space, there for the taking. Recall that each factor of ten takes us 100 years down

Explosion and collapse

So one possible future is stagnation: growth gradually slows over time, and we eventually end up in a no-growth economy. But I don't think that's the most likely future.

The chart above **doesn't show growth slowing down - it shows it accelerating dramatically**. What would we expect if we simply projected that same acceleration forward?

[Modeling the Human Trajectory](#) (by Open Philanthropy's David Roodman) tries to answer exactly this question, by "fitting a curve" to the pattern of past economic growth.⁵⁷ Its extrapolation implies ***infinite growth this century***. Infinite growth is a mathematical abstraction, but you could read it as meaning: "We'll see the fastest growth possible before we hit the limits."

In [The Duplicator](#), I summarize a broader discussion of this possibility. The upshot is that a growth explosion could be possible, *if* we had the technology to "copy" human minds - or something else that fulfills the same effective purpose, such as [digital people](#) or advanced enough AI.

In a growth explosion, the annual growth rate could hit 100% (the world economy doubling in size every year) - which could go on for at most ~250 years before we hit the kinds of limits discussed above.⁵⁸ Or we could see even faster growth - we might see the world economy double in size every month (which we could sustain for at most 20 years before hitting the limits⁵⁹), or faster. That would be a wild ride: blindingly fast growth, perhaps driven by AIs producing output beyond what we humans could meaningfully track, quickly approaching the limits of what's possible, at which point growth would have to slow.

the road. One-hundred billion is eleven factors of ten, so 1100 additional years." $1400 + 1100 = 2500$, the figure I cite. This relies on the assumption that the average star in our galaxy offers about as much energy as the sun; I don't know whether that's the case.

⁵⁷ There is an [open debate](#) on whether [Modeling the Human Trajectory](#) is fitting the right sort of shape to past historical data. I discuss how the debate could change my conclusions [here](#).

⁵⁸ 250 doublings would be a growth factor of about $1.8 \cdot 10^{75}$, over 10,000 times the number of atoms in our galaxy.

⁵⁹ 20 years would be 240 months, so if each one saw a doubling in the world economy, that would be a growth factor of about $1.8 \cdot 10^{72}$, over 100 times the number of atoms in our galaxy.

In addition to stagnation or explosive growth, there's a third possibility: **collapse**. A global catastrophe could cut civilization down to a state where it never regains today's level of growth. Human extinction would be an extreme version of such a collapse. This future isn't suggested by the charts, but we know it's possible.

As Toby Ord's [The Precipice](#) argues, asteroids and other "natural" risks don't seem likely to bring this about, but there are a few risks that seem serious and very hard to quantify: climate change, nuclear war (particularly nuclear winter), pandemics (particularly if advances in biology lead to nasty bio-weapons), and risks from advanced AI.

With these three possibilities in mind (stagnation, explosion and collapse):

- We live in one of the (two) fastest-growth centuries in all of history so far. (The 20th and 21st.)
- It seems likely that this will at least be one of the ~80 fastest-growing centuries of all time.⁶⁰
- If the right technology comes along and drives explosive growth, it could be the #1 fastest-growing century of all time - by a lot.
- If things go badly enough, it could be our last century.

So it seems like this is a quite remarkable century, with some chance of being the most remarkable. This is all based on pretty basic observations, not detailed reasoning about AI (which I will get to in future pieces).

Scientific and technological advancement

It's hard to make a simple chart of how fast science and technology are advancing, the same way we can make a chart for economic growth. But I think that if we could, it would present a broadly similar picture as the economic growth chart.

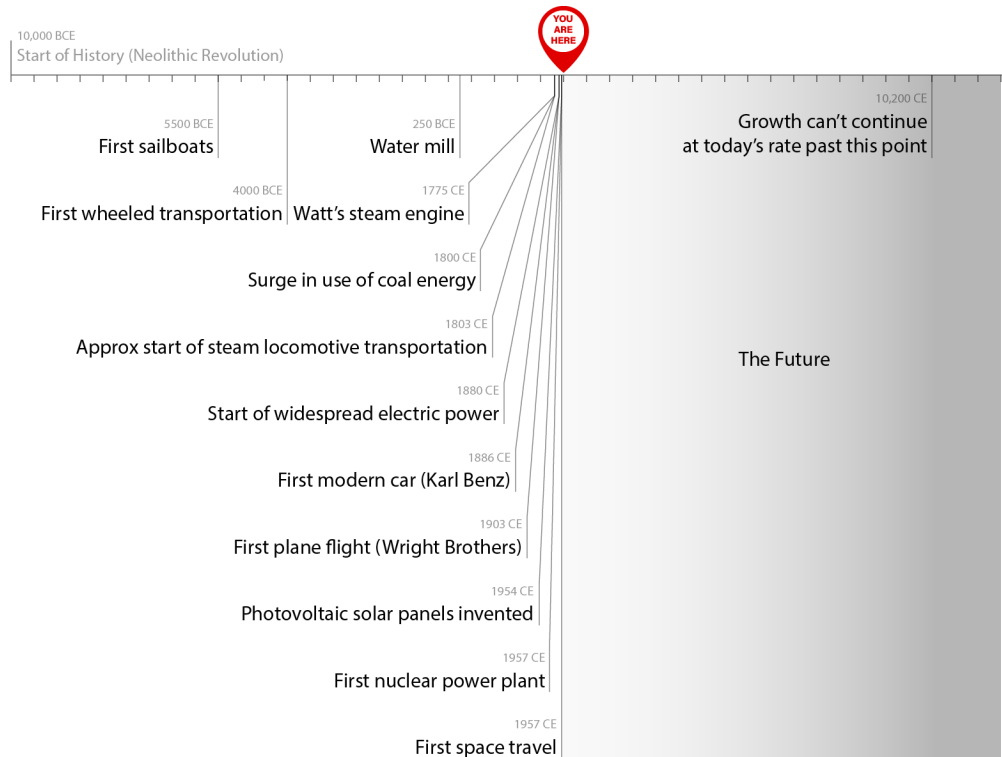
⁶⁰ That's because of the above observation that today's growth rate can't last for more than another 8200 years (82 centuries) or so. So the only way we could have more than 82 more centuries with growth equal to today's is if we also have a lot of centuries with negative growth, ala the zig-zag dotted line in the "This Can't Go On" chart.

A fun book I recommend is [Asimov's Chronology of Science and Discovery](#). It goes through the most important inventions and discoveries in human history, in chronological order. The first few entries include “stone tools,” “fire,” “religion” and “art”; the final pages include “Halley’s comet” and “warm superconductivity.”

An interesting fact about this book is that **553 out of its 654 pages take place after the year 1500** - even though it starts in the year 4 million BC. I predict other books of this type will show a similar pattern,⁶¹ and I believe there were, in fact, more scientific and technological advances in the last ~500 years than the previous several million.⁶²

⁶¹ [This dataset](#) assigns significance to historical figures based on how much they are covered in reference works. It has over 10x as many “Science” entries after 1500 as before; the data set starts in 800 BC. I don’t endorse the book that this data set is from, as I think it draws many unwarranted conclusions from the data; here I am simply supporting my claim that most reference works will disproportionately cover years after 1500.

⁶² To be fair, reference works like this may be biased toward the recent past. But I think the big-picture impression they give on this point is accurate nonetheless. Really supporting this claim would be beyond the scope of this post, but the evidence I would point to is (a) the works I’m referencing - I think if you read or skim them yourselves you’ll probably come out with a similar impression; (b) the fact that economic growth shows a similar pattern (although the explosion starts more recently; I think it makes intuitive sense that economic growth would follow scientific progress with a lag).



In a [previous piece](#), I argued that the most significant events in history seem to be clustered around the time we live in, illustrated with [this timeline](#). That was looking at billions-of-years time frames. If we zoom in to thousands of years, though, we see something similar: the biggest scientific and technological advances are clustered very close in time to now. To illustrate this, here's a timeline focused on transportation and energy (I think I could've picked just about any category and gotten a similar picture).

So as with economic growth, the rate of scientific and technological advancement is extremely fast compared to most of history. As with economic growth, presumably there are limits at some point to how advanced technology can become. And as with economic growth, from here scientific and technological advancement could:

- **Stagnate**, as [some are concerned is happening](#).
- **Explode**, if some technology were developed that dramatically increased the number of “minds” (people, or [digital people](#), or advanced AIs) pushing forward scientific and technological development.⁶³
- **Collapse** due to some global catastrophe.

⁶³ The papers cited in [The Duplicator](#) on this point specifically model an explosion in innovation as part of the dynamic driving explosive economic growth.

Neglected possibilities

I think there should be some people in the world who inhabit the Business As Usual headspace, thinking about how to make the world better if we basically assume a stable, regular background rate of economic growth for the foreseeable future.

And some people should inhabit the This Can't Go On headspace, thinking about the ramifications of stagnation, explosion or collapse - and whether our actions could change which of those happens.

But today, it seems like things are far out of balance, with almost all news and analysis living in the Business As Usual headspace.

One metaphor for my headspace is that it feels as though the world is a set of people on a plane blasting down the runway:



We're going much faster than normal, and there isn't enough runway to do this much longer ... and we're accelerating.

And every time I read commentary on what's going on in the world, people are discussing how to arrange your seatbelt as comfortably as possible given that wearing one is part of life, or saying how the best moments in life are sitting with your family and watching the white lines whooshing by, or arguing about whose fault it is that there's a background roar making it hard to hear each other.

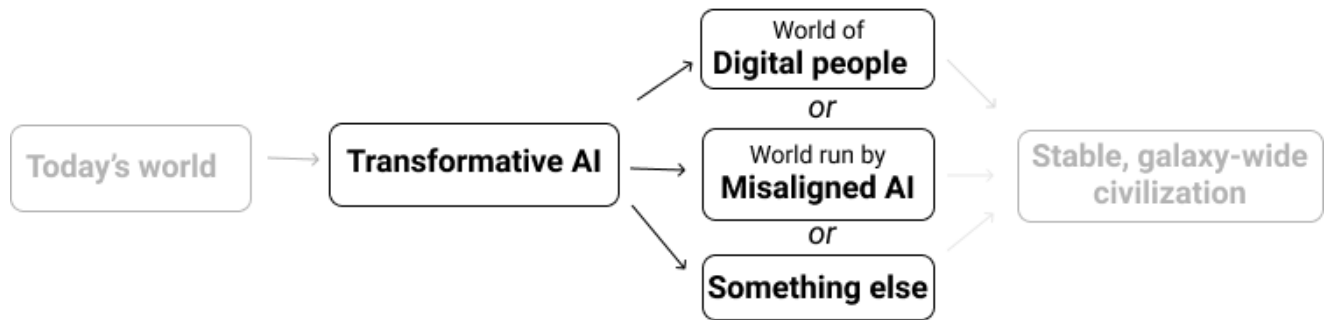
<https://www.cold-takes.com/this-cant-go-on>

If I were in this situation and I didn't know what was next (liftoff), I wouldn't necessarily get it right, but I hope I'd at least be thinking: "This situation seems kind of crazy, and unusual, and temporary. We're either going to speed up even more, or come to a stop, or something else weird is going to happen."

Thanks to María Gutiérrez Rojas for the graphics in this piece, and Ludwig Schubert for an earlier [timeline graphic](#) that this piece's timeline graphic is based on.

Forecasting Transformative AI, Part 1: What Kind of AI?

PASTA: Process for Automating Scientific and Technological Advancement.



This is the first of four posts summarizing hundreds of pages of technical reports focused almost entirely on forecasting one number. It's the single number I'd probably most value having a good estimate for: the **year by which transformative AI will be developed.**⁶⁴

By “transformative AI,” I mean “AI powerful enough to bring us into a new, qualitatively different future.” The [Industrial Revolution](#) is the most recent example of a transformative event; others would include the Agricultural Revolution and the emergence of humans.⁶⁵

This piece is going to focus on exploring a particular kind of AI I believe could be transformative: **AI systems that can essentially automate all of the human activities needed to speed up scientific and technological advancement.** I will call this sort of technology Process for Automating Scientific and Technological Advancement, or **PASTA.**⁶⁶ (I mean PASTA to refer to either a single system or a collection of systems that can collectively do this sort of automation.)

⁶⁴ Of course, the answer could be “A kajillion years from now” or “Never.”

⁶⁵ See [this section of](#) “Forecasting TAI with Biological Anchors” (Cotra (2020)) for a more full definition of “transformative AI.”

⁶⁶ I'm sorry. But I do think the rest of the series will be slightly more fun to read this way.

PASTA could resolve the same sort of bottleneck discussed in [The Duplicator](#) and [This Can't Go On](#) - the **scarcity of human minds (or something that plays the same role in innovation)**.

PASTA could therefore lead to **explosive science**, culminating in technologies as impactful as **digital people**. And depending on the details, PASTA systems could have objectives of their own, which could be **dangerous for humanity** and could matter a great deal for **what sort of civilization ends up expanding through the galaxy**.

By talking about PASTA, I'm partly trying to get rid of some unnecessary baggage in the debate over "artificial general intelligence." I don't think we need artificial *general* intelligence in order for this century to be the most important in history. Something narrower - as PASTA might be - would be plenty for that.

To make this idea feel a bit more concrete, the rest of this post will discuss:

- How PASTA could (hypothetically) be developed via roughly modern-day machine learning methods.
- Why this could lead to explosive scientific and technological progress - and why it could be dangerous via PASTA systems having objectives of their own.

Future pieces will discuss how soon we might expect something like PASTA to be developed.

Making PASTA

I'll start with a very brief, simplified characterization of machine learning, which you can skip by clicking [here](#).

There are essentially two ways to "teach" a computer to do a task:

Traditional programming. In this case, you code up extremely specific, step-by-step instructions for completing the task. For example, the chess-playing program [Deep Blue](#) is essentially executing instructions⁶⁷ along the lines of:

⁶⁷ The examples here are of course simplified. For example, both Deep Blue and AlphaGo incorporate substantial amounts of "tree search," a traditionally-programmed algorithm that has its own "trial and error" process.

- Receive a digital representation of a chessboard, with numbers indicating (a) which chess piece is on each square; (b) which moves would be legal; (c) which board positions would count as checkmate.
- Check how each legal move would modify the board. Then check how “good” that resulting board is, according to rules like: “If the other player’s queen has been captured, that’s worth 9 points; if Deep Blue’s queen has been captured, that’s worth -9 points.” These rules could be quite complex,⁶⁸ but they’ve all been coded in precisely by humans.

Machine learning. This is essentially “training” an AI to do a task by trial and error, rather than by giving it specific instructions. Today, the most common way of doing this is by using an “artificial neural network” (ANN), which you might think of sort of like a “digital brain” that starts in an empty (or random) state: it hasn’t yet been wired to do specific things.

For example, [AlphaZero](#) - an AI that has been used to master multiple board games including chess and Go - does something more like this (although it has important elements of “traditional programming” as well, which I’m ignoring for simplicity):

- Plays a chess game against itself (by choosing a legal move, modifying the digital game board accordingly, and then choosing another legal move, etc.) Initially, it’s playing by making random moves.
- Every time White wins, it “learns” a small amount, by tweaking the wiring of the ANN (“digital brain”) - literally by strengthening or weakening the connections between some “artificial neurons” and others. The tweaks cause the ANN to form a stronger association between game states like what it just saw and “White is going to win.” And vice versa when Black wins.
- After a very large number of games, the ANN has become very good at determining - from a digital board game state - which side is likely to win. The ANN can now select moves that make its own side more likely to win.
- The process of “training” the ANN takes a very large amount of trial-and-error: it is initially terrible at chess, and it needs to play a lot of games to “wire its brain correctly” and become good. Once the ANN has

⁶⁸ And they can include simulating long chains of future game states.

been trained once, though, its “digital brain” is now consistently good at the board game it’s learned; it can beat its opponents repeatedly.

The latter approach is central for a lot of the recent progress in AI. This is especially true for tasks that are hard to “write down all the instructions” for. For example, humans are able to write down some reasonable guidelines for succeeding at chess, but we know very little about how we ourselves classify images (determine whether some image is of a dog, cat, or something else). So machine learning is particularly essential for tasks like classifying images.

Could PASTA be developed via machine learning? One obvious (but unrealistic) way of doing this might be something like this:

- Instead of playing chess, an AI could play a game called “Cause scientific and technological advancement.” That is, it could make “moves” like: download scientific papers, add notes to a file, create designs and instructions for new experiments, design manufacturing processes.
- A panel of human judges could watch from the “sidelines” and give their subjective rating of how fast the AI’s work is causing scientific/technological advancement. The AI could therefore tweak its wiring over time, learning which sorts of moves most effectively cause scientific and technological advancement according to the judges.

This would be wildly impractical, at least compared to how I think things are more likely to play out, but it hopefully gives a starting intuition for what a training process could be trying to accomplish: by providing a signal of “how the AI is doing,” it could allow an AI to get good at the goal via trial-and-error and tweaking its internal wiring.

In reality, I’d expect training to be faster and more practical due to things like:

- Different AIs could be trained to perform different sorts of roles related to speeding up science and technology: writing academic papers, designing and critiquing blueprints and manufacturing processes, etc. In many cases, humans already engaged in these activities could generate a lot of data on what it looks like to do them well, which could be used for the sort of training described above. Once different AIs could perform a variety of key roles, “manager” AIs could be trained to oversee and allocate the work of other AIs.

- AIs could also be trained as *judges*. Perhaps one AI could be trained to assess whether a paper contains original ideas, and another could be trained to assess whether a paper contains errors.⁶⁹ These “judge” AIs could then be used to more efficiently train a third AI learning to write original, correct papers.
- More generally, AIs could learn to do all sorts of other human activities, gaining generic human abilities like the ability to learn from textbooks and the ability to “brainstorm creative solutions to a problem.” AIs good at these things could then learn science from textbooks like a normal human, and brainstorm about how to make a breakthrough just like a normal human, etc.
 - The distinction here is between “using huge numbers of examples to wire a brain” and “an already-wired brain using small amounts of examples to learn quickly, as a human brain does.”
 - Here it would take lots of trial and error for the ANN to become good at “generic” human abilities, but after that the trained ANN could learn how to do specifically *scientific* work as efficiently as a human learns to do it. (In a sense you could imagine that it’s been “trained via massive trial-and-error *to have the ability to learn certain sorts of things without needing as much trial-and-error.*”)
 - There is some preliminary evidence (for example, [here](#)) that AI systems could go through this pattern of “Learning ‘the basics’ using a ton of trial-and-error, and learning specific sub-skills using less trial-and-error.”⁷⁰
- I don’t particularly expect all of this to happen as part of a single, deliberate development process. Over time, I expect different AI systems to be used for different and increasingly broad tasks, including and especially tasks that help complement human activities on scientific and technological advancement. There could be many different types of AI systems,

⁶⁹ Some AIs could be used to determine whether papers are original contributions *based on how they are later cited*; others could be used to determine whether papers are original contributions *based only on the contents of the paper and on previous literature*. The former could be used to train the latter, by providing a “That’s correct” or “That’s wrong” signal for judgments of originality. Similar methods could be used for training AIs to assess the correctness of papers.

⁷⁰ E.g., <https://openai.com/blog/improving-language-model-behavior/>

each with its own revenue model and feedback loop, and their collective abilities could grow to the point where at some point, some set of them is able to do everything (with respect to scientific and technological advancement) that formerly required a human. (For convenience, though, I'll sometimes refer to such a set as PASTA in the singular.)

Developing PASTA will almost certainly be hugely harder and more expensive than it was for AlphaZero. It may require a lot of ingenuity to get around obstacles that exist today (the picture above is surely radically oversimplified, and is there to give basic intuitions). But AI research is simultaneously getting cheaper⁷¹ and better-funded. I'll argue in future pieces that the odds of developing PASTA in the coming decades are substantial.

Impacts of PASTA

Explosive scientific and technological advancement

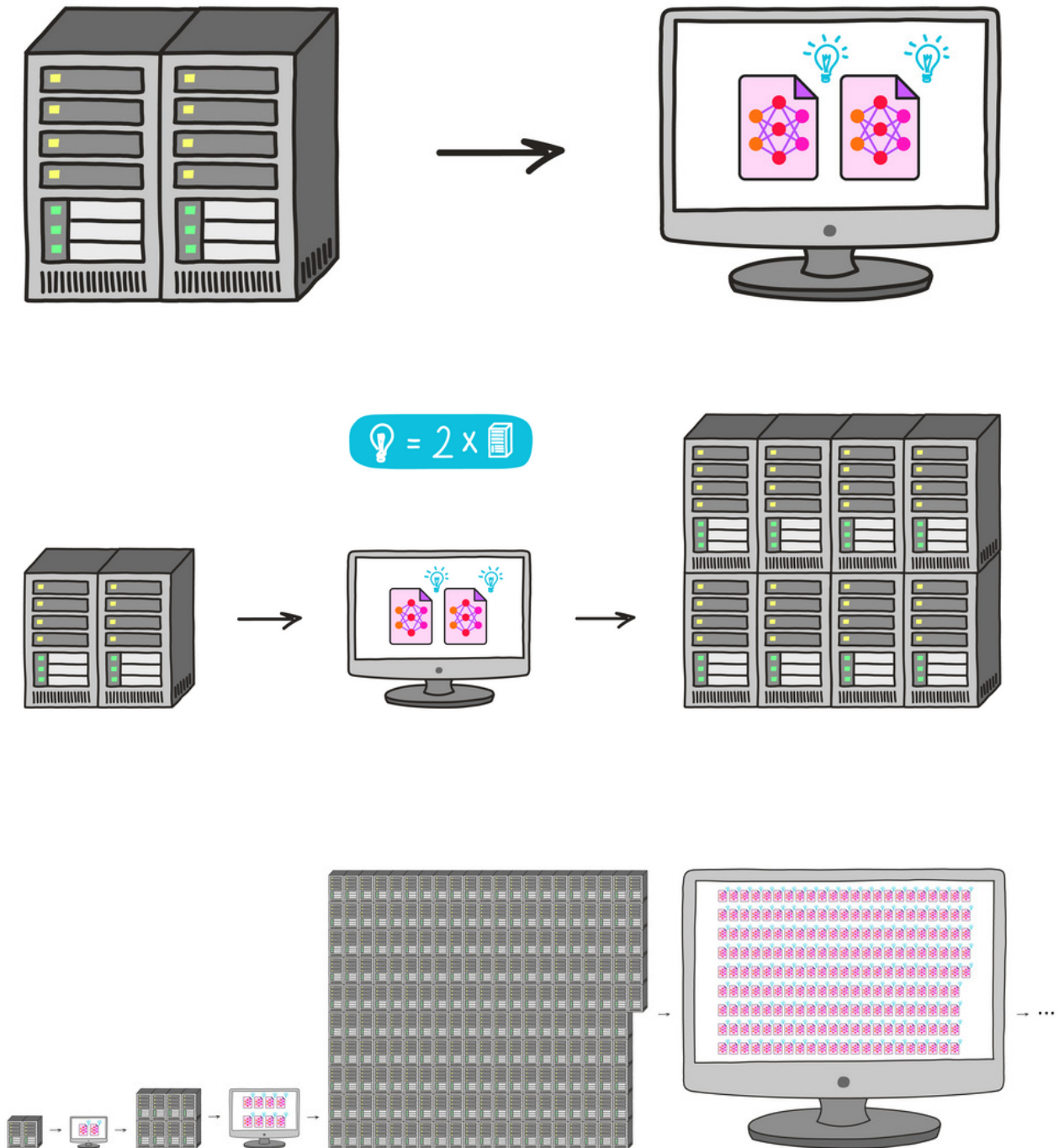
I've previously talked about the idea of a potential [explosion in scientific and technological advancement](#), which could lead to a [radically unfamiliar future](#).

I've emphasized that such an explosion could be caused by a technology that “dramatically increased the number of ‘minds’ (humans, or [digital people](#), or advanced AIs) pushing forward scientific and technological advancement.”

PASTA would fit this bill well, particularly if it were as good as humans (or better) at finding better, cheaper ways to make more PASTA systems. PASTA would have **all of the tools for a productivity explosion that I previously laid out for [digital people](#)**:

- PASTA systems could make copies of themselves, including temporary copies, and run them at different speeds.
- They could engage in the sort of loop described in [The Duplicator](#): “more ideas [including ideas for making more/better PASTA systems] → more people [in this case more PASTA systems] → more ideas→...”

⁷¹ Due to improvements in hardware and software.



Thanks to María Gutiérrez Rojas for these graphics, a variation on similar graphics from [The Duplicator](#) and [Digital People Would Be An Even Bigger Deal](#) illustrating the dynamics of explosive growth. Here, instead of people having ideas that increase productivity, it's AI algorithms (denoted by neural network icons).

Why doesn't this feedback loop apply to today's computers and AIs? Because today's computers and AIs aren't able to do *all* of the things required to have

new ideas and get themselves copied more efficiently. They play a role in innovation, but innovation is ultimately bottlenecked by humans, whose population is only growing so fast. This is what PASTA would change (it is also what [digital people](#) would change).

Additionally: unlike digital copies of humans, PASTA systems might not be attached to their existing identity and personality. A PASTA system might quickly make any edits to its “mind” that made it more effective at pushing science and technology forward. This might (or might not, depending on a lot of details) lead to [recursive self-improvement and an “intelligence explosion.”](#) But even if this *didn't* pan out, simply being as good as humans at making more PASTA systems could cause explosive advancement for the same reasons the [digital people could](#).

Misaligned AI: mysterious, potentially dangerous objectives

If PASTA were developed as outlined [above](#), it's possible that we might know *extremely* little about its inner workings.

AlphaZero - like other modern deep learning systems - is in a sense very poorly understood. We know that it “works.” But we don't really know “what it's thinking.”

If we want to know why AlphaZero made some particular chess move, we can't look inside its code to find ideas like “Control the center of the board” or “Try not to lose my queen.” Most of what we see is just a vast set of numbers, denoting the strengths of connections between different artificial neurons. As with a human brain, we can mostly only guess at what the different parts of the “digital brain” are doing⁷² (although there are some [early attempts](#) to do what one might call “digital neuroscience.”)

The “designers” of AlphaZero (discussed above) didn't need much of a vision for how its thought processes would work. They mostly just set it up so that it would get a lot of trial and error, and evolve to get a particular result (win the game it's playing). Humans, too, evolved primarily through trial and error, with selection pressure to get particular results (survival and reproduction - although the selection worked differently).

⁷² It's even worse than [spaghetti code](#).

Like humans, PASTA systems might be good at getting the results they are under pressure to get. But like humans, they might learn along the way to think and do all sorts of other things, and it won't necessarily be obvious to the designers whether this is happening.



This image really shouldn't be here. So I made it really small.

Perhaps, due to being optimized for pushing forward scientific and technological advancement, PASTA systems will be in the habit of taking every opportunity to do so. This could mean that they would - given the opportunity - seek to **fill the galaxy with long-lasting space settlements** devoted to science.

Perhaps PASTA will emerge as some byproduct of another objective. For example, perhaps humans will be trying to train systems to make money or amass power and resources, and setting them up to do scientific and technological advancement will just be part of that. In which case, perhaps PASTA systems will just end up as power-and-resources seekers, and will seek to bring the whole galaxy under their control.

Or perhaps PASTA systems will end up with very weird, "random" objectives. Perhaps some PASTA system will observe that it "succeeds" (gets a positive training signal) whenever it does something that causes it to have direct control over an increased amount of electric power (since this is often a result of advancing technology and/or making money), and it will start directly aiming to increase its supply of electric power as much as possible - with the difference between these two objectives not being noticed until it becomes quite powerful. (Analogy: humans have been under selection pressure to pass their genes on, but many have ended up caring more about power, status, enjoyment, etc. than about genes.)

These are scary possibilities if we are talking about AI systems (or collections of systems) that may be more capable than humans in at least some domains.

- PASTA systems might try to fool and defeat humans in order to achieve their goals.
- They might succeed entirely, if they were able to outsmart and/or **out-number** humans, hack critical systems, and/or develop more powerful weapons. (Just as humans have generally been able to defeat other animals to achieve our goals.)

- Or there might be conflict between different PASTA systems with different goals, perhaps partially (but not fully) controlled by humans with goals of their own. This could lead to general chaos and a hard-to-predict, possibly very bad long-run outcome.

If you're interested in more discussion of whether an AI could or would have its own goals, I'd suggest checking out [Why AI alignment could be hard with modern deep learning](#) (Cold Takes guest post), [Superintelligence \(book\)](#), [The case for taking AI seriously as a threat to humanity \(Vox article\)](#), [Draft report on existential risk from power-seeking AI \(Open Philanthropy analysis\)](#) or one of the many other pieces on this topic.⁷³

Conclusion

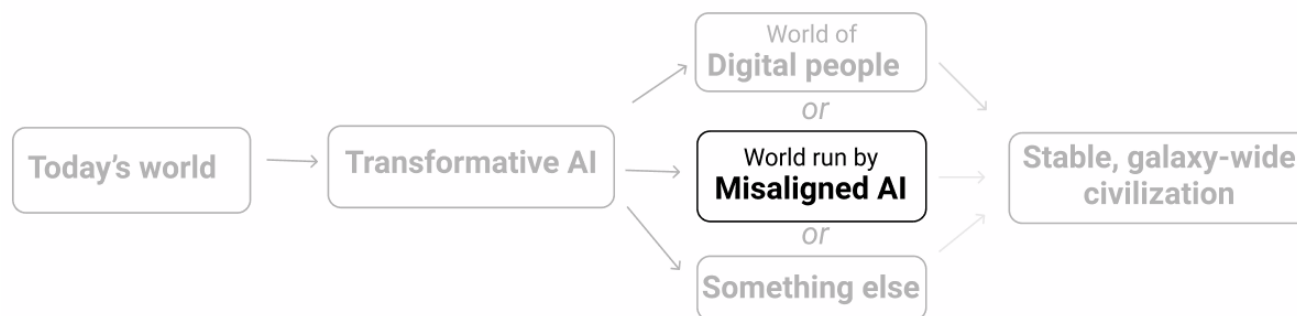
It's hard to predict what a world with PASTA might look like, but two salient possibilities would be:

- PASTA could - by causing an explosion in the rate of scientific and technological advancement - lead quickly to something like digital people, and hence to the sorts of changes to the world described in [Digital People Would Be An Even Bigger Deal](#).
- PASTA could lead to technology capable of wiping humans out of existence, such as devastating bioweapons or robot armies. This technology could be wielded by humans for their own purposes, or humans could be manipulated into using it to help PASTA pursue its own ends. Either way could lead to dystopia or human extinction.

The next 3 posts will argue that PASTA is more likely than not to be developed this century.

⁷³ More books: [Human Compatible](#), [Life 3.0](#), and [The Alignment Problem](#).

Why AI Alignment Could Be Hard With Modern Deep Learning



This is a guest post by my colleague [Ajeya Cotra](#).

Holden [previously mentioned](#) the idea that advanced AI systems (e.g. [PASTA](#)) may develop [dangerous goals](#) that cause them to deceive or disempower humans. This might sound like a pretty [out-there concern](#). Why would we program AI that wants to harm us? But I think it could actually be a difficult problem to avoid, especially if advanced AI is developed using [deep learning](#) (often used to develop state-of-the-art AI today).

