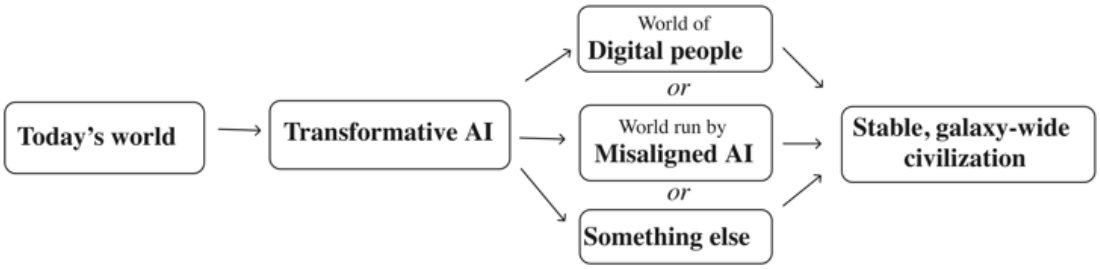
The Most Important Century

The Most Important Century: Sequence Introduction

[*Moderators' note: with permission, I am crossposting Holden Karnofsky's* Most Important Century *series, originally published on his blog. In my (Ruby's) opinion, this may be the most compelling extant write-up that argues we are living in an exceptionally important time.*

Posts in the series will be published every 4th day until we are caught up, then will be posted here as they are published. Links in this roadmap will be updated as posts are published to LessWrong. Eager readers are encouraged to "read ahead" in the original.]



This is a roadmap/"extended table of contents" for a series of posts arguing for a good chance that we're in the most important century of all time.

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.

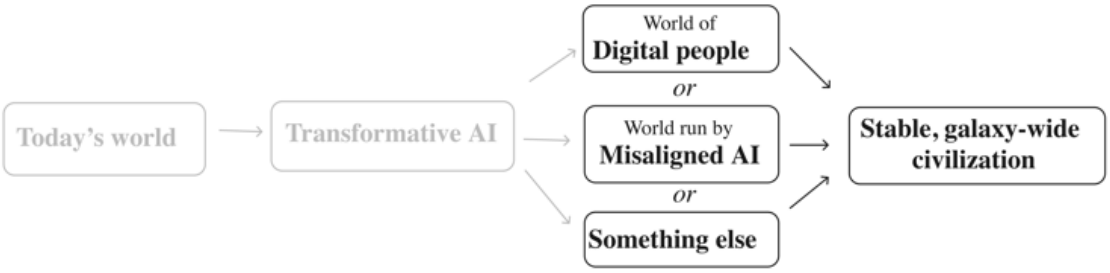
I think it's very unclear whether this would be a good or bad thing. What matters is that it could go a lot of different ways, and we have a chance to affect that.

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.

Roadmap

Our wildly [2] important era

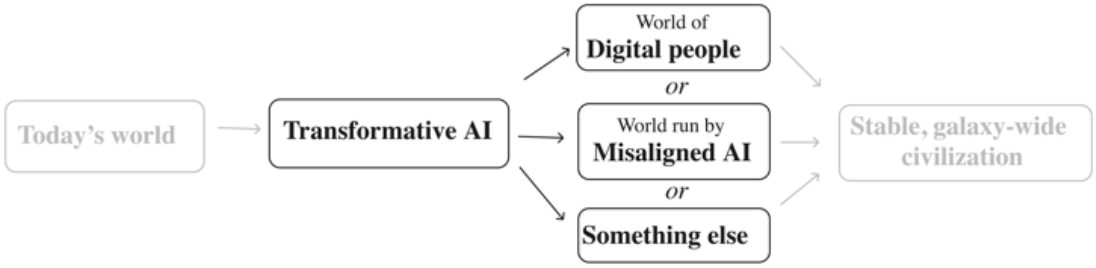


All possible views about humanity's long-term future are wild argues that two simple observations - (a) it appears likely that we will *eventually* be able to spread throughout the galaxy, and (b) it doesn't seem any other life form has done that yet - are sufficient to make the case that we live in 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



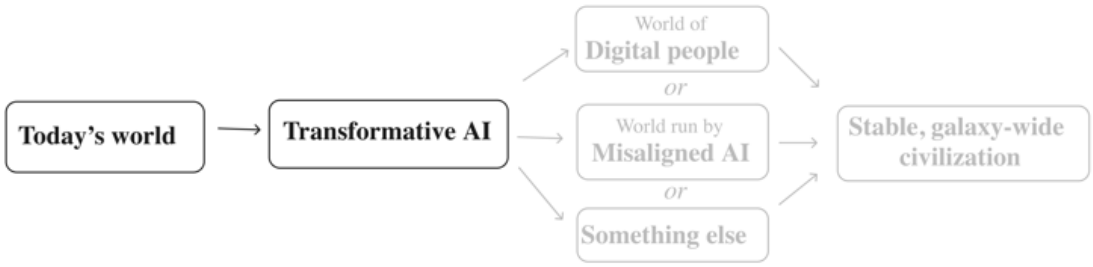
This Can't Go On looks at economic growth and scientific advancement over the course of human history. Over the last few generations, growth has been pretty steady. But zooming out to a longer time frame, it seems that growth has greatly accelerated recently; is near its historical high point; and is 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? (not yet posted on LW) 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.

Forecasting transformative AI this century



Forecasting Transformative AI: What's the Burden of Proof? (not yet posted on LW) 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? (not yet posted on LW) 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.

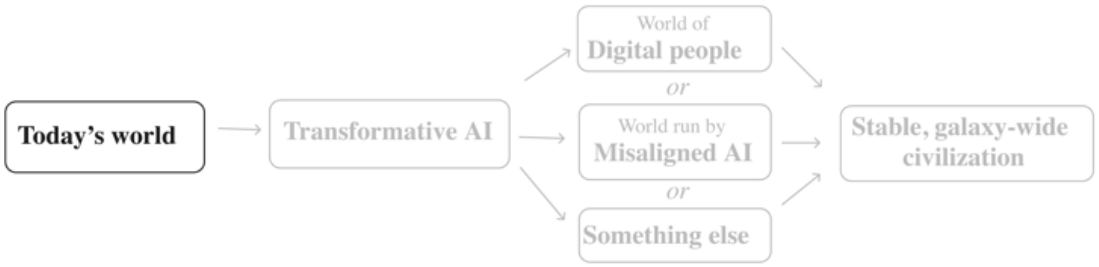
Forecasting transformative AI: the "biological anchors" method in a nutshell (not yet published) summarizes the biological anchors framework for forecasting AI. This framework is the main factor in my specific forecasts.

