Cold Takes

The "most important century" blog post series

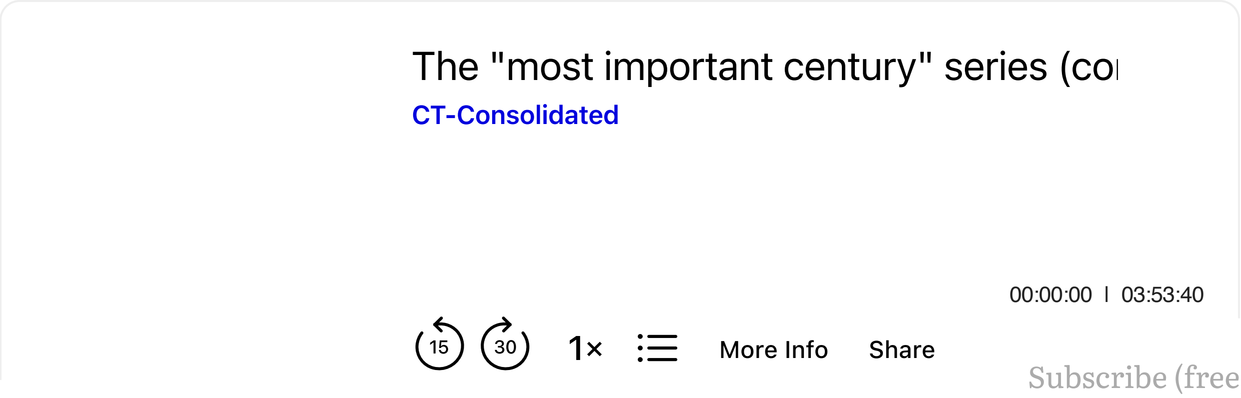
The "most important century" series of blog posts argues that **the 21st century could be the most important century ever for humanity,** via the development of advanced AI systems that could dramatically speed up scientific and technological advancement, getting us more quickly than most people imagine to a deeply unfamiliar future.

You can get the **highlights from the series** via:

- A few-page summary (below)
- Discussion of the series on The Ezra Klein Show (NYT, 90 minutes) or The 80,000 Hours podcast (2 hours)

You can **read the whole series** as:

- A **series of blog posts**: I'd suggest starting with the Roadmap. (Each piece links to the next piece at the end.) This is the original format, where it will be easiest to click around, see graphics at full size, etc.
- An **audio series** available on most podcast platforms (Spotify, Stitcher, Apple Podcasts, etc.):



The "most important century" series (co

CT-Consolidated

- A single printable pdf.
- For Kindle, you can buy a Kindle-formatted version for $0.99 (the minimum price they let me set) or download this AZW3 file for free (see instructions for putting this on your Kindle). There's also a free ePub file for other readers.

The series in a nutshell

I've spent most of my career looking for ways to do as much good as possible, per unit of money or time. I worked on finding evidence-backed charities working on global health and development (co-founding GiveWell), and later moved into philanthropy that takes more risks (co-founding Open Philanthropy).

Over the last few years - thanks to general dialogue with the effective altruism community, and extensive research done by Open Philanthropy's Worldview Investigations team - I've become convinced that humanity as a whole faces huge risks and opportunities this century. Better understanding and preparing for these risks and opportunities is where I am now focused.

This piece will summarize a series of posts on why I believe we could be in the **most important century of all time for humanity**. It gives a short summary, key post(s), and sometimes key graphics for 5 basic points:

- **The long-run future is radically unfamiliar.** Enough advances in technology could lead to a long-lasting, galaxy-wide civilization that could be a radical utopia, dystopia, or anything in between.
- **The long-run future could come much faster than we think**, due to a

- The relevant kind of **AI looks like it will be developed this century** - making this century the one that will initiate, and have the opportunity to shape, a future galaxy-wide civilization.
- These claims seem too "wild" to take seriously. But there are a lot of reasons to think that **we live in a wild time, and should be ready for anything.**
- We, the people living in this century, have the chance to have a huge impact on huge numbers of people to come - if we can make sense of the situation enough to find helpful actions. But right now, **we aren't ready for this.**

This thesis has a wacky, sci-fi feel. It's very far from where I expected to end up when I set out to do as much good as possible.