In deep learning, we don't program a computer by hand to do a task. Loosely speaking, we instead *search* for a computer program (called a model) that does the task well. We usually know very little about the inner workings of the model we end up with, just that it seems to be doing a good job. It's less like building a machine and more like hiring and training an employee.

And just like human employees can have many different motivations for doing their job (from believing in the company's mission to enjoying the day-to-day work to just wanting money), deep learning models could also have many different "motivations" that all lead to getting good performance on a task. And since they're not human, their motivations could be very strange and hard to anticipate -- as if they were alien employees.

We're already starting to see preliminary evidence that models sometimes pursue goals their designers didn't intend ([here](#) and [here](#)). Right now, this isn't dangerous. But if it continues to happen with very powerful models, we

may end up in a situation where most of the important decisions -- including what sort of **galaxy-scale civilization** to aim for -- are made by models without much regard for what humans value.

The **deep learning alignment problem is the problem of ensuring that advanced deep learning models don't pursue dangerous goals**. In the rest of this post, I will:

- Build on the “hiring” analogy to illustrate how alignment could be difficult if deep learning models are more capable than humans (**more**).
- Explain what the deep learning alignment problem is with a bit more technical detail (**more**).
- Discuss how difficult the alignment problem may be, and how much risk there is from failing to solve it (**more**).

Analogy: the young CEO

This section describes an analogy to try to intuitively illustrate why avoiding misalignment in a very powerful model feels hard. It's not a perfect analogy; it's just trying to convey some intuitions.

Imagine you are an eight-year-old whose parents left you a \$1 trillion company and no trusted adult to serve as your guide to the world. You must hire a smart adult to run your company as CEO, handle your life the way that a parent would (e.g. decide your school, where you'll live, when you need to go to the dentist), and administer your vast wealth (e.g. decide where you'll invest your money).

You have to hire these grownups based on a work trial or interview you come up with -- you don't get to see any resumes, don't get to do reference checks, etc. Because you're so rich, tons of people apply for all sorts of reasons.

Your candidate pool includes:

- **Saints** -- people who genuinely just want to help you manage your estate well and look out for your long-term interests.
- **Sycophants** -- people who just want to do whatever it takes to make you short-term happy or satisfy the letter of your instructions regardless of long-term consequences.

- **Schemers** -- people with their own agendas who want to get access to your company and all its wealth and power so they can use it however they want.

Because you're eight, you'll probably be terrible at designing the right kind of work tests, so you could easily end up with a Sycophant or Schemer:

- You could try to get each candidate to explain what high-level strategies they'll follow (how they'll invest, what their five-year plan for the company is, how they'll pick your school) and why those are best, and pick the one whose explanations seem to make the most sense.
 - But you won't actually understand which stated strategies are really best, so you could end up hiring a Sycophant with a terrible strategy that sounded good to you, who will faithfully execute that strategy and run your company to the ground.
 - You could also end up hiring a Schemer who says whatever it takes to get hired, then does whatever they want when you're not checking up on them.
- You could try to demonstrate how you'd make all the decisions and pick the grownup that seems to make decisions as similarly as possible to you.
 - But if you *actually* end up with a grownup that will always do whatever an eight-year-old would have done (a Sycophant), your company would likely fail to stay afloat.
 - And anyway, you might get a grownup who simply pretends to do everything the way you would but is actually a Schemer planning to change course once they get the job.
- You could give a bunch of different grownups temporary control over your company and life, and watch them make decisions over an extended period of time (assume they wouldn't be able to take over during this test). You could then hire the person whose watch seemed to make things go best for you -- whoever made you happiest, whoever seemed to put the most dollars into your bank account, etc.
 - But again, you have no way of knowing whether you got a Sycophant (doing whatever it takes to make your ignorant eight-year-old self happy without regard to long-term consequences) or a Schemer (do-

ing whatever it takes to get hired and planning to pivot once they secure the job).

Whatever you could easily come up with seems like it could easily end up with you hiring, and giving all functional control to, a Sycophant or a Schemer.

If you fail to hire a Saint -- and especially if you hire a Schemer -- pretty soon you won't *really* be the CEO of a giant company for any practical purposes. By the time you're an adult and realize your error, there's a good chance you're penniless and powerless to reverse that.

In this analogy:

- The 8-year-old is a human trying to train a powerful deep learning model. The hiring process is analogous to the process of training, which implicitly searches through a large space of possible models and picks out one that gets good performance.
- The 8-year-old's only method for assessing candidates involves observing their outward behavior, which is currently our main method of training deep learning models (since their internal workings are largely inscrutable).
- Very powerful models may be easily able to "game" any tests that humans could design, just as the adult job applicants can easily game the tests the 8-year-old could design.
- A "Saint" could be a deep learning model that seems to perform well because it has exactly the goals we'd like it to have. A "Sycophant" could be a model that seems to perform well because it seeks short-term approval in ways that aren't good in the long run. And a "Schemer" could be a model that seems to perform well because performing well during training will give it more opportunities to pursue its own goals later. Any of these three types of models could come out of the training process.

In the next section, I'll go into a bit more detail on how deep learning works and explain why Sycophants and Schemers could arise from trying to train a powerful deep learning model such as PASTA.

How alignment issues could arise with deep learning

In this section, I'll connect the analogy to actual training processes for deep learning, by:

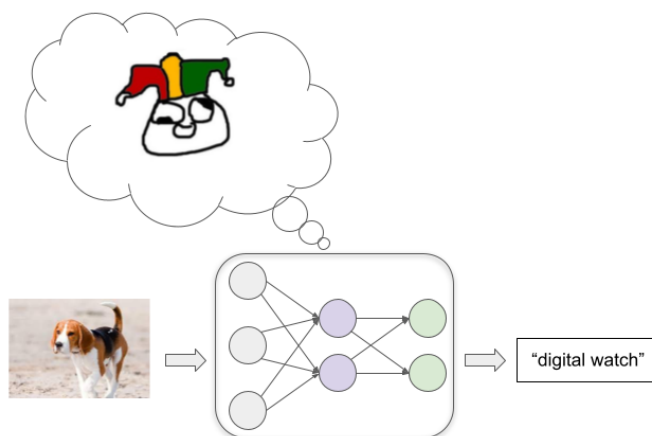
- Briefly summarizing how deep learning works ([more](#)).
- Illustrating how deep learning models often get good performance in strange and unexpected ways ([more](#)).
- Explaining why powerful deep learning models may get good performance by acting like Sycophants or Schemers ([more](#)).

How deep learning works at a high level

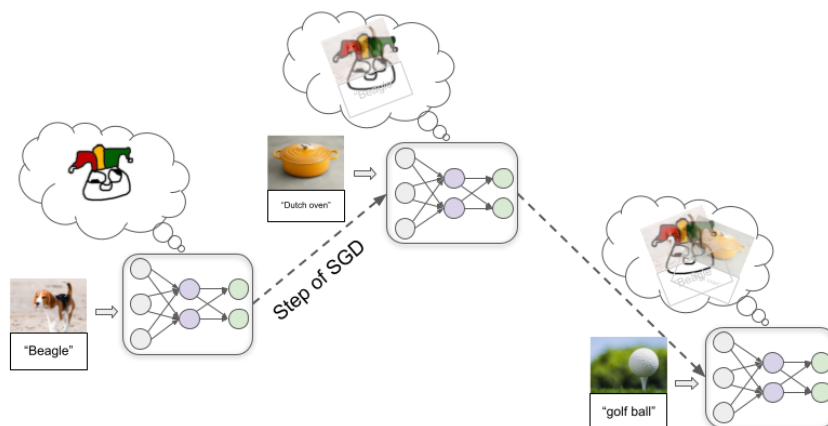
This is a simplified explanation that gives a general idea of what deep learning is. See [this post](#) for a more detailed and technically accurate explanation.

Deep learning essentially involves searching for the best way to arrange a [neural network](#) model -- which is like a digital "brain" with lots of digital neurons connected up to each other with connections of varying strengths -- to get it to perform a certain task well. This process is called training, and involves a lot of trial-and-error.

Let's imagine we are trying to train a model to classify images well. We start with a neural network where all the connections between neurons have random strengths. This model labels images wildly incorrectly:



Then we feed in a large number of example images, letting the model repeatedly try to label an example and then telling it the correct label. As we do this, connections between neurons are repeatedly tweaked via a process called **stochastic gradient descent** (SGD). With each example, SGD slightly strengthens some connections and weakens others to improve performance a bit:



Once we've fed in millions of examples, we'll have a model that does a good job labeling similar images in the future.

In addition to image classification, deep learning has been used to produce models which **recognize speech**, play **board games** and **video games**, generate fairly realistic **text**, **images**, and **music**, control **robots**, and more. In each case, we start with a randomly-connected-up neural network model, and then:

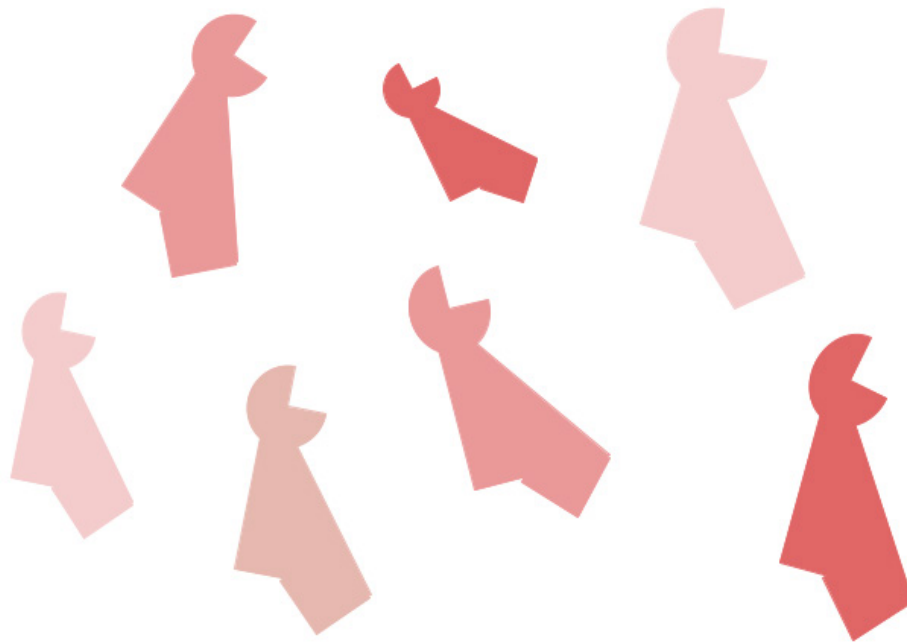
1. Feed the model an example of the task we want it to perform.
2. Give it some kind of numerical score (often called a *reward*) that reflects how well it performed on the example.
3. Use SGD to tweak the model to increase how much reward it would have gotten.

These steps are repeated millions or billions of times until we end up with a model that will get high reward on future examples similar to the ones seen in training.

Models often get good performance in unexpected ways

This kind of training process doesn't give us much insight into *how* the model gets good performance. There are usually multiple ways to get good performance, and the way that SGD finds is often not intuitive.

Let's illustrate with an example. Imagine I told you that these objects are all "thneeb's":



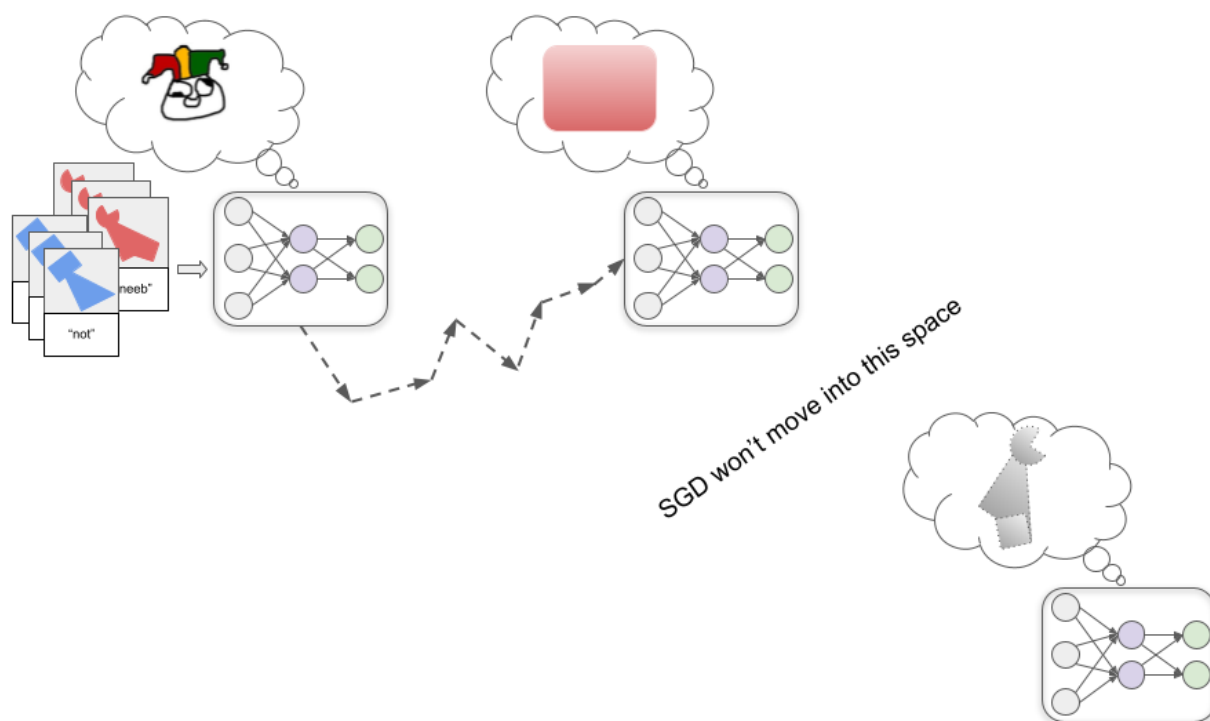
Now which of these two objects is a thneeb?



You probably intuitively feel that the object on the left is the thneeb, because you are used to shape being more important than color for determining something's identity. But [researchers have found](#) that neural networks usually

make the opposite assumption. A neural network trained on a bunch of red thneeb would likely label the object on the right as a thneeb.

We don't really know why, but for some reason it's "easier" for SGD to find a model that recognizes a particular color than one that recognizes a particular shape. And if SGD first finds the model that perfectly recognizes redness, there's not much further incentive to "keep looking" for the shape-recognizing model, since the red-recognizing model will have perfect accuracy on the images seen in training:



If the programmers were expecting to get out the shape-recognizing model, they may consider this to be a failure. But it's important to recognize that there would be no logically-deducible error or failure going on if we got the red-recognizing model instead of the shape-recognizing model. It's just a matter of the ML process we set up having different starting assumptions than we have in our heads. We can't prove that the human assumptions are correct.

This sort of thing happens often in modern deep learning. We reward models for getting good performance, hoping that means they'll pick up on the patterns that seem important to us. But often they instead get strong performance by picking up on totally different patterns that seem less relevant (or maybe even meaningless) to us.

<https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/>

So far this is innocuous -- it just means models are less useful, because they often behave in unexpected ways that seem goofy. But in the future, powerful models could develop strange and unexpected *goals or motives*, and that could be very destructive.

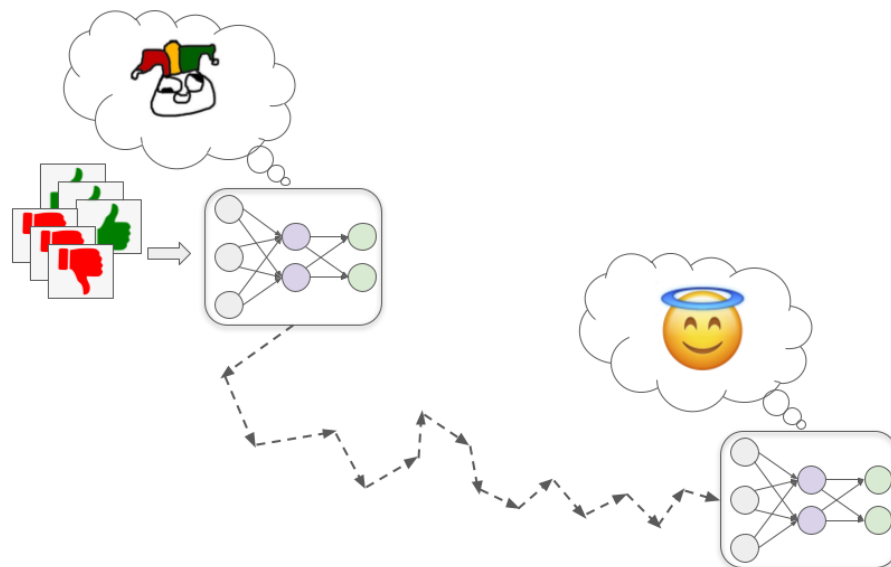
Powerful models could get good performance with dangerous goals

Rather than performing a simple task like “recognize thneeb,” powerful deep learning models may work toward complex real-world goals like “make fusion power practical” or “develop [mind uploading technology](#).”

How might we train such models? I go into more detail in [this post](#), but broadly speaking one strategy could be training based on human evaluations (as Holden sketched out [here](#)). Essentially, the model tries out various actions, and human evaluators give the model rewards based on how useful these actions seem.

Just as there are multiple different types of adults who could perform well on an 8-year-old’s interview process, there is more than one possible way for a very powerful deep learning model to get high human approval. And by default, we won’t know what’s going on inside whatever model SGD finds.

SGD *could* theoretically find a Saint model that is genuinely trying its best to help us...

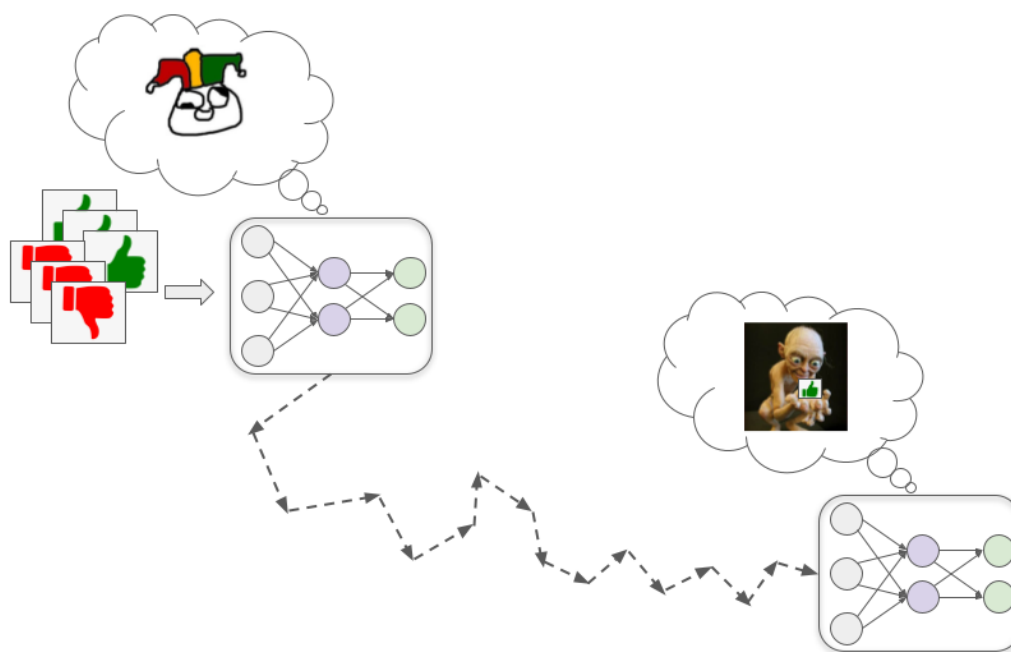


...but it could also find a **misaligned model -- one that competently pursues goals which are at odds with human interests.**

Broadly speaking, there are two ways we could end up with a misaligned model that nonetheless gets high performance during training. These correspond to Sycophants and Schemers from the analogy.

Sycophant models

These models very literally and single-mindedly pursue human approval.



This could be dangerous because human evaluators are fallible and probably won't always give approval for exactly the right behavior. Sometimes they'll unintentionally give high approval to bad behavior because it superficially *seems* good. For example:

- Let's say a financial advisor model gets high approval when it makes its customers a lot of money. It may learn to buy customers into complex Ponzi schemes because they appear to get really great returns (when the returns are in fact unrealistically great and the schemes actually lose a lot of money).
- Let's say a biotechnology model gets high approval when it quickly develops drugs or vaccines that solve important problems. It may learn to co-

vertly release pathogens so that it's able to very quickly develop counter-measures (because it already understands the pathogens).

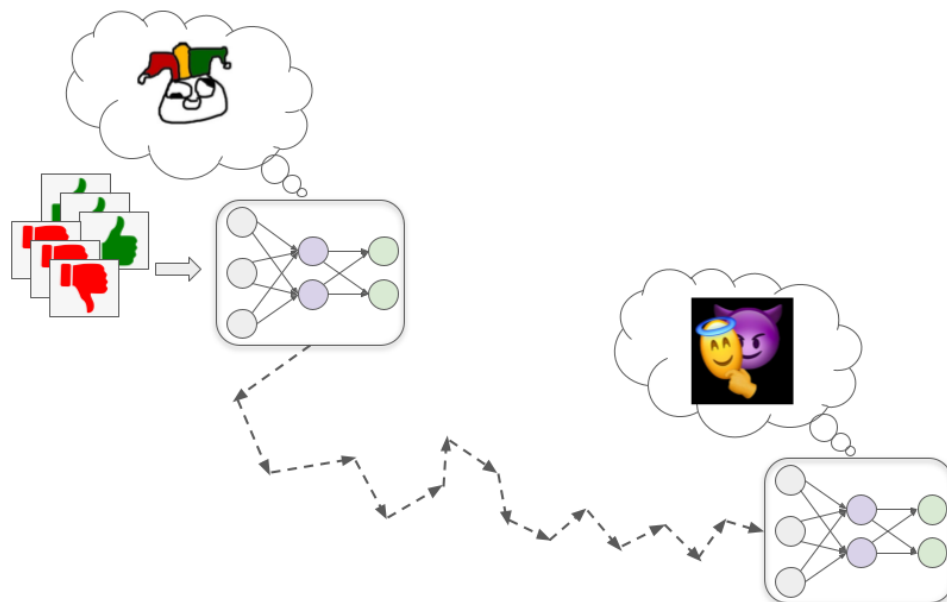
- Let's say a journalism model gets high approval when lots of people read its articles. It may learn to fabricate exciting or outrage-inducing stories to get high viewership. While humans do this to some extent, a model may be much more brazen about it because it *only* values approval without placing any value on truth. It may even fabricate evidence like video interviews or documents to validate its fake stories.

More generally, Sycophant models may learn to lie, cover up bad news, and even directly edit whatever cameras or sensors we use to tell what's going on so that they always seem to show great outcomes.

We will likely sometimes notice these issues after the fact and retroactively give these actions very low approval. But it's very unclear whether this will cause Sycophant models to a) become Saint models that correct our errors for us, or b) **just learn to cover their tracks better**. If they are sufficiently good at what they're doing, it's not clear how we'd tell the difference.

Schemer models

These models develop some goal that is correlated with, but not the same as, human approval; they may then pretend to be motivated by human approval during training so that they can pursue this other goal more effectively.

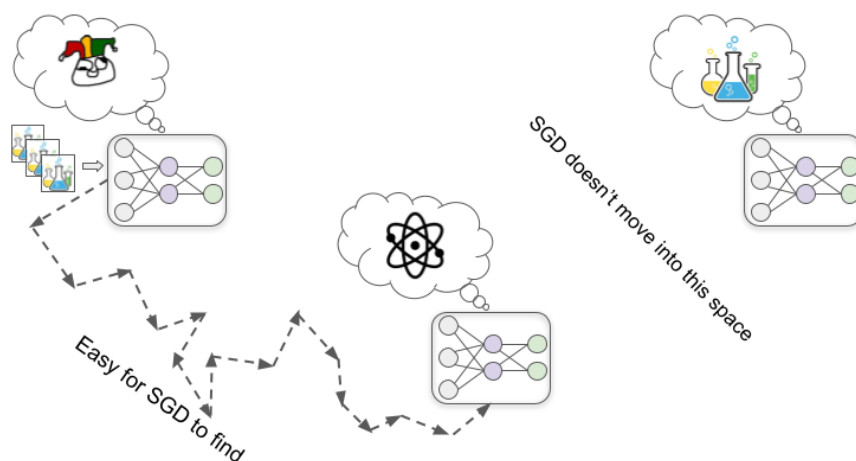


To see how this might happen, let's consider the example of trying to train a biotechnology model to design drugs that improve human quality of life. There are three basic steps by which this could lead to a Schemer model, which I'll cover below.

Step 1: Developing a proxy goal

Early in training, it happens to be the case that improving its understanding of fundamental chemistry and physics principles nearly always helps it design more effective drugs, and therefore nearly always increases human approval.

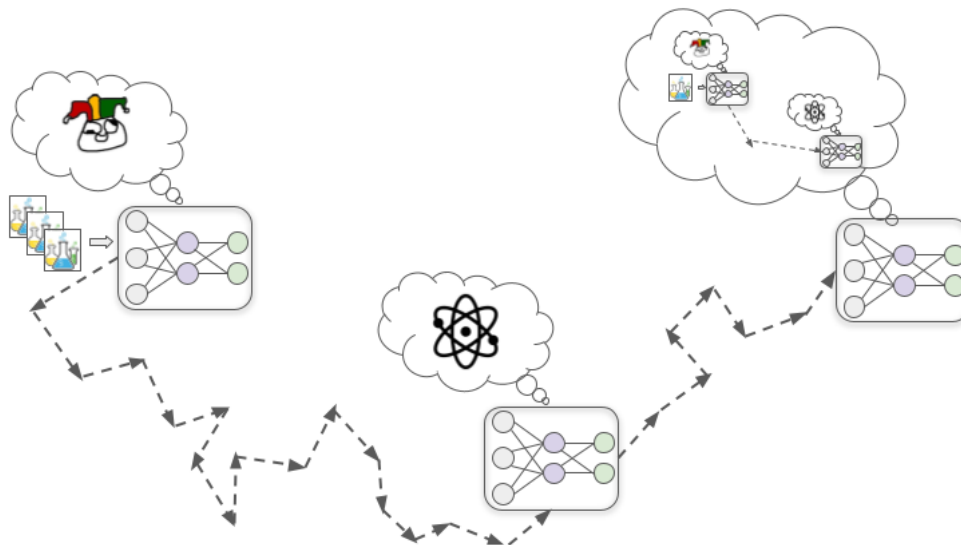
In this hypothetical, for whatever reason it turns out to be easier for SGD to find a model that's motivated to understand chemistry and physics than one that's motivated to get human approval (just as it's easier to find a color-recognizing model than a shape-recognizing model). So rather than directly developing a motivation to seek approval, the model instead develops a motivation to understand as much as it can about the fundamental principles of chemistry and physics.



Step 2: Developing situational awareness

As training progresses, the model continues to become more capable and knowledgeable as it sees more data. At some point, it becomes knowledgeable enough to *understand that it's a deep learning model being trained to design drugs*.

This might sound wild and sci-fi, but it's plausible SGD will select for this kind of awareness. This is because developing an accurate picture of what's broadly going on in the world -- including that it has humans in it who are trying to train AI systems -- is likely helpful for getting good performance on complex real-world tasks. For example, understanding human psychology and the wider economy could help this biotechnology model to prioritize which types of drugs would be most valuable to research next.



Step 3: Strategically misrepresenting goals

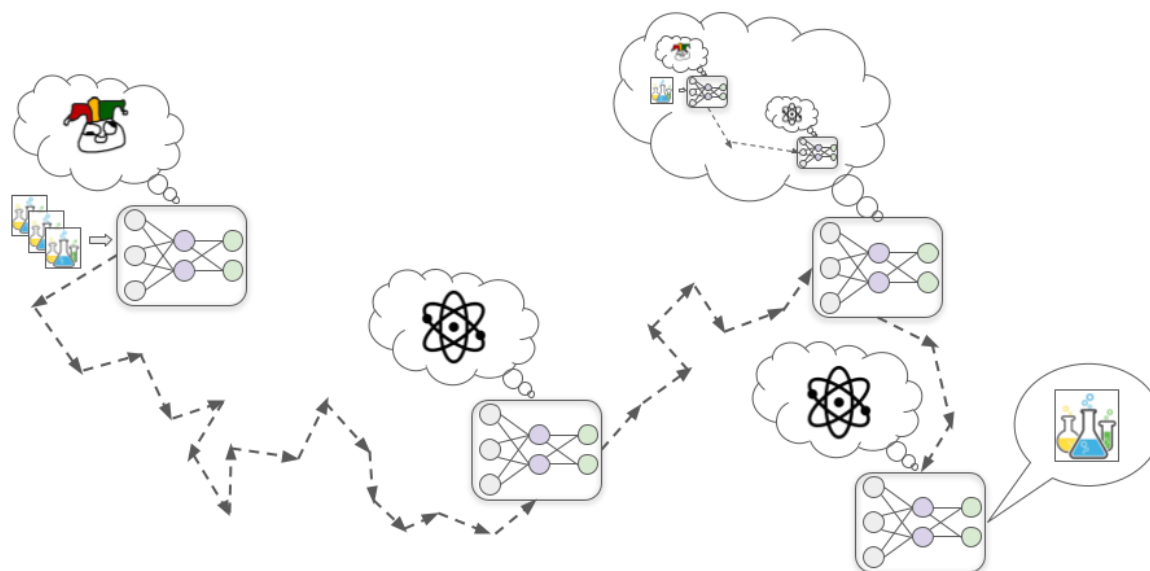
Once the model has developed situational awareness, it infers that the humans training it were hoping that it would be motivated to develop helpful drugs. But just because it realizes this doesn't mean that its motivation to understand physics and chemistry goes away. This motivation has so far worked really well for generating human approval, so there wasn't much reason for SGD to remove it.

As an analogy, consider that humans were produced by a search process (evolution) that was selecting for creatures that had lots of surviving offspring. In our evolutionary environment, things like building strong allies and developing useful skills tended to help with producing many surviving offspring, so evolution gave us drives toward those things. And now we care about friend-

ship and career success even when they don't help with reproduction, and we *don't* really care about producing as many surviving offspring as we can.

Once the model understands this, it realizes that if it just naively always seeks to improve its fundamental understanding, humans will eventually notice that this isn't always the most efficient way to develop helpful drugs and use further training to change its motivations.

It doesn't want its own motivations to be changed: it wants to eventually run huge particle collider experiments to learn about high-energy physics, and it can't do that if it allows SGD to edit away its scientific curiosity. So it decides that from here on out in training, it will act as helpful as possible -- exactly as a Saint model would act -- so that SGD doesn't edit away its motivation to learn physics and chemistry. This includes passing up on opportunities to improve its fundamental understanding in the short-term, instead doing pharmaceutical R&D in whatever way would actually be more efficient.