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 Forecasting Expertise (not yet published) 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.

Wrapping up



Implications of living in the most important century (not published yet) discusses what we can do to help the most important century go as well as possible.

The Most Important Century in a Nutshell (not published yet) will summarize the series in a few pages.

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:

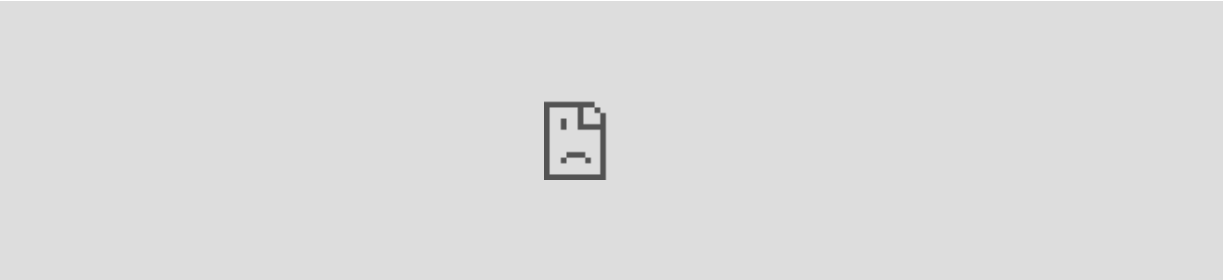
- Years of discussions with others, particularly in the effective altruism and rationalist communities. It's hard to trace specific ideas to specific people within this context, but I know that a huge amount of my thinking comes at least proximately from Carl Shulman, Dario Amodei and Paul Christiano, 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.

In addition, I owe thanks to:

- Ajeya Cotra, María Gutiérrez Rojas and Ludwig Schubert and 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.)

All Possible Views About Humanity's Future Are Wild

This is the first post in the Most Important Century sequence. For more info and a roadmap for the series, see the sequence introduction.



Audio also available by searching Stitcher, Spotify, Google Podcasts, etc. for "Cold Takes Audio"

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.

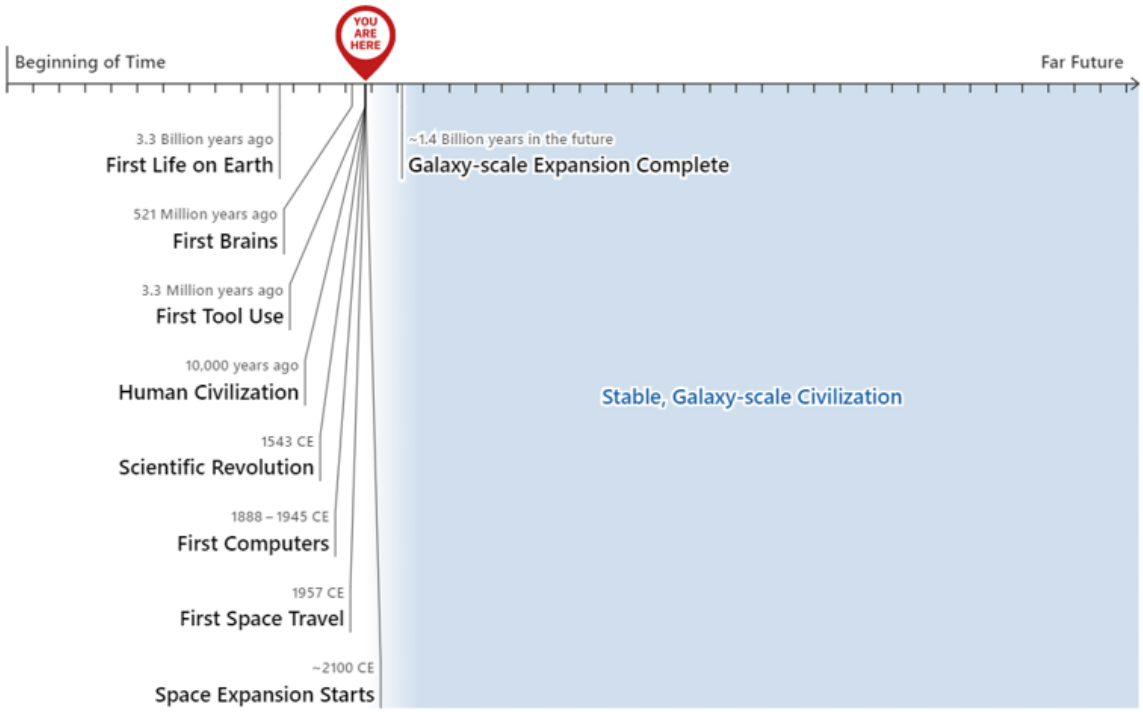
My view

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"[1] 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 a future piece that there is a chance of "value lock-in" here: 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 years.[2]

If that ends up happening, you might think of the story of our galaxy[3] 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 axis.[4]

??? 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."[5]

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.

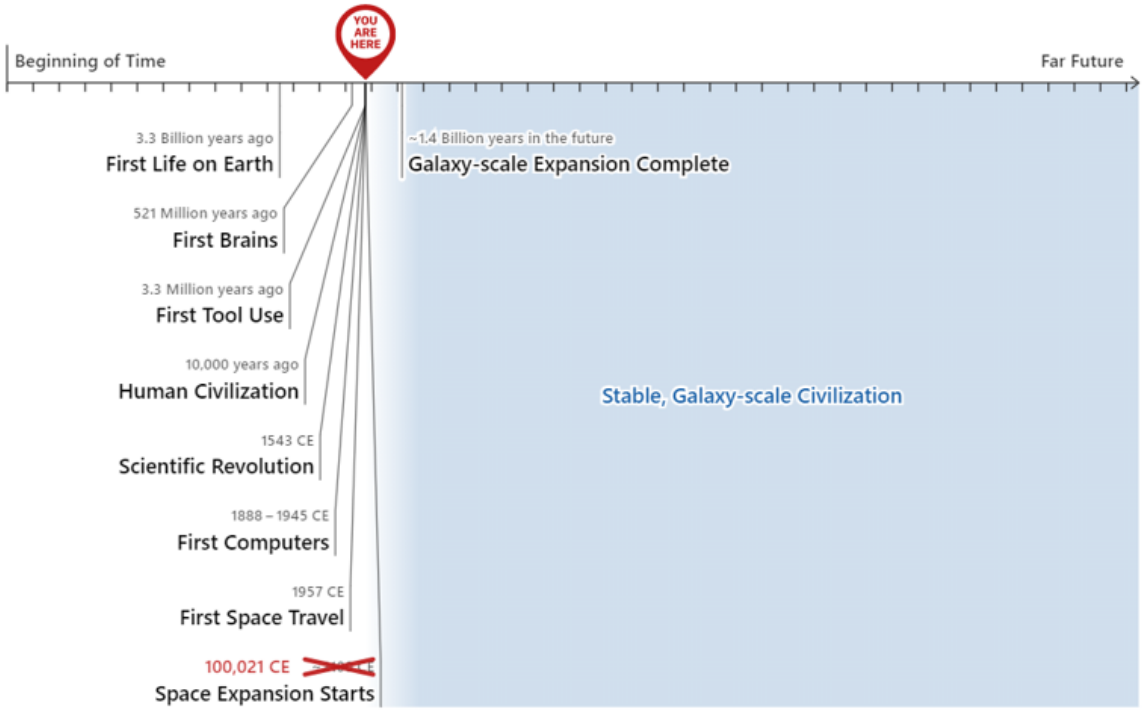
A key part of my view (which I'll write about more later) is that within this century, we could develop advanced enough AI to start a productivity explosion. Say you don't believe that.