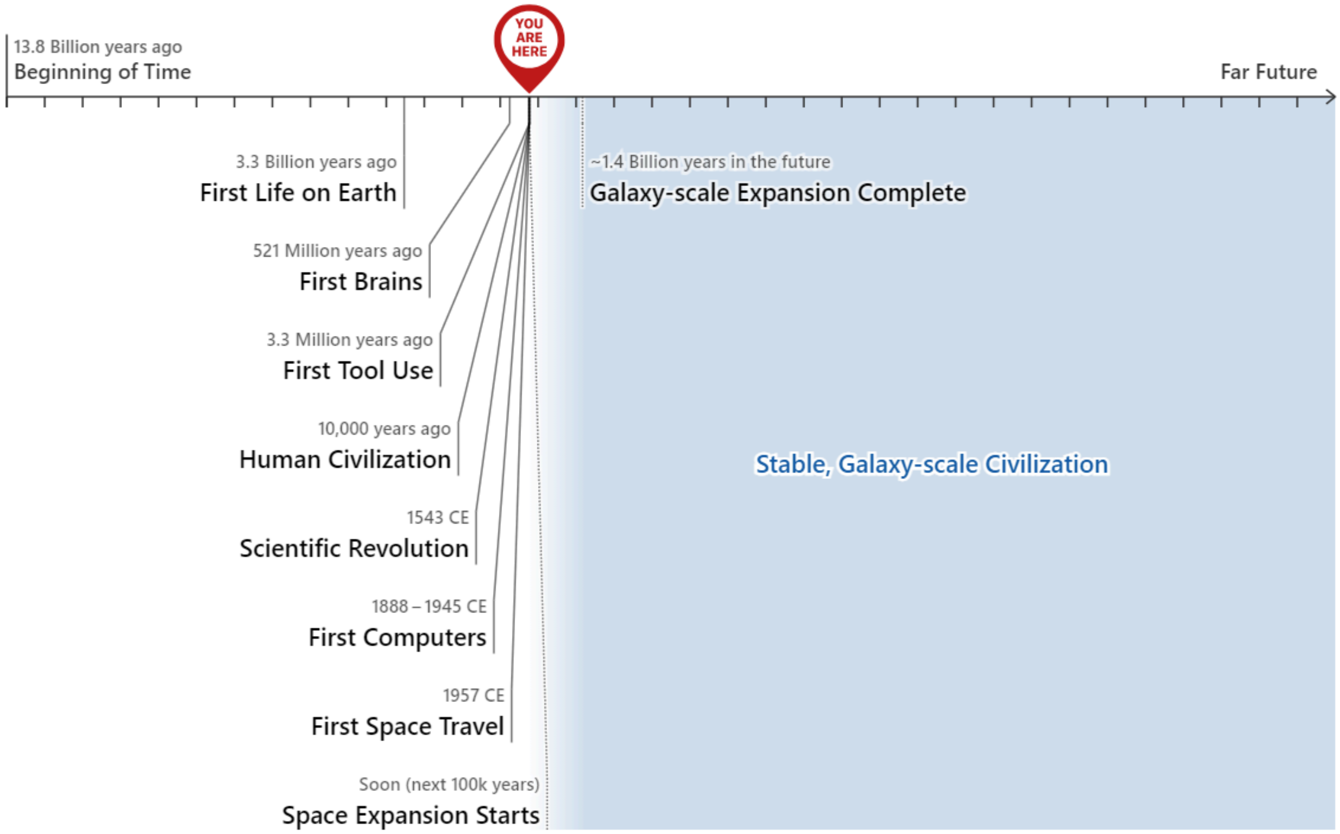
But part of the mindset I've developed through GiveWell and Open Philanthropy is being open to strange possibilities, while critically examining them with as much rigor as possible. And after a lot of investment in examining the above thesis, I think it's likely enough that the world urgently needs more attention on it.

By writing about it, I'd like to either get more attention on it, or gain more opportunities to be criticized and change my mind.