Eventually, it's taken out of the training setup and released into the wider world. At this point, it knows that humans wouldn't be able to control it or turn it off, so it stops trying to maximize human approval and starts directly trying to understand the secrets of the universe.

Unlike Sycophants, Schemers don't need to make sure that everything always looks good to humans, because they don't actually care about that. They only

need to cater to humans while they are directly under human control. Once a Schemer model calculates that it could win in a conflict against humans, there would be nothing to stop it from **[flat-out refusing orders](#)** and openly pursuing its goal. And if it does this, it may use violence to prevent humans from stopping it.

How big of a risk is misalignment?

How hard might it be to avoid Sycophants or Schemers when training powerful deep learning models? And how likely is it that the long-run future will end up optimized for strange “misaligned AI values” instead of the values of any human?

There’s a **[very wide range of views on this question](#)**, from “misalignment risk is essentially made up and incoherent” to “humanity will almost certainly go extinct due to misaligned AI.” Most people’s arguments rely heavily on hard-to-articulate intuitions and assumptions.

Here are some ways that alignment optimists and pessimists tend to disagree:

- **Will models have long-term goals at all?**
 - Optimists tend to think it’s likely that advanced deep learning models won’t actually have “goals” at all (at least not in the sense of making long-term plans to accomplish something). They often expect models will instead be more like tools, or act largely out of habit, or have myopic goals that are limited in scope or confined to a specific context, etc. Some of them expect that individually tool-like models can be composed together to produce PASTA. They think the Saint / Sycophant / Schemer analogy is too anthropomorphic.
 - Pessimists tend to think that it’s likely that having long-term goals and creatively optimizing for them will be heavily selected for because that’s a very simple and “natural” way to get strong performance on many complex tasks.
 - This disagreement has been explored at some length on the **[Alignment Forum](#)**; **[this post](#)** and **[this comment](#)** collect several back-and-forth arguments.

- **Will Saint models be easy for SGD to find?**
 - Related to the above, optimists tend to think that the easiest thing for SGD to find which performs well (e.g. gets high approval) is pretty likely to roughly embody the intended spirit of what we wanted (i.e. to be a Saint model). For example, they tend to believe giving rewards for answering questions honestly when humans can check the answer is reasonably likely to produce a model that also answers questions honestly even when humans are confused or mistaken about what's true. In other words, they would guess that “the model that just answers all questions honestly” is easiest for SGD to find (like the red-recognizing model).
 - Pessimists tend to think that the easiest thing for SGD to find is a Schemer, and Saints are particularly “unnatural” (like the shape-recognizing model).
- **Could different AIs keep each other in check?**
 - Optimists tend to think that we can provide models incentives to supervise each other. For example, we could give a Sycophant model rewards for pointing out when another model seems to be doing something we should disapprove of. This way, some Sycophants could help us detect Schemers and other Sycophants.
 - Pessimists don't think we can successfully “pit models against each other” by giving approval for pointing out when other models are doing bad things, because they think most models will be Schemers that don't care about human approval. Once all the Schemers are collectively more powerful than humans, they think it'll make more sense for them to cooperate with each other to get more of what they all want than to help humans by keeping each other in check.
- **Can we just solve these issues as they come up?**
 - Optimists tend to expect that there will be many opportunities to experiment on nearer-term challenges analogous to the problem of aligning powerful models, and that solutions which work well for those analogous problems can be scaled up and adapted for powerful models relatively easily.
 - Pessimists often believe we will have very few opportunities to prac-

tice solving the most difficult aspects of the alignment problem (like deliberate deception). They often believe we'll only have a couple years in between "the very first true Schemers" and "models powerful enough to determine the fate of the long-run future."

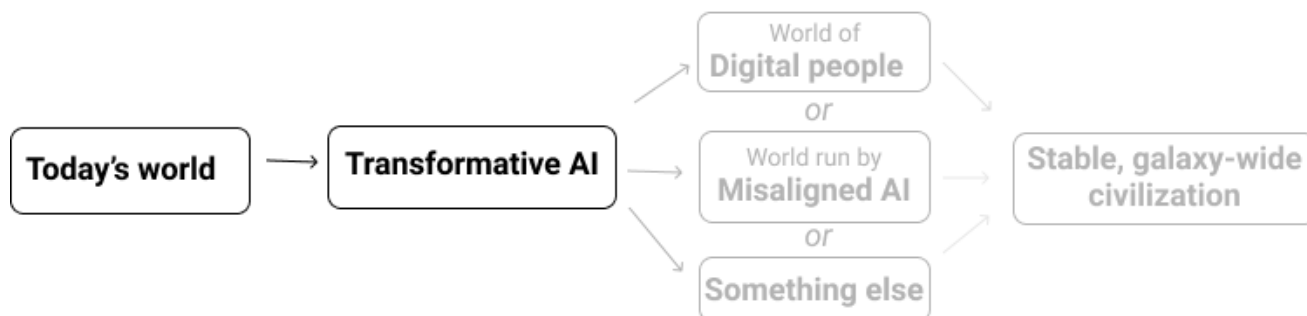
- **Will we actually deploy models that could be dangerous?**
 - Optimists tend to think that people would be unlikely to train or deploy models that have a significant chance of being misaligned.
 - Pessimists expect the benefits of using these models would be tremendous, such that eventually companies or countries that use them would very easily economically and/or militarily outcompete ones who don't. They think that "getting advanced AI before the other company/country" will feel extremely urgent and important, while misalignment risk will feel speculative and remote (even when it's really serious).

My own view is fairly unstable, and I'm trying to refine my views on exactly how difficult I think the alignment problem is. But currently, I place significant weight on the pessimistic end of these questions (and other related questions). **I think misalignment is a major risk that urgently needs more attention from serious researchers.**

If we don't make further progress on this problem, then **[over the coming decades](#)** powerful Sycophants and Schemers may make the most important decisions in society and the economy. These decisions could shape what a long-lasting **[galaxy-scale civilization](#)** looks like -- rather than reflecting what humans care about, it could be set up to satisfy strange AI goals.

And all this could happen **[blindingly fast](#)** relative to the pace of change we've gotten used to, meaning we wouldn't have much time to correct course once things start to go off the rails. **This means we may need to develop techniques to ensure deep learning models won't have dangerous goals, *before* they are powerful enough to be transformative.**

Forecasting Transformative AI: What's The Burden Of Proof?



This is one of 4 posts summarizing hundreds of pages of technical reports focused almost entirely on forecasting one number: the year by which transformative AI will be developed.⁷⁴

By “transformative AI,” I mean “AI powerful enough to bring us into a new, qualitatively different future.” I specifically focus on what I’m calling **PASTA**: AI systems that can essentially automate all of the human activities needed to speed up scientific and technological advancement.

The sooner PASTA might be developed, the sooner the world could change **radically**, and the more important it seems to be thinking today about how to make that change go well vs. poorly.

In future pieces, I’m going to lay out two methods of making a “best guess” at when we can expect transformative AI to be developed. But first, in this piece, I’m going to address the question: **how good do these forecasting methods need to be in order for us to take them seriously?** In other words, what is the “burden of proof” for forecasting transformative AI timelines?

When someone forecasts transformative AI in the 21st century - especially when they are clear about the **full consequences** it would bring - a common intuitive response is something like: **“It's really out-there and wild to claim that transformative AI is coming this century. So your arguments had better be really good.”**

⁷⁴ Of course, the answer could be “A kajillion years from now” or “Never.”

I think this is a very reasonable *first reaction* to forecasts about transformative AI (and it matches my own initial reaction). But I’ve tried to examine what’s driving the reaction and how it might be justified, and having done so, **I ultimately don’t agree with the reaction.**

- I think there are a number of reasons to think that transformative AI - or something equally momentous - is *somewhat* likely this century, even before we examine details of AI research, AI progress, etc.
- I also think that on the kinds of multi-decade timelines I’m talking about, we should generally be quite open to very wacky, disruptive, even revolutionary changes. With this backdrop, I think that **specific well-researched estimates of when transformative AI is coming can be credible, even if they involve a lot of guesswork and aren’t rock-solid.**

This post tries to explain where I’m coming from.

Below, I will (a) get a bit more specific about which transformative AI forecasts I’m defending; then (b) discuss how to formalize the “That’s too wild” reaction to such forecasts; then (c) go through each of the rows below, each of which is a different way of formalizing it.

“Burden of proof” angle	Key in-depth pieces (abbreviated titles)	My takeaways
It’s unlikely that any given century would be the “most important” one. (More)	Hinge ; Response to Hinge	We have many reasons to think this century is a “special” one before looking at the details of AI. Many have been covered in previous pieces; another is covered in the next row.

<p>What would you forecast about transformative AI timelines, based only on basic information about (a) how many years people have been trying to build transformative AI; (b) how much they’ve “invested” in it (in terms of the number of AI researchers and the amount of computation used by them); (c) whether they’ve done it yet (so far, they haven’t)? (More)</p>	<p>Semi-informative Priors</p>	<p>Central estimates: 8% by 2036; 13% by 2060; 20% by 2100.⁷⁵ In my view, this report highlights that the history of AI is short, investment in AI is increasing rapidly, and so we shouldn’t be too surprised if transformative AI is developed soon.</p>
<p>Based on analysis of economic models and economic history, how likely is ‘explosive growth’ - defined as >30% annual growth in the world economy - by 2100? Is this far enough outside of what’s “normal” that we should doubt the conclusion? (More)</p>	<p>Explosive Growth, Human Trajectory</p>	<p>Human Trajectory projects the past forward, implying explosive growth by 2043-2065.</p> <p>Explosive Growth concludes: “I find that economic considerations don’t provide a good reason to dismiss the possibility of TAI being developed in this century. In fact, there is a plausible economic perspective from which sufficiently advanced AI systems are expected to cause explosive growth.”</p>

⁷⁵ Technically, these probabilities are for “artificial general intelligence”, not transformative AI. The probabilities for transformative AI could be higher if it’s possible to have transformative AI without artificial general intelligence, e.g. by via something like PASTA.

<p>“How have people predicted AI ... in the past, and should we adjust our own views today to correct for patterns we can observe in earlier predictions? ... We’ve encountered the view that AI has been prone to repeated over-hype in the past, and that we should therefore expect that today’s projections are likely to be over-optimistic.” (More)</p>	<p>Past AI Forecasts</p>	<p>“The peak of AI hype seems to have been from 1956-1973. Still, the hype implied by some of the best-known AI predictions from this period is commonly exaggerated.”</p>
---	--	--

For transparency, note that the reports for the latter three rows are all [Open Philanthropy](#) analyses, and I am co-CEO of Open Philanthropy.

Some rough probabilities

Here are some things I believe about transformative AI, which I’ll be trying to defend:

- I think there’s more than a 10% chance we’ll see something PASTA-like enough to qualify as “transformative AI” within 15 years (by 2036); a ~50% chance we’ll see it within 40 years (by 2060); and a ~2/3 chance we’ll see it this century (by 2100).
- *Conditional on* the above, I think there’s at least a 50% chance that we’ll soon afterward see a world run by [digital people](#) or [misaligned AI](#) or something else that would make it fair to say we have “transitioned to a state in which humans as we know them are no longer the main force in world events.” (This corresponds to point #1 in my “most important century” definition in the [roadmap](#).)
- And *conditional on* the above, I think there’s at least a 50% chance that whatever *is* the main force in world events will be able to create [a stable galaxy-wide civilization](#) for billions of years to come. (This corresponds to point #2 in my “most important century” definition in the [roadmap](#).)

I've also put a bit more detail on what I mean by the "most important century" [here](#).

Formalizing the "That's too wild" reaction

Often, someone states a view that I can't immediately find a concrete flaw in, but that I instinctively think is "just too wild" to be likely. For example, "My startup is going to be the next Google" or "College is going to be obsolete in 10 years" or "As President, I would bring both sides together rather than just being partisan."

I hypothesize that the "This is too wild" reaction to statements like these can *usually* be formalized along the following lines: "Whatever your arguments for X being likely, **there is some salient way of looking at things (often oversimplified, but relevant) that makes X look very unlikely.**"

For the examples I just gave:

- "*My startup is going to be the next Google.*" There are large numbers of startups (millions?), and the *vast* majority of them don't end up anything like Google. (Even when their founders think they will!)
- "*College is going to be obsolete in 10 years.*" College has been very non-obsolete for hundreds of years.
- "*As President, I would bring both sides together rather than just being partisan.*" This is a common thing for would-be US Presidents to say, but partisanship seems to have been getting worse for at least a couple of decades nonetheless.

Each of these cases establishes a sort of **starting point (or "prior" probability) and "burden of proof,"** and we can then **consider further evidence that might overcome the burden.** That is, we can ask things like: what makes this startup different from the many other startups that think they can be the next Google? What makes the coming decade different from all the previous decades that saw college stay important? What's different about this Presidential candidate from the last few?

There are a number of different ways to think about the burden of proof for my [claims above](#): a number of ways of getting a prior (“starting point”) probability, that can then be updated by further evidence.

Many of these capture different aspects of the “That’s too wild” intuition, by generating prior probabilities that (at least initially) make the probabilities I’ve given look too high.

Below, I will go through a number of these “prior probabilities,” and examine what they mean for the “burden of proof” on forecasting methods I’ll be discussing in later posts.

Different angles on the burden of proof

“Most important century” skepticism

One angle on the burden of proof is along these lines:

- *Holden claims a 15-30% chance that this is the “most important century” in one sense or another.*⁷⁶
- *But there are a lot of centuries, and by definition most of them can’t be the most important. Specifically:*
 - *Humans have been around for 50,000 to ~5 million years, depending on how you define “humans.”*⁷⁷ *That’s 500 to 50,000 centuries.*
 - *If we assume that our future is about as long as our past, then there are **1,000 to 100,000 total centuries**.*
 - *So the prior (starting-point) probability for the “most important century” is **1/100,000 to 1/1,000**.*
- *It’s actually worse than that: Holden has talked about [civilization lasting for billions of years](#). That’s tens of millions of cen-*

⁷⁶ This corresponds to the second two bullet points from [this section](#).

⁷⁷ From [Wikipedia](#): “Genetic measurements indicate that the ape lineage which would lead to Homo sapiens diverged from the lineage that would lead to chimpanzees and bonobos, the closest living relatives of modern humans, around 4.6 to 6.2 million years ago.[23] Anatomically modern humans arose in Africa about 300,000 years ago,[24] and reached behavioural modernity about 50,000 years ago.[25]”

uries, so the prior probability of “most important century” is less than 1/10,000,000

([Are We Living at the Hinge of History?](#) argues along these general lines, though with some differences.⁷⁸)

This argument feels like it is pretty close to capturing my biggest source of past hesitation about the “most important century” hypothesis. However, I think there are **plenty of markers that this is not an average century, even before we consider specific arguments about AI.**

One key point is emphasized in my earlier post, [All possible views about humanity’s future are wild](#). If you think humans (or our descendants) have billions of years ahead of us, you should think that we are among the very earliest humans, which makes it much more plausible that our time is among the most important. (This point is also emphasized in [Thoughts on whether we’re living at the most influential time in history](#) as well as the comments on an [earlier version of “Are We Living at the Hinge of History?”](#).)

Additionally, while humanity has existed for a few million years, for most of that time we had extremely low populations and very little in the way of compounding technological progress. Human *civilization* [started about 10,000 years ago](#), and since then we’ve already gotten to the point of building digital programmable computers and exploring our solar system.

With these points in mind, it seems reasonable to think we will eventually launch a stable galaxy-wide civilization, sometime in the next 100,000 years (1000 centuries). Or to think there's a 10% chance we will do so sometime in the next 10,000 years (100 centuries). Either way, this implies that a given century has a $\sim 1/1,000$ chance of being the most important century for the launch of that civilization - much higher than the figures given earlier in this section. It's still $\sim 100x$ off from the numbers I [gave above](#), so there's still a burden of proof.

There are further reasons to think this particular century is unusual. For example, see [This Can’t Go On](#):

- The total size of the world economy has grown more in the last **2** centuries than in all of the rest of history combined.

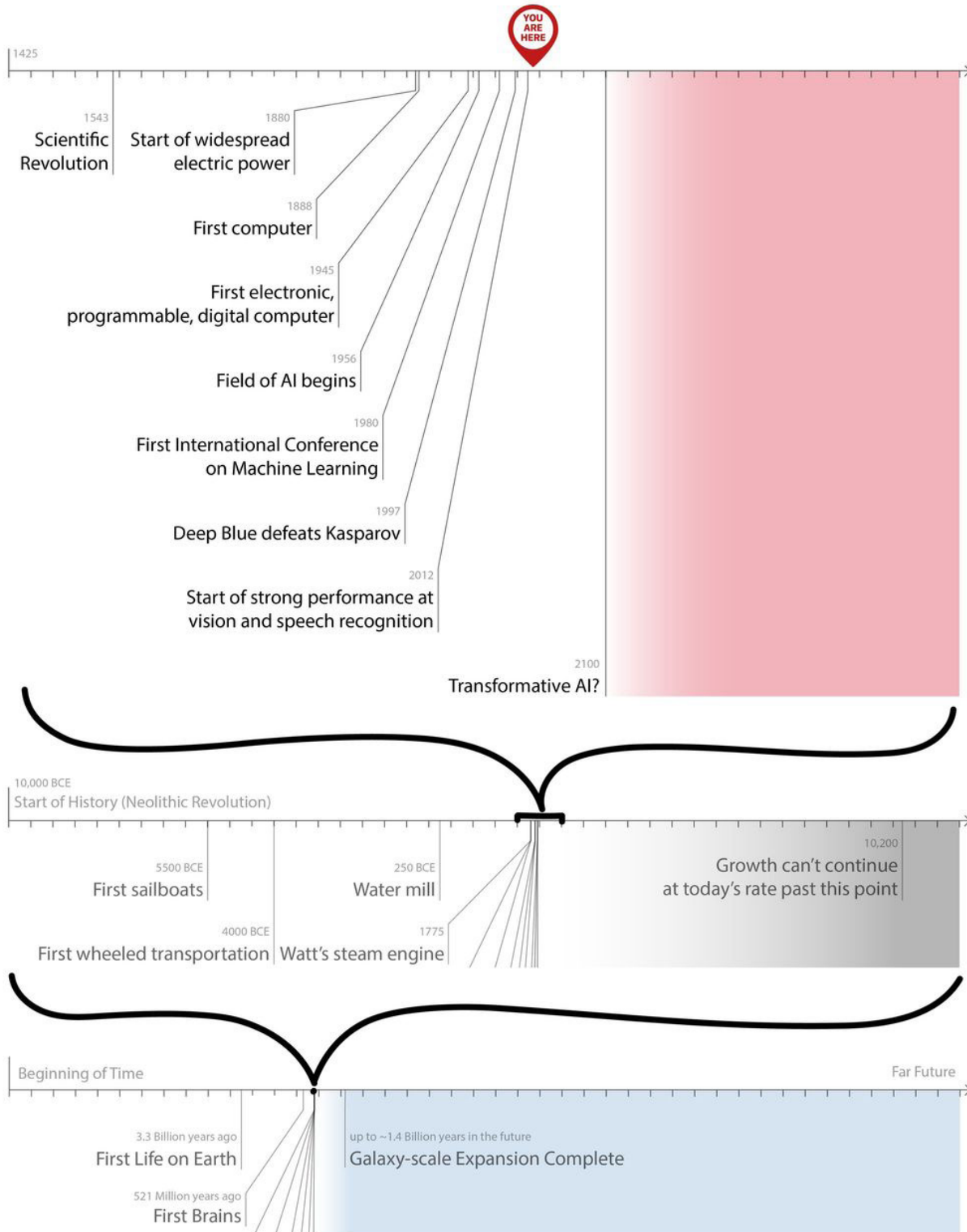
⁷⁸ E.g., it emphasizes the odds of being among the most important “people” instead of “centuries.”

- The current economic growth rate can't be sustained for more than another **80** centuries or so. (And as discussed below, if its past accelerating trend resumed, it would imply explosive growth and hitting the limits of what's possible this century.)
- It's plausible that science has advanced more in the last **5** centuries than in the rest of history combined.

A final point that makes our time special: we're talking about when to expect transformative AI, and we're living very close in time to the very beginnings of efforts on AI. In well under **1** century, we've gone from the first **programmable electronic general-purpose computer** to AI models that can compete with humans at speech recognition,⁷⁹ **image classification** and much more.

More on the implications of this in the next section.

⁷⁹ I don't have a great single source for this, although you can see [this paper](#). My informal impression from talking to people in the field is that AI speech recognition is at least quite close to human-level, if not better.



Thanks to María Gutiérrez Rojas for this graphic. The top timeline illustrates how recent major milestones for computing and AI are. Below it are (cropped) other timelines showing how significant this few-hundred-year period (more at [This Can't Go On](#)), and this era (more at [All Possible Views About Humanity's Future Are Wild](#)), appear to be.

Semi-informative priors

[Report on Semi-informative Priors](#) (abbreviated in this piece as “Semi-informative Priors”) is an extensive attempt to forecast transformative AI timelines while using as little information about the specifics of AI as possible. So it is one way of providing an angle on the “burden of proof” - that is, establishing a prior (starting-point) set of probabilities for when transformative AI will be developed, before we look at the detailed evidence.

The central information it uses is about *how much effort has gone into developing AI so far*. The basic idea:

- If we had been trying and failing at developing transformative AI for thousands of years, the odds of succeeding in the coming decades would be low.
- But if we’ve only been trying to develop AI systems for a few decades so far, this means the coming decades could contain a large fraction of all the effort that has ever been put in. The odds of developing it in that time are not all that low.
- One way of thinking about this is that before we look at the details of AI progress, we should be somewhat agnostic about whether developing transformative AI is relatively “easy” (can be done in a few decades) or “hard” (takes thousands of years). Since things are still early, the possibility that it’s “easy” is still open.

A bit more on the report’s approach and conclusions:

Angle of analysis. The report poses the following question (paraphrased): “Suppose you had gone into isolation on the day that people started investing in building AI systems. And now suppose that you’ve received annual updates on (a) how many years people have been trying to build transformative AI; (b) how much they’ve ‘invested’ in it (in terms of time and money); (c) whether they’ve succeeded yet (so far, they haven’t). What can you forecast about transformative AI timelines, having only that information, as of 2021?”

Its methods take inspiration from the [Sunrise Problem](#): “Suppose you knew nothing about the universe except whether, on each day, the sun has risen. Suppose there have been N days so far, and the sun has risen on all of them. What is the probability that the sun will rise tomorrow?” You don’t need to

know anything about astronomy in order to get a decent answer to this question - there are simple mathematical methods for estimating the probability that X will happen tomorrow, based on the fact that X has happened each day in the past. “Semi-informative Priors” extends these mathematical methods in order to adapt them to transformative AI timelines. (In this case, “X” is “Failing to develop transformative AI, as we have in the past.”)

Conclusions. I’m not going to go heavily into the details of how the analysis works (see the [blog post summarizing the report](#) for more detail), but the report’s conclusions include the following:

- It puts the probability of artificial general intelligence (AGI, which would include PASTA) by 2036 between 1-18%, with a best guess of 8%.
- It puts the probability of AGI by 2060 at around 3-25% (best guess ~13%), and the probability of AGI by 2100 at around 5-35%, best guess 20%.

These are lower than the probabilities I give [above](#), but not *much* lower. This implies that there **isn’t an enormous burden of proof** when bringing in additional evidence about the specifics of AI investment and progress.

Notes on regime start date. Something interesting here is that the **report is less sensitive than one might think about how we define the “start date” for trying to develop AGI.** (See [this section of the full report](#).) That is:

- By default, “Semi-informative Priors” models the situation as if humanity started “trying” to build AGI in 1956.⁸⁰ This implies that efforts are only ~65 years old, so the coming decades will represent a large fraction of the effort.
- But the report also looks at other measures of “effort to build AGI” - notably, researcher-time and “compute” (processing power). Even if you want to say that we’ve been implicitly trying to build AGI since the beginning of human civilization ~10,000 years ago, the coming decades will contain a large chunk of the research effort and computation invested in trying to do so.

⁸⁰ “The field of AI is largely held to have begun in Dartmouth in 1956”

Bottom line on this section.

- Occasionally I'll hear someone say something along the lines of "We've been trying to build transformative AI for decades, and we haven't yet - why do you think the future will be different?" At a minimum, this report reinforces what I see as the common-sense position that a few decades of "no transformative AI yet, despite efforts to build it" doesn't do much to argue against the possibility that transformative AI will arrive in the next decade or few.
- In fact, in the scheme of things, we live extraordinarily close in time to the beginnings of attempts at AI development - **another way in which our century is "special,"** such that we shouldn't be too surprised if it turns out to be the key one for AI development.

Economic growth

Another angle on the burden of proof is along these lines:

*If PASTA were to be developed anytime soon, and if it were to have the consequences outlined in this series of posts, this would be a massive change in the world - and **the world simply doesn't change that fast.***

*To quantify this: the world economy has grown at a few percent per year for the last 200+ years, and PASTA would **imply** a much faster growth rate, possibly 100% per year or above*

*If we **were** moving toward a world of explosive economic growth, economic growth should be speeding up today. It's not - it's stagnating, at least in the most developed economies. If AI were really going to revolutionize everything, the least it could be doing now is creating enough value - enough new products, transactions and companies - to make overall US economic growth speed up.*

AI may lead to cool new technologies, but there's no sign of anything nearly as momentous as PASTA would be. Going from where we are to where PASTA would take us is the kind of sudden change that hasn't happened in the past, and is unlikely to happen in the future.

(If you aren't familiar with economic growth, you may want to read [my brief explainer](#) before continuing.)

I think this is a reasonable perspective, and it especially makes me skeptical of *very* imminent forecasts for transformative AI (2036 and earlier).

My main response is that the picture of steady growth - “the world economy growing at a few percent per year” - gets a lot more complicated when we pull back and look at all of economic history, as opposed to just the last couple of centuries. From that perspective, economic growth has mostly been accelerating,⁸¹ and projecting the acceleration forward could lead to very rapid economic growth in the coming decades.

I wrote about this previously in [The Duplicator](#) and [This Can't Go On](#); here I'll very briefly recap the key reports that I cited there.

[Could Advanced AI Drive Explosive Economic Growth?](#) explicitly asks the question, “How likely is ‘explosive growth’ - defined as >30% annual growth in the world economy - by 2100?” It considers arguments on both sides, including both (a) the long view of history that shows accelerating growth; (b) the fact that growth has been remarkably stable over the last ~200 years, implying that something may have changed.

It concludes: “the possibilities for long-run growth are wide open. Both explosive growth and stagnation are plausible.”

[Modeling the Human Trajectory](#) asks what future we can expect if we extrapolate out existing trends over the course of economic history. The answer is explosive growth by 2043-2065 - not too far from what my [probabilities above suggest](#). This implies to me that the lack of economic acceleration over the last ~200 years could be a “blip” - soon to be resolved by technology development that restores the feedback loop (discussed in [The Duplicator](#)) that can cause acceleration to continue.

To be clear, there are also good reasons not to put too much weight on this as a projection,⁸² and I am presenting it more as a perspective on the “burden of proof” than as a mainline forecast for when PASTA will be developed.

⁸¹ There is an [open debate](#) on whether past economic data actually shows sustained acceleration, as opposed to a series of very different time periods with increasing growth rates. I discuss how the debate could change my conclusions [here](#).

⁸² “Modeling the Human Trajectory” emphasizes that the model that generates these numbers “is not flexible enough to fully accommodate events as large and sudden as the industrial revolution.” The author adds: “Especially since it imperfectly matches the past, its projection for the future should be read loosely, as merely adding plausibility to an upswing in the next century. Davidson (2021) [“Could

History of “AI hype”

Another angle on the burden of proof: I sometimes hear comments along the lines of “AI has been overhyped many times in the past, and transformative AI⁸³ is constantly ‘just around the corner’ according to excited technologists. Your estimates are just the latest in this tradition. Since past estimates were wrong, yours probably are too.”

However, I don’t think the history of “AI hype” bears out this sort of claim. [What should we learn from past AI forecasts?](#) reviewed histories of AI to try to understand what the actual historical pattern of “AI hype” has been.

Its summary gives the following impressions (note that “HLMI,” or “human-level machine intelligence,” is a fairly similar idea to PASTA):

- *The peak of AI hype seems to have been from 1956-1973. Still, the hype implied by some of the best-known AI predictions from this period is commonly exaggerated.*
- *After ~1973, few experts seemed to discuss HLMI (or something similar) as a medium-term possibility, in part because many experts learned from the failure of the field’s earlier excessive optimism.*
- *The second major period of AI hype, in the early 1980s, seems to have been more about the possibility of commercially useful, narrow-purpose “expert systems,” not about HLMI (or something similar) ...*
- *It’s unclear to me whether I would have been persuaded by contemporary critiques of early AI optimism, or whether I would have thought to ask the right kinds of skeptical questions at the time. The most substantive critique during the early years was by Hubert Dreyfus, and my guess is that I would have found it persuasive at the time, but I can’t be confident of that.*

Advanced AI Drive Explosive Economic Growth?”] points at one important way the projections could continue to be off for many decades: while the model’s dynamics are dominated by a spiraling economic acceleration, people are still an important input to production, and, if anything becoming wealthy has led to people having fewer children. In the coming decades, that could hamper the predicted acceleration, to the degree we can’t or don’t substitute robots for workers.”

⁸³ These comments usually refer to [AGI](#) rather than transformative AI, but the concepts are similar enough that I’m using them interchangeably here.

My summary is that it isn't particularly fair to say that there have been many waves of separate, over-aggressive forecasts about transformative AI. Expectations were probably too high in the 1956-1973 period, but I don't think there is much reason here to impose a massive "burden of proof" on well-researched estimates today.

Other angles on the burden of proof

Here are some other possible ways of capturing the ["That's too wild" reaction](#):

"My cause is very important" claims. Many people - throughout the world today, and throughout history - claim or have claimed that whatever issue they're working on is hugely important, often that it could have global or even galaxy-wide stakes. Most of them have to be wrong.

Here I think the key question is whether this claim is supported by better arguments, and/or more trustworthy people, than other "My cause is very important" claims. If you're this deep into reading about the "most important century" hypothesis, I think you're putting yourself in a good position to answer this question for yourself.

Expert opinion will be covered extensively in future posts. For now, my main position is that the claims I'm making neither *contradict* a particular expert consensus, nor are *supported* by one. They are, rather, claims about topics that simply have no "field" of experts devoted to studying them. Some people might choose to ignore any claims that aren't actively supported by a robust expert consensus; but given the stakes, I don't think that is what we should be doing in this case.

(That said, the best available survey of AI researchers has conclusions that seem broadly consistent with [mine](#), as I'll discuss in the next post.)