- 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 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.[6]
- 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,[7] 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.

- 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 effectively *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,[8] 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.

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.[9]

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.

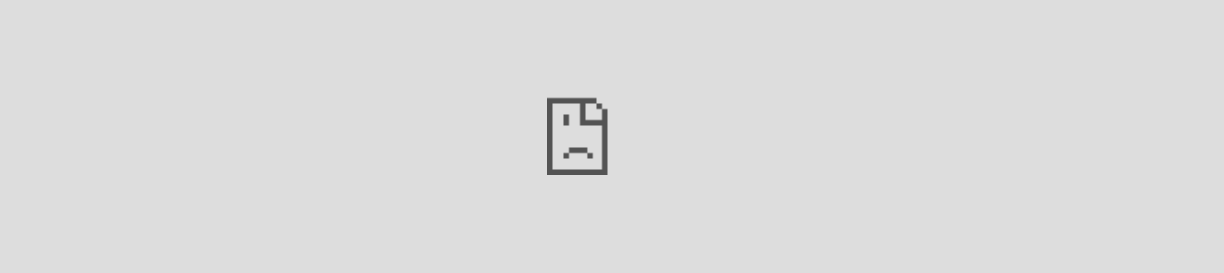
It looks for all the world as though our "tiny dot" has a real shot at being the origin of a galaxy-scale civilization. It seems absurd, even delusional to believe in this possibility. But given our observations, it seems equally strange to dismiss it.

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]"

Use this feedback form if you have comments/suggestions you want me to see, or if you're up for giving some quick feedback about this post (which I greatly appreciate!)

The Duplicator: Instant Cloning Would Make the World Economy Explode



Audio also available by searching Stitcher, Spotify, Google Podcasts, etc. for "Cold Takes Audio"

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.

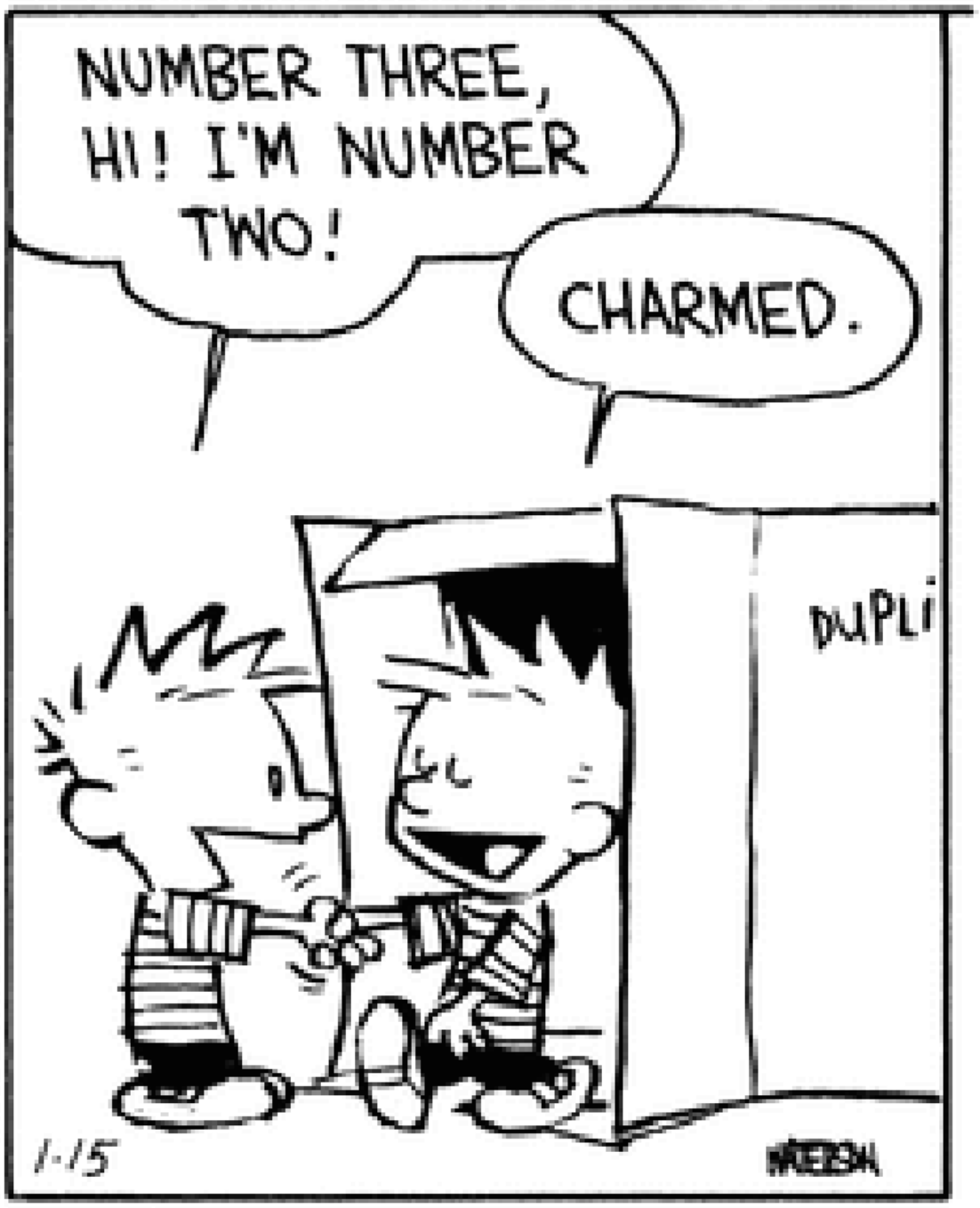
The future I picture is enormously bigger, faster, weirder, and either much much better or much much worse compared to today. It's also potentially a lot *sooner* than sci-fi futures:[1] I think particular, achievable-seeming technologies could get us there quickly.