We live in a wild time, and should be ready for anything

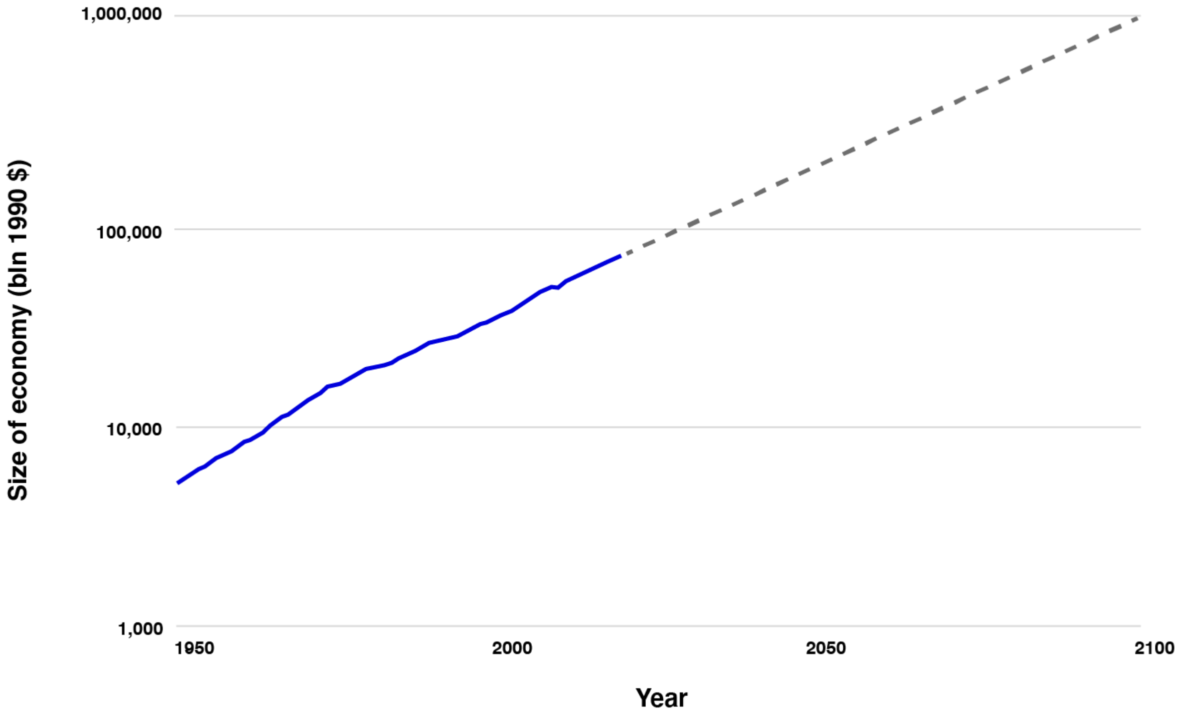
Many people find the "most important century" claim too "wild": a radical future with advanced AI and civilization spreading throughout our galaxy may happen *eventually*, but it'll be more like 500 years from now, or 1,000 or 10,000. (Not this century.)

These longer time frames would put us in a *less* wild position than if we're in the "most important century." But in the scheme of things, **even if galaxy-wide expansion begins 100,000 years from now, that still means we live in an** nearly lifeless to largely populated. It means 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.