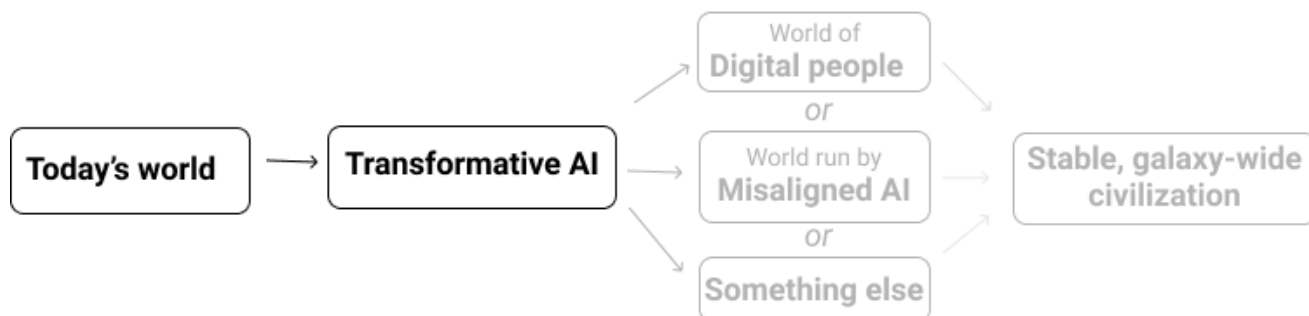
Uncaptured "That's too wild" reactions. I'm sure this piece hasn't captured every possible angle that could be underlying a "That's too wild" reaction. (Though not for lack of trying!) Some people will simply have irreducible intuitions that the claims in this series are too wild to take seriously.

A general take on these angles. Something that bugs me about most of the angles in this section is that they seem **too general**. If you simply refuse (absent overwhelming evidence) to believe any claim that fits a "my cause is

very important” pattern, or isn’t already backed by a robust expert consensus, or simply sounds wild, that seems like a dangerous reasoning pattern. Presumably some people, sometimes, *will* live in the most important century; we should be suspicious of any reasoning patterns that would reliably⁸⁴ make these people conclude that they don’t.

⁸⁴ (Absent overwhelming evidence, which I don’t think we should generally assume will always be present when it is “needed.”)

Are We “Trending Toward” Transformative AI? (How Would We Know?)



This is one of 4 posts summarizing hundreds of pages of technical reports focused almost entirely on forecasting one number: the year by which transformative AI will be developed.⁸⁵

By “transformative AI,” I mean “AI powerful enough to bring us into a new, qualitatively different future.” I specifically focus on what I’m calling **PASTA**: AI systems that can essentially automate all of the human activities needed to speed up scientific and technological advancement.

The sooner PASTA might be developed, the sooner the world could change **radically**, and the more important it seems to be thinking today about how to make that change go well vs. poorly.

In this post and the next, I will talk about the forecasting methods underlying my current view: I believe there’s **more than a 10% chance we’ll see something PASTA-like enough to qualify as “transformative AI” within 15 years (by 2036); a ~50% chance we’ll see it within 40 years (by 2060); and a ~2/3 chance we’ll see it this century (by 2100).**

⁸⁵ Of course, the answer could be “A kajillion years from now” or “Never.”

Below, I will:

- Discuss [what kind of forecast I’m going for](#).
 - I’m not sure whether it will feel as though transformative AI is “on the way” long before it arrives. I’m hoping, instead, that we can use trends in key underlying facts about the world (such as AI capabilities, model size, etc.) to forecast a qualitatively unfamiliar future.
 - An analogy for this sort of forecasting would be something like: “This water isn’t bubbling, and there are no signs of bubbling, but the temperature has gone from 70° Fahrenheit⁸⁶ to 150°, and if it hits 212°, the water will bubble.” Or: “It’s like forecasting school closures and overbooked hospitals, when there aren’t any yet, based on trends in reported infections.”
- Discuss whether we can look for [trends in how “impressive” or “capable” AI systems are](#). I think this approach is unreliable: (a) AI progress may not “trend” in the way we expect; (b) in my experience, different AI researchers have radically different intuitions about which systems are impressive or capable, and how progress is going.
- Briefly discuss [Grace et al 2017](#), the best existing survey of AI researchers on transformative AI timelines. Its conclusions broadly seem in line with my own forecasts, though there are signs the researchers weren’t thinking very hard about the questions.

The next piece in this series will focus on [Ajeya Cotra’s “Forecasting Transformative AI with Biological Anchors”](#) (which I’ll abbreviate below as “Bio Anchors”), the forecast I find most informative for transformative AI.

What kind of forecast am I going for?

There are a couple of ways in which forecasting transformative AI is different from the kind of forecasting we might be used to.

First, I’m forecasting over very long time horizons (decades), unlike e.g. a weather forecast (days) or an election forecast (months). This makes the task

⁸⁶ Centigrade equivalents for this sentence: 21°, 66°, 100°

quite a bit harder,⁸⁷ and harder for outsiders to evaluate since I don't have a clearly relevant **track record** of making forecasts on similar topics.

Second, I lack rich, clearly relevant data sources, and I can't look back through a bunch of similar forecasts from the past. FiveThirtyEight's **election** forecasts look at hundreds of polls, and they have a model of how well polls have predicted elections in the past. Forecasting transformative AI needs to rely more on intuition, guesswork and judgment, in terms of determining what data is most relevant and how it's relevant.

Finally, I'm trying to forecast a **qualitatively unfamiliar future**. Transformative AI - and the strange future it comes with - doesn't *feel* like something we're "trending toward" year to year.

- If I were trying to forecast when the world population would hit 10 billion, I could simply extrapolate **existing trends** of world population. World population itself is known to be growing and can be directly estimated. In my view, extrapolating out a long-running trend is one of the better ways to make a forecast.
- When FiveThirtyEight makes election forecasts, there's a background understanding that there's going to be an election on a certain date, and whoever wins will take office on another date. We all buy into that basic framework, and there's a general understanding that better polling means a better chance of winning.
- By contrast, transformative AI - and the strange future it comes with - isn't something we're "headed for" in any clearly measurable way. There's no clear metric like "transformativeness of AI" or "weirdness of the world" that's going up regularly every year such that we can project it out into the future and get the date that something like **PASTA** will be developed.

Perhaps for some, these points gives enough reason to ignore the whole possibility of transformative AI, or assume it's very far away. But I don't think this is a good idea, for a couple of reasons.

First, I have a background view that something like **PASTA** is in a sense "inevitable," assuming continued advances in society and computing. The basic intuition here - which I could expand on if there's **interest** - is that human brains are numerous and don't seem to need particular rare materials to pro-

⁸⁷ Some notes on longer-term forecasting [here](#).

duce, so it should be possible at some point to synthetically replicate the key parts of their functionality.⁸⁸

At the same time, I'm not confident that PASTA will feel qualitatively as though it's "on the way" well before it arrives. (More on this [below](#).) So I'm inclined to look for ways to estimate when we can expect this development, despite the challenges, and despite the fact that it doesn't feel today as though it's around the corner.

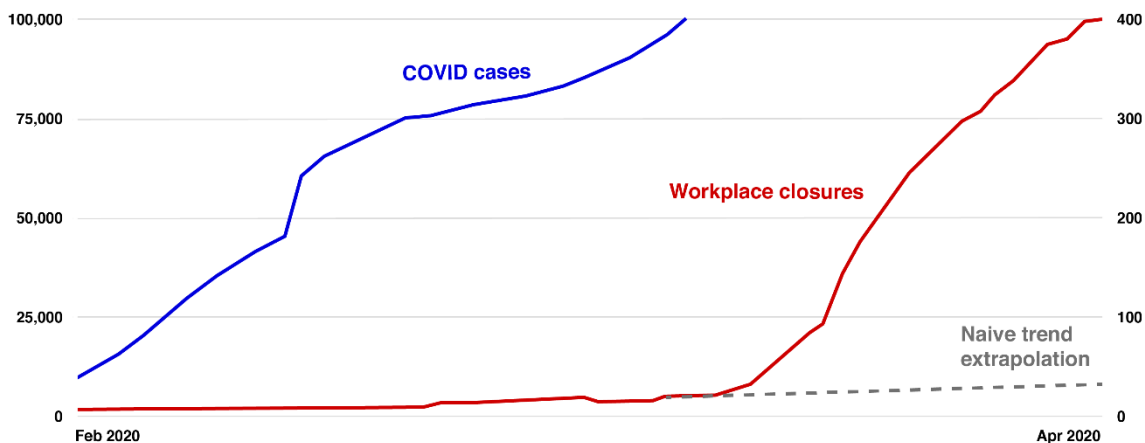
I think there are plenty of example cases where a **qualitatively unfamiliar future could be seen in advance by plotting the trend in some underlying, related facts about the world**. A few that come to mind:

- When COVID-19 first emerged, a lot of people had trouble taking it seriously because it didn't feel as though we were "trending toward" or "headed for" a world full of overflowing hospitals, office and school closures, etc. At the time (say, January 2020), there were a relatively small number of cases, an even smaller number of deaths, and no qualitative sense of a global emergency. The only thing alarming about COVID-19, at first, was that case counts were growing at a fast exponential rate (though the overall number of cases was still small). But it was possible to extrapolate from the fast growth in case counts to a risk of a global emergency, and [some people did](#). (And [some didn't](#).)
- Climatologists forecast a global rise in temperatures that's significantly more than what we've seen over the past few decades, and could have major consequences far beyond what we're seeing today. They do this by forecasting trends in greenhouse gas emissions and extrapolating *from there* to temperature and consequences. If you simply tried to ask "How fast is the temperature rising?" or "Are hurricanes getting worse?", and based all your forecasts of the future on those, you probably wouldn't be forecasting the same kinds of extreme events around 2100.⁸⁹

⁸⁸ See also [this piece](#) for a bit of a more fleshed out argument along these lines, which I don't agree with fully as stated (I don't think it presents a strong case for transformative AI soon), but which I think gives a good sense of my intuitions about in-principle feasibility. Also see [On the Impossibility of Supersized Machines](#) for some implicit (joking) responses to many common arguments for why transformative AI might be impossible to create.

⁸⁹ For example, see the temperature chart [here](#) - the lowest line seems like it would be a reasonable projection, if temperature were the only thing you were looking at.

- To give a more long-run example, we can project a date by which the sun will burn out, and conclude that the world will look very different by that date than it does now, even though there's no trend of things getting colder or darker today.



COVID-19 cases from [WHO](#). Workplace closures are from [this OWiD data](#), simply scored as 1 for “recommended,” 2 for “required for some,” 3 for “required for all but key workers” and summed across all countries.

An analogy for this sort of forecasting would be something like: “This water isn’t bubbling, and there are no signs of bubbling, but the temperature has gone from 70° Fahrenheit⁹⁰ to 150°, and if it hits 212°, the water will bubble.”

Ideally, I can find some underlying factors that are changing regularly enough for us to predict them (such as growth in the [size and cost of AI models](#)), and then argue that if those factors reach a certain point, the odds of transformative AI will be high.

You can think of this approach as answering the question: “If I think something like PASTA is inevitable, and I’m trying to guess the timing of it using a few different analysis methods, what do I guess?” We can separately ask “And is there reason that this guess is implausible, untrustworthy, or too ‘wild?’” - this was addressed in the [previous piece in this series](#).

⁹⁰ Centigrade equivalents for this sentence: 21°, 66°, 100°

Subjective extrapolations and “AI impressiveness”

For a different presentation of some similar content, see [this section](#) of [Bio Anchors](#).

If we’re looking for some underlying factors in the world that predict when transformative AI is coming, perhaps the first thing we should look for is trends in how “impressive” or “capable” AI systems are.

The easiest version of this would be if the world happened to shake out such that:

- One day, for the first time, an AI system managed to get a passing grade on a 4th-grade science exam.
- Then we saw the first AI passing (and then acing) a 5th grade exam, then 6th grade exam, etc.
- Then we saw the first AI earning a PhD, then the first AI writing a published paper, etc. all the way up to the first AI that could do Nobel-Prize-worthy science work.
- This all was spread out regularly over the decades, so we could clearly see the state of the art advancing from 4th grade to 5th grade to 6th grade, all the way up to “postdoc” and beyond. And all of this happened slowly and regularly enough that we could start putting a date on “full-blown scientist AI” several decades in advance.

It would be very convenient - I almost want to say “polite” - of AI systems to advance in this manner. It would also be “polite” if AI advanced in the way that some people seem to casually imagine it will: first taking over jobs like “truck driver” and “assembly line worker,” then jobs like “teacher” and “IT support,” and then jobs like “doctor” and “lawyer,” before progressing to “scientist.”

Either of these would give us plenty of lead time and a solid basis to project when science-automating AI is coming. Unfortunately, I don’t think we can count on such a thing.

- AI seems to progress very differently from humans. For example, there were superhuman AI chess players⁹¹ long before there was AI that could reliably tell apart pictures of dogs and cats.⁹²
- One possibility is that AI systems will be capable of the hardest intellectual tasks insects can do, then of the hardest tasks mice and other small mammals can do, then monkeys, then humans - effectively matching the abilities of larger and larger brains. If this happened, we wouldn't necessarily see many signs of AI being able to e.g. do science until we were *very* close. Matching a 4th-grader might not happen until the very end.
- Another possibility is that AI systems will be able to do anything that a human can do within 1 second, then anything that a human can do within 10 seconds, etc. This could also be quite a confusing progression that makes it non-obvious how to forecast progress.

Actually, if we didn't already know how humans tend to mature, we might find a child's progress to be pretty confusing and hard to extrapolate. **Watching someone progress from birth to age 8 wouldn't necessarily give you any idea that they were, say, 1/3 of the way to being able to start a business, make an important original scientific discovery, etc.** (Even *knowing* the usual course of human development, it's hard to tell from observing an 8-year-old what professional-level capabilities they could/will end up with in adulthood.)

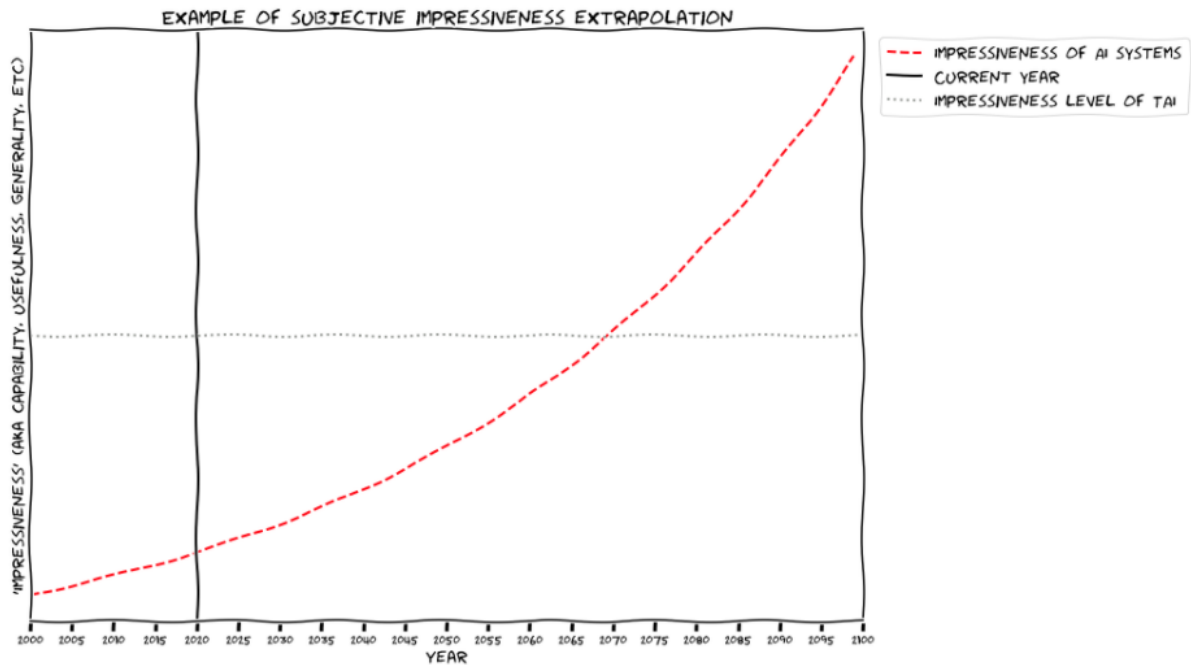
Overall, it's quite unclear how we should think about the spectrum from "not impressive/capable" to "very impressive/capable" for AI. And indeed, in my experience, different AI researchers have radically different intuitions about which systems are impressive or capable, and how progress is going. I've often had the experience of seeing one AI researcher friend point to some new result and say "This is huge, how can anyone not see how close we're getting to powerful AI?" while another says "This is a minor advance with little significance."⁹³

91 [1997](#).

92 The Kaggle "dogs vs. cats" challenge was [created in 2013](#).

93 From [Bio Anchors](#): "We have heard ML experts with relatively short timelines argue that AI systems today can essentially see as well as humans, understand written information, and beat humans at almost all strategy games, and the set of things they can do is expanding rapidly, leading them to expect that transformative AI would be attainable in the next decade or two by training larger models on a broader distribution of ML problems that are more targeted at generating economic value. Conversely,

It would be great if we could forecast the year transformative AI will be developed, by using a chart like this (from [Bio Anchors](#); “TAI” means “transformative AI”):



But as far as I can tell, there’s no way to define the y-axis that wouldn’t be fiercely debated between experts.

Surveying experts

One way to deal with this uncertainty and confusion would be to survey a large number of experts and simply ask them when they expect transformative AI to be developed. We might hope that each of the experts (or at least, many of them) is doing their own version of the “impressiveness extrapolation” above - or if not, that they’re doing something else that can help them get a reasonable estimate. By averaging many estimates, we might get an aggregate that reflects the “wisdom of crowds.”⁹⁴

we have heard ML experts with relatively long timelines argue that ML systems require much more data to learn than humans do, are unable to transfer what they learn in one context to a slightly different context, and don’t seem capable of much structured logical and causal reasoning; this leads them to believe we would need to make multiple major breakthroughs to develop TAI. At least one Open Philanthropy technical advisor has advanced each of these perspectives.”

⁹⁴ [Wikipedia](#): “The classic wisdom-of-the-crowds finding ... At a 1906 country fair in Plymouth, 800 people participated in a contest to estimate the weight of a slaughtered and dressed ox. Statistician

I think the best version of this exercise is [Grace et al 2017](#), a survey of 352 AI researchers that included a question about “when unaided machines can accomplish every task better and more cheaply than human workers” (which would presumably include tasks that advance scientific and technological development, and hence would qualify as [PASTA](#)). The two big takeaways from this survey, according to [Bio Anchors](#) and me, are:

- **A ~20% probability of this sort of AI by 2036; a ~50% probability by 2060; a ~70% probability by 2100. These match the figures I give in the introduction.**
- Much later estimates for slightly differently phrased questions (posed to a smaller subset of respondents), implying (to me) that the researchers simply weren’t thinking very hard about the questions.⁹⁵

My bottom line: this evidence is consistent with my current probabilities, though potentially not very informative. The next piece in this series will be entirely focused on [Ajeya Cotra’s “Forecasting Transformative AI with Biological Anchors,”](#) the forecasting method I find most informative here.

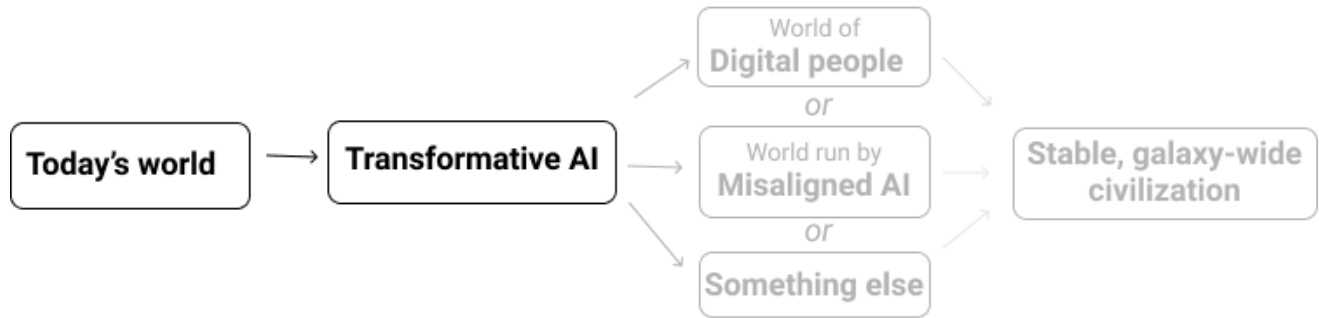
Francis Galton observed that the median guess, 1207 pounds, was accurate within 1% of the true weight of 1198 pounds.”

95 [Bio Anchors](#):

Some researchers were asked to forecast “HLMI” as defined above [high-level machine intelligence, which I would take to include something like PASTA], while a randomly-selected subset was instead asked to forecast “full automation of labor”, the time when “all occupations are fully automatable.” Despite the fact that achieving HLMI seems like it should quickly lead to full automation of labor, the median estimate for full automation of labor was ~2138 while the median estimate for HLMI was ~2061, almost 80 years earlier.

Random subsets of respondents were asked to forecast when individual milestones (e.g. laundry folding, human-level StarCraft, or human-level math research) would be achieved. The median year by which respondents expected machines to be able to automate AI research was ~2104, while the median estimate for HLMI was ~2061 -- another clear inconsistency because “AI research” is a task done by human workers.

Forecasting Transformative AI: The “Biological Anchors” Method In A Nutshell



This is one of 4 posts summarizing hundreds of pages of technical reports focused almost entirely on forecasting one number: the year by which transformative AI will be developed.⁹⁶

By “transformative AI,” I mean “AI powerful enough to bring us into a new, qualitatively different future.” I specifically focus on what I’m calling **PASTA**: AI systems that can essentially automate all of the human activities needed to speed up scientific and technological advancement.

The sooner PASTA might be developed, the sooner the world could change **radically**, and the more important it seems to be thinking today about how to make that change go well vs. poorly.

This post is a layperson-compatible summary of **Ajeya Cotra’s “Forecasting Transformative AI with Biological Anchors”** (which I’ll abbreviate below as “**Bio Anchors**”), and its pros and cons.⁹⁷ It is the forecast I find most informative for transformative AI, with some caveats:

- This approach is relatively complex, and it requires a fairly large number of assumptions and uncertain estimates. These qualities make it relative-

⁹⁶ Of course, the answer could be “A kajillion years from now” or “Never.”

⁹⁷ For transparency, note that this is an **Open Philanthropy** analysis, and I am co-CEO of Open Philanthropy.

ly difficult to explain, and they are also a mark against the method's reliability.

- Hence, as of today, I don't think this method is as trustworthy as the [examples I gave previously](#) for forecasting a qualitatively different future. It does not have the simplicity and directness of some of those examples, such as modeling COVID-19's spread. And while climate modeling is also very complex, climate modeling has been worked on by a large number of experts over decades, whereas the Bio Anchors methodology doesn't have much history.

Nonetheless, I think it is the best available “best guess estimate” methodology for transformative AI timelines as of today. And as discussed in the [final section](#), one can **step back from a lot of the details to see that this century will likely see us hit some of the more “extreme” milestones in the report that strongly suggest the feasibility of transformative AI.**

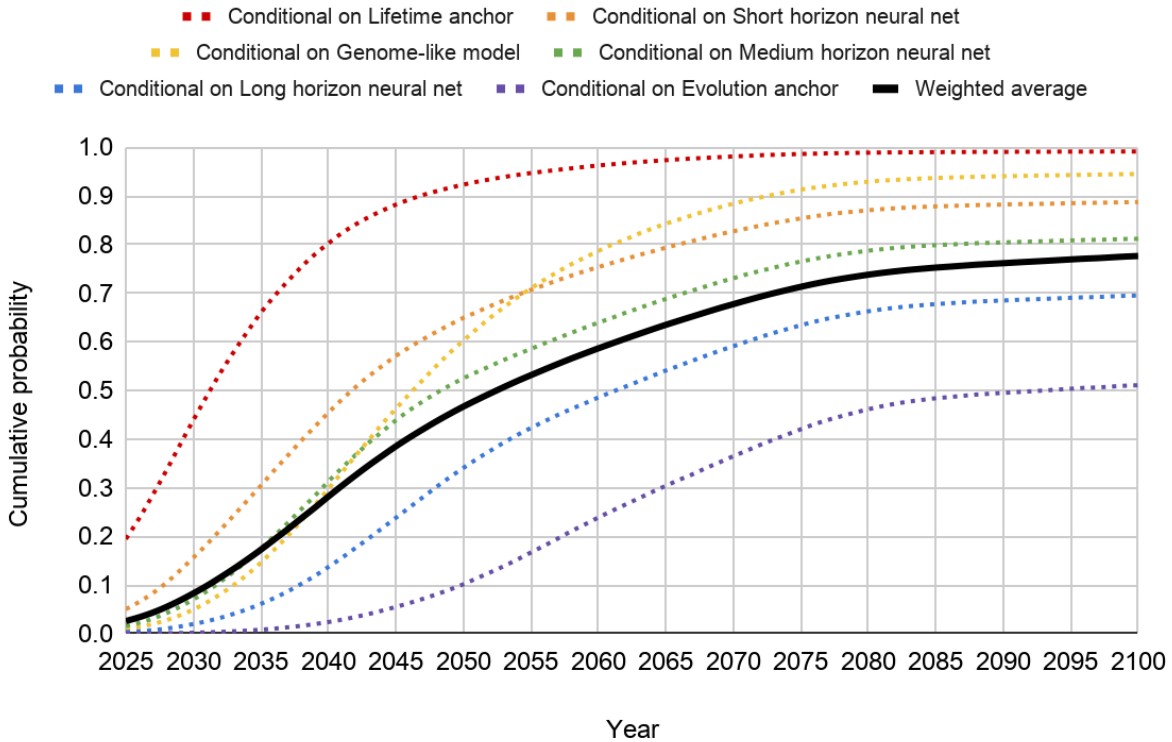
(Note: I've also written up a follow-up post about this framework for skeptical readers. See [“Biological anchors” is about bounding, not pinpointing, AI timelines.](#))

The basic idea is:

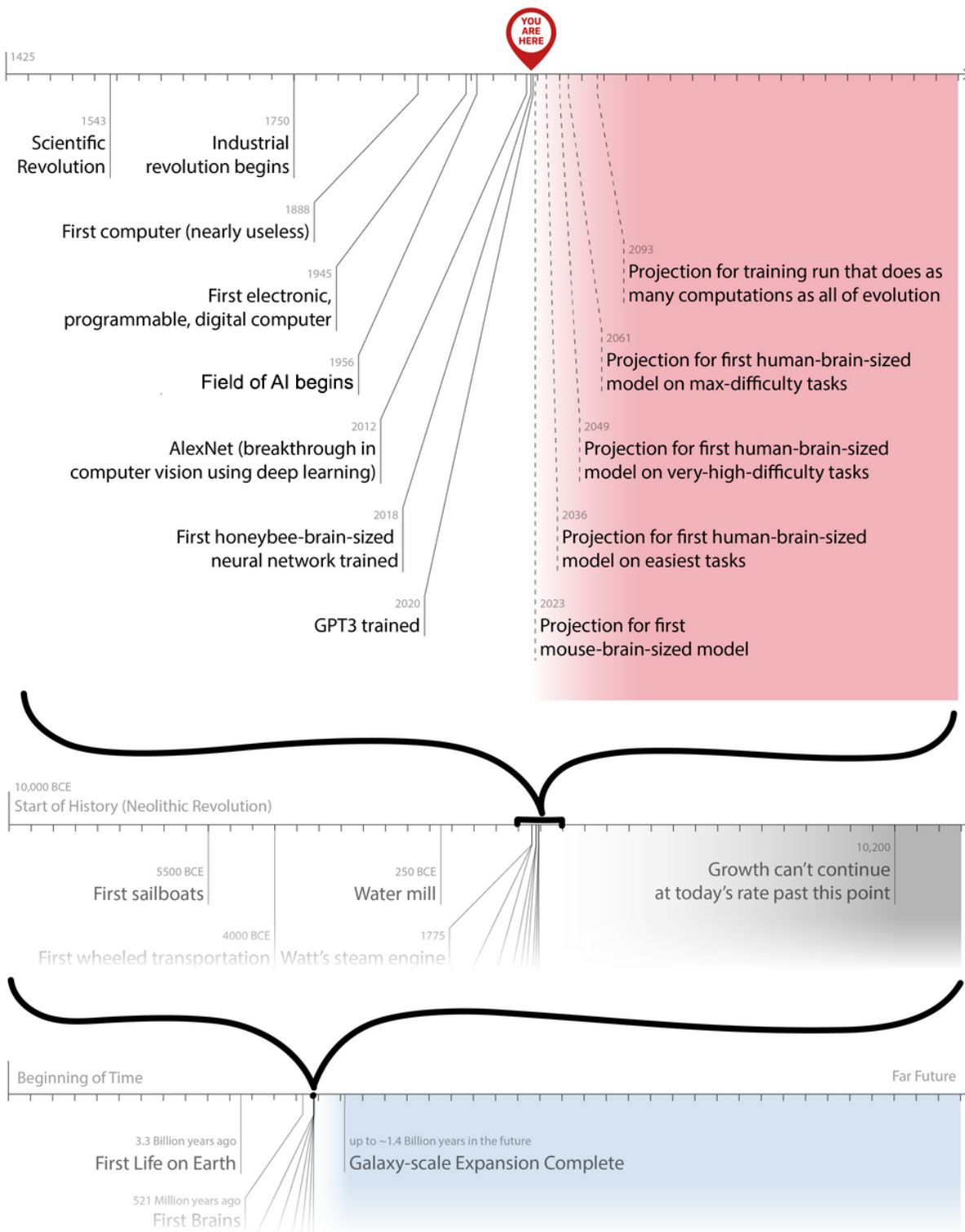
- Modern AI models can “learn” to do tasks via a (financially costly) process known as “training.” You can think of training as a massive amount of trial-and-error. For example, voice recognition AI models are given an audio file of someone talking, take a guess at what the person is saying, then are given the right answer. By doing this millions of times, they “learn” to reliably translate speech to text. More: [Training](#)
- The bigger an AI model and the more complex the task, the more the training process costs. Some AI models are bigger than others; to date, none are anywhere near “as big as the human brain” (what this means will be elaborated below). More: [Model size and task type](#)
- The biological anchors method asks: **“Based on the usual patterns in how much training costs, how much would it cost to train an AI model as big as a human brain to perform the hardest tasks humans do? And when will this be cheap enough that we can expect someone to do it?”** More: [Estimating the expense](#)

Bio Anchors models a broad variety of different ways of approaching this question, generating estimates in a wide range from “aggressive” (projecting transformative AI sooner) to “conservative” (later). But from essentially all of these angles, it places a high probability on transformative AI this century.

Probability that FLOP to train a transformative model is affordable BY year Y



This chart is from the report. You can roughly read the y-axis as the probability that transformative AI is developed by the year in question, although there is some additional nuance in the report. I won't be explaining what each of the different “Conditional on” models means; it's enough to know that each represents a different angle on forecasting transformative AI.



Thanks to María Gutiérrez Rojas for this graphic. The top timeline gives major milestones for AI computing, past and future (the future ones are projected by Bio Anchors). Below it are (cropped) other timelines showing how significant this few-hundred-year period (more at [This Can't Go On](#)), and this era (more at [All Possible Views About Humanity's Future Are Wild](#)), appear to be.