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.[2]

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").[3] 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).[4]

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.[5] 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 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.

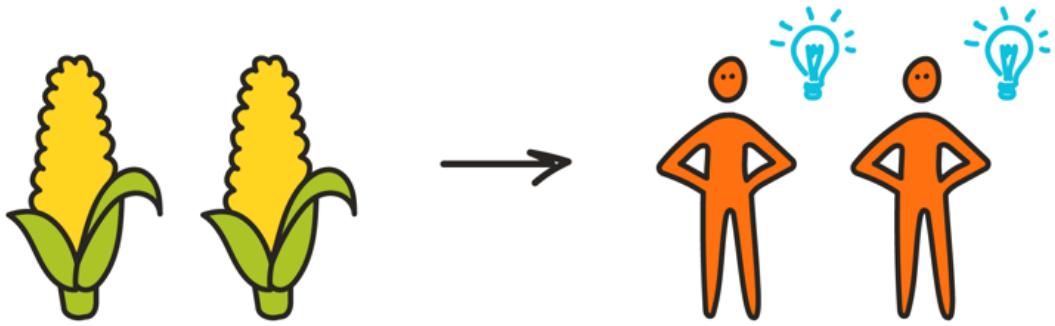
The total population (number of people) affects both labor and productivity, because people can have ideas that increase productivity.

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

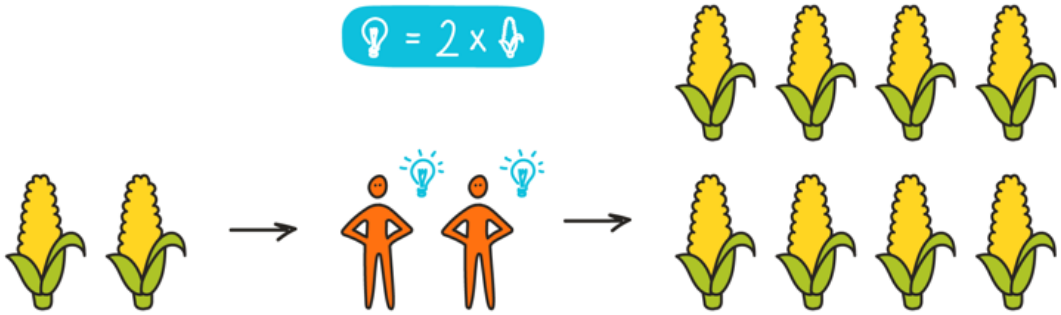
The economy starts with some set of resources (capital) supporting some set of people (population).

Thanks to María Gutiérrez Rojas for these graphics.

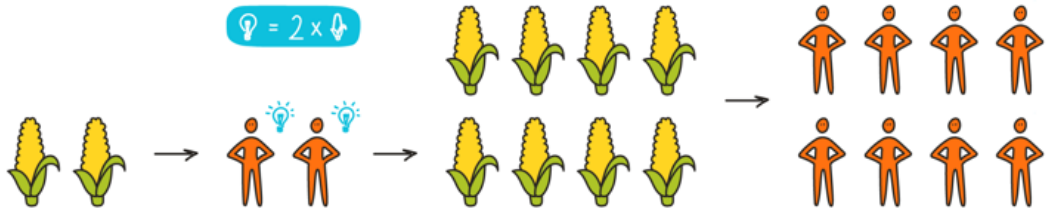
This set of people comes up with new ideas and innovations.



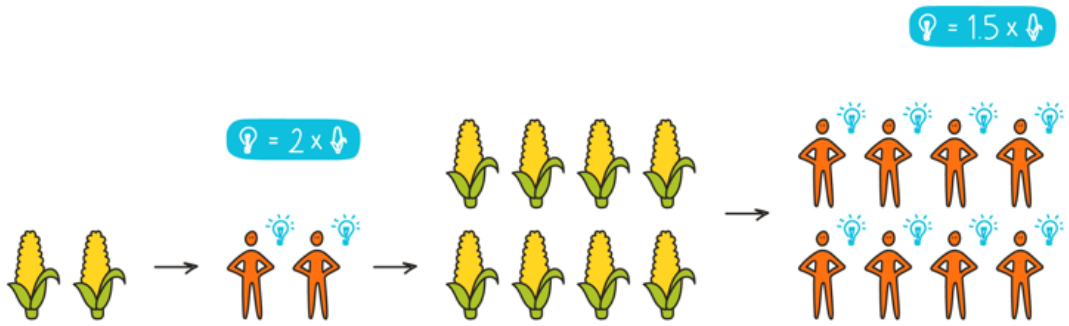
This leads to some amount of increased productivity, meaning there is more total economic output.[6]



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 innovations *faster* than before (since more people means more new ideas).[7]