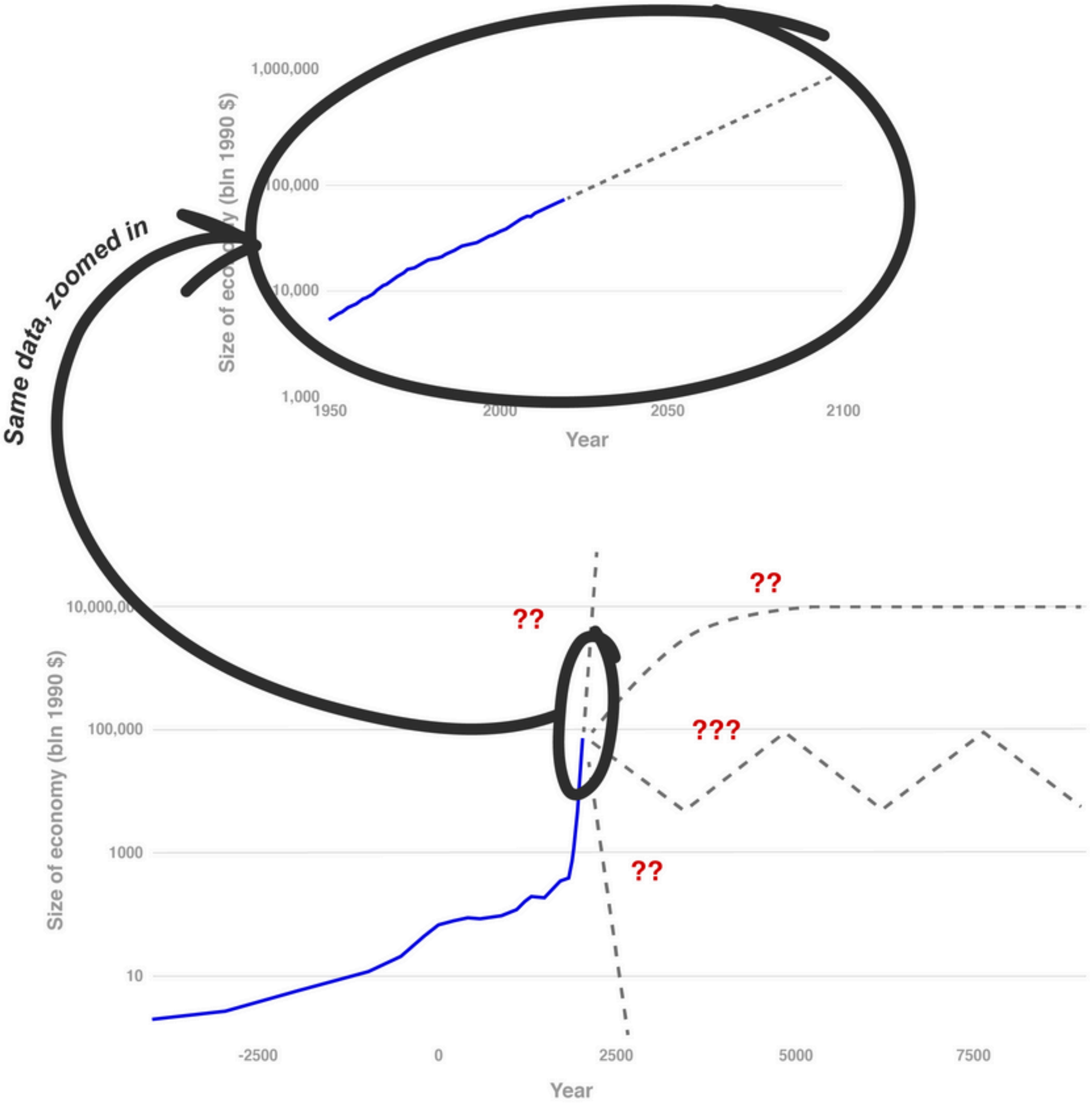


More at [All Possible Views About Humanity's Future Are Wild](https://www.cold-takes.com/all-possible-views-about-humanitys-future-are-wild/)

Zooming in, we live in a special century, not just a special era. We can see this by looking at how fast the economy is growing. It doesn't *feel* like anything special is going on, because for as long as any of us have been alive, the world economy has grown at a few percent per year:



However, when we zoom out to look at history in greater context, we see a picture of an unstable past and an uncertain future:

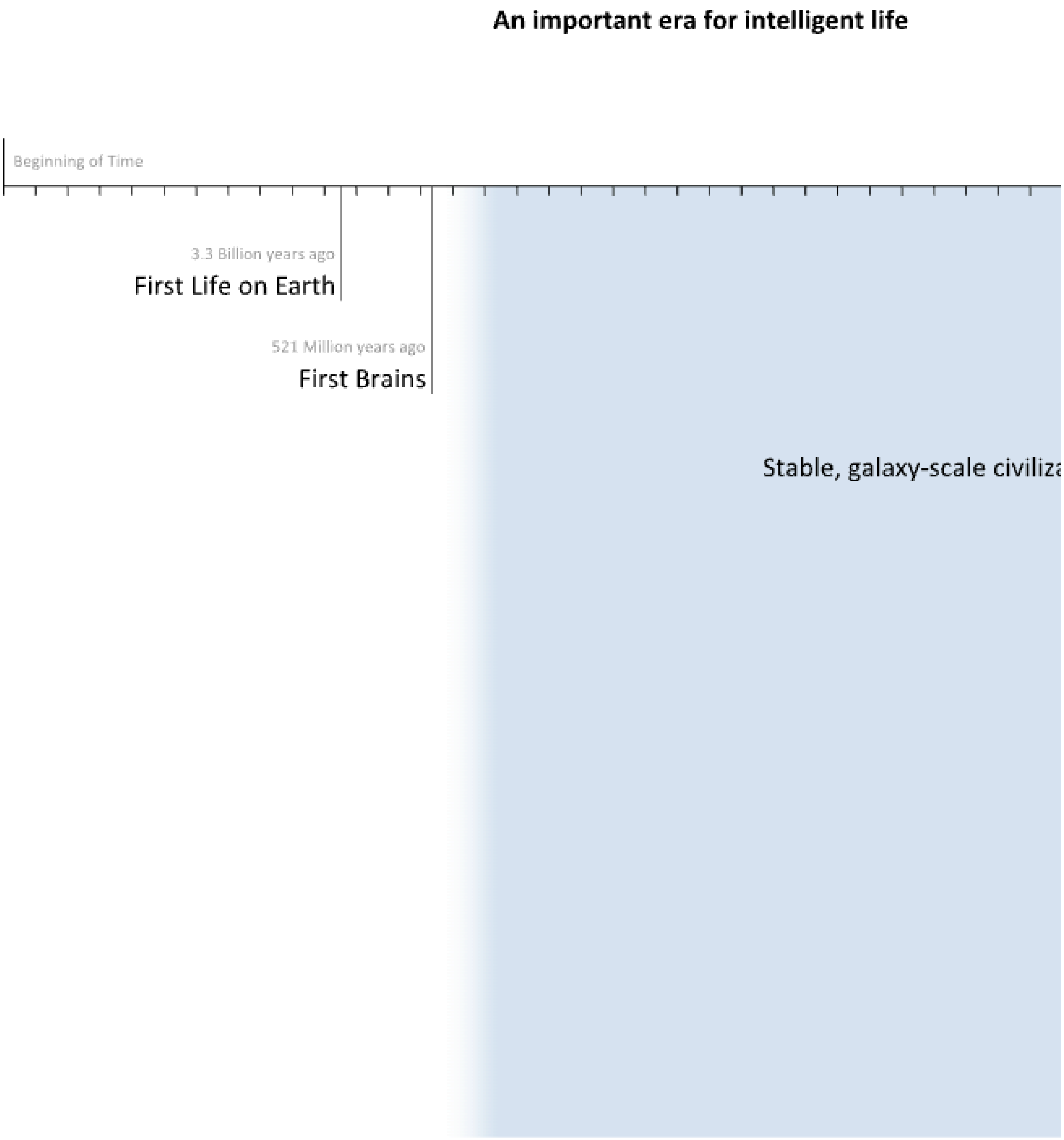


More at [This Can't Go On](https://www.cold-takes.com/this-cant-go-on/)

We're currently living through the fastest-growing time in history. This rate of growth hasn't gone on long, and can't go on indefinitely (there aren't enough atoms in the galaxy to sustain this rate of growth for even another 10,000 years). And if we get *further acceleration* in this rate of growth - in line with historical acceleration - we could reach the limits of what's possible more

To recap:

- The last few millions of years - with the start of our species - have been more eventful than the previous several billion.
- The last few hundred years have been more eventful than the previous several million.
- If we see another accelerator (as I think AI could be), the next few decades could be the most eventful of all.



More info about these timelines at All Possible Views About Humanity's Future Are Wild, This Can't Go On, and **Forecasting Transformative AI: Biological Anchors**, respectively.

Cold Takes - The "most important century" blog post series

Given the times we live in, we need to be open to possible ways in which the world could change quickly and radically. Ideally, we'd be a bit *over*-attentive to such things, like putting safety first when driving. But today, such possibilities get little attention.

Key pieces:

- All Possible Views About Humanity's Future Are Wild
- This Can't Go On

The long-run future is radically unfamiliar

Technology tends to increase people's control over the environment. For a concrete, easy-to-visualize example of what things could look like if technology goes far enough, we might imagine a technology like "digital people": fully conscious people "made out of software" who inhabit virtual environments such that they can experience anything at all and can be copied, run at different speeds and even "reset."

A world of digital people could be radically dystopian (virtual environments used to entrench some people's absolute power over others) or utopian (no disease, material poverty or non-consensual violence, and far greater wisdom and self-understanding than is possible today). Either way, digital people could enable a civilization to spread throughout the galaxy and last for a long time.