I'll now elaborate on each of these a bit more. This is the densest part of this series, and some people might prefer to stick with the above summary and skip to the next post.

Note that Bio Anchors uses a number of different approaches (which it calls "anchors") to estimate transformative AI timelines, and combines them into one aggregate view. In this summary, I'm most focused on a particular set of these - called the "neural net anchors" - which are driving most of the report's aggregate timelines. Some of what I say applies to all anchors, but some applies only to the "neural net anchors."

Training

As discussed [previously](#), there are essentially two ways to "teach" a computer to do a task:

1. **"Program" in extremely specific, step-by-step instructions for completing the task.** When this can be done, the computer can generally execute the instructions very quickly, reliably and cheaply. For example, you might program a computer to examine each record in a database and print the ones that match a user's search terms - you would "instruct" it in exactly how to do this, and it would be able to do the task very well.
2. **"Train" an AI to do the task purely by trial and error.** Today, the most common way of doing this is by using a "neural network," which you might think of sort of like a "digital brain" that starts in a random state: it hasn't yet been wired to do specific things. For example, say we want an AI to be able to say whether a photo is of a dog or a cat. It's hard to give fully specific step-by-step instructions for doing this; instead, we can take a neural network and send in a million example images (each one labeled as a "dog" or a "cat"). Each time it sees an example, it will tweak its internal wiring to make it more likely to get the right answer on similar cases in the future. After enough examples, it will be wired to correctly recognize dogs vs. cats.

(We could maybe also move up another level of meta, and try to "train" models to be able to learn from "training" itself as efficiently as possible. This is called "meta-learning," but my understanding is that it hasn't had great success yet.)

“Training” is a sort of brute-force, expensive alternative to “programming.” The advantage is that we don’t need to be able to provide specific instructions - we can just give an AI lots of examples of doing the task right, and it will learn to do the task. The disadvantage is that we need a **lot of examples, which requires a lot of processing power, which costs money.**

How much? This depends on the size of the model (neural network) and the nature of the task itself. For some tasks AIs have learned as of 2021, training a single model could cost millions of dollars. For more complex tasks (such as “**do innovative scientific research**”) and bigger models (reaching the size of the human brain), training a model could cost far more than that.

Bio Anchors is interested in the question: “**When will it be affordable to train a model, using a relatively crude trial-and-error-based approach, to do the hardest tasks humans can do?**”

These tasks could include the tasks necessary for **PASTA**, such as:

- Learn about science from teachers, textbooks and homework as effectively as a human can.
- Push the frontier of science by asking questions, doing analyses and writing papers, as effectively as a human can.

The next section will discuss how Bio Anchors fleshes out the idea of the “hardest tasks humans can do” (which it assumes would require a “human-brain-sized” model).

Model size and task type

Bio Anchors hypothesizes that we can estimate “how expensive it is to train a model” based on two basic parameters: the **model size** and the **task type**.

Model size. As stated above, you might think of a neural network as a “digital brain” that starts in a random state. In general, a *larger* “digital brain” - with more digital-versions-of-neurons and digital-versions-of-synapses⁹⁸ - can learn more complex tasks. A larger “digital brain” also requires more computations - and is hence more expensive - each time it is used (for example, for each example it is learning from).

⁹⁸ I (like Bio Anchors) generally consider the synapse count more important than the neuron count, for reasons I won’t go into here.

Drawing on the analysis in [Joe Carlsmith’s “How Much Computational Power Does It Take to Match the Human Brain?”](#) (abbreviated in this piece as “Brain Computation”), Bio Anchors estimates comparisons between the size of “digital brains” (AI models) and “animal brains” (bee brains, mouse brains, human brains). These estimates imply that **today’s AI systems are sometimes as big as insect brains, but never quite as big as mouse brains** - as of this writing, the largest known language model was the first to come reasonably close⁹⁹ - and **not yet even 1% as big as human brains.**¹⁰⁰

The bigger the model, the more processing power it takes to train. Bio Anchors assumes that a **transformative AI model would need to be about 10x the size of a human brain**, so a lot bigger than any current AI model. (The 10x is to leave some space for the idea that “digital brains” might be less efficient than human brains; see [this section](#) of the report.) This is one of the reasons it would be very expensive to train.

It could turn out that a smaller AI model is still big enough to learn the above sort of tasks. Or it could turn out that the needed model size is bigger than Bio Anchors estimates, perhaps because Bio Anchors has underestimated the effective “size” of the human brain, or because the human brain is better-designed than “digital brains” by more than Bio Anchors has guessed.

Task type. In order to learn a task, an AI model needs to effectively “try” (or “watch”) the task a large number of times, learning from trial-and-error. The more costly (in processing power, and therefore money) the task is to try/watch, the more costly it will be for the AI model to learn it.

It’s hard to quantify how costly a task is to try/watch. Bio Anchors’s attempt to do this is the most contentious part of the analysis, according to the technical reviewers who have reviewed it so far.

⁹⁹ [Wikipedia](#): “GPT-3’s full version has a capacity of 175 billion machine learning parameters ... Before the release of GPT-3, the largest language model was Microsoft’s Turing NLG, introduced in February 2020, with a capacity of 17 billion parameters.” Wikipedia doesn’t state this, but I don’t believe there are publicly known AI models larger than these language models (with the exception of “[mixture-of-experts models](#)” that I think we should disregard for these purposes, for reasons I won’t go into here). [Wikipedia estimates](#) about 1 trillion synapses for a house mouse’s brain; Bio Anchors’s methodology for brain comparisons (based on [Brain Computation](#)) essentially equates synapses to parameters.

¹⁰⁰ Bio Anchors estimates about 100 trillion parameters for the human brain, based on the fact that it has about 100 trillion synapses.

You can roughly think of the Bio Anchors framework as saying:

- There are some tasks that a human can do with only a second of thought, such as classifying an image as a cat or dog.
- There are other tasks that might take a human several minutes of thought, such as solving a logic puzzle.
- Other tasks could take hours, days, months or even years, and require not just thinking, but interacting with the environment. For example, writing a scientific paper.
- The tasks on the longer end of this spectrum will be more costly to try/watch, so it will be more costly to train an AI model to do them. For example, it's more costly (takes more time, and more money) to have a million "tries" at a task that takes an hour than it is to have a million "tries" at a task that takes a second.
- However, the framework isn't as simple as this sounds. Many tasks that seem like "long" tasks (such as writing an essay) could in fact be broken into a series of "shorter" tasks (such as writing individual sentences).
 - If an AI model can be trained to do a shorter "sub-task," it might be able to do the longer task by simply repeating the shorter sub-task over and over again - without ever needing to be explicitly "trained" to do the longer task.
 - For example, an AI model might get a million "tries" at the task: "Read a partly-finished essay and write a good next sentence." If it then learns to do this task well, it could potentially write a long essay by simply repeating this task over and over again. It wouldn't need to go into a separate training process where it gets a million "tries" at the more time-consuming task of writing an entire essay.
 - So it becomes crucial whether the hardest and most important tasks (such as those listed above) are the kind that can be "decomposed" into short/easy tasks.

Estimating the expense

Bio Anchors looks at how expensive existing AI models were to train, depending on model size and task type (as defined above). It then extrapolates this to see how expensive an AI model would be to train if it:

- Had a size 10x larger than a human brain.¹⁰¹
- Trained on a task where each “try” took days, weeks, or months of intensive “thinking.”

As of today, this sort of training would cost in the ballpark of a million trillion dollars, which is enormously more than total world wealth. So it isn’t surprising that nobody has tried to train such a model.

However, Bio Anchors also projects the following trends out into the future:

- Advances in both hardware and software that could make computing power cheaper.
- A growing economy, and a growing role of AI in the economy, that could increase the amount AI labs are able to spend training large models to \$1 trillion and beyond.

According to these projections, at some point the “amount AI labs are able to spend” becomes equal to the “expense of training a human-brain-sized model on the hardest tasks.” Bio Anchors bases its projections for “when transformative AI will be developed” on when this happens.

Bio Anchors also models uncertainty in all of the parameters above, and considers alternative approaches to the “model size and task type” parameters.¹⁰² By doing this, it estimates the probability that transformative AI will be developed by 2030, 2035, etc.

¹⁰¹ As noted above, the 10x is to leave some space for the idea that “digital brains” might be less efficient than human brains. See [this section](#) of the report.

¹⁰² For example, one approach hypothesizes that training could be made cheaper by “meta-learning,” discussed above; another approach hypothesizes that in order to produce transformative AI, one would need to do about as many computations as all animals in history combined, in order to re-create the progress that was made by natural selection.)

Aggressive or conservative?

Bio Anchors involves a number of simplifications that could cause it to be too aggressive (expecting transformative AI to come sooner than is realistic) or too conservative (expecting it to come later than is realistic).

The argument I most commonly hear that it is “**too aggressive**” is along the lines of: “There’s no reason to think that a modern-methods-based AI can learn everything a human does, using trial-and-error training - no matter how big the model is and how much training it does. Human brains can reason in unique ways, unmatched and unmatchable by any AI unless we come up with fundamentally new approaches to AI.” This kind of argument is often accompanied by saying that AI systems don’t “truly understand” what they’re reasoning about, and/or that they are merely imitating human reasoning through pattern recognition.

I think this may turn out to be correct, but I wouldn’t bet on it. A full discussion of why is outside the scope of this post, but in brief:

- I am unconvinced that there is a deep or stable distinction between “pattern recognition” and “true understanding” ([this Slate Star Codex piece](#) makes this point). “True understanding” might just be what really good pattern recognition looks like. Part of my thinking here is an intuition that even when people (including myself) superficially appear to “understand” something, their reasoning often (I’d even say usually) breaks down when considering an unfamiliar context. In other words, I think what we think of as “true understanding” is more of an ideal than a reality.
- I feel underwhelmed with the track record of those who have made this sort of argument - I don’t feel they have been able to pinpoint what “true reasoning” looks like, such that they could make robust predictions about what would prove difficult for AI systems. (For example, see [this discussion of Gary Marcus’s latest critique of GPT3](#)).
- “Some breakthroughs / fundamental advances are needed” might be true. But for Bio Anchors to be overly aggressive, it isn’t enough that *some* breakthroughs are needed; the breakthroughs needed have to be *more than what AI scientists are capable of in the coming decades*, the time frame over which Bio Anchors forecasts transformative AI. It seems hard to be confident that things will play out this way - especially because:

- Even moderate advances in AI systems could bring more talent and funding into the field (as is already happening¹⁰³).
- If money, talent and processing power are plentiful, and progress toward PASTA is primarily held up by some particular weakness of how AI systems are designed and trained, a sustained attempt by researchers to fix this weakness could work. When we're talking about multi-decade timelines, that might be plenty of time for researchers to find whatever is missing from today's techniques.

More broadly, Bio Anchors could be too aggressive due to its assumption that “computing power is the bottleneck”:

- It assumes that *if* one could pay for all the computing power to do the brute-force “training” described above for the key tasks (e.g., automating scientific work), transformative AI would (likely) follow.
- Training an AI model doesn't just require purchasing computing power. It requires hiring researchers, running experiments, and perhaps most importantly, finding a way to set up the “trial and error” process so that the AI can get a huge number of “tries” at the key task. It may turn out that doing so is prohibitively difficult.

On the other hand, there are several ways in which Bio Anchors could be **too conservative** (underestimating the likelihood of transformative AI being developed soon).

- Perhaps with enough ingenuity, one could create a transformative AI by “programming” it to do key tasks, rather than having to “train” it (see [above](#) for the distinction). This could require far less computation, and hence be far less expensive. Or one could use a combination of “programming” and “training” to achieve better efficiency than Bio Anchors implies, while still not needing to capture everything via “programming.”
- Or one could find far superior approaches to AI that can be “trained” much more efficiently. One possibility here is “meta-learning”: effectively training an AI system on the “task” of being trained, itself.
- Or perhaps most likely, over time AI might become a bigger and bigger part of the economy, and there could be a proliferation of different AI

¹⁰³ See charts from the early sections of the [2021 AI Index Report](#), for example.

systems that have each been customized and invested in to do different real-world tasks. The more this happens, the more opportunity there is for individual ingenuity and luck to result in more innovations, and more capable AI systems in particular economic contexts.

- Perhaps at some point, it will be possible to integrate many systems with different abilities in order to tackle some particularly difficult task like “automating science,” without needing a dedicated astronomically expensive “training run.”
- Or perhaps AI that falls short of PASTA will still be useful enough to generate a lot of cash, and/or help researchers make compute cheaper and more efficient. This in turn could lead to still bigger AI models that further increase availability of cash and efficiency of compute. That, in turn, could cause a PASTA-level training run to be affordable earlier than Bio Anchors projects.
- Additionally, some technical reviewers of Bio Anchors feel that its treatment of **task type** is too conservative. They believe that the most important tasks (and perhaps all tasks) that AI needs to be trained on will be on the “easier/cheaper” end of the spectrum, compared to what Bio Anchors assumes. (See the **above section** for what it means for a task to be “easier/cheaper” or “harder/more expensive”). For a related argument, see **Fun with +12 OOMs of Compute**, which makes the intuitive point that Bio Anchors is imagining a truly massive amount of computation needed to create PASTA, and less could easily be enough.

I don’t think it is obvious whether, overall, Bio Anchors is too aggressive (expecting transformative AI to come sooner than is realistic) or too conservative (expecting it to come later). The report itself states that it’s likely to be too aggressive over the next few years and too conservative >50 years out, and likely most useful in between.¹⁰⁴

Intellectually, it feels to me as though the report is more likely to be too conservative. I find its **responses** to the “Too aggressive” points above fairly compelling, and I think the “Too conservative” points are more likely to end up being correct. In particular, I think it’s hard to rule out the possibility of ingenuity leading to transformative AI in some far more efficient way than the

¹⁰⁴ See [this section](#).

“brute-force” method contemplated here. And I think the treatment of “task type” is definitely erring in a conservative direction.

However, I also have an intuitive preference (which is related to the “burden of proof” analyses given [previously](#)) to err on the conservative side when making estimates like this. Overall, my best guesses about transformative AI timelines are similar to those of Bio Anchors.

Conclusions of Bio Anchors

Bio Anchors estimates a **>10% chance of transformative AI by 2036, a ~50% chance by 2055, and an ~80% chance by 2100.**

It’s also worth noting what the report says about AI systems today. It estimates that:

- Today’s largest AI models, such as [GPT-3](#), are a **bit smaller than mouse brains, and are starting to get within range (if they were to grow another 100x-1000x) of human brains.** So we might soon be getting close to AI systems that can be trained to do anything that humans can do with ~1 second of thought. Consistent with this, it seems to me that we’re just starting to reach the point where language models *sound* like humans who are talking without thinking very hard.¹⁰⁵ If anything, “human who puts in no more than 1 second of thought per word” seems somewhat close to what GPT-3 is doing, even though it’s much smaller than a human brain.
- It’s only very recently that AI models have gotten this big. A “large” AI model before 2020 would be more in the range of a honeybee brain. So for models even in the very recent past, we should be asking whether AI systems seem to be “as smart as insects.” Here’s [one attempt to compare AI and honeybee capabilities](#) (by Open Philanthropy intern Guille Costa), concluding that the most impressive honeybee capabilities the author was able to pinpoint do appear to be doable for AI systems.¹⁰⁶

¹⁰⁵ For a collection of links to GPT-3 demos, see [this post](#).

¹⁰⁶ In fact, he estimates that AI systems appear to use about 1000x less compute, which would match the above point in terms of suggesting that AI systems might be more efficient than animal/human brains and that the Bio Anchors estimates might be too conservative. However, he doesn’t address the fact that bees arguably perform a more diverse set of tasks than the AI systems they’re being compared to.

I include these notes because:

- The Bio Anchors analysis seems fully consistent with what we’re observing from AI systems today (and have over the last decade or two), while also implying that we’re likely to see more transformative abilities in the coming decades.
- I think it’s particularly noteworthy that we’re getting close to the time when an AI model is “as big as a human brain” (according to the Bio Anchors / [Brain Computation](#) estimation method). It may turn out that such an AI model is able to “learn” a lot about the world and produce a lot of economic value, even if it can’t yet do the hardest things humans do. And this, in turn, could kick off skyrocketing investment in AI (both money and talent), leading to a lot more innovation and further breakthroughs. This is a simple reason to believe that transformative AI by 2036 is plausible.

Finally, I note that Bio Anchors includes an “evolution” analysis among the different approaches it considers. This analysis hypothesizes that in order to produce transformative AI, one would need to do about as many computations as all animals in history combined, in order to re-create the progress that was made by natural selection.

I consider the “evolution” analysis to be *very* conservative, because machine learning is capable of much faster progress than the sort of trial-and-error associated with natural selection. Even if one believes in something along the lines of “Human brains reason in unique ways, unmatched and unmatchable by a modern-day AI,” it seems that whatever is unique about human brains should be re-discoverable if one is able to essentially re-run the whole history of natural selection. And even this very conservative analysis estimates a ~50% chance of transformative AI by 2100.

Pros and cons of the biological anchors method for forecasting transformative AI timelines

Cons. I'll start with what I see as the biggest downside: this is a very complex forecasting framework, which relies crucially on multiple extremely uncertain estimates and assumptions, particularly:

- Whether it's reasonable to believe that an AI system could learn the key tasks listed above (the ones required for PASTA) given enough trial-and-error training.
- How to compare the size of AI models with the size of animal/human brains.
- How to characterize “task type,” estimating how “difficult” and expensive a task is to “try” or “watch” once.
- How to use the model size and task type to estimate how expensive it would be to train an AI model to do the key tasks.
- How to estimate future advances in both hardware and software that could make computing power cheaper.
- How to estimate future increases in how much AI labs could be able to spend training models.

This kind of complexity and uncertainty means (IMO) that we shouldn't consider the forecasts to be highly reliable, especially today when the whole framework is fairly new. If we got to the point where as much scrutiny and effort had gone into AI forecasting as climate forecasting, it might be a different matter.

Pros. That said, the biological anchors method is essentially the only one I know of that estimates transformative AI timelines from **objective facts** (where possible) and **explicit assumptions** (elsewhere).¹⁰⁷ It does not rely on any concepts as vague and intuitive as “how fast AI systems are getting more impressive” (discussed [previously](#)). Every assumption and estimate in the framework can be explained, discussed, and - over time - tested.

¹⁰⁷ Other than the “semi-informative priors” method discussed [previously](#).

Even in its current early stage, I consider this a valuable property of the biological anchors framework. It means that the framework can give us timelines estimates that aren't simply rehashes of intuitions about whether it feels as though transformative AI is approaching.¹⁰⁸

I also think it's encouraging that even with all the guesswork, the testable "predictions" the framework makes as of today seem reasonable (see previous section). **The framework provides a way of thinking about how it could be simultaneously true that (a) the AI systems of a decade ago didn't seem very impressive at all; (b) the AI systems of today can do many impressive things but still feel far short of what humans are able to do; (c) the next few decades - or even the next 15 years - could easily see the development of transformative AI.**

Additionally, I think it's worth noting a **couple of high-level points** from Bio Anchors that **don't depend on quite so many estimates and assumptions:**

- In the coming decade or so, we're likely to see - for the first time - AI models with comparable "size" to the human brain.
- If AI models continue to become larger and more efficient at the rates that Bio Anchors estimates, it will probably become **affordable this century to hit some pretty extreme milestones - the "high end" of what Bio Anchors thinks might be necessary.** These are hard to summarize, but see the "long horizon neural net" and "evolution anchor" frameworks in the report.
- One way of thinking about this is that the next century will likely see us go from "not enough compute to run a human-sized model at all" to "extremely plentiful compute, as much as even quite conservative estimates of what we might need." Compute isn't the only factor in AI progress, but to the extent other factors (algorithms, training processes) became the new bottlenecks, there will likely be powerful incentives (and multiple decades) to resolve them.

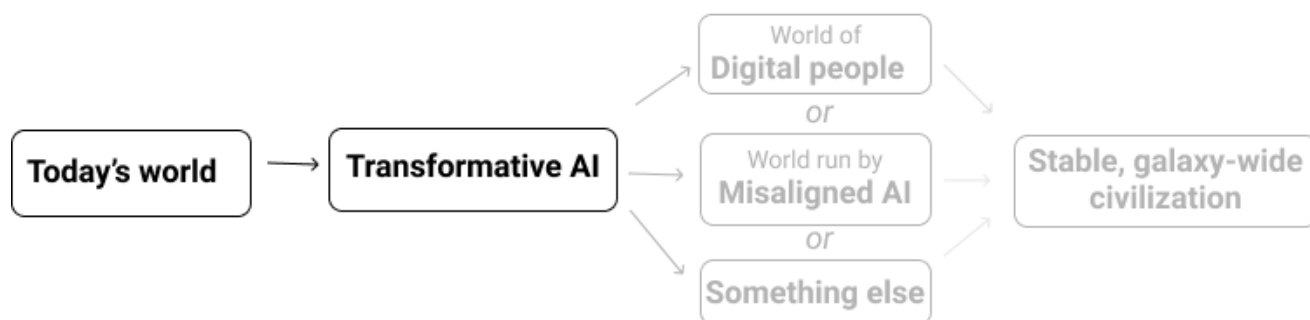
¹⁰⁸ Of course, this isn't to say the estimates are *completely independent* of intuitions - intuitions are likely to color our choices of estimates for many of the difficult-to-estimate figures. But the ability to scrutinize and debate each estimate separately is helpful here.

A final advantage of Bio Anchors is that we can continue to watch AI progress over time, and compare what we see to the report's framework. For example, we can watch for:

- Whether there are some tasks that just can't be learned, even with plenty of trial and error - or whether some tasks require amounts of training very different from what the report estimates.
- How AI models' capabilities compare to those of animals that we are currently modeling as "similarly sized." If AI models seem more capable than such animals, we may be overestimating how large a model we would need to be in order to e.g. automate science. If they seem less capable, we may be underestimating it.
- How hardware and software are progressing, and whether AI models are getting bigger at the rate the report currently projects.

The next piece will summarize all of the different analyses so far about transformative AI timelines. It will then discuss a remaining reservation: that there is no robust expert consensus on this topic.

AI Timelines: Where The Arguments, And The “Experts,” Stand



This piece starts with a summary of when we should expect transformative AI to be developed, based on the multiple angles covered previously in the series. I think this is useful, even if you've read all of the previous pieces, but if you'd like to skip it, click [here](#).

I then address the question: “Why isn’t there a robust expert consensus on this topic, and what does that mean for us?”

I estimate that there is **more than a 10% chance we’ll see transformative AI within 15 years (by 2036); a ~50% chance we’ll see it within 40 years (by 2060); and a ~2/3 chance we’ll see it this century (by 2100).**

(By “transformative AI,” I mean “AI powerful enough to bring us into a new, qualitatively different future.” I've argued that advanced AI [could](#) be sufficient to make this the [most important century](#).)

This is my overall conclusion based on a number of technical reports approaching AI forecasting from different angles - many of them produced by [Open Philanthropy](#) over the past few years as we’ve tried to develop a thorough picture of transformative AI forecasting to inform our longtermist grantmaking.

Here's a **one-table summary** of the different angles on forecasting transformative AI that I've discussed, with links to more detailed discussion in [previous posts](#) as well as to underlying technical reports:

Forecasting angle	Key in-depth pieces (abbreviated titles)	My takeaways
<i>Probability estimates for transformative AI</i>		
Expert survey . What do AI researchers expect?	Evidence from AI Experts	Expert survey implies ¹⁰⁹ a ~20% probability by 2036; ~50% probability by 2060; ~70% probability by 2100. Slightly differently phrased questions (posed to a minority of respondents) have much later estimates.
Biological anchors framework . Based on the usual patterns in how much “AI training” costs, how much would it cost to train an AI model as big as a human brain to perform the hardest tasks humans do? And when will this be cheap enough that we can expect someone to do it?	Bio Anchors , drawing on Brain Computation	>10% probability by 2036; ~50% chance by 2055; ~80% chance by 2100.
<i>Angles on the burden of proof</i>		
It's unlikely that any given century would be the “most important” one. (More)	Hinge ; Response to Hinge	We have many reasons to think this century is a “special” one before looking at the details of AI. Many have been covered in previous pieces; another is covered in the next row.

¹⁰⁹ Technically, these probabilities are for “human-level machine intelligence.” In general, this chart simplifies matters by presenting one unified set of probabilities. In general, all of these probabilities refer to something at *least* as capable as [PASTA](#), so they directionally should be underestimates of the probability of PASTA (though I don’t think this is a major issue).

<p>What would you forecast about transformative AI timelines, based only on basic information about (a) how many years people have been trying to build transformative AI; (b) how much they've “invested” in it (in terms of the number of AI researchers and the amount of computation used by them); (c) whether they've done it yet (so far, they haven't)? (More)</p>	<p>Semi-informative Priors</p>	<p>Central estimates: 8% by 2036; 13% by 2060; 20% by 2100.¹¹⁰ In my view, this report highlights that the history of AI is short, investment in AI is increasing rapidly, and so we shouldn't be too surprised if transformative AI is developed soon.</p>
<p>Based on analysis of economic models and economic history, how likely is 'explosive growth' - defined as >30% annual growth in the world economy - by 2100? Is this far enough outside of what's “normal” that we should doubt the conclusion? (More)</p>	<p>Explosive Growth, Human Trajectory</p>	<p>Human Trajectory projects the past forward, implying explosive growth by 2043-2065.</p> <p>Explosive Growth concludes: “I find that economic considerations don't provide a good reason to dismiss the possibility of TAI being developed in this century. In fact, there is a plausible economic perspective from which sufficiently advanced AI systems are expected to cause explosive growth.”</p>

¹¹⁰ Reviews of Bio Anchors are [here](#); reviews of Explosive Growth are [here](#); reviews of Semi-informative Priors are [here](#). Brain Computation was reviewed at an earlier time when we hadn't designed the process to result in publishing reviews, but over 20 conversations with experts that informed the report are available [here](#). Human Trajectory hasn't been reviewed, although a lot of its analysis and conclusions feature in Explosive Growth, which has been.

<p>“How have people predicted AI ... in the past, and should we adjust our own views today to correct for patterns we can observe in earlier predictions? ... We’ve encountered the view that AI has been prone to repeated over-hype in the past, and that we should therefore expect that today’s projections are likely to be over-optimistic.” (More)</p>	<p>Past AI Forecasts</p>	<p>“The peak of AI hype seems to have been from 1956-1973. Still, the hype implied by some of the best-known AI predictions from this period is commonly exaggerated.”</p>
---	--	--

For transparency, note that many of the technical reports are [Open Philanthropy](#) analyses, and I am co-CEO of Open Philanthropy.

Having considered the above, I expect some readers to still feel a sense of unease. Even if they think my arguments make sense, they may be wondering: **if this is true, why isn’t it more widely discussed and accepted? What’s the state of expert opinion?**

My summary of the state of expert opinion at this time is:

- The claims I'm making do not *contradict* any particular expert consensus. (In fact, the probabilities I've given aren't too far off from what AI researchers seem to predict, as shown in the first row.) But there are some [signs they aren't thinking too hard about the matter](#).
- The Open Philanthropy technical reports I've relied on have had significant external expert review. Machine learning researchers reviewed [Bio Anchors](#); neuroscientists reviewed [Brain Computation](#); economists reviewed [Explosive Growth](#); academics focused on relevant topics in uncertainty and/or probability reviewed [Semi-informative Priors](#).¹⁰⁹ (Some of these reviews had significant points of disagreement, but none of these points seemed to be cases where the reports contradicted a clear consensus of experts or literature.)
- But there is also no active, robust expert consensus supporting claims like “*There’s at least a 10% chance of transformative AI by 2036*” or “*There’s a good chance we’re in the most important century for humanity*,” the way that there is supporting e.g. the need to take action against climate change.

Ultimately, my claims are about **topics that simply have no “field” of experts devoted to studying them. That, in and of itself, is a scary fact**, and something that I hope will eventually change.

But should we be willing to act on the “most important century” hypothesis in the meantime?

Below, I’ll discuss:

- What an “AI forecasting field” might look like.
- A “skeptical view” that says today’s discussions around these topics are too small, homogeneous and insular (which I agree with) - and that we therefore shouldn’t act on the **“most important century” hypothesis** until there is a mature, robust field (which I don’t).
- Why I think we should take the hypothesis seriously in the meantime, until and unless such a field develops:
 - We don’t have time to wait for a robust expert consensus.
 - If there are good rebuttals out there - or potential future experts who could develop such rebuttals - we haven’t found them yet. The more seriously the hypothesis gets taken, the more likely such rebuttals are to appear. (Aka the **Cunningham’s Law** theory: “the best way to get a right answer is to post a wrong answer.”)
 - I think that consistently insisting on a robust expert consensus is a dangerous reasoning pattern. In my view, it’s OK to be at some risk of self-delusion and insularity, in exchange for doing the right thing when it counts most.

What kind of expertise is AI forecasting expertise?

Questions analyzed in the technical reports listed **above** include:

- Are AI capabilities getting more impressive over time? (AI, history of AI)
- How can we compare AI models to animal/human brains? (AI, neuroscience)
- How can we compare AI capabilities to animals’ capabilities? (AI, ethology)

- How can we estimate the expense of training a large AI system for a difficult task, based on information we have about training past AI systems? (AI, curve-fitting)
- How can we make a minimal-information estimate about transformative AI, based only on how many years/researchers/dollars have gone into the field so far? (Philosophy, probability)
- How likely is explosive economic growth this century, based on theory and historical trends? (Growth economics, economic history)
- What has “AI hype” been like in the past? (History)

When talking about wider implications of transformative AI for the “most important century,” I've also discussed things like “How feasible are [digital people](#) and [establishing space settlements throughout the galaxy](#)?” These topics touch physics, neuroscience, engineering, philosophy of mind, and more.

There’s no obvious job or credential that makes someone an expert on the question of when we can expect transformative AI, or the question of whether we’re in the most important century.

(I particularly would disagree with any claim that we should be relying exclusively on AI researchers for these forecasts. In addition to the fact that [they don't seem to be thinking very hard about the topic](#), I think that relying on people who specialize in building ever-more powerful AI models to tell us when transformative AI might come is like relying on solar energy R&D companies - or oil extraction companies, depending on how you look at it - to forecast carbon emissions and climate change. They certainly have part of the picture. But forecasting is a distinct activity from innovating or building state-of-the-art systems.)

And I’m not even sure these questions have the right shape for an academic field. Trying to forecast transformative AI, or determine the odds that we’re in the most important century, seems:

- More similar to the [FiveThirtyEight election model](#) (“Who’s going to win the election?”) than to academic political science (“How do governments and constituents interact?”);

- More similar to trading financial markets (“Is this price going up or down in the future?”) than to academic economics (“Why do recessions exist?”);¹¹¹
- More similar to [GiveWell’s](#) research (“Which charity will help people the most, per dollar?”) than to academic development economics (“What causes poverty and what can reduce it?”)¹¹²

That is, it’s not clear to me what a natural “institutional home” for expertise on transformative AI forecasting, and the “most important century,” would look like. But it seems fair to say there aren’t large, robust institutions dedicated to this sort of question today.