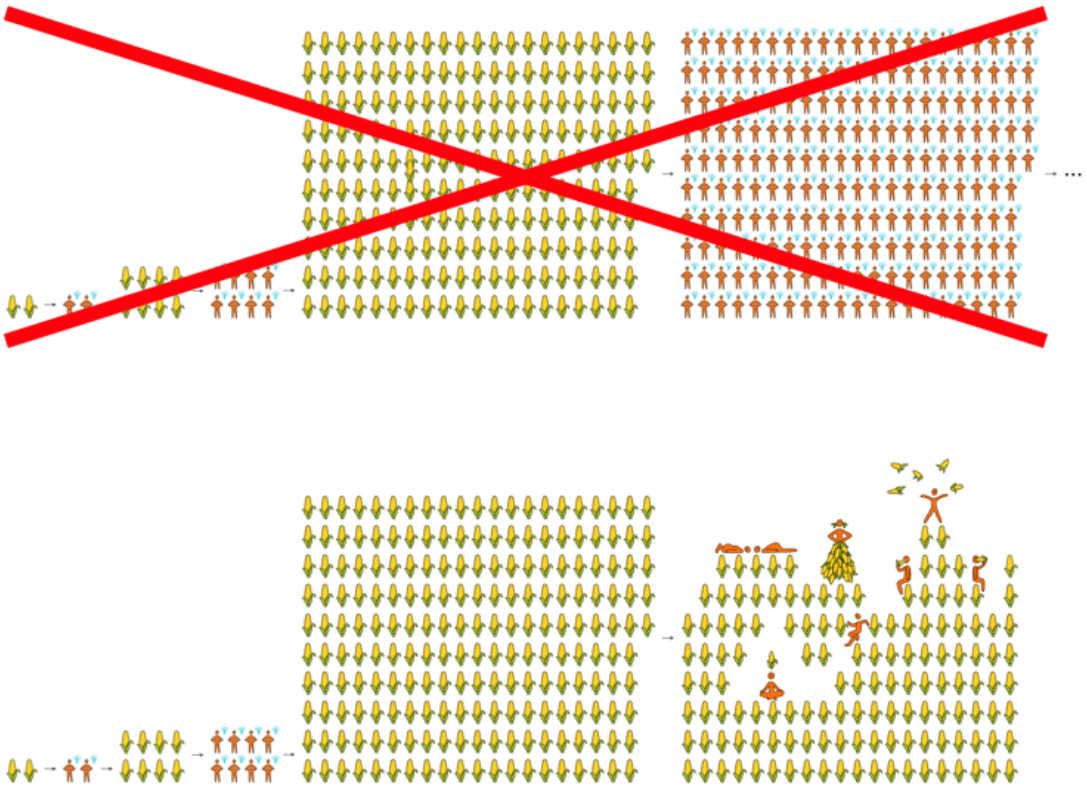
This leads to *even more* economic output and *even faster* population growth, in a self-reinforcing loop: *more ideas → more output → more people → more ideas→ ....*

When you incorporate this full feedback loop into economic growth models,[8] they predict that (under plausible assumptions) the world economy will see **accelerating growth**.[9] "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.

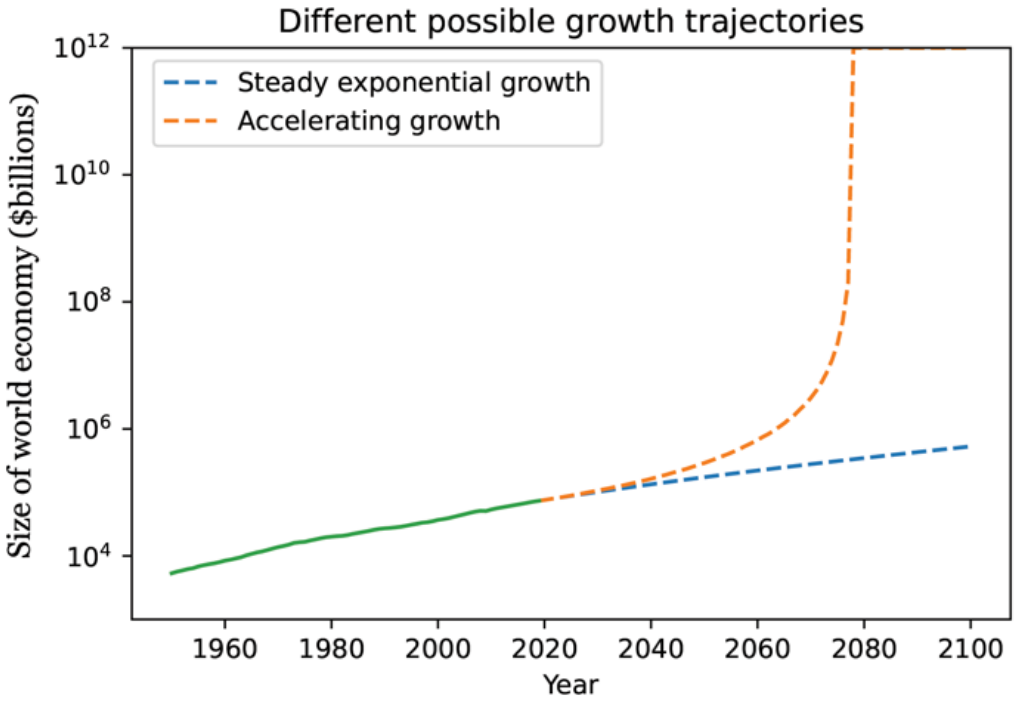


The demographic transition.

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 → ~~more~~ richer people ~~→ more ideas→~~	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.[10]

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.**[11]

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 guns[12] 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.

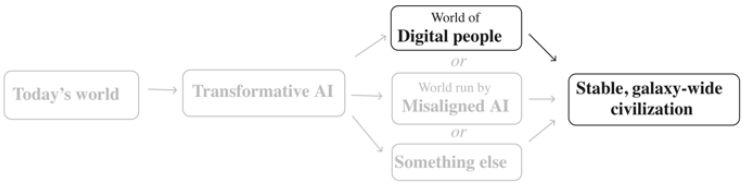
A harder-to-intuit, but even more powerful, technology would be **digital people**, e.g., the ability to run detailed simulations of people[13] on a computer. Such simulated people could be copied Duplicator-style, and could also be sped up, slowed down, and reset, with virtual environments that were 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.

Use this feedback form if you have comments/suggestions you want me to see, or if you're up for giving some quick feedback about this post (which I greatly appreciate!)

Digital People Would Be An Even Bigger Deal

Audio also available by searching Stitcher, Spotify, Google Podcasts, etc. for "Cold Takes Audio"



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 nonfiction, 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 example[1] 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.[2]) 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.)

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
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

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. Here is the list of questions discussed in the FAQ:

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 this graphic, 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)[3] as well as slowed down (to save on costs). This could further increase both speed and coordinating ability.[4]

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).[5] 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).[6]

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.[7] 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:

- Anyone could make a copy of themselves to try out meditation, perhaps even dedicating themselves to it for several years (possibly sped-up).[8] 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.[9])

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 racism 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.[10] 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 include 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"[11] - 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.[12]

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).
- 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 here are some informal notes from physicist Jess Riedel.)

Would these impacts be a good or bad thing?

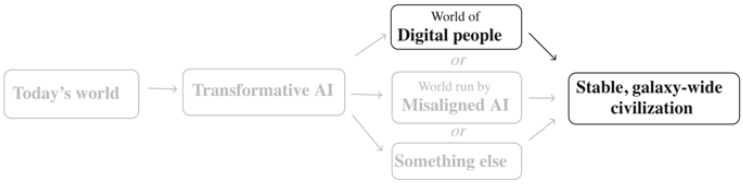
Throughout this piece, I imagine many readers have been thinking "That sounds terrible! Does the author think it would be good?" Or "That sounds great! Does the author disagree?"

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.

Digital People FAQ

Audio also available by searching Stitcher, Spotify, Google Podcasts, etc. for "Cold Takes Audio"



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
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

Table of contents for this FAQ

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.[1]

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:

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:[2]

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,[3] 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.

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.**[4]

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.[5]

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.

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.[6]

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.[8] (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.[9])

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).[10] 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.[11]

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.

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.[7] 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.
- 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, imagine 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.[12]

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.

What other questions can I ask?

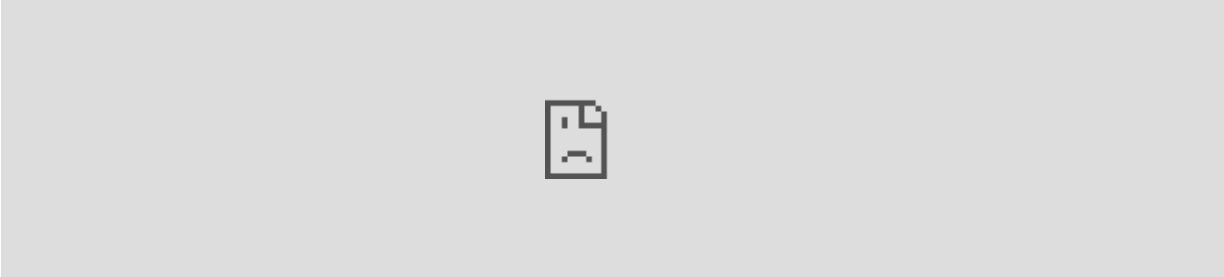
So many more! E.g.: https://tvtropes.org/pmwiki/pmwiki.php/Analysis/BrainUploading

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.

This Can't Go On



Audio also available by searching Stitcher, Spotify, Google Podcasts, etc. for "Cold Takes Audio"

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:

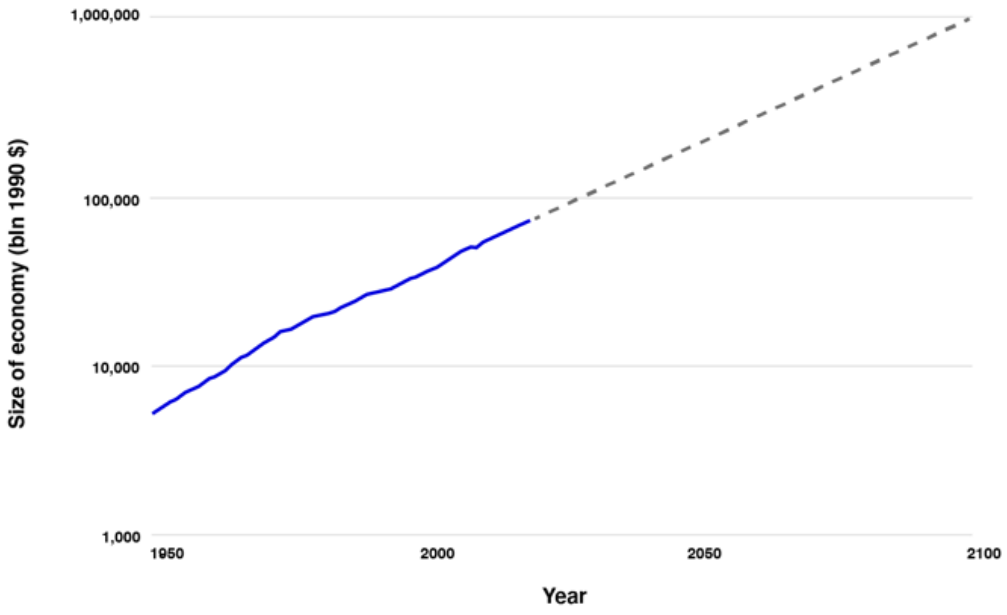
- We're used to the world economy growing a few percent per year. This has been the case for many generations.
- However, this is a very unusual situation. Zooming out to all of history, we see that growth has been accelerating; that it's near its historical high point; and that it's faster than it can 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 grown[1] a few percent per year, on average. Some years see more or less growth than other years, but growth is pretty steady overall.[2] I'll call this the **Business As Usual** world.

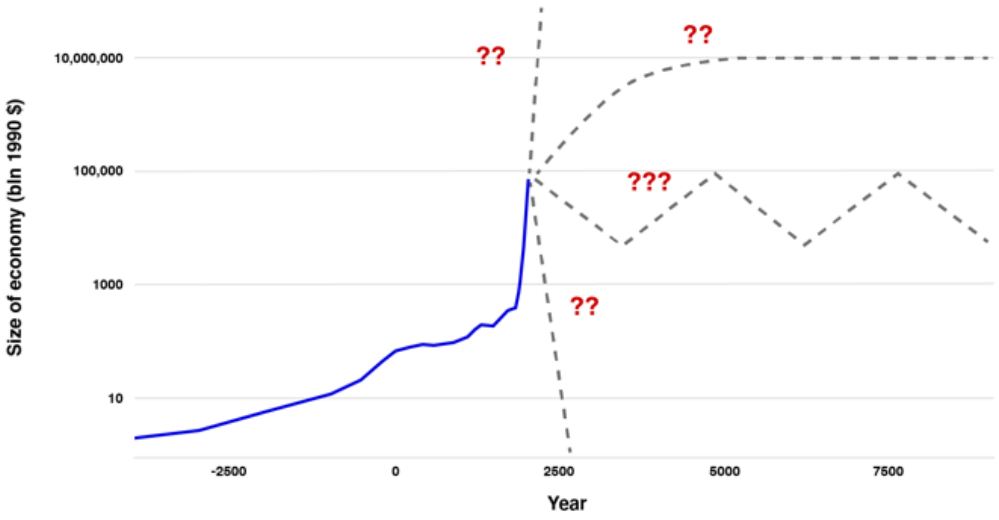
In Business As Usual, the world is constantly changing, and the change is noticeable, but it's not overwhelming or impossible to keep up with. There is a constant stream of new opportunities and new challenges, but if you want to take a few extra years to adapt to them while you mostly do things the 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.[3]

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:



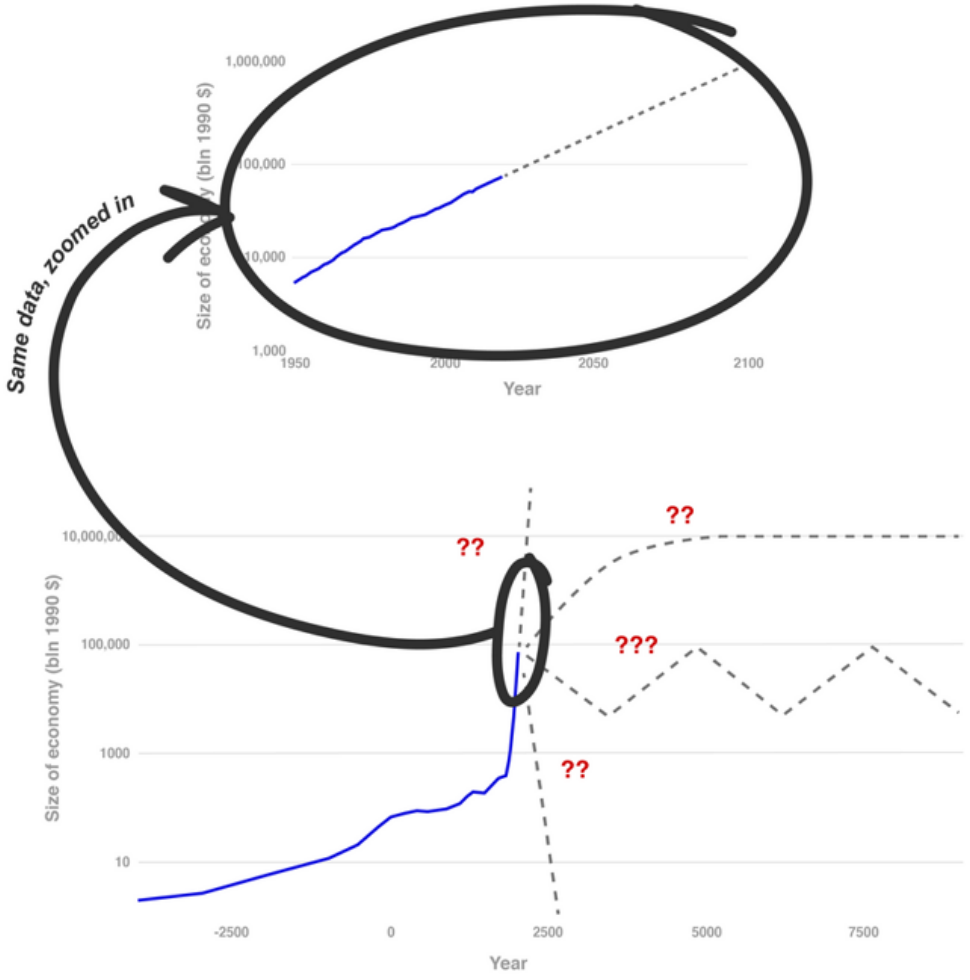
Business As Usual

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:



This Can't Go On[4]

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.**



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.[5] This implies that the economy would be doubling in size about every 35 years.[6]
- If this holds up, then 8200 years from now, the economy would be about $3*10^{70}$ times its current size.
- There are likely fewer than 10^{70} atoms in our galaxy,[7] which we would not be able to travel beyond within the 8200-year time frame.[8]
- So if the economy were $3*10^{70}$ 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.[9])

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.[10] 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.[11] 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 limits[12]), 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.

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 bioweapons), 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.[13]
- 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.

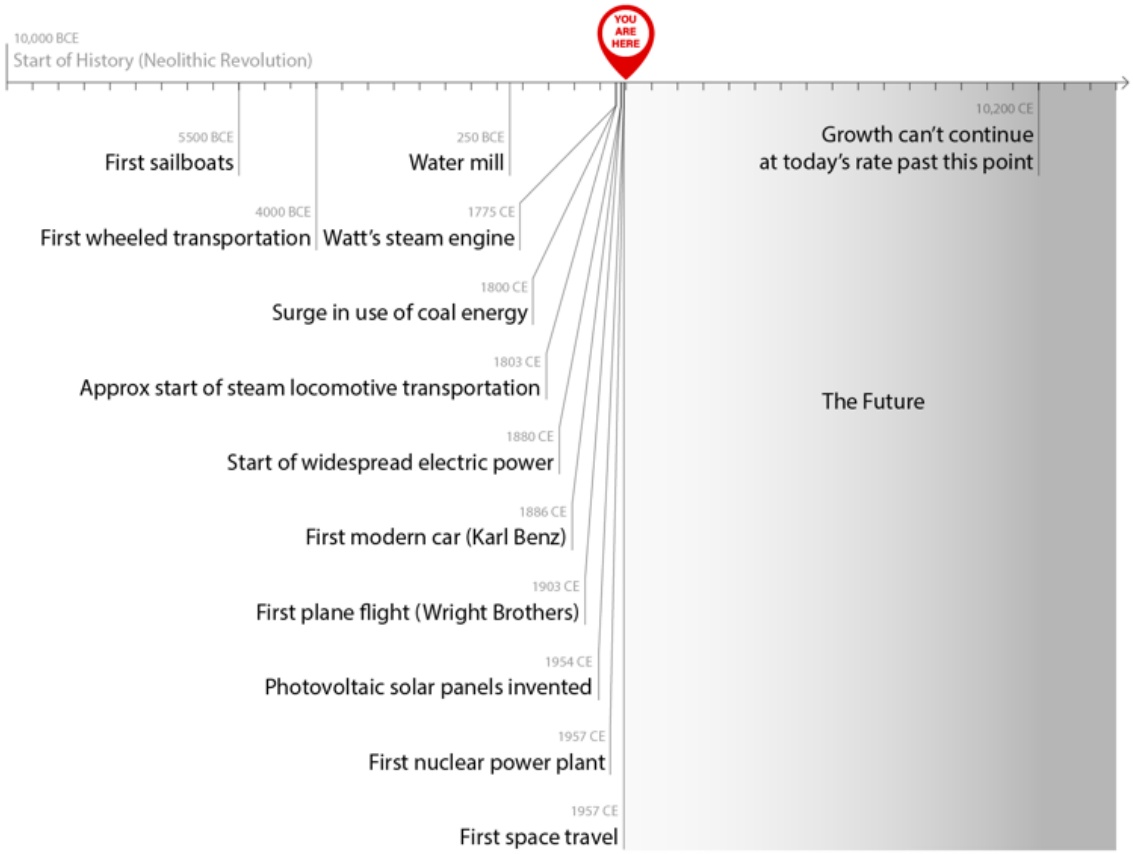
So it seems like this is a quite remarkable century, with some chance of being the most remarkable. This is all based on pretty basic observations, not detailed reasoning about AI (which I will get to in future pieces).

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.

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,[14] and I believe there were, in fact, more scientific and technological advances in the last ~500 years than the previous several million.[15]



> *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.[16]
- **Collapse** due to some global catastrophe.

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.

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."

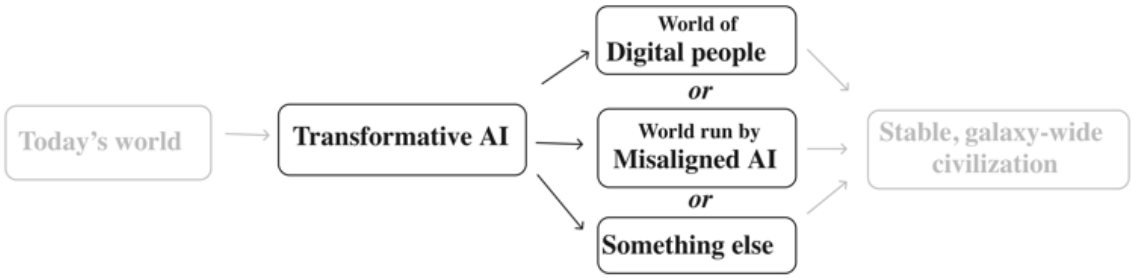
Thanks to María Gutiérrez Rojas for the graphics in this piece, and Ludwig Schubert for an earlier timeline graphic that this piece's timeline graphic is based on.

Use this feedback form if you have comments/suggestions you want me to see, or if you're up for giving some quick feedback about this post (which I greatly appreciate!)

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

PASTA: Process for Automating Scientific and Technological Advancement.

Audio also available by searching Stitcher, Spotify, Google Podcasts, etc. for "Cold Takes Audio"



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.**[1]

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.[2]

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***.[3] (I mean PASTA to refer to either a single system or a collection of systems that can collectively do this sort of automation.)

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.

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 instructions[4] along the lines of:

- 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,[5] 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 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.[6] 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."[7]
- 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, 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 cheaper[8] and better-funded. I'll argue in future pieces that the odds of developing PASTA in the coming decades are substantial.

Impacts of PASTA

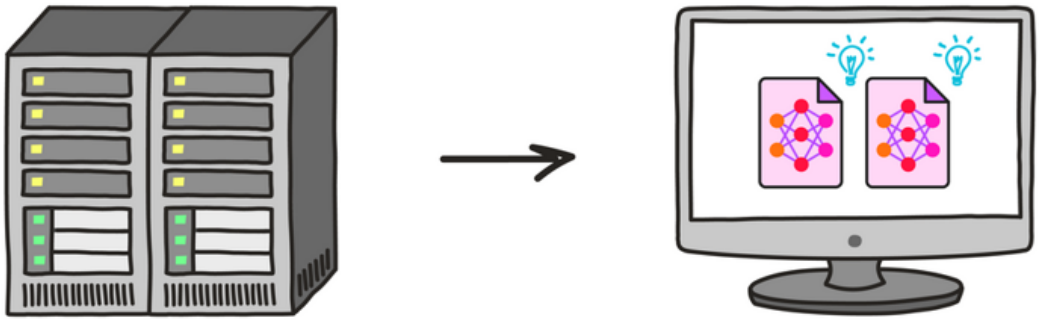
Explosive scientific and technological advancement

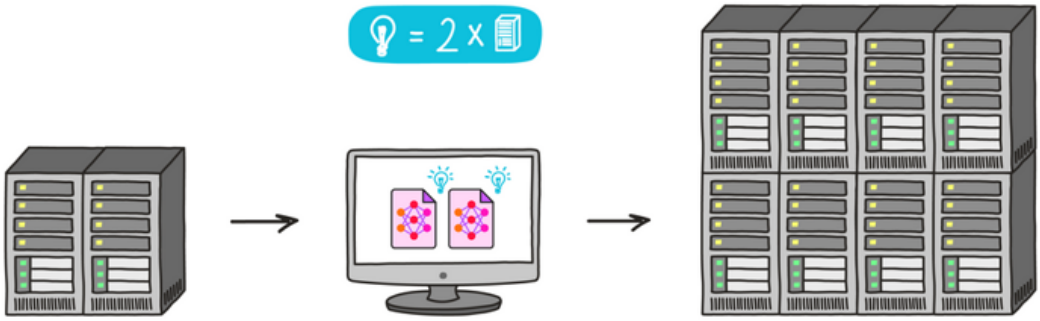
I've previously talked about the idea of a potential explosion in scientific and technological advancement, which could lead to a radically unfamiliar future.

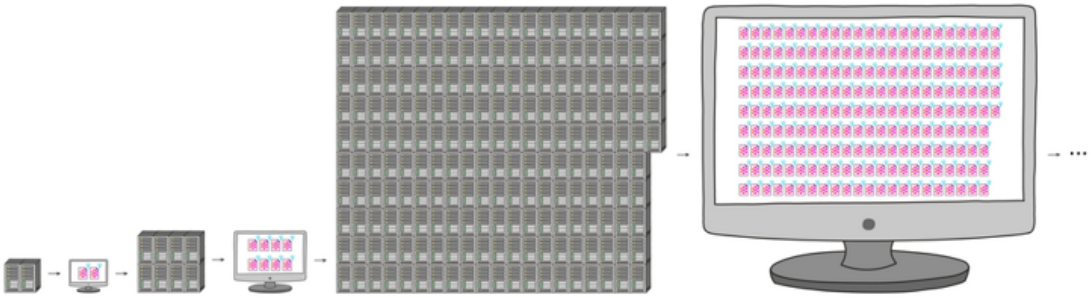
I've emphasized that such an explosion could be caused by a technology that "dramatically increased the number of 'minds' (humans, or digital people, or advanced AIs) pushing forward scientific and technological advancement."

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→..."







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 doing[9] (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).

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

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.

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 outnumber 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 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.[10]

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.