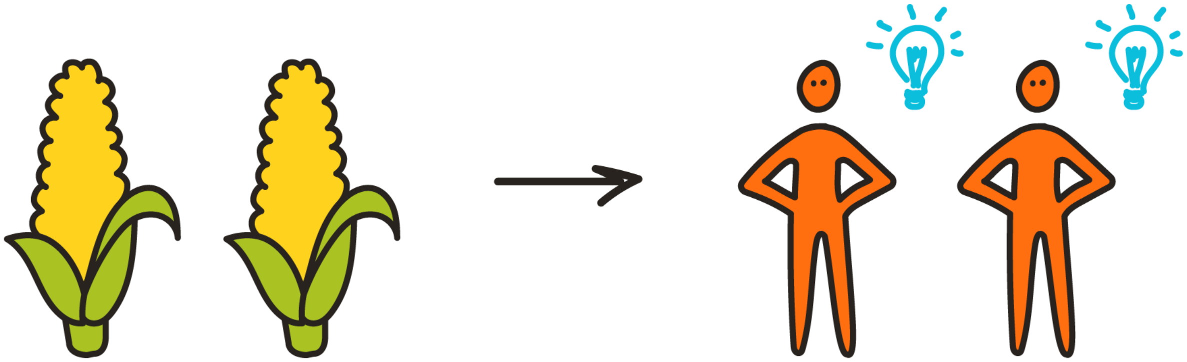
Many people think this sort of large, stable future civilization is where we could be headed eventually (whether via digital people or other technologies that increase control over the environment), but don't bother to discuss it because it seems so far off.

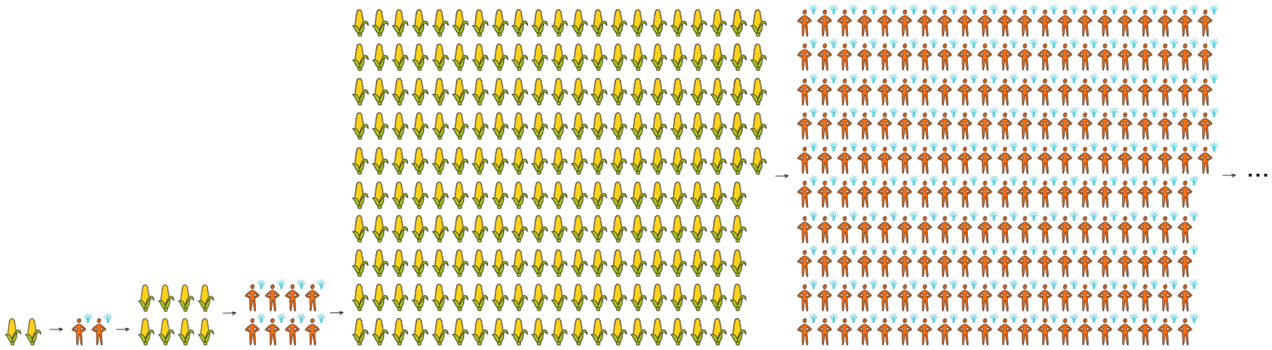
Key piece: Digital People Would Be An Even Bigger Deal

The long-run future could come much faster than we think

Standard economic growth models imply that **any technology that could fully automate innovation would cause an "economic singularity":** productivity going to infinity this century. This is because it would create a powerful feedback loop: more resources -> more ideas and innovation -> more resources -> more ideas and innovation ...