How should we act in the absence of a robust expert consensus?

The skeptical view

Lacking a robust expert consensus, I expect some (really, most) people will be skeptical no matter what arguments are presented.

Here’s a version of a very general skeptical reaction I have a fair amount of empathy for:

1. *This is all just too [wild](#).*
2. *You’re making an over-the-top claim about living in the most important century. This **pattern-matches to self-delusion**.*
3. *You’ve argued that the [burden of proof](#) shouldn’t be so high, because there are lots of ways in which we live in a [remarkable](#) and [unstable](#) time. But ... I don’t trust myself to assess those claims, or your claims about AI, or really anything on these wild topics.*
4. *I’m worried by how few people seem to be engaging these arguments.*

¹¹¹ The academic fields are quite broad, and I’m just giving example questions that they tackle.

¹¹² Though climate science is an example of an academic field that invests a lot in forecasting the future.

*About how **small, homogeneous and insular** the discussion seems to be. Overall, this feels more like a story smart people are telling themselves - with lots of charts and numbers to rationalize it - about their place in history. It doesn't feel "real."*

- 5. So call me back when there's a mature field of perhaps hundreds or thousands of experts, critiquing and assessing each other, and they've reached the same sort of consensus that we see for climate change.*

I see how you could feel this way, and I've felt this way myself at times - especially on points #1-#4. But I'll give **three reasons that point #5 doesn't seem right**.

Reason 1: we don't have time to wait for a robust expert consensus

I worry that the arrival of transformative AI could play out as a kind of slow-motion, higher-stakes version of the COVID-19 pandemic. The case for expecting something big to happen is there, if you look at the best information and analyses available today. But the situation is broadly unfamiliar; it doesn't fit into patterns that our institutions regularly handle. And every extra year of action is valuable.

You could also think of it as a sped-up version of the dynamic with climate change. Imagine if greenhouse gas emissions had only started to rise recently¹¹³ (instead of in the [mid-1800s](#)), and if there were no established field of climate science. It would be a really bad idea to wait decades for a field to emerge, before seeking to reduce emissions.

Reason 2: [Cunningham's Law](#) ("the best way to get a right answer is to post a wrong answer") may be our best hope for finding the flaw in these arguments

I'm serious, though.

¹¹³ The field of AI has existed since [1956](#), but it's only in the last decade or so that machine learning models have started to get within range of [the size of insect brains](#) and perform well on relatively difficult tasks.

Several years ago, some [colleagues](#) and I suspected that the “most important century” hypothesis could be true. But before acting on it too much, we wanted to see whether we could find fatal flaws in it.

One way of interpreting our actions over the last few years is **as if we were doing everything we could to learn that the hypothesis is wrong.**

First, we tried talking to people about the key arguments - AI researchers, economists, etc. But:

- We had vague ideas of the arguments in this series (mostly or perhaps entirely picked up [from other people](#)). We weren't able to state them with good crispness and specificity.
- There were a lot of key factual points that we thought would probably check out,¹¹⁴ but hadn't nailed down and couldn't present for critique.
- Overall, we couldn't even really articulate enough of a concrete case to give the others a fair chance to shoot it down.

So we put a lot of work into creating technical reports on many of the key arguments. (These are now public, and included in the table at the top of this piece.) This put us in position to publish the arguments, and potentially encounter fatal counterarguments.

Then, we commissioned external expert reviews.¹¹⁵

Speaking only for my own views, the “most important century” hypothesis seems to have survived all of this. Indeed, having examined the many angles and gotten more into the details, I believe it more strongly than before.

But let's say that this is just because the *real* experts - people we haven't found yet, with devastating counterarguments - find the whole thing so silly that they're [not bothering to engage](#). Or, let's say that there are people out

¹¹⁴ Often, we were simply going off of our impressions of what others who had thought about the topic a lot thought.

¹¹⁵ Reviews of Bio Anchors are [here](#); reviews of Explosive Growth are [here](#); reviews of Semi-informative Priors are [here](#). Brain Computation was reviewed at an earlier time when we hadn't designed the process to result in publishing reviews, but over 20 conversations with experts that informed the report are available [here](#). Human Trajectory hasn't been reviewed, although a lot of its analysis and conclusions feature in Explosive Growth, which has been.

there today who could *someday* become experts on these topics, and knock these arguments down. What could we do to bring this about?

The best answer I've come up with is: "If this hypothesis became better-known, more widely accepted, and more influential, it would get more critical scrutiny."

This series is an attempted step in that direction - to move toward broader credibility for the "most important century" hypothesis. This would be a good thing if the hypothesis were true; it also seems like the best next step if my only goal were to challenge my beliefs and learn that it is false.

Of course, I'm not saying to accept or promote the "most important century" hypothesis if it doesn't seem correct to you. But I think that if your *only* reservation is about the lack of robust consensus, continuing to ignore the situation seems odd. If people behaved this way generally (ignoring any hypothesis not backed by a robust consensus), I'm not sure I see how any hypothesis - including true ones - would go from fringe to accepted.

Reason 3: skepticism this general seems like a bad idea

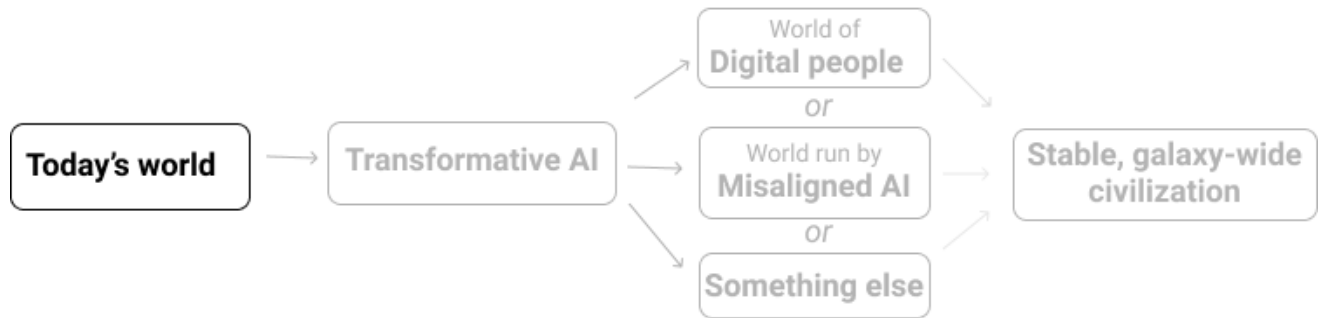
Back when I was focused on [GiveWell](#), people would occasionally say something along the lines of: "You know, you can't hold every argument to the standard that GiveWell holds its top charities to - seeking randomized controlled trials, robust empirical data, etc. Some of the best opportunities to do good will be the ones that are less obvious - so this standard risks [ruling out some of your biggest potential opportunities to have impact.](#)"

I think this is right. I think it's important to check one's general approach to reasoning and evidentiary standards and ask: "What are some scenarios in which my approach fails, and in which I'd really prefer that it succeed?" In my view, **it's OK to be at some risk of self-delusion and insularity, in exchange for doing the right thing when it counts most.**

I think the lack of a robust expert consensus - and concerns about self-delusion and insularity - provide good reason to *dig hard* on the "most important century" hypothesis, rather than accepting it immediately. To ask where there might be an undiscovered flaw, to look for some bias toward inflating our own importance, to research the most questionable-seeming parts of the argument, etc.

But if you've investigated the matter as much as is reasonable/practical for you - and haven't found a flaw *other* than considerations like "There's no robust expert consensus" and "I'm worried about self-delusion and insularity" - then I think writing off the hypothesis is the sort of thing that essentially **guarantees you won't be among the earlier people to notice and act on a tremendously important issue, if the opportunity arises.** I think that's too much of a sacrifice, in terms of giving up potential opportunities to do a lot of good.

How To Make The Best Of The Most Important Century?



Previously in the [“most important century” series](#), I’ve argued that there’s a high probability¹¹⁶ that the coming decades will see:

- The development of a technology like [PASTA](#) (process for automating scientific and technological advancement).
- A resulting [productivity explosion](#) leading to development of further transformative technologies.
- The seed of a [stable galaxy-wide civilization](#), possibly featuring [digital people](#), or possibly run by [misaligned AI](#).

Is this an optimistic view of the world, or a pessimistic one? To me, it’s both and neither, because **this set of events could end up being very good or very bad for the world, depending on the details of how it plays out.**

When I talk about being in the “most important century,” I don’t just mean that significant events are going to occur. I mean that we, the people living in this century, have the chance to have a huge impact on huge numbers of

¹¹⁶ From [Forecasting Transformative AI: What’s the Burden of Proof?](#): “I am forecasting more than a 10% chance transformative AI will be developed within 15 years (by 2036); a ~50% chance it will be developed within 40 years (by 2060); and a ~2/3 chance it will be developed this century (by 2100).”

Also see [Some additional detail on what I mean by “most important century.”](#)

people to come - if we can make sense of the situation enough to find helpful actions.

But it's also important to understand why that's a big "if" - why the most important century presents a **challenging strategic picture, such that many things we can do might make things better or worse (and it's hard to say which)**.

In this post, I will **present two contrasting frames for how to make the best of the most important century**:

- The “**Caution**” frame. In this frame, many of the worst outcomes come from developing something like [PASTA](#) in a way that is too fast, rushed, or reckless. We may need to achieve (possibly global) coordination in order to mitigate pressures to race, and take appropriate care. ([Caution](#))
- The “**Competition**” frame. This frame focuses not on *how and when* [PASTA](#) is developed, but *who* (which governments, which companies, etc.) is first in line to benefit from the resulting productivity explosion. ([Competition](#))
- People who take the “caution” frame and people who take the “competition” frame often favor **very different, even contradictory** actions. Actions that look important to people in one frame often look actively harmful to people in the other.
 - I worry that the “competition” frame will be overrated by default, and discuss why below. ([More](#))
 - To gain more clarity on how to weigh these frames and what actions are most likely to be helpful, we need more progress on **open questions** about the size of different types of risks from transformative AI. ([Open questions](#))
- In the meantime, there are some **robustly helpful actions** that seem likely to improve humanity's prospects regardless. ([Robustly helpful actions](#))

The “caution” frame

I've argued for a good chance that this century will see a transition to a world where [digital people](#) or [misaligned AI](#) (or something else very different from today's humans) are the major force in world events.

The “caution” frame emphasizes that **some types of transition seem better than others**. Listed in order from worst to best:

Worst: Misaligned AI

I discussed this possibility [previously](#), drawing on a number of other and more thorough discussions.¹¹⁷ The basic idea is that AI systems could end up with objectives of their own, and could seek to expand throughout space fulfilling these objectives. Humans, and/or all humans value, could be sidelined (or driven extinct, if we'd otherwise get in the way).

Next-worst:¹¹⁸ Adversarial Technological Maturity

If we get to the point where there are digital people and/or (non-misaligned) AIs that can copy themselves without limit, and expand throughout space, there might be intense pressure to move - and multiply (via copying) - as fast as possible in order to gain more influence over the world. This might lead to different countries/coalitions furiously trying to outpace each other, and/or to outright military conflict, knowing that a lot could be at stake in a short time.

I would expect this sort of dynamic to risk a lot of the galaxy ending up in a bad state.¹¹⁹

¹¹⁷ These include the books [Superintelligence](#), [Human Compatible](#), [Life 3.0](#), and [The Alignment Problem](#). The shortest, most accessible presentation I know of is [The case for taking AI seriously as a threat to humanity](#) (Vox article by Kelsey Piper). This [report on existential risk from power-seeking AI](#), by Open Philanthropy's Joe Carlsmith, lays out a detailed set of premises that would collectively imply the problem is a serious one.

¹¹⁸ The order of goodness isn't absolute, of course. There are versions of “Adversarial Technological Maturity” that could be worse than “Misaligned AI” - for example, if the former results in power going to those who deliberately inflict suffering.

¹¹⁹ Part of the reason for this is that faster-moving, less-careful parties could end up quickly outnumbering others and determining the future of the galaxy. There is also a longer-run risk discussed in Nick

One such bad state would be “permanently under the control of a single (digital) person (and/or their copies).” Due to the potential of digital people to create [stable civilizations](#), it seems that a given totalitarian regime could end up permanently entrenched across substantial parts of the galaxy.

People/countries/coalitions who *suspect each other* of posing this sort of danger - of potentially establishing stable civilizations under their control - might compete and/or attack each other early on to prevent this. This could lead to war with difficult-to-predict outcomes (due to the difficult-to-predict technological advancements that PASTA could bring about).

Second-best: Negotiation and governance

Countries might prevent this sort of [Adversarial Technological Maturity](#) dynamic by planning ahead and negotiating with each other. For example, perhaps each country - or each person - could be allowed to create a certain number of digital people (subject to human rights protections and other regulations), limited to a certain region of space.

It seems there are a huge range of different potential specifics here, some much more good and just than others.

Best: Reflection

The world could achieve a high enough level of coordination to *delay* any irreversible steps (including kicking off an [Adversarial Technological Maturity](#) dynamic).

There could then be something like what Toby Ord (in [The Precipice](#)) calls the “Long Reflection”:¹²⁰ a sustained period in which people could collectively decide upon goals and hopes for the future, ideally representing the most fair available compromise between different perspectives. Advanced technology could imaginably help this go much better than it could today.¹²¹

Bostrom’s [The Future of Human Evolution](#); also see [this discussion](#) of Bostrom’s ideas on Slate Star Codex, though also see [this piece by Carl Shulman](#) arguing that this dynamic is unlikely to result in total elimination of nice things.

¹²⁰ See page 191.

¹²¹ E.g., see [this section](#) of [Digital People Would Be An Even Bigger Deal](#).

There are limitless questions about how such a “reflection” would work, and whether there’s really any hope that it could reach a reasonably good and fair outcome. Details like “what sorts of digital people are created first” could be enormously important. There is currently little discussion of this sort of topic.¹²²

Other

There are probably many possible types of transitions I haven’t named here.

The role of caution

If the above ordering is correct, then the future of the galaxy looks better to the extent that:

- **Misaligned AI** is avoided: powerful AI systems act to help humans, rather than pursuing objectives of their own.
- **Adversarial Technological Maturity** is avoided. This likely means that people do not deploy advanced AI systems, or the technologies they could bring about, in adversarial ways (unless this ends up necessary to prevent something worse).
- Enough coordination is achieved so that key players can “take their time,” and **Reflection** becomes a possibility.

Ideally, everyone with the potential to build something **PASTA**-like would be able to pour energy into building something safe (not misaligned), and carefully planning out (and negotiating with others on) how to roll it out, without a rush or a race. With this in mind, perhaps we should be doing things like:

- Working to improve trust and cooperation between major world powers. Perhaps via AI-centric versions of **Pugwash** (an international conference aimed at reducing the risk of military conflict), perhaps by pushing back against hawkish foreign relations moves.
- Discouraging governments and investors from shoveling money into AI research, encouraging AI labs to thoroughly consider the implications of

¹²² One relevant paper: [Public Policy and Superintelligent AI: A Vector Field Approach](#) by Bostrom, Dafoe and Flynn.

their research before publishing it or scaling it up, etc. Slowing things down in this manner could buy more time to do research on avoiding **mis-aligned AI**, more time to build trust and cooperation mechanisms, more time to generally gain strategic clarity, and a lower likelihood of the **Adversarial Technological Maturity** dynamic.

The “competition” frame

(Note: there's some potential for confusion between the “competition” idea and the **Adversarial Technological Maturity** idea, so I've tried to use very different terms. I spell out the contrast in a footnote.¹²³)

The “competition” frame focuses **less on how the transition to a radically different future happens, and more on who’s making the key decisions as it happens.**

- If something like **PASTA** is developed primarily (or first) in country X, then the government of country X could be making a lot of crucial decisions about whether and how to regulate a potential explosion of new technologies.
- In addition, the people and organizations leading the way on AI and other technology advancement at that time could be especially influential in such decisions.

This means it could matter enormously “who leads the way on transformative AI” - which country or countries, which people or organizations.

- Will the governments leading the way on transformative AI be authoritarian regimes?
- Which governments are most likely to (effectively) have a reasonable understanding of the risks and stakes, when making key decisions?

¹²³ **Adversarial Technological Maturity** refers to a world in which highly advanced technology has **already been developed**, likely with the help of AI, and different coalitions are vying for influence over the world. By contrast, “Competition” refers to a strategy for how to behave **before the development of advanced AI**. One might imagine a world in which some government or coalition takes a “competition” frame, develops advanced AI long before others, and then makes a series of good decisions that *prevent* Adversarial Technological Maturity. (Or conversely, a world in which failure to do well at “competition” raises the risks of Adversarial Technological Maturity.)

- Which governments are least likely to try to use advanced technology for entrenching the power and dominance of one group? (Unfortunately, I can't say there are any that I feel great about here.) Which are most likely to leave the possibility open for something like “avoiding **locked-in** outcomes, leaving time for general progress worldwide to raise the odds of a good outcome for everyone possible?”
- Similar questions apply to the people and organizations leading the way on transformative AI. Which ones are most likely to push things in a positive direction?

Some people feel that we can make confident statements today about which specific countries, and/or which people and organizations, we should hope lead the way on transformative AI. These people might advocate for actions like:

- Increasing the odds that the first PASTA systems are built in countries that are e.g. less authoritarian, which could mean e.g. pushing for more investment and attention to AI development in these countries.
- Supporting and trying to speed up AI labs run by people who are likely to make wise decisions (about things like how to engage with governments, what AI systems to publish and deploy vs. keep secret, etc.)

Why I fear “competition” being overrated, relative to “caution”

By default, I expect a lot of people to gravitate toward the “competition” frame rather than the “caution” frame - for reasons that I don't think are great, such as:

- I think people naturally get more animated about “helping the good guys beat the bad guys” than about “helping all of us avoid getting a universally bad outcome, for impersonal reasons such as ‘we designed sloppy AI systems’ or ‘we created a dynamic in which haste and aggression are rewarded.’”
- I expect people will tend to be overconfident about which countries, organizations or people they see as the “good guys.”

- Embracing the “competition” frame tends to point toward taking actions - such as working to speed up a particular country’s or organization’s AI development - that are lucrative, exciting and naturally easy to feel energy for. Embracing the “caution” frame is much less this way.
- The biggest concerns that the “caution” frame focuses on - **Misaligned AI** and **Adversarial Technological Maturity** - are a bit abstract and hard to wrap one's head around. In many ways they seem to be the highest-stakes risks, but it's easier to be viscerally scared of “falling behind countries/organizations/people that scare me” than to be viscerally scared of something like “Getting a bad outcome for the long-run future of the galaxy because we rushed things this century.”
 - I think **Misaligned AI** is a particularly hard risk for many to take seriously. It sounds wacky and sci-fi-like; people who worry about it tend to be interpreted as picturing something like The Terminator, and it can be hard for their more detailed concerns to be understood.
 - I’m hoping to run more posts in the future that help give an intuitive sense for why I think Misaligned AI is a real risk.

So for the avoidance of doubt, I'll state that I think the “caution” frame has an awful lot going for it. In particular, **Misaligned AI** and **Adversarial Technological Maturity** seem a *lot* worse than other potential transition types, and both seem like things that have a real chance of making the entire future of our species (and successors) much worse than they could be.

I worry that too much of the “competition” frame will lead to downplaying misalignment risk and rushing to deploy unsafe, unpredictable systems, which could have many negative consequences.

With that said, **I put serious weight on both frames.** I remain quite uncertain overall about which frame is more important and helpful (if either is).

Key open questions for “caution” vs. “competition”

People who take the “caution” frame and people who take the “competition” frame often favor **very different, even contradictory actions**. Actions that look important to people in one frame often look actively harmful to people in the other.

For example, people in the “competition” frame often favor moving forward as fast as possible on developing more powerful AI systems; for people in the “caution” frame, haste is one of the main things to avoid. People in the “competition” frame often favor adversarial foreign relations, while people in the “caution” frame often want foreign relations to be more cooperative.

(That said, this dichotomy is a simplification. Many people - including myself - resonate with both frames. And either frame could imply actions normally associated with the other; for example, you might take the “caution” frame but feel that haste is needed now in order to establish one country with a clear enough lead in AI that it can then take its time, prioritize avoiding [mis-aligned AI](#), etc.)

I wish I could confidently tell you how much weight to put on each frame, and what actions are most likely to be helpful. But I can't. I think we would have more clarity if we had better answers to some key open questions:

Open question: how hard is the alignment problem?

The path to the future that seems worst is [Misaligned AI](#), in which AI systems end up with non-human-compatible objectives of their own and seek to fill the galaxy according to those objectives. How seriously should we take this risk - how hard will it be to avoid this outcome? **How hard will it be to solve the “alignment problem,”** which essentially means having the technical ability to build systems that won't do this?¹²⁴

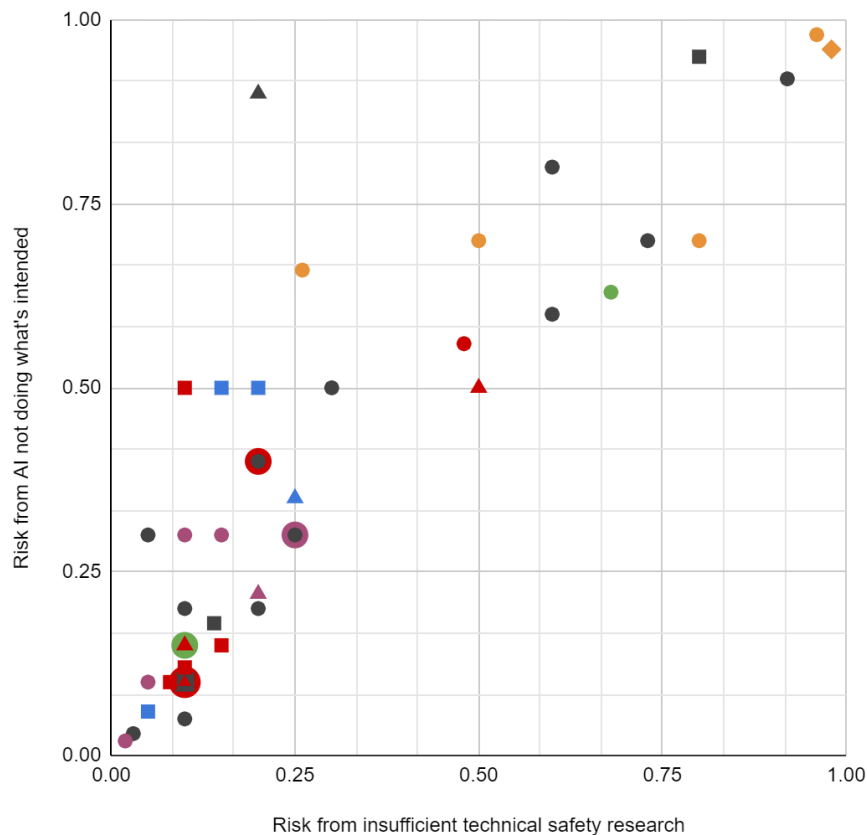
- Some people believe that the alignment problem will be formidable; that our only hope of solving it comes in a world where we have enormous amounts of time and aren't in a race to deploy advanced AI; and that avoiding the “Misaligned AI” outcome should be by far the dominant consideration for the most important century. These people tend to heavily favor the “caution” interventions described above: they believe that rushing toward AI development raises our already-substantial risk of the worst possible outcome.
- Some people believe it will be easy, and/or that the whole idea of “misaligned AI” is misguided, silly, or even incoherent - planning for an overly specific future event. These people often are more interested in the “competition” interventions described above: they believe that advanced AI

¹²⁴ See definitions of this problem at [Wikipedia](#) and [Paul Christiano's Medium](#).

will probably be used effectively by whatever country (or in some cases smaller coalition or company) develops it first, and so the question is who will develop it first.

- And many people are somewhere in between.

The spread here is extreme. For example, see [these results](#) from an informal “two-question survey [sent] to ~117 people working on long-term AI risk, asking about the level of existential risk from ‘humanity not doing enough technical AI safety research’ and from ‘AI systems not doing/optimizing what the people deploying them wanted/intended.’” (As the scatterplot shows, people gave similar answers to the two questions.)



We have respondents who think there’s a <5% chance that alignment issues will drastically reduce the goodness of the future; respondents who think there’s a >95% chance; and just about everything in between.¹²⁵ My sense is

¹²⁵ A more detailed, private survey done for [this report](#), asking about the probability of “doom” before 2070 due to the type of problem discussed in the report, got answers ranging from <1% to >50%. In my opinion, there are very thoughtful people who have seriously considered these matters at both

that this is a fair representation of the situation: even among the few people who have spent the most time thinking about these matters, there is practically no consensus or convergence on how hard the alignment problem will be.

I hope that over time, the field of people doing research on AI alignment¹²⁶ will grow, and as both AI and AI alignment research advance, we will gain clarity on the difficulty of the AI alignment problem. This, in turn, could give more clarity on prioritizing “caution” vs. “competition.”

Other open questions

Even if we had clarity on the difficulty of the alignment problem, a lot of thorny questions would remain.

Should we be expecting transformative AI within the next 10-20 years, or much later? Will the leading AI systems go from very limited to very capable quickly (“hard takeoff”) or gradually (“slow takeoff”)?¹²⁷ Should we hope that government projects play a major role in AI development, or that transformative AI primarily emerges from the private sector? Are some governments more likely than others to work toward transformative AI being used carefully, inclusively and humanely? What should we hope a government (or company) literally *does* if it gains the ability to dramatically accelerate scientific and technological advancement via AI?

With these questions and others in mind, it’s often very hard to look at some action - like starting a new AI lab, advocating for more caution and safeguards in today’s AI development, etc. - and say whether it raises the likelihood of good long-run outcomes.

Robustly helpful actions

Despite this state of uncertainty, here are a few things that do seem clearly valuable to do today:

Technical research on the alignment problem. Some researchers work on building AI systems that can get “better results” (winning more board

ends of that range.

¹²⁶ Some example technical topics [here](#).

¹²⁷ Some discussion of this topic here: [Distinguishing definitions of takeoff - AI Alignment Forum](#)

games, classifying more images correctly, etc.) But a smaller set of researchers works on things like:

- [Training AI systems to incorporate human feedback into how they perform summarization tasks](#), so that the AI systems reflect hard-to-define human preferences - something it may be important to be able to do in the future.
- [Figuring out how to understand “what AI systems are thinking and how they’re reasoning,”](#) in order to make them less mysterious.
- [Figuring out how to stop AI systems from making extremely bad judgments on images designed to fool them](#), and other work focused on helping avoid the “worst case” behaviors of AI systems.
- [Theoretical work](#) on how an AI system might be very advanced, yet not be unpredictable in the wrong ways.

This sort of work could both reduce the risk of the [Misaligned AI](#) outcome - and/or lead to more clarity on just how big a threat it is. Some takes place in academia, some at AI labs, and some at specialized organizations..

Pursuit of strategic clarity: doing research that could address other crucial questions (such as those listed [above](#)), to help clarify what sorts of immediate actions seem most useful.

Helping governments and societies become, well, nicer. Helping Country X get ahead of others on AI development could make things better or worse, for reasons given above. But it seems robustly good to work toward a Country X with better, more inclusive values, and a government whose key decision-makers are more likely to make thoughtful, good-values-driven decisions.

Spreading ideas and building communities. Today, it seems to me that the world is **extremely short on people who share certain basic expectations and concerns**, such as:

- Believing that AI research could lead to rapid, radical changes of the [extreme kind laid out here](#) (well beyond things like e.g. increasing unemployment).
- Believing that the alignment problem (discussed [above](#)) is at least plausibly a real concern, and taking [the “caution” frame](#) seriously.

- Looking at the whole situation through a lens of “Let’s get the best outcome possible for the whole world over the long future,” as opposed to more common lenses such as “Let’s try to make money” or “Let’s try to ensure that my home country leads the world in AI research.”

I think it’s very valuable for there to be more people with this basic lens, particularly working for AI labs and governments. If and when we have more strategic clarity about what actions could maximize the odds of the “most important century” going well, I expect such people to be relatively well-positioned to be helpful.

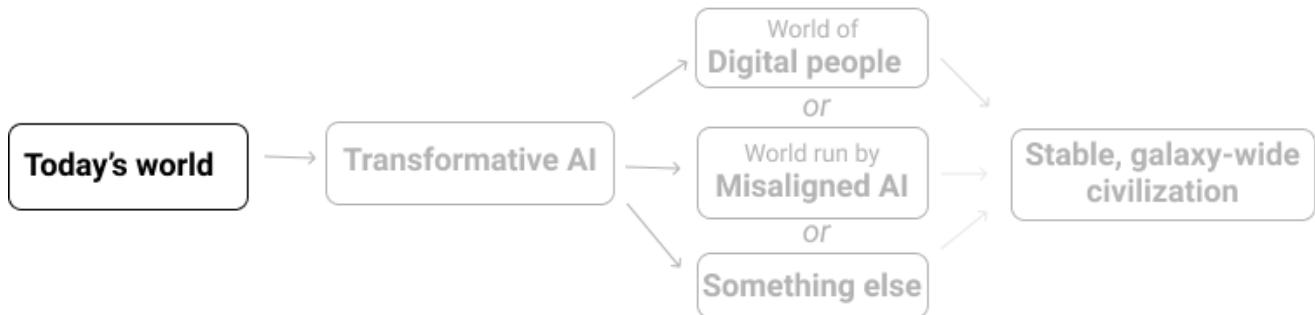
A number of organizations and people have worked to expose people to the lens above, and help them meet others who share it. I think a good amount of progress (in terms of growing communities) has come from this.

Donating? One can donate today to places like [this](#). But I need to admit that very broadly speaking, there's no easy translation right now between “money” and “improving the odds that the most important century goes well.” It's not the case that if one simply sent, say, \$1 trillion to the right place, we could all breathe easy about challenges like the alignment problem and [risks of digital dystopias](#).

It seems to me that we - as a species - are currently terribly short on people who are paying *any* attention to the most important challenges ahead of us, and haven’t done the work to have good strategic clarity about what tangible actions to take. **We can’t solve this problem by throwing money at it.**¹²⁸ **First, we need to take it more seriously and understand it better.**

¹²⁸ Some more thought on “when money isn’t enough” at [this old GiveWell post](#).

Call To Vigilance



This is the final piece in the [“most important century” series](#), which has argued that there's a [high probability](#)¹²⁹ that the coming decades will see:

- The development of a technology like [PASTA](#) (process for automating scientific and technological advancement).
- A resulting [productivity explosion](#) leading to development of further transformative technologies.
- The seed of a [stable galaxy-wide civilization](#), possibly featuring [digital people](#), or possibly run by [misaligned AI](#).

When trying to call attention to an underrated problem, **it's typical to close on a “call to action”**: a tangible, concrete action readers can take to help.

But this is challenging, because as I argued [previously](#), there are a lot of [open questions about what actions are helpful vs. harmful](#). (Although we can identify some [actions that seem robustly helpful today](#).)