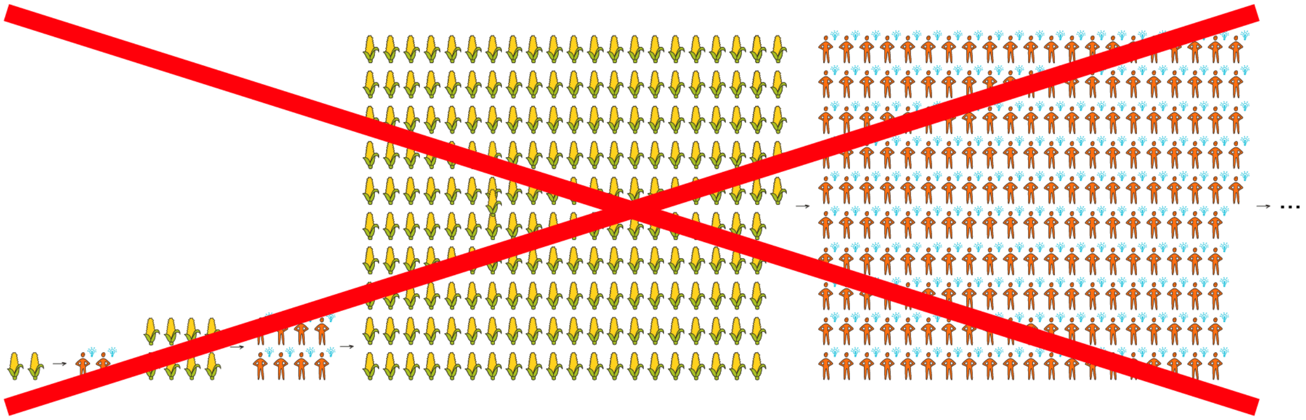
This loop would not be unprecedented. I think it is in some sense the "default" way the economy operates - for most of economic history up until a couple hundred years ago.





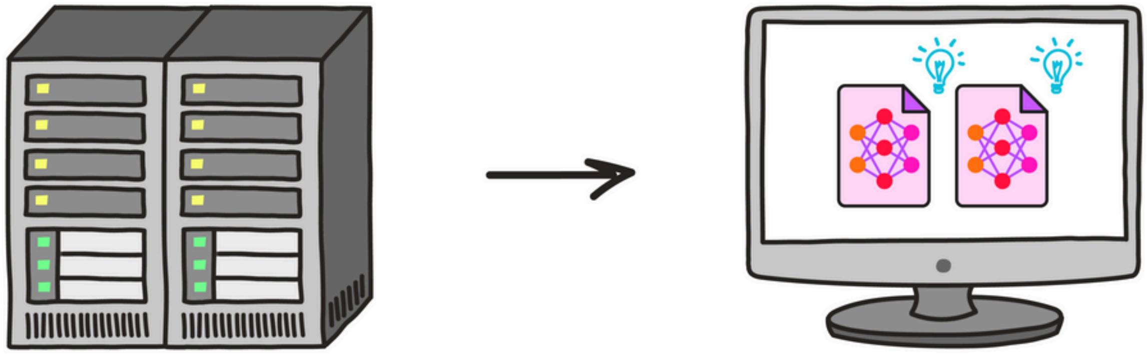
Cold Takes - The "most important century" blog post series

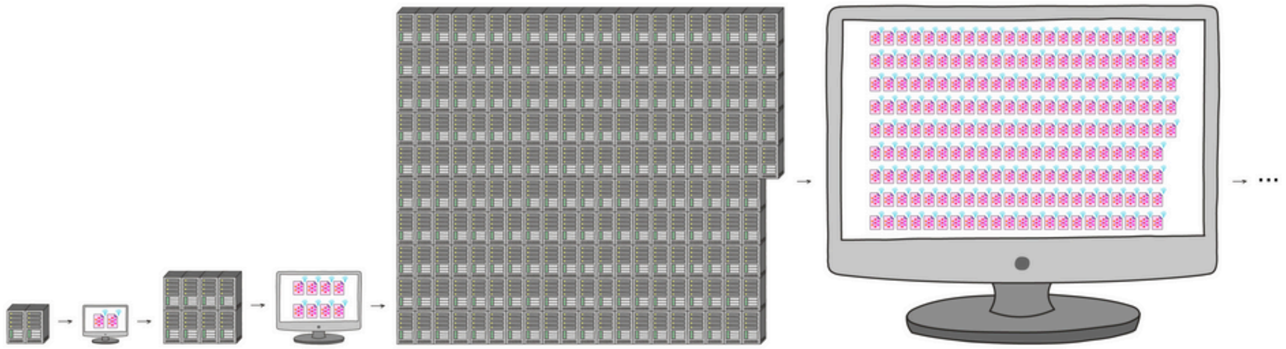
But in the "demographic transition" a couple hundred years ago, the "more resources -> more people" step of that loop stopped. Population growth leveled off, and more resources led to richer people instead of more people:



Today's economy: more resources -> ~~more~~ richer people -> same pace of ideas -> ...

The feedback loop could come back if some other technology restored the "more resources -> more ideas" dynamic. One such technology could be the right kind of AI: what I call PASTA, or Process for Automating Scientific and Technological





Possible future: more resources -> more AIs -> more ideas -> more resources ...

That means that **our radical long-run future could be upon us very fast** after PASTA is developed (if it ever is).

It also means that if PASTA systems are *misaligned* - pursuing their own non-human-compatible objectives - things could very quickly go sideways.

Key pieces:

- The Duplicator: Instant Cloning Would Make the World Economy Explode
- Forecasting Transformative AI, Part 1: What Kind of AI?

PASTA looks like it will be developed this century

It's not controversial to say a highly general AI system, such as PASTA, would be momentous. The question is, when (if ever) will such a thing exist?

Over the last few years, a team at Open Philanthropy has investigated this question from multiple angles.

One forecasting method observes that:

- No AI model to date has been even 1% as "big" (in terms of computations performed) as a human brain, and until recently this wouldn't have been affordable - but that will change relatively soon.
- And by the end of this century, it will be affordable to train enormous AI models many times over; to train human-brain-sized models on enormously difficult, expensive tasks; and even perhaps to perform as many computations as have been done "by evolution" (by all animal brains in history to date).

This method's predictions are in line with the latest survey of AI researchers: something like PASTA is more likely than not this century.

A number of other angles have been examined as well.

One challenge for these forecasts: there's **no "field of AI forecasting"** and no expert consensus comparable to the one around climate change.

It's hard to be confident when the discussions around these topics are small and limited. But I think we should take the "most important century" hypothesis seriously based on what we know now, until and unless a "field of AI forecasting" develops.

- AI Timelines: Where the Arguments, and the "Experts," Stand (recaps the others, and discusses how we should reason about topics like this where it's unclear who the "experts" are)
- Forecasting Transformative AI: What's the Burden of Proof?
- Are we "trending toward" transformative AI?
- Forecasting transformative AI: the "biological anchors" method in a nutshell

We're not ready for this

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 people to come - if we can make sense of the situation enough to find helpful actions.

But that's a big "if." Many things we can do might make things better or worse (and it's hard to say which).

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 ignorant about what to do next.

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."

- If you're convinced by the arguments in this series, 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.
- For those looking for a quick action that will make future action more likely, see this section of "Call to Vigilance."

Key pieces:

- Making the Best of the Most Important Century
- Call to Vigilance.

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:

And every time I read commentary on what's going on in the world, people are 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.

I don't know where we're actually heading, or what we can do about it. But I feel pretty solid in saying that we as a civilization are not ready for what's coming, and we need to start by taking it more seriously.

SUBSCRIBE FEEDBACK

Acknowledgements

I have few-to-no claims to originality. The vast bulk of the claims, observations and insights in this series came from some combination of:

- Years of discussions with others, particularly in the effective altruism and rationalist communities. It's hard to trace specific ideas to specific people within this context, but I know that a huge amount of my thinking comes at least proximately from Carl Shulman, Dario Amodei and Paul Christiano, and that Nick Bostrom's and Eliezer Yudkowsky's work has been very influential generally. (I also understand that earlier futurists and transhumanists influenced these people and communities, though I haven't engaged directly much with their works.)
- In-depth analyses by the Open Philanthropy Longtermist Worldview Investigations team: Ajeya Cotra and Tom Davidson (especially) as well as Nick Beckstead, Joe Carlsmith, and David Roodman. I've also drawn heavily on reports by Katja Grace and Luke Muehlhauser.

- Ajeya Cotra, María Gutiérrez Rojas and Ludwig Schubert for help with visualizations.
- A number of people for feedback on earlier drafts:
  - My sister Daliya Karnofsky, my wife Daniela Amodei, and Elie Hassenfeld: special thanks for reading the earliest (least readable) drafts and often giving detailed feedback on multiple iterations.
  - People who served as "beta readers" and gave significant amounts of feedback, particularly on what was and wasn't making sense for them: Alexander Berger, Damon Binder, Lukas Gloor, Derek Hopf, Mike Levine, Eli Nathan, Sella Nevo, Julian Sancton, Simon Shifrin, Tracy Williams. (Plus a number of people already mentioned above.)

Powered by Ghost