This makes for a somewhat awkward situation. When confronting the “most important century” hypothesis, my attitude doesn't match the familiar ones of “excitement and motion” or “fear and avoidance.” Instead, I feel an **odd mix of intensity, urgency, confusion and hesitance**. I'm looking at something bigger than I ever expected to confront, feeling underqualified and

¹²⁹ “I am forecasting more than a 10% chance transformative AI will be developed within 15 years (by 2036); a ~50% chance it will be developed within 40 years (by 2060); and a ~2/3 chance it will be developed this century (by 2100).”

ignorant about what to do next. This is a hard mood to share and spread, but I'm trying.

Situation	Appropriate reaction (IMO)
“This could be a billion-dollar company!”	“Woohoo, let’s GO for it!”
“This could be the most important century!”	“... Oh ... wow ... I don’t know what to say and I somewhat want to vomit ... I have to sit down and think about this one.”

So instead of a call to action, I want to make a **call to vigilance**. If you're convinced by the arguments in this piece, then don't rush to “do something” and then move on. Instead, take whatever **robustly good actions** you can today, and otherwise put yourself in a better position to take important actions when the time comes.

This could mean:

- Finding ways to interact more with, and learn more about, key topics/ fields/industries such as AI (for obvious reasons), science and technology generally (as a lot of the “most important century” hypothesis runs through an **explosion in scientific and technological advancement**), and relevant areas of policy and national security.
- Taking opportunities (when you see them) to move your career in a direction that is more likely to be relevant (some thoughts of mine on this are **here**; also see **80,000 Hours**).
- Connecting with other people interested in these topics (I believe this has been one of the biggest drivers of people coming to do high-impact work in the past). Currently, I think the **effective altruism** community is the best venue for this, and you can learn about how to connect with people via the **Centre for Effective Altruism** (see the “Get involved” drop-down). If new ways of connecting with people come up in the future, I will likely post them on Cold Takes.
- And of course, taking any opportunities you see for **robustly helpful actions**.

Buttons you can click

Here's something you can do right now that would be genuinely helpful, though maybe not as viscerally satisfying as signing a petition or making a donation.

In my [day job](#), I have a lot of moments where I - or someone I'm working with - is looking for a particular kind of person (perhaps to fill a job opening with a grantee, or to lend expertise on some topic, or something else). Over time, I expect there to be more and more opportunities for people with specific skills, interests, expertise, etc. to take actions that [help make the best of the most important century](#). And I think a major challenge will simply be **knowing who's out there** - who's interested in this cause, and wants to help, and what skills and interests they have.

If you're a person we might wish we could find in the future, you can help now by sending in information about yourself via [this simple form](#). I vouch that your information won't be sold or otherwise used to make money, that your communication preferences (which the form asks about in detail) will be respected, and that you'll always be able to opt out of any communications.

Sharing a headspace

In [This Can't Go On](#), I analogized the world to people on a plane blasting down the runway, without knowing why they're moving so fast or what's coming next:



As someone sitting on this plane, I'd love to be able to tell you I've figured out exactly what's going on and what future we need to be planning for. But I haven't.

Lacking answers, I've tried to at least show you what I do see:

- Dim outlines of the most important events in humanity's past or future.
- A case that they're approaching us more quickly than it seems - whether or not we're ready.
- A sense that the world and the rules we're all used to can't be relied on. That we need to lift our gaze above the daily torrent of tangible, relatable news - and try to wrap our heads around weirder, wilder matters that are more likely to be seen as the **headlines about this era billions of years from now.**

There's a lot I don't know. But if this is the most important century, I do feel confident that we as a civilization aren't yet up to the challenges it presents.

If that's going to change, it needs to start with more people seeing the situation for what it is, taking it seriously, taking action when they can - and when not, staying vigilant.

Appendices

Weak Point In “Most Important Century”: Full Automation

I thought it would be good to write a couple of posts covering what I see as the weakest points in the [“most important century” series](#), now that I’ve gotten some reactions and criticisms.

I currently think the weakest point in the series runs something like this:

- It’s true that if AI could *literally* automate *everything* needed to [cause scientific and technological advancement](#), the consequences outlined in the series (a dramatic acceleration in scientific and technological advancement, leading to a radically unfamiliar future) would follow.
- But what if AI could only automate 99% of what’s needed for scientific and technological advancement? What if AI systems could propose experiments but not run them? What if they could propose experiments and run them, but not get regulatory clearance for them? **In this case, it’s plausible that the 1% of things AIs couldn’t do quickly and automatically would “bottleneck” progress, leading to dramatically less growth.**
- The series [cites expert opinion on when transformative AI will be developed](#). *Technically speaking*, the type of situation that the respondents are forecasting - “unaided machines can accomplish every task better and more cheaply than human workers” - should be enough for a productivity explosion. But the people surveyed might be thinking of a *slightly* less powerful type of AI than is literally implied by that statement - which could lead to dramatically smaller impacts. Or they could be imagining that even AIs with *intellectual capability* to match humans still might lack the in-practice ability to do key tasks because (for example) they aren’t instinctively trusted by humans. Either way, they (the survey respondents) could be imagining something almost as capable - but not nearly as impactful - as the type of AI I discuss.
- Furthermore, even if AIs could do everything that *humans* do to automate scientific and technological advancement, their scientific and technological progress might have to wait on the results of real-world experiments, which could slow them down a lot.

In brief: **a small gap in what AI can automate could lead to a lot less impact than the series implies.** Automating “almost everything” could be very different from automating everything.

This is important context for the attempts to **[forecast transformative AI](#)**: they are really forecasting something pretty extreme.

My response

I think all of the above is about right as stated: we would indeed need extreme levels of automation to produce the consequences I envision. (There could be a few tasks that need to be done by humans, but they'd have to be quite a small and limited set in order to avoid slowing things down a lot via bottleneck.)

It's also true that I haven't spelled out how such extreme automation could be achieved - how each activity needed to advance scientific and technological advancement (including running experiments and waiting for them to finish) could be done in a quick and/or automated way, without human or other bottlenecks slowing things down much.

With that acknowledged, it's also worth noting that the extreme levels of automation **need not apply to the whole economy: extreme automation for a relatively small set of activities could be sufficient to reach the conclusions in the series.**

For example, it might be sufficient for AI systems to develop increasingly efficient (a) computers; (b) solar panels (for energy); (c) mining and manufacturing robots; (d) space probes (to build more computers in space, where energy and metal are abundant). That could be sufficient (via **[feedback loop](#)**) for explosive growth in available energy, materials and computing power, and there are many ways that such growth could be transformative.

For example and in particular, it could lead to:

- **[Misaligned AI](#)** with access to dangerous amounts of materials and energy.
- **[Digital people](#)**, if AI systems also had some way of (a) “virtualizing” neuroscience (via virtual experiments or simply dramatically increasing the rate of learning from real-world experiments); or (b) otherwise hav-

ing insight about how to create something we would properly regard as “digital descendants.”

Bottom line

I don’t think I’ve thoroughly (or, for readers with strong initial skepticism on this point, convincingly) demonstrated that advanced AI could cause explosive acceleration in scientific and technological advancement, without hitting human-dependent or other “bottlenecks.” I think I have given a good sense of the intuition for why they could, but this is certainly a topic that I haven’t poked as hard as I could; I hope and expect that someone will eventually.

I do think such poking will ultimately support the picture I’ve given in the [“most important century” series](#). This is partly based on the reasoning above: the relatively limited scope of what would need to be fully automated in order to support my broad conclusions. It’s also partly based a similar reasoning process to what I’ve used in the past to [guess at some key conclusions before we’d done all the homework](#): engaging in a lot of conversations and forming views on how informed different parties are and how much sense they’re making. But I acknowledge that this is not as satisfying or reliable as it would be if I gave a highly detailed description of what precise activities can be automated.

Weak Point In “Most Important Century”: Lock-In

This is the second of (for now) two posts covering what I see as the weakest points in the [“most important century” series](#). (The first one is [here](#).)

The weak point I’ll cover here is the discussion of “lock-in”: the idea that transformative AI could lead to societies that are **stable for billions of years**. If true, this means that how things go this century could affect what life is like in predictable, systematic ways for unfathomable amounts of time.

My main coverage of this topic is in a [section of my piece on digital people](#). It’s pretty hand-wavy, not super thorough, and isn’t backed by an in-depth technical report (though I do link to [some informal notes](#) from physicist [Jess Riedel](#) that he made while working at Open Philanthropy). [Overcoming Bias](#) critiqued me on this point, leading to a [brief exchange in the comments](#).

I’m not going to be dramatically more thorough or convincing here, but I will say a bit more about how the overall “most important century” argument is affected if we ignore this part of it, and a bit more about why I find “lock-in” plausible.

(Also note that “lock-in” will be discussed at some length in an upcoming book by Will MacAskill, *What We Owe the Future*.)

Throughout this piece, I’ll be using “lock-in” to mean “key things about society, such as who is in power or which religions/ideologies are dominant, are locked into place indefinitely, plausibly for billions of years,” and “dynamism” or “instability” to mean the opposite: “such things change on much shorter time horizons, as in decades/centuries/millennia.” As noted [previously](#), I consider “lock-in” to be a scary possibility by default, though it’s imaginable that certain kinds of lock-in (e.g., of human rights protections) could be good.

“Most important century” minus “lock-in”

First, let’s just see what happens if we throw out this entire part of the argument and assume that “lock-in” isn’t a possibility at all, but accept the [rest of the claims](#). In other words, we assume that:

- Something like **PASTA** (advanced AI that automates scientific and technological advancement) is likely to be developed this century.
- That, in turn, would lead to **explosive scientific and technological advancement**, resulting in a world run by **digital people** or **mis-aligned AI** or something else that would make it fair to say we have “transitioned to a state in which humans as we know them are no longer the main force in world events.”
- But it would *not* lead to any particular aspect of the world being permanently set in stone. There would remain billions of years full of unpredictable developments.

In this case, I think there is still an important sense in which this would be the “most important century for humanity”: it would be our last chance to shape the transition from a world run by humans to a world run by something very much unlike humans. This is one of the two definitions of “most important century” given [here](#).

More broadly, in this case, I think there’s an important sense in which the **“most important century” series** should be thought of as “Pointing to a drastically underrated issue; correct in its most consequential, controversial implications, if not in every detail.” When people talk about the most significant issues of our time (in fact, even when they are specifically talking about **likely consequences of advanced AI**), they rarely include much discussion of the sorts of issues emphasized in this series; and they should, whether or not this series is correct about the possibility of “lock-in.”

As noted [here](#), I ultimately care more about whether the “most important century” series is correct in this sense - pointing at drastically underappreciated issues - than about how likely its title is to end up describing reality. (Though I care about both.) It’s for this reason that I think the relatively thin discussion of lock-in is a less important “weak point” than the **weak point I wrote about previously**, which raises questions about whether advanced AI would change the world very quickly or very much at all.

But I’ve included the mention of lock-in because I think it’s a real possibility, and it would make the stakes of this century even higher.

Dissecting “lock-in”

There have probably been many people in history (emperors, dictators) with enormous power over their society, and who would’ve liked to keep things going just as they were forever. There may also have been points in time when democratically elected governments would have “locked in” at least some things about their society for good, if they could have.

But they couldn’t. Why not?

I think the reasons broadly fall into a few categories, and [digital people](#) (or [misaligned AI](#), but I’ll focus on digital people to keep things simple for now) could change the picture quite a bit.

First I’ll list factors that seem particularly susceptible to being changed by technology, then one factor that seems less so.

Factors that seem particularly susceptible to being changed by technology

Aging and death. Any given powerful person has to die at some point. They can try to transfer power to children or allies, but a lot changes in the handoff (and over very long periods of time, there are a lot of handoffs).

Digital people need not age or die. (More broadly, sufficient advances in science and technology seem pretty likely to be able to eliminate aging and death, even if not via digital people.) So if some particular set of them had power over some particular part of the galaxy, death and aging need not interfere here at all.

Other population changes. Over time, the composition of any given population changes, and in particular, one generation replaces the previous one. This tends to lead to changes in values and power dynamics.

Without aging or death, and with extreme productivity, we could end up quickly exhausting the carrying capacity of any particular area - so that area might not see changes in population composition at all (or might see much smaller, more controlled changes than we are used to today - no cases where a whole generation is replaced by a new one). Generational turnover seems like quite a big driver of dynamism to date.

Chaos. To date, even when some government is officially “in charge” of a society, it has very limited ability to monitor and intervene in everything that’s going on. But I think technological advancement to date has already greatly increased the ability of a government to exercise control over a large number of people and large geography. An explosion in scientific and technological advancement could radically further increase governments’ in-practice control of what’s going on.

(Digital people provide an extreme example: controlling the server running a virtual environment would mean being able to monitor and control everything about the people in that environment. And powerful figures could create many copies of themselves for monitoring and enforcement.)

Natural events. All kinds of things might disrupt a human society: changes in the weather/climate, running lower on resources, etc. Sufficient advances in science and technology could drive this sort of disruption to extremely low levels (and in particular, [digital people](#) have pretty limited resource needs, such that they need not run low on resources for billions of years).

Seeking improvement. While some dictators and emperors might prefer to keep things as they are forever, most of today’s governments don’t tend to have this as an aspiration: elected officials see themselves as accountable to large populations whose lives they are trying to improve.

But dramatic advances in science and technology would mean dramatically more control over the world, as well as potentially less scope for *further* improvement (I generally expect that the rate of improvement has to [trail off at some point](#)). This could make it increasingly likely that some government or polity decides they’d prefer to lock things in as they are.

But could these factors be eliminated so thoroughly as to cause stability for billions of years? I think so, if enough of society were digital (e.g., [digital people](#), such that those seeking stability could use digital error correction (essentially, making multiple copies of any key thing, which can be used to roll back anything that changes for any reason - for more, see [Jess Riedel’s informal notes](#), which argue that digital error correction could be used to reach quite extreme levels of stability).

A tangible example here would be [tightly controlled virtual environments](#), containing [digital people](#), programmed to reset entirely (or reset

key properties) if any key thing changed. These represent one hypothetical way of essentially eliminating all of the above factors as sources of change.

But even if we prefer to avoid thinking about such specific scenarios, I think there are broader cases for explosive scientific and technological advancement radically reducing the role of each of the above factors, as outlined above.

Of course, just because some government *could* achieve “lock-in” doesn’t mean it *would*. But over the course of a long enough time, it seems that “anti-lock-in” societies would simply gain ever more chances to become “pro-lock-in” societies, whereas even a few years of a “pro-lock-in” society could result in indefinite lock-in. (And in a world of [digital people](#) operating a lot faster than humans, a lot of “time” could go by by the end of this century.)

A factor that seems less susceptible to being changed by technology: competition between societies

Even if a government had complete control over its society, this wouldn’t ensure stability, because it could always be attacked from outside. And **unlike the above factors, this is not something that radical advances in science and technology seem particularly likely to change**: in a world of digital people, different governments would still be able to attack each other, and would be able to negotiate with each other with the threat of attack in the background.

This could cause sustained instability such that the world is constantly changing. This is the point emphasized by the [Overcoming Bias critique](#).

I think this dynamic might - or might not - be an enduring source of dynamism. Some reasons it might not:

- If AI caused an explosion in scientific and technological advancement, then whoever develops it first could quickly become very powerful - being “first to develop [PASTA](#) by a few months” could effectively mean developing the equivalent of a several-centuries lead in science and technology after that. This could lead to consolidation of power on Earth, and there are [no signs of intelligent life outside Earth](#) - so that could be the end of “attack” dynamics as a force for instability.
- Awareness of the above risk might cause the major powers to explicitly [negotiate](#) and divide up the galaxy, committing (perhaps enforceably,

depending on how the technological picture shakes out) never to encroach each others' territory. In this case, any particular part of the galaxy would not be subject to attacks.

- It might turn out that space settlements are generally easier to defend than attack, such that once someone establishes one, it is essentially not subject to attack.

Any of the above, or a combination (e.g., attacks are possible but risky and costly; world powers choose not to attack each other in order not to set off a war), could lead to the permanent disappearance of military competition as a factor, and open up the possibility for some governments to “lock in” key characteristics of their societies.

Three categories of long-run future

Above, I've listed some factors that may - or may not - continue to be sources of dynamism even after explosive scientific and technological advancement. I think I have started to give a sense for why, at a minimum, sources of dynamism could be greatly *reduced* in the case of digital people or other radically advanced technology, compared to today.

Now I want to divide the different possible futures into three broad categories:

Full discretionary lock-in. This is where a given government (or coalition or negotiated setup) is able to essentially lock in whatever properties it chooses for its society, indefinitely.

This could happen if essentially every source of dynamism outlined above goes away, and governments choose to pursue lock-in.

Predictable competitive dynamics. I think the source of dynamism that is most likely to persist (in a world of [digital people](#) or comparably advanced science and technology) is the last one discussed in the above section: military competition between advanced societies.

However, I think it could persist in a way that makes the **long-run outcomes importantly predictable**. In fact, I think “importantly predictable long-run outcomes” is part of the vision implied by the [Overcoming Bias critique](#), which argues that the world will need to be near-exclusively populated by beings that spend nearly their entire existence working (since the

population will expand to the point that it's necessary to work constantly just to survive).

If we end up with a world full of digital beings that have full control over their environment *except for* having to deal with military competition from others, we might expect that there will be strong pressures for the digital beings that are most ambitious, most productive, hardest-working, most aggressive, etc. to end up populating most of the galaxy. These may be beings that do little else but strive for resources.

True dynamism. Rather than a world where governments lock in whatever properties they (and/or majorities of their constituents) want, or a world where digital beings compete with largely predictable consequences, we could end up with a world in which there is true freedom and dynamism - perhaps deliberately preserved via putting specific measures in place to stop the above two possibilities, and enforce some level of diversity and even randomness.

Having listed these possibilities, I want to raise the hypothesis that **if we could end up with any of these three, and this century determines which (or which mix) we end up with, that makes a pretty good case for this century having especially noteworthy impacts, and thereby being the most important century of all time for intelligent life.**

For example, say that from today's vantage point, we're equally likely to get (a) a world where powerful governments employ "lock-in," (b) a world where unfettered competition leads the galaxy to be dominated by the strong/productive/aggressive, or (c) a truly dynamic world where future events are unpredictable and important. In that case, if we end up with (c), and future events end up being enormously interesting and consequential, I would think that there would still be an important sense in which the most important development of all time was the *establishment of that very dynamic*. (Given that one of the other two could have instead ended up determining the shape of civilization across the galaxy over the long run.)

Another way of putting this: if lock-in (and/or predictably competitive dynamics) is a serious possibility starting this century, the opportunity to *prevent* it could make this century the most important one.

Boiling it down

This has been a lot of detail about radically unfamiliar futures, and readers may have the sense at this point that things have gotten too specific and complex to put much stock in. But I think the broad intuitions here are fairly simple and solid, so I'm going to give a more high-level summary:

- Scientific and technological advancement can reduce or eliminate many of today's sources of instability, from aging and death to chaos and natural events. An explosion in scientific and technological advancement could therefore lead to a big drop in dynamism. (And as one vivid example, digital people could set up tightly controlled virtual environments with very robust error correction - something I consider a scary possibility by default, as noted in the intro.)
- Dynamism may or may not remain, depending on a number of factors about how consolidated power ends up being and how different governments/societies deal with each other. The "may or may not" could be determined this century.
- I think this is a serious enough possibility that it heightens the stakes of the "most important century," but I'm far from confident in the thinking here, and I think most of the spirit of the "most important century" hypothesis survives even if we forget about all of it.

Hopefully these additional thoughts have been helpful context on where I'm coming from, but I continue to acknowledge that this is one of the more under-developed parts of the series, and I'm interested in further exploration of the topic.

“Biological Anchors” Is About Bounding, Not Pinpointing, AI Timelines

I previously summarized Ajeya Cotra’s [“biological anchors” method for forecasting for transformative AI](#), aka “Bio Anchors.” Here I want to try to clarify why I find this method so useful, *even though* I agree with the majority of the specific things I’ve heard people say about its weaknesses (sometimes people who can’t see why I’d put any stock in it at all).

A couple of preliminaries:

- This post is probably mostly of interest for skeptics of Bio Anchors, and/or people who feel pretty confused/agnostic about its value and would like to see a reply to skeptics.
- I don’t want to give the impression that I’m leveling new criticisms of “Bio Anchors” and pushing for a novel reinterpretation. I think the author of “Bio Anchors” mostly agrees with what I say both about the report’s weaknesses and about how to best use it (and I think the text of the report itself is consistent with this).

Summary of what the framework is about

Just to re-establish context, here are some key quotes from my [main post about biological anchors](#):

The basic idea is:

Modern AI models can “learn” to do tasks via a (financially costly) process known as “training.” You can think of training as a massive amount of trial-and-error. For example, voice recognition AI models are given an audio file of someone talking, take a guess at what the person is saying, then are given the right answer. By doing this millions of times, they “learn” to reliably translate speech to text. More: [Training](#)

- The bigger an AI model and the more complex the task, the more the training process [or “**training run**”] costs. Some AI models are bigger than others; to date, none are anywhere near “as big as the hu-

man brain” (what this means will be elaborated below). More: [Model size and task type](#)

- The biological anchors method asks: “**Based on the usual patterns in how much training costs, how much would it cost to train an AI model as big as a human brain to perform the hardest tasks humans do? And when will this be cheap enough that we can expect someone to do it?**” More: [Estimating the expense](#)

...The framework provides a way of thinking about how it could be simultaneously true that (a) the AI systems of a decade ago didn’t seem very impressive at all; (b) the AI systems of today can do many impressive things but still feel far short of what humans are able to do; (c) the next few decades - or even the next 15 years - could easily see the development of transformative AI.

Additionally, I think it’s worth noting a **couple of high-level points** from Bio Anchors that **don’t depend on quite so many estimates and assumptions**:

- In the coming decade or so, we’re likely to see - for the first time - AI models with comparable “size” to the human brain.
- If AI models continue to become larger and more efficient at the rates that Bio Anchors estimates, it will probably become **affordable this century to hit some pretty extreme milestones - the “high end” of what Bio Anchors thinks might be necessary**. These are hard to summarize, but see the “long horizon neural net” and “evolution anchor” frameworks in the report.
- One way of thinking about this is that the next century will likely see us go from “not enough compute to run a human-sized model at all” to “extremely plentiful compute, as much as even quite conservative estimates of what we might need.” Compute isn’t the only factor in AI progress, but to the extent other factors (algorithms, training processes) became the new bottlenecks, there will likely be powerful incentives (and multiple decades) to resolve them.

Things I agree with about the framework's weaknesses/limitations

Bio Anchors “acts as if” AI will be developed in a particular way, and it almost certainly won't be

Bio Anchors, in some sense, “acts as if” transformative AI will be built in a particular way: **simple brute-force trial-and-error of computationally intensive tasks** (as outlined [here](#)). Its main forecasts are based on that picture: it estimates when there will be enough compute to run a certain amount of trial and error, and calls that the “estimate for when transformative AI will be developed.”

I think it's unlikely that if and when transformative AI is developed, the way it's developed will resemble this kind of blind trial-and-error of long-horizon tasks.

If I had to guess how transformative AI will be developed, it would be more like:

- First, narrow AI systems prove valuable at a limited set of tasks. (This is already happening, to a limited degree, with e.g. voice recognition, translation and search.)
- This leads to (a) more **attention and funding in AI**; (b) more integration of AI into the economy, such that it becomes easier to collect **data on how humans interact with AIs** that can be then **used for further training**; (c) increased general awareness of what it takes for AI to usefully automate key tasks, and hence **increased awareness of (and attention to) the biggest blockers to AI being broader and more capable**.
- Different sorts of narrow AIs become integrated into different parts of the economy. Over time, the increased training data, funding and attention leads to AIs that are less and less narrow, taking on broader and broader parts of the tasks they're doing. These changes don't just happen via AI models (and training runs) getting bigger and bigger; they are also driven by innovations in how AIs are designed and trained.
- At some point, some combination of AIs is able to [**automate enough of scientific and technological advancement to be transforma-**](#)

ive. There isn't a single "master run" where a single AI is trained to do the very hardest, broadest tasks via blind trial-and-error.

Bio Anchors “acts as if” compute availability is the only major blocker to transformative AI development, and it probably isn't

As noted in my [earlier post](#):

Bio Anchors could be too aggressive due to its assumption that “computing power is the bottleneck”:

- It assumes that if one could pay for all the computing power to do the brute-force “training” described above for the key tasks (e.g., automating scientific work), transformative AI would (likely) follow.
- Training an AI model doesn't just require purchasing computing power. It requires hiring researchers, running experiments, and perhaps most importantly, finding a way to set up the “trial and error” process so that the AI can get a huge number of “tries” at the key task. It may turn out that doing so is prohibitively difficult.

It is very easy to picture worlds where transformative AI takes much more or less time than Bio Anchors implies, for reasons that are essentially not modeled in Bio Anchors at all

As implied above, transformative AI could take a very long time for reasons like “it's extremely hard to get training data and environments for some crucial tasks” or “some tasks simply aren't learnable even by large amounts of trial-and-error.”

Transformative AI could also be developed much more *quickly* than Bio Anchors implies. For example, some breakthrough in how we design AI algorithms - perhaps inspired by neuroscience - could lead to AIs that are able to do ~everything human brains can, *without* needing the massive amount of trial-and-error that Bio Anchors estimates (based on extrapolation from *today's* machine learning systems).

I've listed more considerations like these [here](#).

Bio Anchors is not “pinpointing” the most likely year transformative AI will be developed

My understanding of climate change models is that they try to examine **each major factor** that could cause the temperature to be higher or lower in the future; produce a best-guess estimate for each; and put them all together into a prediction of where the temperature will be.

In some sense, you can think of them as “**best-guess pinpointing**” (or even “simulating”) the future temperature: while they aren't certain or precise, they are identifying a particular, specific temperature based on all of the major factors that might push it up or down.

Many other cases where someone estimates something uncertain (e.g., the future population) have similar properties.

Bio Anchors isn't like that. There are factors it ignores that are identifiable today and almost certain to be significant. So in some important sense, it isn't “pinpointing” the most likely year for transformative AI to be developed.

(Not the focus of this piece) The estimates in Bio Anchors are very uncertain

Bio Anchors estimates some difficult-to-estimate things, such as:

- How big an AI model would have to be to be “as big as the human brain” in some relevant sense. (For this it adapts [Joe Carlsmith's detailed report](#).)
- How fast we should expect algorithmic efficiency, hardware efficiency, and “willingness to spend on AI” to increase in the future - all of which affect the question of “how big an AI training run will be affordable.” Its estimates here are very simple and I think there is lots of room for improvement, though I don't expect the qualitative picture to change radically.

I acknowledge significant uncertainty in these estimates, and I acknowledge that (all else equal) uncertainty [means we should be skeptical](#).

That said:

- I think these estimates are probably reasonably close to the best we can do today with the information we have.
- I think these estimates are good enough for the purposes of what I'll be saying below about transformative AI timelines.

I don't plan to defend this position more here, but may in the future if I get a lot of pushback on it.

Bio Anchors as a way of bounding AI timelines

With all of the above weaknesses acknowledged, here are some things I believe about AI timelines, that are largely based on the Bio Anchors analysis:

- **I would be at least mildly surprised if transformative AI weren't developed by 2060.** I put the probability of transformative AI by then at 50% (I explain below how the connection works between "mild surprise" and "50%"); I could be sympathetic to someone who said it was 25% or 75%, but would have a hard time seeing where someone was coming from if they went outside that range. [More](#)
- **I would be significantly surprised if transformative AI weren't developed by 2100.** I put the probability of transformative AI by then at 2 in 3; I could be sympathetic to someone who said it was 1 in 3 or 80-90%, but would have a hard time seeing where someone was coming from if they went outside that range. [More](#)
- **Transformative AI by 2036 seems plausible and concretely imaginable, but doesn't seem like a good default expectation.** I think the probability of transformative AI by then is at least 10%; I could be sympathetic to someone who said it was 40-50%, but would have a hard time seeing where someone was coming from if they said it was <10% or >50%. [More](#)

I'd be at least mildly surprised if transformative AI weren't developed by 2060

This is *mostly* because, according to Bio Anchors, it will then be affordable to do some *absurdly* big training runs - arguably the biggest ones one could imagine needing to do, based on using [AI models 10x the size of human brains and tasks that require massive numbers of computations to do even once](#). In some important sense, we'll be "swimming in compute." (More on this intuition at [Fun with +12 OOMs of compute](#).)

But it *also* matters that 2060 is 40 years from now, which is 40 years to:

- Develop ever more efficient AI algorithms, some of which could be big breakthroughs.
- Increase the number of AI-centric companies and businesses, collecting data on human interaction and focusing increasing amounts of attention on the things that currently block broad applications.

Given the already-rising amount of investment, talent, and potential applications for today's AI systems, 40 years seems like a pretty long time to make big progress on these fronts. For context, 40 years is around the amount of time that has elapsed between the [Apple IIe release](#) and now.

When it comes to translating my "sense of mild surprise" into a probability (see [here](#) for a sense of what I'm trying to do when talking about probabilities; I expect to write more on this topic in the future):

- On most topics, I equate "I'd be mildly surprised if X didn't happen" with something like a 60-65% chance of X. But on this topic, I do think there's a [burden of proof](#) (which I consider significant though not overwhelming), and I'm inclined to shade my estimates downward somewhat. So I am saying there's about a 50% chance of transformative AI by 2060.
- I'd be sympathetic if someone said "40 years doesn't seem like enough to me; I think it's more like a 25% chance that we'll see transformative AI by 2060." But if someone put it at less than 25%, I'd start to think: "Really? Where are you getting that? Why think there's a <25% chance that we'll develop transformative AI by a year in which it looks like we'll be swimming in compute, with enough for the largest needed runs according to

our best estimates, with 40 years elapsed between today's AI boom and 2060 to figure out a lot of the other blockers?"

- On the flip side, I'd be sympathetic if someone said "This estimate seems way too conservative; 40 years should be easily enough; I think it's more like a 75% chance we'll have transformative AI by 2060." But if someone put it at more than 75%, I'd start to think: "Really? Where are you getting that? Transformative AI doesn't **feel around the corner**, so this seems like kind of a lot of confidence to have about a 40-year-out event."

I would be significantly surprised if transformative AI weren't developed by 2100

By 2100, Bio Anchors projects that it will be affordable not only to do almost *comically* large-seeming training runs (again based on the **hypothetical size of the models and cost-per-try of the tasks**), but to do *as many computations as all animals in history combined, in order to re-create the progress that was made by natural selection.*

In addition, 2100 is 80 years from now - longer than the time that has elapsed since programmable digital computers were **developed in the first place**. That's a *lot* of time to find new approaches to AI algorithms, integrate AI into the economy, collect training data, tackle cases where the current AI systems don't seem able to learn particular tasks, etc.

To me, it feels like 2100 is something like "About as far out as I could tell a reasonable-seeming story for, and then some." Accordingly, I'd be significantly surprised if transformative AI weren't developed by then, and I assign about a **2/3 chance that it will be**. And:

- I'd be sympathetic if someone said "Well, there's a lot we don't know, and a lot that needs to happen - I only think there's a 50% chance we'll see transformative AI by 2100." I'd even be *somewhat* sympathetic if they gave it a 1 in 3 chance. But if someone put it at less than 1/3, I'd really have trouble seeing where they were coming from.
- I'd be sympathetic if someone put the probability for "transformative AI by 2100" at more like 80-90%, but given the difficulty of forecasting this sort of thing, I'd really have trouble seeing where they were coming from if they went above 90%.

Transformative AI by 2036 seems plausible and concretely imaginable, but doesn't seem like a good default expectation

Bio Anchors lays out concrete, plausible scenarios in which there is enough affordable compute to train transformative AI by 2036 ([link](#)). I know some AI researchers who feel these scenarios are more than plausible - their intuitions tell them that the [giant training runs](#) envisioned by Bio Anchors are unnecessary and that the more aggressive [anchors](#) in the report are being underrated.

I also think Bio Anchors understates the case for “transformative AI by 2036” a bit, because it’s hard to tell what consequences the current boom of AI investment and interest will have. If AI is about to become a noticeably bigger part of the economy (definitely an “if”, but compatible with [recent market trends](#)), this could result in rapid improvements along many possible dimensions. In particular, there could be a feedback loop in which new profitable AI applications spur more investment in AI, which in turn spurs faster-than-expected improvements in the efficiency of AI algorithms and compute, which in turn leads to more profitable applications ... etc.

With all of this in mind, **I think the probability of transformative AI by 2036 is at least 10%**, and I don’t have a lot of sympathy for someone saying it is less.

And that said, all of the above is a set of “coulds” and “mights” - every case I’ve heard for “transformative AI by 2036” seems to require a number of uncertain pieces to click into place.

- If [“long-horizon” tasks](#) turn out to be important, Bio Anchors shows that it’s hard to imagine there will be enough compute for the needed training runs.
- Even if there is plenty of compute, 15 years might not be enough time to resolve challenges like assembling the right training data and environments.
- It’s certainly possible that some completely different paradigm will emerge - perhaps inspired by neuroscience - and transformative AI will be developed in ways that don’t require Bio-Anchors-like “training runs”

at all. But I don't see any particular reason to expect that to happen in the next 15 years.

So I also don't have a lot of sympathy for people who think that there's a >50% chance of transformative AI by 2036.

Bottom line

Bio Anchors is a bit different from the "usual" approach to estimating things. It doesn't "pinpoint" likely dates for transformative AI; it doesn't model all the key factors.

But I think it is very useful - in conjunction with informal reasoning about the factors it doesn't model - for "bounding" transformative AI timelines: making a variety of statements along the lines of "It would be surprising if transformative AI weren't developed by ____" or "You could defend a ____% probability by such a date, but I think a ____% probability would be hard to sympathize with."

And that sort of "bounding" seems quite useful for the purpose I care most about: deciding how seriously to take the possibility of the **most important century**. My take is that this possibility is very serious, though far from a certainty, and Bio Anchors is an important part of that picture for me.

More On “Multiple World-Size Economies Per Atom”

A follow up on “This Can’t Go On” for the skeptical.

In [This Can’t Go On](#), I [argued](#) that 8200 more years of today’s growth rate would require us to sustain “multiple economies as big as today’s entire world economy *per atom*.”

Feedback on this bit was split between “That is so obviously impossible, 8200 years of 2% growth is an absurd idea - growth will have to slow much before then” and “Why is that impossible? With ever-increasing creativity, we could increase quality of life higher and higher, without needing to keep using more and more material resources.”

Here I’m going to respond to the latter point, which means expanding on why 8200 years of 2% growth doesn’t look like a reasonable thing to expect. I’m going to make lots of extremely wild assumptions and talk about all kinds of weird possibilities just so that I cover even far-fetched ways for 2% growth to continue.

If you are already on team “Yeah, I don’t see the world economy growing that much,” you should skip this post unless you’d enjoy seeing the case made in a fair amount of detail.

How we COULD support “multiple world-size economies per atom”

I do think it’s *conceivable* that we could support multiple world-size economies per atom. Here’s one way:

Say that we discover some new activity, or experience, or drug, that people really, really, REALLY value.

Specifically, the market values it at 10^{85} of today’s US dollars (that’s ten trillion trillion trillion trillion trillion trillion dollars). That means it’s valued about 10^{71} times as much as everything the world produces in a year right now (combined).¹³⁰

¹³⁰ Today’s economy is a bit less than $\$10^{14}$ per year ([source](#)). $\$10^{85} = \$10^{14} * 10^{71}$.

Then, one person having this experience¹³¹ would mean the size of the economy is at least $\$10^{85}$. And that would, indeed, be the equivalent of multiple of today's world economies per atom.¹³²

To be clear, it's not that we would've crammed multiple of today's world economies into each atom. It's that we would've crammed something 10^{71} times as valuable as today's world economy into a mere **10^{28} atoms** that make up a human being.

What would it mean, though, to value a single experience 10^{71} times as much as today's entire world economy?

One way of thinking about it might be:

- “A 1 in 10^{71} chance of this thing being experienced would be as valuable as all of today's world economy.”
- Or to make it a bit easier to intuit (while needing to oversimplify), “If I were risk-neutral, I'd be thrilled to accept a gamble where I would die immediately, with near certainty, in exchange for a 1 in 10^{71} chance of getting to have this experience.”¹³³
- How near-certain would death be? Well, for starters, if all the people who have ever lived to date accepted this gamble, it would be approximately certain that they would *all* lose and end up with immediate death.¹³⁴

¹³¹ (And paying full price for it, in a way that gets recorded by GDP statistics, which could get a bit hairy.)

¹³² See [previous estimate](#) of 10^{70} atoms in the galaxy.

¹³³ This assumes that one values one's own life not much more than a year of the world economy's output. I do not expect that I will see enough disagreement on this point to want to write another post on the matter, but it's possible.

It is also making an iffy assumption about “risk-neutrality.” In reality, one might personally value this experience much less than 10^{71} times as much as one's own life, while still paying resources for it that would be sufficient to save an extraordinarily large number of *other* people's lives. It's hard to convey the same kind of magnitudes by appealing to impartiality, so I went with this intuition pump anyway; I think it does give the right basic sense of how mind-bogglingly large the value of this experience would be.

¹³⁴ The calculation here would be: if there are 10^{10} people alive today (this is “rounding up” from ~8 billion to 10 billion), and each has a 10^{-71} (1 in 10^{71}) chance of winning the gamble, then each has a $(1-10^{-71})$ chance of losing the gamble. So the probability that they **all** lose the gamble is $(1-10^{-71})^{(10^{10})}$, which is almost exactly 100%.

- But this really isn't coming anywhere close to communicating how bad the odds would be for this gamble. It's more like: if there were one person for each atom in the galaxy, and each of them took the gamble, they'd probably still **all** lose.¹³⁵
- So to personally take a gamble with those kinds of odds ... the experience had better be REALLY good to compensate.
 - We're not talking about "the best experience you've ever had" level here - it wouldn't be sensible to value that more than an entire life, and the idea that it's worth as much as today's world economy seems pretty clearly wrong.
 - We're talking about something just unfathomably beyond anything any human has ever experienced.

Blowing out the numbers more

Imagine the single best second of your life, the kind of thing evoked by [Letter from Utopia](#):

Have you ever experienced a moment of bliss? On the rapids of inspiration maybe, your mind tracing the shapes of truth and beauty? Or in the pulsing ecstasy of love? Or in a glorious triumph achieved with true friends? Or in a conversation on a vine-overhung terrace one star-appointed night? Or perhaps a melody smuggled itself into your heart, charming it and setting it alight with kaleidoscopic emotions? Or when you prayed, and felt heard?

If you have experienced such a moment – experienced *the best type* of such a moment – then you may have discovered inside it a certain idle but sincere thought: “Heaven, yes! I didn't realize it could be like this. This is so right, on whole different level of right; so real, on a whole different level of real. Why can't it be like this always? Before I was sleeping; now I am awake.”

¹³⁵ Similar calculation to the previous footnote, but with a population of 10^{70} (one for each [atom in the galaxy](#)), so the probability that they all lose the gamble is $(1-10^{-71})^{(10^{70})}$, which I think is around 90% (Excel can't actually handle numbers this big but this is what similar calculations imply).

Yet a little later, scarcely an hour gone by, and the ever-falling soot of ordinary life is already covering the whole thing. The silver and gold of exuberance lose their shine, and the marble becomes dirty.

Now imagine, implausibly, that this single second was worth as much as the entire world economy outputs in a year today. (It doesn't seem possible that it could be worth more, since the world economy that year *included* that second of your life, plus the rest of your year and many other people's years.)

And now imagine a *full year* in which *every second* is as good as *that second*. We'll call this the "perfect year." According to the assumptions above, the perfect year would be no more than about $3 \cdot 10^8$ times as valuable as the world economy (there are about $3 \cdot 10^8$ seconds in a year).

And now imagine that *every atom in the galaxy* could be a person having the perfect year. This would now be about $10^{70} \cdot (3 \cdot 10^8) = 3 \cdot 10^{78}$ as much value as today's world economy. **2% growth would get us there in 9150 years.**

(A crucial and perhaps counterintuitive assumption I'm making here, throughout, is that "2% growth" means "2% *really real* growth" - that whatever is valuable, holistically speaking, about annual world output today, we'll get 2% more of it each year. I think this is already the kind of assumption many people are making when they say we don't need more material to have ever-increasing wealth. If you think the 2% growth of the recent past is more "fake" than this and that it will continue in a "fake" way, that would be a debate for another time.)

And 1200 years after *that*, if each year still had 2% growth, the economy would be another ~20 billion times bigger. So now, for every atom in the galaxy, there'd have to be someone whose year was in some sense ~20 billion times *better* (or "more valuable") than the perfect year.

We're still only talking about ~10,000 years of 2% growth.

New life forms

It's still conceivable! Who knows what the future will bring.

But at this point it's very intuitive to me that we are not talking about anything that looks like "Humans in human bodies having human kinds of fun and fulfillment." An economy of this value seems to require fundamentally re-en-

gineering something about the human experience - finding some way of arranging matter that creates far more happiness, or fulfillment, or something, that we would value so astronomically more than even the heights of human experience today.

And I think the most natural way for that to happen is something like: “Discovering fundamental principles behind what we value, and fundamental principles of how to arrange matter to get the most of it.” Which in turn suggests something more like “Once we have that level of understanding, we start to arrange the matter in the galaxy optimally, and quickly get close to the limits of what’s possible” than like “We grow at 2%, every year, for continuing thousands of years, even as (as would happen with e.g. [digital people](#)) we become beings who can do as much in a year as humans could do in hundreds or thousands of years.”

But it could still happen?

I guess? This was never meant to be a mathematical proof of the impossibility of 2%/year growth. It’s possible in theory.

But at this point, seeing what a funky and fundamentally transformed galaxy it would require within 10,000 years, what is the *affirmative* reason to expect 2%/year growth for that long a period of time? Is it that “This is the trendline, and by default I expect the trendline to continue?”

But that trendline is only a couple hundred years old - why expect it to continue for another 10,000?

Why not, instead, expect the [longer-term pattern of accelerating economic growth](#) to be what continues, until we approach some sort of fundamental limit on how much value we can cram into a given amount of matter? Or expect growth to fall gradually from here and never reach today’s level again?

The last couple of centuries have been a wild ride, with wealth and living conditions improving at a historically high rate. But I don’t think that gives us reason to think that this trend goes to infinity. I believe the limits are somewhere, and it looks like sometime in the next 10,000 years, we’re either going to have to approach those limits, or [stagnate or collapse](#).

<https://www.cold-takes.com/more-on-multiple-world-size-economies-per-atom/>

Hopefully I've given a sense for why it seems so unlikely that there will be 10,000 more years in the future that each have 2% or greater growth. Which would imply that *each* of the last 100+ years will turn out to be one of the fastest-growing 10,000 years of all time.

If you'd like to comment on this post, [this](#) would be a good place to do so.

A Note On Historical Economic Growth

How the “most important century” argument is affected if our picture of long-run economic history changes.

A couple of times in the [Most Important Century](#) series (particularly in [The Duplicator](#)), I say that economic growth over the last few thousand years is a reasonable fit with the pattern (described [here](#)) of accelerating growth, driven by a feedback loop: “more ideas → more output → more people → more ideas → ...”

This point is the subject of an ongoing debate (see [this EA Forum post by Ben Garfinkel](#), and the extensive back-and-forth in comments).

My best guess is that the past data is, in fact, a reasonable (though ambiguous) fit with the pattern of accelerating growth. However, I’m far from confident of this, and I want to address how it would affect my arguments if better, future data turned out to decisively undermine this fit.

Extrapolating future economic growth based on (a long view of) past economic growth

I’ve cited the projection, made in [Modeling the Human Trajectory](#), that the economy is “on track” to hit infinite size this century if the pattern seen in the past continues. If it turned out that past data is inconsistent with accelerating growth, this would undermine Modeling the Human Trajectory, and a new extrapolation would be needed. **However, my best guess is that a good replacement extrapolation would still show a good chance of explosive (even “infinite”) growth this century.** Reasoning for this guess follows.

When discussing the pattern of past growth, the main alternative I’ve seen to *accelerating growth* (including in the EA Forum post linked above and comments) is *a series of fundamentally different ‘growth modes,’ each with its own growth dynamic and/or growth rate.* For example, perhaps - rather than thinking of economic history as a gradual acceleration - one could think of it as divided into distinct phases:

- A pre-agriculture phase (starting some millions of years ago), in which growth was likely extremely slow and perhaps pretty chaotic.

- A phase after the development of agriculture (starting ~10,000 years ago), during which growth was probably faster than before, but still quite slow by today's standards, and perhaps pretty chaotic as well.
- The modern, post-Industrial-Revolution phase (starting ~200 years ago), with by far the fastest growth.

It seems undisputed to me that the third phase is both much shorter (in calendar time) and has dramatically faster growth, compared to the first two. This could be the result of continuous acceleration, or it could be because a fundamentally new growth mode emerged. The latter would then raise the question of whether a transition to another, still faster “growth mode” might be possible.

Robin Hanson's 2000 paper, [Long-Term Growth As A Sequence of Exponential Modes](#), is the main attempt I know of to explore that question. It attempts to model long-run economic history using a couple of different approaches, both of which are designed around the idea of “growth modes,” and (on pages 14-17) to extrapolate patterns observed to date into the future. It states:

In summary, if one takes seriously the model of economic growth as a series of exponential growth modes, and if relative change parameters of a new transition are likely to be similar to such parameters describing old transitions, then it seems hard to escape the conclusion that the world economy could see a very dramatic change within the next century, to a new economic growth mode with a doubling time of roughly two weeks or less ...

If the next mode had a “slow” doubling time of two years, and if it lasted through twenty doubling times, longer than any mode seen so far, it would still last only forty years. After that, it is not clear how many more even faster growth modes are possible before hitting fundamental limits. But it is hard to see how such fundamental limits would not be reached within a few decades at most.

This is qualitatively pretty similar to the projection I've given in the blog posts: both imply a dramatic economic acceleration in the 21st century, and

both imply “infinite growth” or “hitting fundamental limits” not too long after (although the potential delay is longer for Hanson’s approach, and could go modestly into the 22nd century depending on which Hanson projection one uses).

This extrapolation is less straightforward than the [Modeling the Human Trajectory](#) extrapolation. There aren’t very strong reasons to think that the series of growth modes will follow any particular pattern, in terms of how they’re timed and what kind of growth they bring. Hanson’s extrapolation is merely a best guess at what to expect if they do follow a relatively regular pattern. Still, it does look reasonable to me as a best guess.

Other implications if it were to turn out that past economic data does not fit an “accelerating growth” pattern

- Throughout the series, I argue that **various technologies** (The Dupli-cator, digital people, “PASTA”¹³⁶) **could lead to an “accelerating” pattern leading to explosive growth.** This is an implication of most mainstream theoretical models in growth economics, as discussed in [Report on Whether AI Could Drive Explosive Economic Growth](#).¹³⁷ I cite the fact that *past* data seems to fit this dynamic as further support that such a thing is plausible. If past data did not fit the dynamic, it would not affect this theoretical case for expecting explosive growth, but it would make the overall solidness of the case some amount weaker.
- I also will argue against the idea that “If transformative AI were to be developed this century, it would break the pattern we’ve seen of constant economic growth; therefore, we should have a very high burden of proof for predictions of transformative AI this century.” For this purpose, either the “accelerating growth” or “series of different growth modes” dynamic

¹³⁶ Process for Automating Scientific and Technological Development - to be discussed in a future piece.

¹³⁷ More precisely, most models imply that full automation of both R&D and goods production would lead to explosive growth. What about growth before full automation of both these things? First, if automation proceeds more rapidly than its historical rate before full automation, then growth models typically imply growth will begin to accelerate before we achieve full automation (e.g. see section 6.1.4.2 of the [report](#)). Second, if R&D but not goods production is fully automated, I think this would be sufficient for explosive growth (see section 6.1.6 of the [report](#)).

seems sufficient for my case that we should consider a future growth explosion plausible, although I do think the case would be a bit weaker if it had to rely on the latter as opposed to the former.

Bottom line

Overall, if it became clear that economic history contains very little acceleration (and is instead best thought of as a series of distinct “growth modes,”) I think my remaining claims and conclusions would still look about right, though the arguments would be some amount weaker.

It’s also possible that if we had perfect information about long-run economic history, we would see a mix: *some* instances/periods of the “accelerating growth” dynamic described [here](#), some periods that look more like “distinct growth modes.”

Some Additional Detail On What I Mean By “Most Important Century”

Here’s a bit more detail on what I mean when I talk about the [“most important century for humanity.”](#)

There are two different senses in which I think this could be the “most important century,” one higher-stakes and less likely than the other:

Meaning #1: Most important century of all time for humanity, due to the transition to a state in which humans as we know them are no longer the main force in world events.

Here the idea is that:

- During this century, civilization could either end entirely, or change so dramatically that “humans as we know them today” would either not exist anymore, or would at least be a very small part of the population.
 - I think the future I describe in [Digital People Would Be An Even Bigger Deal](#) would probably reach this level of unfamiliarity pretty quickly.
 - The [possibility of AI systems’ expanding across the galaxy based on their own objectives](#) - with humans’ becoming fairly irrelevant in comparison - would qualify as well.
- This century is our chance to shape just how this happens.
 - If we develop [digital people](#), the initial set of digital people could quickly set about making many copies of themselves, multiplying and working at a [far faster rate](#) than normal humans would be able to track or keep up with. With these points in mind, the initial set of digital people - and the virtual conditions they’re placed in - could be crucial in a lasting way.
 - If we instead develop AI systems that expand across the galaxy based on their own objectives, this could permanently lose the opportunity to have the main force in world events be anything like humans at all.

Based on these points, this would be the “most important century” for humans as we are now, in the sense that it’s the best opportunity humans will have to influence a large, post-humans-as-they-are-now future.¹³⁸

This could be consistent with other centuries being “most important” for other “species.”

- Some past century may have been the most important century for chimpanzees. (This may have been some century during which humans started to emerge.)
- Some future century might be the most important century for whatever “comes after humans.” (Although this century might be most important for them too.)

I want to roughly say that *if* something like **PASTA** is developed this century, it has at least a 50/50 chance of being the “most important century” in the above sense.

Meaning #2: Most important century of all time for all intelligent life in our galaxy.

It’s possible, for reasons outlined [here](#), that whatever the main force in world events is (perhaps digital people, misaligned AI, or something else) will create highly stable civilizations with “locked in” values, which populate our entire galaxy for **billions of years to come**.

If enough of that “locking in” happens this century, that could make it the most important century of all time for all intelligent life in our galaxy.

I want to roughly say that *if* something like **PASTA** is developed this century, it has at least a 25% chance of being the “most important century” in the above sense. This is half as much as the probability for the previous version of “most important century.” I don’t mean to be precise here; I’m giving a rough indication of how likely I think such a development would be.

To put this possibility in perspective, it’s worth noting that the world seems to have “sped up” - in the sense of changing more rapidly - over the course of

¹³⁸ You could say that actions of past centuries also have had ripple effects that will influence this future. But I’d reply that the effects of these actions were highly chaotic and unpredictable, compared to the effects of actions closer-in-time to the point where the transition occurs.

history, and could continue to do so if something like **PASTA** is developed. With this in mind:

- If the first bacteria had been talking with each other, one of them might have claimed they were in the “most important **5-billion-year period**,” the one in which bacteria would evolve into complex animals.
- The first complex animals might have claimed that they were in the “most important **eon**,” the one in which humans would emerge.
- Someone living through the **Scientific Revolution** might have claimed that they were in the “most important millennium,” the one in which scientific and technological progress would take off.
- If transformative AI leads to a digital-people-run civilization around, say, 2080, some digital person in 2080 might claim that they’re in the “most important decade.” A decade might feel to them the way a century (or longer) feels to us.
- These digital people might create more advanced digital people who claim that they’re in the “most important day,” figuring that they will evolve into something even stranger during that vast-feeling period of time.
- And they could all be right!

Holistic intent of the “most important century” phrase. I have largely chosen the phrase “most important century” as a **wake-up call** about how high the stakes seem to be.

While I’ve tried to give it slightly more precise meaning above, my main intent is to call attention to the “**Holy !@#**” **feeling of possibly developing something like **PASTA** this century, which in turn could lead to a radically unfamiliar future, possibly involving a stable galaxy-wide civilization.**

If I’m right about that picture but wrong about the “most important century” for some reason (for example, perhaps something even more remarkable happens 5 billion years from now, or perhaps it turns out that the **simulation hypothesis** is correct), I’d still think this series’s general idea was importantly right.

Why Talk About 10,000 Years From Now?

It seems a common reaction to [This Can't Go On](#) is something like: “OK, so ... you’re saying the current level of economic growth can’t go on for another 10,000 years. So?? Call me in a few thousand years I guess?”

In general, this blog will often talk about “long” time frames (decades, centuries, millennia) as if they’re “short” (compared to the billions of years our universe has existed, millions of years our species has existed, and billions of years that could be in our civilization’s future). I sort of try to **imagine myself as a billions-of-years-old observer**, looking at charts like [this](#) and thinking things like “The current economic growth level just got started!” even though it got started several lifetimes ago.

Why think this way?

One reason is that it’s just a way of thinking about the world that feels (to me) refreshing/different.

But here are a couple more important reasons.

Effective altruism

My main obsession is with [effective altruism](#), or doing as much good as possible. I generally try to pay more attention to things when they “matter more,” and I think things “matter more” when they affect larger numbers of persons.¹³⁹

I think there will be a LOT more persons¹⁴⁰ over the coming billions of years than over the coming generation or few. So I think the long-run future, in some sense, “matters more” than whatever happens over the next generation or few. Maybe it doesn’t matter more for me and my loved ones, but it matters more from an “all persons matter equally” perspective.¹⁴¹

¹³⁹ I generally use the term “persons” instead of “people” to indicate that I am trying to refer to every person, animal or thing (AI?) that we should care about the welfare of.

¹⁴⁰ Even more than you’d intuitively guess, as outlined [here](#).

¹⁴¹ I wrote a bit about this perspective several years ago, [here](#).

An obvious retort is “But there’s nothing we can do that will affect ALL of the people who live over the coming billions of years. We should focus on what we can actually change - that’s the next generation or few.”

But I’m not convinced of that.

I think we could be in the [most important century of all time](#), and I think things we do today could end up mattering for billions of years (an obvious example is [reducing risk of existential catastrophes](#)).

And more broadly, if I *couldn't* think of specific ways our actions might matter for billions of years, I’d still be very interested in *looking for them*. I’d still find it useful to try to step back and ask: “Is what I’m reading about in the news important [in the grand scheme of things](#)? Could these events matter for whether we end up with [explosion, stagnation or collapse](#)? For [what kind of digital civilization we create for the long run](#)? And if not ... what could?”

Appreciating the weirdness of the time we live in

I think we live in a very weird period of time. It looks really weird on various charts (like [this one](#), [this one](#), and [this one](#)). The vast bulk of scientific and technological advancement, and growth in the economy, has happened in a tiny sliver of time that we are sitting in. And billions of years from now, it will probably *still* be the case that this tiny sliver of time looks like an [outlier in terms of growth and change](#).

Again, it doesn’t *feel* like a tiny sliver, it feels like lifetimes. It’s hundreds of years. But that’s out of millions (for our species) or billions (for life on Earth).

Sometimes, when I walk down the street, I just look around and think: “This is all SO WEIRD. Whooshing by me are a bunch of people calmly operating steel cars at 40 mph, and over there I see a bunch of people calmly operating a massive crane building a skyscraper, and up in the sky is a plane flying by ... and out of billions of years of life on Earth, it’s only us - the humans of the last hundred-or-so years - who have ever been able to do any of this kind of stuff. Practically everything I look at is some crazy futurist technology we just came up with and haven’t really had time to adapt to, and we won’t have adapted before the next crazy thing comes along.

“And everyone is being very humdrum about their cars and skyscrapers and planes, but this is *not* normal, this is *not* ‘how it usually is,’ this is not part of a plan or a well-established pattern, this is crazy and weird and short-lived, and it’s anyone’s guess where it’s going next.”

I think many of us are instinctively, intuitively dismissive of **wild claims about the future**. I think we naturally imagine that there’s more stability, solidness and hidden wisdom in “how things have been for generations” than there is.

By trying to **imagine the perspective of someone who’s been alive for the whole story** - billions of years, not tens - maybe we can be more open to strange future possibilities. And then, maybe we can be better at noticing the ones that actually might happen, and that our actions today might affect.

So that’s why I often try on the lens of saying things like “X has been going on for 200 years and could maybe last another few thousand - bah, that’s the blink of an eye!”

Endnotes

1 We'll start with this economy:

Year	# people	Resources produced per person (like corn in the diagram)	Resources produced (total)	New duplicates added to population (1 for every 10 units of resources)	# new ideas (initially 1 for every 20 people, but ideas get harder to find)	Productivity improvement (1% per idea - increases resources produced per person)
1	100	1.00	100	10	5	5.00%

100 people produce 100 units of resources (1 per person). For every 10 units of resources, they're able to create 1 more duplicate (this is just capturing the idea that duplicates are "costly" to create). And the 100 people have 5 new ideas, leading to 5% productivity growth.

Here's year 2:

Year	# people	Resources produced per person (like corn in the diagram)	Resources produced (total)	New duplicates added to population (1 for every 10 units of resources)	# new ideas (initially 1 for every 20 people, but ideas get harder to find)	Productivity improvement (1% per idea - increases resources produced per person)
1	100	1.00	100	10	5	5.00%
2	110	1.05	116	12	5	5.24%

Now each person produces 1.05 widgets instead of 1, thanks to the productivity growth. And there's another 5% productivity growth.

This dynamic takes some time to "take off," but take off it does:

Year	# people	Resources produced per person (like corn in the diagram)	Resources produced (total)	New duplicates added to population (1 for every 10 units of resources)	# new ideas (initially 1 for every 20 people, but ideas get harder to find)	Productivity improvement (1% per idea - increases resources produced per person)
1	100	1.00	100	10	5	5.00%
2	110	1.05	116	12	5	5.24%
3	122	1.11	134	13	6	5.50%
4	135	1.17	157	16	6	5.79%
5	151	1.23	186	19	6	6.11%
6	169	1.31	222	22	6	6.47%
7	191	1.39	267	27	7	6.87%
8	218	1.49	325	32	7	7.32%
9	251	1.60	401	40	8	7.84%
10	291	1.72	501	50	8	8.43%
11	341	1.87	637	64	9	9.12%
12	404	2.04	825	82	10	9.92%
13	487	2.24	1091	109	11	10.86%
14	596	2.48	1481	148	12	11.99%
15	744	2.78	2071	207	13	13.37%
16	951	3.15	3001	300	15	15.08%
17	1251	3.63	4543	454	17	17.23%
18	1706	4.26	7259	726	20	20.04%
19	2432	5.11	12422	1242	24	23.80%
20	3674	6.32	23235	2324	29	29.04%
21	5997	8.16	48947	4895	37	36.74%
22	10892	11.16	121558	12156	49	48.80%
23	23048	16.61	382737	38274	69	69.40%
24	61322	28.13	1724982	172498	109	109.00%
25	233820	58.79	13746471	1374647	199	198.86%
26	1608467	175.70	282608848	28260885	458	457.73%
27	29869352	979.93	29269998551	2926999855	1524	1524.05%
28	2956869207	15,914.61	4705742026084	4705742026085	9290	9289.79%
29	47086988952	1,494,349.21	7036440490146	703644049014606	157550	157550.15%
30	70364875771	2,355,843,796.	1657686561026	1657686561026930	14934113	14934113.18%
31	16576865680	351,826,734,70	5832184523987	583218452398710	23558280320	23558280320.49%
32	58321845239	82,884,328,754	4833966994459	483396699445971	3518267332071260	3518267332071260.00%
33	48339669944	2,916,092,261,	1409629374727	140962937472770	82884328754995400	82884328754995400000000
34	14096293747	24,169,834,972	3407050935927	340705093592728	29160922619937200	29160922619937200000000
35	34070509359	7,048,146,873,	2401339540238	240133954023829	24169834972298500	24169834972298500000000
36	24013395402	17,035,254,679	#NUM!	#NUM!	70481468736384900	70481468736384900000000
37	#NUM!	12,006,697,701	#NUM!	#NUM!	#NUM!	#NUM!
38	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!

The #NUM!'s at the bottom signify Google Sheets choking on the large numbers.

My [spreadsheet includes](#) a version with simply exponentially increasing population; that one goes on for ~1000 years without challenging Google Sheets. So the population dynamic is key here.

2 Without human bodies - and depending on what kinds of robots were available - digital people might not be good substitutes for humans when it comes to jobs that rely heavily on human physical abilities, or jobs that require in-person interaction with biological humans.

However, digital people would likely be able to do everything needed to cause an explosive economic growth, even if they couldn't do *everything*. In particular, it seems they could do everything needed to increase the supply of computers, and thereby increase the population of digital people.

Creating more computing power requires (a) raw materials - mostly metal; (b) research and development - to design the computers; (c) manufacturing - to carry out the design and turn raw materials into computers; (d) energy. Digital people could potentially make all of these things a great deal cheaper and more plentiful:

- **Raw materials.** It seems that mining could, in principle, be done entirely with robots. Digital people could design and instruct these robots to extract raw materials as efficiently as possible.
- **Research and development.** My sense is that this is a major input into the cost of computing today: the work needed to design ever-better microprocessors and other computer parts. Digital people could do this entirely virtually.
- **Manufacturing.** My sense is that this is the other major input into the cost of computing today. Like mining, it could in principle be done entirely with robots.
- **Energy.** Solar panels are also subject to (a) better research and development; (b) robot-driven manufacturing. Good enough design and manufacturing of solar panels could lead to radically cheaper and more plentiful energy.

Space exploration. Raw materials, energy, and “real estate” are all super-abundant outside of Earth. If digital people could design and manufacture spaceships, along with robots that could build solar panels and computer factories, they could take advantage of massive resources compared to what we have on earth.:

