Machines of Loving Grace

How AI Could Transform the World for the Better

October 2024

I think and talk a lot about the risks of powerful AI. The company I'm the CEO of, Anthropic, does a lot of research on how to reduce these risks. Because of this, people sometimes draw the conclusion that I'm a pessimist or "doomer" who thinks AI will be mostly bad or dangerous. I don't think that at all. In fact, one of my main reasons for focusing on risks is that they're the only thing standing between us and what I see as a fundamentally positive future. **I think that most people are underestimating just how radical the upside of AI could be**, just as I think most people are underestimating how bad the risks could be.

In this essay I try to sketch out what that upside might look like—what a world with powerful AI might look like if everything goes *right*. Of course no one can know the future with any certainty or precision, and the effects of powerful AI are likely to be even more unpredictable than past technological changes, so all of this is unavoidably going to consist of guesses. But I am aiming for at least educated and useful guesses, which capture the flavor of what will happen even if most details end up being wrong. I'm including lots of details mainly because I think a concrete vision does more to advance discussion than a highly hedged and abstract one.

First, however, I wanted to briefly explain why I and Anthropic haven't talked that much about powerful AI's upsides, and why we'll probably continue, overall, to talk a lot about risks. In particular, I've made this choice out of a desire to:

- **Maximize leverage**. The basic development of AI technology and many (not all) of its benefits seems inevitable (unless the risks derail everything) and is fundamentally driven by powerful market forces. On the other hand, the risks are not predetermined and our actions can greatly change their likelihood.
- **Avoid perception of propaganda**. AI companies talking about all the amazing benefits of AI can come off like propagandists, or as if they're attempting to distract

from downsides. I also think that as a matter of principle it's bad for your soul to spend too much of your time "talking your book".

- **Avoid grandiosity**. I am often turned off by the way many AI risk public figures (not to mention AI company leaders) talk about the post-AGI world, as if it's their mission to single-handedly bring it about like a prophet leading their people to salvation. I think it's dangerous to view companies as unilaterally shaping the world, and dangerous to view practical technological goals in essentially religious terms.
- **Avoid "sci-fi" baggage**. Although I think most people underestimate the upside of powerful AI, the small community of people who do discuss radical AI futures often does so in an excessively "sci-fi" tone (featuring e.g. uploaded minds, space exploration, or general cyberpunk vibes). I think this causes people to take the claims less seriously, and to imbue them with a sort of unreality. To be clear, the issue isn't whether the technologies described are possible or likely (the main essay discusses this in granular detail)—it's more that the "vibe" connotatively smuggles in a bunch of cultural baggage and unstated assumptions about what kind of future is desirable, how various societal issues will play out, etc. The result often ends up reading like a fantasy for a narrow subculture, while being off-putting to most people.

Yet despite all of the concerns above, I really do think it's important to discuss what a good world with powerful AI could look like, while doing our best to avoid the above pitfalls. In fact I think it is critical to have a genuinely inspiring vision of the future, and not *just* a plan to fight fires. Many of the implications of powerful AI are adversarial or dangerous, but at the end of it all, there has to be something we're fighting *for*, some positive-sum outcome where everyone is better off, something to rally people to rise above their squabbles and confront the challenges ahead. Fear is one kind of motivator, but it's not enough: we need hope as well.

The list of positive applications of powerful AI is extremely long (and includes robotics, manufacturing, energy, and much more), but I'm going to focus on a small number of areas that seem to me to have the greatest potential to directly improve the quality of human life. The five categories I am most excited about are:

1. Biology and physical health

2. Neuroscience and mental health
3. Economic development and poverty
4. Peace and governance
5. Work and meaning

My predictions are going to be radical as judged by most standards (other than sci-fi "singularity" visions), but I mean them earnestly and sincerely. Everything I'm saying could very easily be wrong (to repeat my point from above), but I've at least attempted to ground my views in a semi-analytical assessment of how much progress in various fields might speed up and what that might mean in practice. I am fortunate to have professional experience in both biology and neuroscience, and I am an informed amateur in the field of economic development, but I am sure I will get plenty of things wrong. One thing writing this essay has made me realize is that it would be valuable to bring together a group of domain experts (in biology, economics, international relations, and other areas) to write a much better and more informed version of what I've produced here. It's probably best to view my efforts here as a starting prompt for that group.

Basic assumptions and framework

To make this whole essay more precise and grounded, it's helpful to specify clearly what we mean by powerful AI (i.e. the threshold at which the 5-10 year clock starts counting), as well as laying out a framework for thinking about the effects of such AI once it's present.

What powerful AI (I dislike the term AGI) will look like, and when (or if) it will arrive, is a huge topic in itself. It's one I've discussed publicly and could write a completely separate essay on (I probably will at some point). Obviously, many people are skeptical that powerful AI will be built soon and some are skeptical that it will ever be built at all. I think it could come as early as 2026, though there are also ways it could take much longer. But for the purposes of this essay, I'd like to put these issues aside, assume it will come reasonably soon, and focus on what happens in the 5-10 years after that. I also want to assume a definition of what such a system *will look like,* what its

capabilities are and how it interacts, even though there is room for disagreement on this.

By *powerful AI*, I have in mind an AI model—likely similar to today's LLMs in form, though it might be based on a different architecture, might involve several interacting models, and might be trained differently—with the following properties:

- In terms of pure intelligence, it is smarter than a Nobel Prize winner across most relevant fields – biology, programming, math, engineering, writing, etc. This means it can prove unsolved mathematical theorems, write extremely good novels, write difficult codebases from scratch, etc.
- In addition to just being a "smart thing you talk to", it has all the "interfaces" available to a human working virtually, including text, audio, video, mouse and keyboard control, and internet access. It can engage in any actions, communications, or remote operations enabled by this interface, including taking actions on the internet, taking or giving directions to humans, ordering materials, directing experiments, watching videos, making videos, and so on. It does all of these tasks with, again, a skill exceeding that of the most capable humans in the world.
- It does not just passively answer questions; instead, it can be given tasks that take hours, days, or weeks to complete, and then goes off and does those tasks autonomously, in the way a smart employee would, asking for clarification as necessary.
- It does not have a physical embodiment (other than living on a computer screen), but it can control existing physical tools, robots, or laboratory equipment through a computer; in theory it could even design robots or equipment for itself to use.
- The resources used to train the model can be repurposed to *run* millions of instances of it (this matches projected cluster sizes by ~2027), and the model can absorb information and generate actions at roughly 10x-100x human speed. It may however be limited by the response time of the physical world or of software it interacts with.
- Each of these million copies can act independently on unrelated tasks, or if needed can all work together in the same way humans would collaborate, perhaps with

different subpopulations fine-tuned to be especially good at particular tasks.

We could summarize this as a "country of geniuses in a datacenter".

Clearly such an entity would be capable of solving very difficult problems, very fast, but it is not trivial to figure out how fast. Two "extreme" positions both seem false to me. First, you might think that the world would be instantly transformed on the scale of seconds or days ("the Singularity"), as superior intelligence builds on itself and solves every possible scientific, engineering, and operational task almost immediately. The problem with this is that there are real physical and practical limits, for example around building hardware or conducting biological experiments. Even a new country of geniuses would hit up against these limits. Intelligence may be very powerful, but it isn't magic fairy dust.

Second, and conversely, you might believe that technological progress is saturated or rate-limited by real world data or by social factors, and that better-than-human intelligence will add very little. This seems equally implausible to me—I can think of hundreds of scientific or even social problems where a large group of really smart people would drastically speed up progress, especially if they aren't limited to analysis and can make things happen in the real world (which our postulated country of geniuses can, including by directing or assisting teams of humans).

I think the truth is likely to be some messy admixture of these two extreme pictures, something that varies by task and field and is very subtle in its details. I believe we need new frameworks to think about these details in a productive way.

Economists often talk about "factors of production": things like labor, land, and capital. The phrase "marginal returns to labor/land/capital" captures the idea that in a given situation, a given factor may or may not be the limiting one – for example, an air force needs both planes and pilots, and hiring more pilots doesn't help much if you're out of planes. I believe that in the AI age, we should be talking about *the marginal returns to intelligence*, and trying to figure out what the other factors are that are complementary to intelligence and that become limiting factors when intelligence is very high. We are not used to thinking in this way—to asking "how much does being

smarter help with this task, and on what timescale?"—but it seems like the right way to conceptualize a world with very powerful AI.

My guess at a list of factors that limit or are complementary to intelligence includes:

- **Speed of the outside world**. Intelligent agents need to operate interactively in the world in order to accomplish things and also to learn. But the world only moves so fast. Cells and animals run at a fixed speed so experiments on them take a certain amount of time which may be irreducible. The same is true of hardware, materials science, anything involving communicating with people, and even our existing software infrastructure. Furthermore, in science many experiments are often needed in sequence, each learning from or building on the last. All of this means that the speed at which a major project—for example developing a cancer cure—can be completed may have an irreducible minimum that cannot be decreased further even as intelligence continues to increase.
- **Need for data**. Sometimes raw data is lacking and in its absence more intelligence does not help. Today's particle physicists are very ingenious and have developed a wide range of theories, but lack the data to choose between them because particle accelerator data is so limited. It is not clear that they would do drastically better if they were superintelligent—other than perhaps by speeding up the construction of a bigger accelerator.
- **Intrinsic complexity**. Some things are inherently unpredictable or chaotic and even the most powerful AI cannot predict or untangle them substantially better than a human or a computer today. For example, even incredibly powerful AI could predict only marginally further ahead in a chaotic system (such as the three-body problem) in the general case, as compared to today's humans and computers.
- **Constraints from humans**. Many things cannot be done without breaking laws, harming humans, or messing up society. An aligned AI would not want to do these things (and if we have an unaligned AI, we're back to talking about risks). Many human societal structures are inefficient or even actively harmful, but are hard to change while respecting constraints like legal requirements on clinical trials, people's willingness to change their habits, or the behavior of governments. Examples of advances that work well in a technical sense, but whose impact has

been substantially reduced by regulations or misplaced fears, include nuclear power, supersonic flight, and even elevators.

- **Physical laws**. This is a starker version of the first point. There are certain physical laws that appear to be unbreakable. It's not possible to travel faster than light. Pudding does not unstir. Chips can only have so many transistors per square centimeter before they become unreliable. Computation requires a certain minimum energy per bit erased, limiting the density of computation in the world.

There is a further distinction based on *timescales*. Things that are hard constraints in the short run may become more malleable to intelligence in the long run. For example, intelligence might be used to develop a new experimental paradigm that allows us to learn *in vitro* what used to require live animal experiments, or to build the tools needed to collect new data (e.g. the bigger particle accelerator), or to (within ethical limits) find ways around human-based constraints (e.g. helping to improve the clinical trial system, helping to create new jurisdictions where clinical trials have less bureaucracy, or improving the science itself to make human clinical trials less necessary or cheaper).

Thus, we should imagine a picture where intelligence is initially heavily bottlenecked by the other factors of production, but over time intelligence itself increasingly routes around the other factors, even if they never fully dissolve (and some things like physical laws are absolute). The key question is how fast it all happens and in what order.

With the above framework in mind, I'll try to answer that question for the five areas mentioned in the introduction.

1. Biology and health

Biology is probably the area where scientific progress has the greatest potential to directly and unambiguously improve the quality of human life. In the last century some of the most ancient human afflictions (such as smallpox) have finally been vanquished, but many more still remain, and defeating them would be an enormous humanitarian accomplishment. Beyond even curing disease, biological science can in principle

improve the *baseline* quality of human health, by extending the healthy human lifespan, increasing control and freedom over our own biological processes, and addressing everyday problems that we currently think of as immutable parts of the human condition.

In the "limiting factors" language of the previous section, the main challenges with directly applying intelligence to biology are data, the speed of the physical world, and intrinsic complexity (in fact, all three are related to each other). Human constraints also play a role at a later stage, when clinical trials are involved. Let's take these one by one.

Experiments on cells, animals, and even chemical processes are limited by the speed of the physical world: many biological protocols involve culturing bacteria or other cells, or simply waiting for chemical reactions to occur, and this can sometimes take days or even weeks, with no obvious way to speed it up. Animal experiments can take months (or more) and human experiments often take years (or even decades for long-term outcome studies). Somewhat related to this, data is often lacking—not so much in quantity, but quality: there is always a dearth of clear, unambiguous data that isolates a biological effect of interest from the other 10,000 confounding things that are going on, or that intervenes causally in a given process, or that directly measures some effect (as opposed to inferring its consequences in some indirect or noisy way). Even massive, quantitative molecular data, like the proteomics data that I collected while working on mass spectrometry techniques, is noisy and misses a lot (which types of cells were these proteins in? Which part of the cell? At what phase in the cell cycle?).

In part responsible for these problems with data is intrinsic complexity: if you've ever seen a diagram showing the biochemistry of human metabolism, you'll know that it's very hard to isolate the effect of any part of this complex system, and even harder to intervene on the system in a precise or predictable way. And finally, beyond just the intrinsic time that it takes to run an experiment on humans, actual clinical trials involve a lot of bureaucracy and regulatory requirements that (in the opinion of many people, including me) add unnecessary additional time and delay progress.

Given all this, many biologists have long been skeptical of the value of AI and "big data" more generally in biology. Historically, mathematicians, computer scientists, and

physicists who have applied their skills to biology over the last 30 years have been quite successful, but have not had the truly transformative impact initially hoped for. Some of the skepticism has been reduced by major and revolutionary breakthroughs like AlphaFold (which has just deservedly won its creators the Nobel Prize in Chemistry) and AlphaProteo, but there's still a perception that AI is (and will continue to be) useful in only a limited set of circumstances. A common formulation is "AI can do a better job analyzing your data, but it can't produce more data or improve the quality of the data. Garbage in, garbage out".

But I think that pessimistic perspective is thinking about AI in the wrong way. If our core hypothesis about AI progress is correct, then the right way to think of AI is not as a method of data analysis, but as a virtual biologist who performs *all* the tasks biologists do, including designing and running experiments in the real world (by controlling lab robots or simply telling humans which experiments to run – as a Principal Investigator would to their graduate students), inventing new biological methods or measurement techniques, and so on. It is by speeding up *the whole research process* that AI can truly accelerate biology. **I want to repeat this because it's the most common misconception that comes up when I talk about AI's ability to transform biology: I am *not* talking about AI as merely a tool to analyze data. In line with the definition of powerful AI at the beginning of this essay, I'm talking about using AI to perform, direct, and improve upon nearly everything biologists do.**

To get more specific on where I think acceleration is likely to come from, a surprisingly large fraction of the progress in biology has come from a truly tiny number of discoveries, often related to broad measurement tools or techniques that allow precise but generalized or programmable intervention in biological systems. There's perhaps ~1 of these major discoveries per year and collectively they arguably drive >50% of progress in biology. These discoveries are so powerful precisely because they cut through intrinsic complexity and data limitations, directly increasing our understanding and control over biological processes. A few discoveries per decade have enabled both the bulk of our basic scientific understanding of biology, and have driven many of the most powerful medical treatments.

Some examples include:

- CRISPR: a technique that allows live editing of any gene in living organisms (replacement of any arbitrary gene sequence with any other arbitrary sequence). Since the original technique was developed, there have been constant improvements to target specific cell types, increasing accuracy, and reducing edits of the wrong gene—all of which are needed for safe use in humans.
- Various kinds of microscopy for watching what is going on at a precise level: advanced light microscopes (with various kinds of fluorescent techniques, special optics, etc), electron microscopes, atomic force microscopes, etc.
- Genome sequencing and synthesis, which has dropped in cost by several orders of magnitude in the last couple decades.
- Optogenetic techniques that allow you to get a neuron to fire by shining a light on it.
- mRNA vaccines that, in principle, allow us to design a vaccine against anything and then quickly adapt it (mRNA vaccines of course became famous during COVID).
- Cell therapies such as CAR-T that allow immune cells to be taken out of the body and "reprogrammed" to attack, in principle, anything.
- Conceptual insights like the germ theory of disease or the realization of a link between the immune system and cancer.

I'm going to the trouble of listing all these technologies because I want to make a crucial claim about them: **I think their rate of discovery could be increased by 10x or more if there were a lot more talented, creative researchers**. Or, put another way, **I think the returns to intelligence are high for these discoveries**, and that everything else in biology and medicine mostly follows from them.

Why do I think this? Because of the answers to some questions that we should get in the habit of asking when we're trying to determine "returns to intelligence". First, these discoveries are generally made by a tiny number of researchers, often the same people repeatedly, suggesting skill and not random search (the latter might suggest lengthy experiments are the limiting factor). Second, they often "could have been made" years earlier than they were: for example, CRISPR was a naturally occurring component of

the immune system in bacteria that's been known since the 1980s, but it took another 25 years for people to realize it could be repurposed for general gene editing. They also are often delayed many years by lack of support from the scientific community for promising directions (see this profile on the inventor of mRNA vaccines; similar stories abound). Third, successful projects are often scrappy or were afterthoughts that people didn't initially think were promising, rather than massively funded efforts. This suggests that it's not just massive resource concentration that drives discoveries, but ingenuity.

Finally, although some of these discoveries have "serial dependence" (you need to make discovery A first in order to have the tools or knowledge to make discovery B)—which again might create experimental delays—many, perhaps most, are independent, meaning many at once can be worked on in parallel. Both these facts, and my general experience as a biologist, strongly suggest to me that there are hundreds of these discoveries waiting to be made if scientists were smarter and better at making connections between the vast amount of biological knowledge humanity possesses (again consider the CRISPR example). The success of AlphaFold/AlphaProteo at solving important problems much more effectively than humans, despite decades of carefully designed physics modeling, provides a proof of principle (albeit with a narrow tool in a narrow domain) that should point the way forward.

Thus, it's my guess that powerful AI could at least 10x the rate of these discoveries, giving us the next 50-100 years of biological progress in 5-10 years. Why not 100x? Perhaps it is possible, but here both serial dependence and experiment times become important: getting 100 years of progress in 1 year requires a lot of things to go right the first time, including animal experiments and things like designing microscopes or expensive lab facilities. I'm actually open to the (perhaps absurd-sounding) idea that we could get *1000* years of progress in 5-10 years, but very skeptical that we can get 100 years in 1 year. Another way to put it is I think there's an unavoidable constant delay: experiments and hardware design have a certain "latency" and need to be iterated upon a certain "irreducible" number of times in order to learn things that can't be deduced logically. But massive parallelism may be possible on top of that.

What about clinical trials? Although there is a lot of bureaucracy and slowdown associated with them, the truth is that a lot (though by no means all!) of their slowness ultimately derives from the need to rigorously evaluate drugs that barely work or ambiguously work. This is sadly true of most therapies today: the average cancer drug increases survival by a few months while having significant side effects that need to be carefully measured (there's a similar story for Alzheimer's drugs). This leads to huge studies (in order to achieve statistical power) and difficult tradeoffs which regulatory agencies generally aren't great at making, again because of bureaucracy and the complexity of competing interests.

When something works really well, it goes much faster: there's an accelerated approval track and the ease of approval is much greater when effect sizes are larger. mRNA vaccines for COVID were approved in 9 months—much faster than the usual pace. That said, even under these conditions clinical trials are still too slow—mRNA vaccines arguably *should* have been approved in ~2 months. But these kinds of delays (~1 year end-to-end for a drug) combined with massive parallelization and the need for some but not too much iteration ("a few tries") are very compatible with radical transformation in 5-10 years. Even more optimistically, it is possible that AI-enabled biological science will reduce the need for iteration in clinical trials by developing better animal and cell experimental models (or even simulations) that are more accurate in predicting what will happen in humans. This will be particularly important in developing drugs against the aging process, which plays out over decades and where we need a faster iteration loop.

Finally, on the topic of clinical trials and societal barriers, it is worth pointing out explicitly that in some ways biomedical innovations have an unusually *strong* track record of being successfully deployed, in contrast to some other technologies. As mentioned in the introduction, many technologies are hampered by societal factors despite working well technically. This might suggest a pessimistic perspective on what AI can accomplish. But biomedicine is unique in that although the process of developing drugs is overly cumbersome, once developed they generally are successfully deployed and used.

To summarize the above, my basic prediction is that AI-enabled biology and medicine will allow us to compress the progress that human biologists would have achieved over the next 50-100 years into 5-10 years. I'll refer to this as the "compressed 21st century": the idea that after powerful AI is developed, we will in a few years make all the progress in biology and medicine that we would have made in the whole 21st century.

Although predicting what powerful AI can do in a few years remains inherently difficult and speculative, there is some concreteness to asking "what could humans do unaided in the next 100 years?". Simply looking at what we've accomplished in the 20th century, or extrapolating from the first 2 decades of the 21st, or asking what "10 CRISPRs and 50 CAR-Ts" would get us, all offer practical, grounded ways to estimate the general level of progress we might expect from powerful AI.

Below I try to make a list of what we might expect. This is not based on any rigorous methodology, and will almost certainly prove wrong in the details, but it's trying to get across the general *level* of radicalism we should expect:

- **Reliable prevention and treatment of nearly all natural infectious disease.** Given the enormous advances against infectious disease in the 20th century, it is not radical to imagine that we could more or less "finish the job" in a compressed 21st. mRNA vaccines and similar technology already point the way towards "vaccines for anything". Whether infectious disease is *fully eradicated from the world* (as opposed to just in some places) depends on questions about poverty and inequality, which are discussed in Section 3.
- **Elimination of most cancer**. Death rates from cancer have been dropping ~2% per year for the last few decades; thus we are on track to eliminate most cancer in the 21st century at the current pace of human science. Some subtypes have already been largely cured (for example some types of leukemia with CAR-T therapy), and I'm perhaps even more excited for very selective drugs that target cancer in its infancy and prevent it from ever growing. AI will also make possible treatment regimens very finely adapted to the individualized genome of the cancer—these are possible today, but hugely expensive in time and human expertise, which AI should allow us to scale. Reductions of 95% or more in both mortality and incidence seem

possible. That said, cancer is extremely varied and adaptive, and is likely the hardest of these diseases to fully destroy. It would not be surprising if an assortment of rare, difficult malignancies persists.

- **Very effective prevention and effective cures for genetic disease**. Greatly improved embryo screening will likely make it possible to prevent most genetic disease, and some safer, more reliable descendant of CRISPR may cure most genetic disease in existing people. Whole-body afflictions that affect a large fraction of cells may be the last holdouts, however.
- **Prevention of Alzheimer's**. We've had a very hard time figuring out what causes Alzheimer's (it is somehow related to beta-amyloid protein, but the actual details seem to be very complex). It seems like exactly the type of problem that can be solved with better measurement tools that isolate biological effects; thus I am bullish about AI's ability to solve it. There is a good chance it can eventually be prevented with relatively simple interventions, once we actually understand what is going on. That said, damage from already existing Alzheimer's may be very difficult to reverse.
- **Improved treatment of most other ailments**. This is a catch-all category for other ailments including diabetes, obesity, heart disease, autoimmune diseases, and more. Most of these seem "easier" to solve than cancer and Alzheimer's and in many cases are already in steep decline. For example, deaths from heart disease have already declined over 50%, and simple interventions like GLP-1 agonists have already made huge progress against obesity and diabetes.
- **Biological freedom**. The last 70 years featured advances in birth control, fertility, management of weight, and much more. But I suspect AI-accelerated biology will greatly expand what is possible: weight, physical appearance, reproduction, and other biological processes will be fully under people's control. We'll refer to these under the heading of *biological freedom:* the idea that everyone should be empowered to choose what they want to become and live their lives in the way that most appeals to them. There will of course be important questions about global equality of access; see Section 3 for these.
- **Doubling of the human lifespan.** This might seem radical, but life expectancy increased almost 2x in the 20th century (from ~40 years to ~75), so it's "on trend"

that the “compressed 21st” would double it again to 150. Obviously the interventions involved in slowing the actual aging process will be different from those that were needed in the last century to prevent (mostly childhood) premature deaths from disease, but the magnitude of change is not unprecedented. Concretely, there already exist drugs that increase maximum lifespan in rats by 25-50% with limited ill effects. And some animals (e.g. some types of turtle) already live 200 years, so humans are manifestly not at some theoretical upper limit. At a guess, the most important thing that is needed might be reliable, non-Goodhart-able biomarkers of human aging, as that will allow fast iteration on experiments and clinical trials. Once human lifespan is 150, we may be able to reach “escape velocity”, buying enough time that most of those currently alive today will be able to live as long as they want, although there’s certainly no guarantee this is biologically possible.

It is worth looking at this list and reflecting on how different the world will be if all of it is achieved 7-12 years from now (which would be in line with an aggressive AI timeline). It goes without saying that it would be an unimaginable humanitarian triumph, the elimination all at once of most of the scourges that have haunted humanity for millennia. Many of my friends and colleagues are raising children, and when those children grow up, I hope that any mention of disease will sound to them the way scurvy, smallpox, or bubonic plague sounds to us. That generation will also benefit from increased biological freedom and self-expression, and with luck may also be able to live as long as they want.

It’s hard to overestimate how surprising these changes will be to everyone except the small community of people who expected powerful AI. For example, thousands of economists and policy experts in the US currently debate how to keep Social Security and Medicare solvent, and more broadly how to keep down the cost of healthcare (which is mostly consumed by those over 70 and especially those with terminal illnesses such as cancer). The situation for these programs is likely to be radically improved if all this comes to pass, as the ratio of working age to retired population will change drastically. No doubt these challenges will be replaced with others, such as how to ensure widespread access to the new technologies, but it is worth reflecting on how much the world will change even if biology is the *only* area to be successfully accelerated by AI.

2. Neuroscience and mind

In the previous section I focused on *physical* diseases and biology in general, and didn't cover neuroscience or mental health. But neuroscience is a subdiscipline of biology and mental health is just as important as physical health. In fact, if anything, mental health affects human well-being even more directly than physical health. Hundreds of millions of people have very low quality of life due to problems like addiction, depression, schizophrenia, low-functioning autism, PTSD, psychopathy, or intellectual disabilities. Billions more struggle with everyday problems that can often be interpreted as much milder versions of one of these severe clinical disorders. And as with general biology, it may be possible to go beyond addressing problems to improving the baseline quality of human experience.

The basic framework that I laid out for biology applies equally to neuroscience. The field is propelled forward by a small number of discoveries often related to tools for measurement or precise intervention – in the list of those above, optogenetics was a neuroscience discovery, and more recently CLARITY and expansion microscopy are advances in the same vein, in addition to many of the general cell biology methods directly carrying over to neuroscience. I think the rate of these advances will be similarly accelerated by AI and therefore that the framework of "100 years of progress in 5-10 years" applies to neuroscience in the same way it does to biology and for the same reasons. As in biology, the progress in 20th century neuroscience was enormous – for example we didn't even understand how or why neurons fired until the 1950s. Thus, it seems reasonable to expect AI-accelerated neuroscience to produce rapid progress over a few years.

There is one thing we should add to this basic picture, which is that some of the things we've learned (or are learning) about AI itself in the last few years are likely to help advance neuroscience, even if it continues to be done only by humans. Interpretability is an obvious example: although biological neurons superficially operate in a completely different manner from artificial neurons (they communicate via spikes and often spike rates, so there is a time element not present in artificial neurons, and a bunch of details relating to cell physiology and neurotransmitters modifies their operation substantially), the basic question of "how do distributed, trained networks of

simple units that perform combined linear/non-linear operations work together to perform important computations" is the same, and I strongly suspect the details of individual neuron communication will be abstracted away in most of the interesting questions about computation and circuits. As just one example of this, a computational mechanism discovered by interpretability researchers in AI systems was recently rediscovered in the brains of mice.

It is much easier to do experiments on artificial neural networks than on real ones (the latter often requires cutting into animal brains), so interpretability may well become a tool for improving our understanding of neuroscience. Furthermore, powerful AIs will themselves probably be able to develop and apply this tool better than humans can.

Beyond just interpretability though, what we have learned from AI about how intelligent systems are *trained* should (though I am not sure it *has* yet) cause a revolution in neuroscience. When I was working in neuroscience, a lot of people focused on what I would now consider the wrong questions about learning, because the concept of the scaling hypothesis / bitter lesson didn't exist yet. The idea that a simple objective function plus a lot of data can drive incredibly complex behaviors makes it more interesting to understand the objective functions and architectural biases and less interesting to understand the details of the emergent computations. I have not followed the field closely in recent years, but I have a vague sense that computational neuroscientists have still not fully absorbed the lesson. My attitude to the scaling hypothesis has always been "aha – this is an explanation, at a high level, of how intelligence works and how it so easily evolved", but I don't think that's the average neuroscientist's view, in part because the scaling hypothesis as "the secret to intelligence" isn't fully accepted even within AI.

I think that neuroscientists should be trying to combine this basic insight with the particularities of the human brain (biophysical limitations, evolutionary history, topology, details of motor and sensory inputs/outputs) to try to figure out some of neuroscience's key puzzles. Some likely are, but I suspect it's not enough yet, and that AI neuroscientists will be able to more effectively leverage this angle to accelerate progress.

I expect AI to accelerate neuroscientific progress along four distinct routes, all of which can hopefully work together to cure mental illness and improve function:

- **Traditional molecular biology, chemistry, and genetics**. This is essentially the same story as general biology in Section 1, and AI can likely speed it up via the same mechanisms. There are many drugs that modulate neurotransmitters in order to alter brain function, affect alertness or perception, change mood, etc., and AI can help us invent many more. AI can probably also accelerate research on the genetic basis of mental illness.
- **Fine-grained neural measurement and intervention**. This is the ability to measure what a lot of individual neurons or neuronal circuits are doing, and intervene to change their behavior. Optogenetics and neural probes are technologies capable of both measurement and intervention in live organisms, and a number of very advanced methods (such as molecular ticker tapes to read out the firing patterns of large numbers of individual neurons) have also been proposed and seem possible in principle.
- **Advanced computational neuroscience**. As noted above, both the specific insights and the *gestalt* of modern AI can probably be applied fruitfully to questions in systems neuroscience, including perhaps uncovering the real causes and dynamics of complex diseases like psychosis or mood disorders.
- **Behavioral interventions**. I haven't much mentioned it given the focus on the biological side of neuroscience, but psychiatry and psychology have of course developed a wide repertoire of behavioral interventions over the 20th century; it stands to reason that AI could accelerate these as well, both the development of new methods and helping patients to adhere to existing methods. More broadly, the idea of an "AI coach" who always helps you to be the best version of yourself, who studies your interactions and helps you learn to be more effective, seems very promising.

It's my guess that these four routes of progress working together would, as with physical disease, be on track to lead to the cure or prevention of most mental illness in the next 100 years even if AI was not involved – and thus might reasonably be completed in 5-10 AI-accelerated years. Concretely my guess at what will happen is something like:

- **Most mental illness can probably be cured**. I'm not an expert in psychiatric disease (my time in neuroscience was spent building probes to study small groups of neurons) but it's my guess that diseases like PTSD, depression, schizophrenia, addiction, etc. can be figured out and very effectively treated via some combination of the four directions above. The answer is likely to be some combination of "something went wrong biochemically" (although it could be very complex) and "something went wrong with the neural network, at a high level". That is, it's a systems neuroscience question—though that doesn't gainsay the impact of the behavioral interventions discussed above. Tools for measurement and intervention, especially in live humans, seem likely to lead to rapid iteration and progress.
- **Conditions that are very "structural" may be more difficult, but not impossible**. There's some evidence that psychopathy is associated with obvious neuroanatomical differences – that some brain regions are simply smaller or less developed in psychopaths. Psychopaths are also believed to lack empathy from a young age; whatever is different about their brain, it was probably always that way. The same may be true of some intellectual disabilities, and perhaps other conditions. Restructuring the brain sounds hard, but it also seems like a task with high returns to intelligence. Perhaps there is some way to coax the adult brain into an earlier or more plastic state where it can be reshaped. I'm very uncertain how possible this is, but my instinct is to be optimistic about what AI can invent here.
- **Effective genetic prevention of mental illness seems possible**. Most mental illness is partially heritable, and genome-wide association studies are starting to gain traction on identifying the relevant factors, which are often many in number. It will probably be possible to prevent most of these diseases via embryo screening, similar to the story with physical disease. One difference is that psychiatric disease is more likely to be polygenic (many genes contribute), so due to complexity there's an increased risk of unknowingly selecting against positive traits that are correlated with disease. Oddly however, in recent years GWAS studies seem to suggest that these correlations might have been overstated. In any case, AI-accelerated neuroscience may help us to figure these things out. Of course, embryo screening for complex traits raises a number of societal issues and will be controversial, though I

would guess that most people would support screening for severe or debilitating mental illness.

- **Everyday problems that we don't think of as clinical disease will also be solved**. Most of us have everyday psychological problems that are not ordinarily thought of as rising to the level of clinical disease. Some people are quick to anger, others have trouble focusing or are often drowsy, some are fearful or anxious, or react badly to change. Today, drugs already exist to help with e.g. alertness or focus (caffeine, modafinil, ritalin) but as with many other previous areas, much more is likely to be possible. Probably many more such drugs exist and have not been discovered, and there may also be totally new modalities of intervention, such as targeted light stimulation (see optogenetics above) or magnetic fields. Given how many drugs we've developed in the 20th century that tune cognitive function and emotional state, I'm very optimistic about the "compressed 21st" where everyone can get their brain to behave a bit better and have a more fulfilling day-to-day experience.
- **Human baseline experience can be much better**. Taking this one step further, many people have experienced extraordinary moments of revelation, creative inspiration, compassion, fulfillment, transcendence, love, beauty, or meditative peace. The character and frequency of these experiences differs greatly from person to person and within the same person at different times, and can also sometimes be triggered by various drugs (though often with side effects). All of this suggests that the "space of what is possible to experience" is very broad and that a larger fraction of people's lives could consist of these extraordinary moments. It is probably also possible to improve various cognitive functions across the board. This is perhaps the neuroscience version of "biological freedom" or "extended lifespans".

One topic that often comes up in sci-fi depictions of AI, but that I intentionally haven't discussed here, is "mind uploading", the idea of capturing the pattern and dynamics of a human brain and instantiating them in software. This topic could be the subject of an essay all by itself, but suffice it to say that while I think uploading is almost certainly possible in principle, in practice it faces significant technological and societal challenges, even with powerful AI, that likely put it outside the 5-10 year window we are discussing.

In summary, AI-accelerated neuroscience is likely to vastly improve treatments for, or even cure, most mental illness as well as greatly expand "cognitive and mental freedom" and human cognitive and emotional abilities. It will be every bit as radical as the improvements in physical health described in the previous section. Perhaps the world will not be visibly different on the outside, but the world as experienced by humans will be a much better and more humane place, as well as a place that offers greater opportunities for self-actualization. I also suspect that improved mental health will ameliorate a lot of other societal problems, including ones that seem political or economic.

3. Economic development and poverty

The previous two sections are about *developing* new technologies that cure disease and improve the quality of human life. However an obvious question, from a humanitarian perspective, is: "will everyone have access to these technologies?"

It is one thing to develop a cure for a disease, it is another thing to eradicate the disease from the world. More broadly, many existing health interventions have not yet been applied everywhere in the world, and for that matter the same is true of (non-health) technological improvements in general. Another way to say this is that living standards in many parts of the world are still desperately poor: GDP per capita is ~$2,000 in Sub-Saharan Africa as compared to ~$75,000 in the United States. If AI further increases economic growth and quality of life in the developed world, while doing little to help the developing world, we should view that as a terrible moral failure and a blemish on the genuine humanitarian victories in the previous two sections. Ideally, powerful AI should help the developing world *catch up to* the developed world, even as it revolutionizes the latter.

I am not as confident that AI can address inequality and economic growth as I am that it can invent fundamental technologies, because technology has such obvious high returns to intelligence (including the ability to route around complexities and lack of data) whereas the economy involves a lot of constraints from humans, as well as a large dose of intrinsic complexity. I am somewhat skeptical that an AI could solve the famous "socialist calculation problem" and I don't think governments will (or should)

turn over their economic policy to such an entity, even if it could do so. There are also problems like how to convince people to take treatments that are effective but that they may be suspicious of.

The challenges facing the developing world are made even more complicated by pervasive corruption in both private and public sectors. Corruption creates a vicious cycle: it exacerbates poverty, and poverty in turn breeds more corruption. AI-driven plans for economic development need to reckon with corruption, weak institutions, and other very human challenges.

Nevertheless, I do see significant reasons for optimism. Diseases *have* been eradicated and many countries *have* gone from poor to rich, and it is clear that the decisions involved in these tasks exhibit high returns to intelligence (despite human constraints and complexity). Therefore, AI can likely do them better than they are currently being done. There may also be targeted interventions that get around the human constraints and that AI could focus on. More importantly though, *we have* to try. Both AI companies and developed world policymakers will need to do their part to ensure that the developing world is not left out; the moral imperative is too great. So in this section, I'll continue to make the optimistic case, but keep in mind everywhere that success is not guaranteed and depends on our collective efforts.

Below I make some guesses about how I think things may go in the developing world over the 5-10 years after powerful AI is developed:

- **Distribution of health interventions**. The area where I am perhaps most optimistic is distributing health interventions throughout the world. Diseases have actually been eradicated by top-down campaigns: smallpox was fully eliminated in the 1970s, and polio and guinea worm are nearly eradicated with less than 100 cases per year. Mathematically sophisticated epidemiological modeling plays an active role in disease eradication campaigns, and it seems very likely that there is room for smarter-than-human AI systems to do a better job of it than humans are. The logistics of distribution can probably also be greatly optimized. One thing I learned as an early donor to GiveWell is that some health charities are way more effective than others; the hope is that AI-accelerated efforts would be more effective still. Additionally, some biological advances actually make the logistics of distribution

much easier: for example, malaria has been difficult to eradicate because it requires treatment each time the disease is contracted; a vaccine that only needs to be administered once makes the logistics much simpler (and such vaccines for malaria are in fact currently being developed). Even simpler distribution mechanisms are possible: some diseases could in principle be eradicated by targeting their animal carriers, for example releasing mosquitoes infected with a bacterium that blocks their ability to carry a disease (who then infect all the other mosquitos) or simply using gene drives to wipe out the mosquitos. This requires one or a few centralized actions, rather than a coordinated campaign that must individually treat millions. Overall, I think 5-10 years is a reasonable timeline for a good fraction (maybe 50%) of AI-driven health benefits to propagate to even the poorest countries in the world. A good goal might be for the developing world 5-10 years after powerful AI to at least be substantially healthier than the developed world is today, even if it continues to lag behind the developed world. Accomplishing this will of course require a huge effort in global health, philanthropy, political advocacy, and many other efforts, which both AI developers and policymakers should help with.

- **Economic growth**. Can the developing world quickly catch up to the developed world, not just in health, but across the board economically? There is some precedent for this: in the final decades of the 20th century, several East Asian economies achieved sustained ~10% annual real GDP growth rates, allowing them to catch up with the developed world. Human economic planners made the decisions that led to this success, not by directly controlling entire economies but by pulling a few key levers (such as an industrial policy of export-led growth, and resisting the temptation to rely on natural resource wealth); it's plausible that "AI finance ministers and central bankers" could replicate or exceed this 10% accomplishment. An important question is how to get developing world governments to adopt them while respecting the principle of self-determination—some may be enthusiastic about it, but others are likely to be skeptical. On the optimistic side, many of the health interventions in the previous bullet point are likely to organically increase economic growth: eradicating AIDS/malaria/parasitic worms would have a transformative effect on productivity, not to mention the economic benefits that some of the neuroscience interventions (such as improved mood and focus) would have in developed and developing worlds alike. Finally,

non-health AI-accelerated technology (such as energy technology, transport drones, improved building materials, better logistics and distribution, and so on) may simply permeate the world naturally; for example, even cell phones quickly permeated sub-Saharan Africa via market mechanisms, without needing philanthropic efforts. On the more negative side, while AI and automation have many potential benefits, they also pose challenges for economic development, particularly for countries that haven't yet industrialized. Finding ways to ensure these countries can still develop and improve their economies in an age of increasing automation is an important challenge for economists and policymakers to address. Overall, a dream scenario—perhaps a goal to aim for—would be 20% annual GDP growth rate in the developing world, with 10% each coming from AI-enabled economic decisions and the natural spread of AI-accelerated technologies, including but not limited to health. If achieved, this would bring sub-Saharan Africa to the current per-capita GDP of China in 5-10 years, while raising much of the rest of the developing world to levels higher than the current US GDP. Again, this is a dream scenario, not what happens by default: it's something all of us must work together to make more likely.

- **Food security**. Advances in crop technology like better fertilizers and pesticides, more automation, and more efficient land use drastically increased crop yields across the 20th century, saving millions of people from hunger. Genetic engineering is currently improving many crops even further. Finding even more ways to do this —as well as to make agricultural supply chains even more efficient—could give us an AI-driven second Green Revolution, helping close the gap between the developing and developed world.
- **Mitigating climate change**. Climate change will be felt much more strongly in the developing world, hampering its development. We can expect that AI will lead to improvements in technologies that slow or prevent climate change, from atmospheric carbon-removal and clean energy technology to lab-grown meat that reduces our reliance on carbon-intensive factory farming. Of course, as discussed above, technology isn't the only thing restricting progress on climate change—as with all of the other issues discussed in this essay, human societal factors are important. But there's good reason to think that AI-enhanced research will give us

the means to make mitigating climate change far less costly and disruptive, rendering many of the objections moot and freeing up developing countries to make more economic progress.

- **Inequality within countries**. I've mostly talked about inequality as a global phenomenon (which I do think is its most important manifestation), but of course inequality also exists *within* countries. With advanced health interventions and especially radical increases in lifespan or cognitive enhancement drugs, there will certainly be valid worries that these technologies are "only for the rich". I am more optimistic about within-country inequality especially in the developed world, for two reasons. First, markets function better in the developed world, and markets are typically good at bringing down the cost of high-value technologies over time. Second, developed world political institutions are more responsive to their citizens and have greater state capacity to execute universal access programs—and I expect citizens to demand access to technologies that so radically improve quality of life. Of course it's not predetermined that such demands succeed—and here is another place where we collectively have to do all we can to ensure a fair society. There is a separate problem in inequality of *wealth* (as opposed to inequality of access to life-saving and life-enhancing technologies), which seems harder and which I discuss in Section 5.
- **The opt-out problem**. One concern in both developed and developing worlds alike is people *opting out* of AI-enabled benefits (similar to the anti-vaccine movement, or Luddite movements more generally). There could end up being bad feedback cycles where, for example, the people who are least able to make good decisions opt out of the very technologies that improve their decision-making abilities, leading to an ever-increasing gap and even creating a dystopian underclass (some researchers have argued that this will undermine democracy, a topic I discuss further in the next section). This would, once again, place a moral blemish on AI's positive advances. This is a difficult problem to solve as I don't think it is ethically okay to coerce people, but we can at least try to increase people's scientific understanding—and perhaps AI itself can help us with this. One hopeful sign is that historically anti-technology movements have been more bark than bite: railing against modern technology is popular, but most people adopt it in the end, at least

when it's a matter of individual choice. Individuals tend to adopt most health and consumer technologies, while technologies that are truly hampered, like nuclear power, tend to be collective political decisions.

Overall, I am optimistic about quickly bringing AI's biological advances to people in the developing world. I am hopeful, though not confident, that AI can also enable unprecedented economic growth rates and allow the developing world to at least surpass where the developed world is now. I am concerned about the "opt out" problem in both the developed and developing worlds, but suspect that it will peter out over time and that AI can help accelerate this process. It won't be a perfect world, and those who are behind won't fully catch up, at least not in the first few years. But with strong efforts on our part, we may be able to get things moving in the right direction—and fast. If we do, we can make at least a down payment on the promises of dignity and equality that we owe to every human being on earth.

4. Peace and governance

Suppose that everything in the first three sections goes well: disease, poverty, and inequality are significantly reduced and the baseline of human experience is raised substantially. It does not follow that all major causes of human suffering are solved. Humans are still a threat to each other. Although there is a trend of technological improvement and economic development leading to democracy and peace, it is a very loose trend, with frequent (and recent) backsliding. At the dawn of the 20th century, people thought they had put war behind them; then came the two world wars. Thirty years ago Francis Fukuyama wrote about "the End of History" and a final triumph of liberal democracy; that hasn't happened yet. Twenty years ago US policymakers believed that free trade with China would cause it to liberalize as it became richer; that very much didn't happen, and we now seem headed for a second Cold War with a resurgent authoritarian bloc. And plausible theories suggest that internet technology may actually advantage authoritarianism, not democracy as initially believed (e.g. in the "Arab Spring" period). It seems important to try to understand how powerful AI will intersect with these issues of peace, democracy, and freedom.

Unfortunately, I see no strong reason to believe AI will preferentially or structurally advance democracy and peace, in the same way that I think it will structurally advance human health and alleviate poverty. Human conflict is adversarial and AI can in principle help both the "good guys" and the "bad guys". If anything, some structural factors seem worrying: AI seems likely to enable much better propaganda and surveillance, both major tools in the autocrat's toolkit. It's therefore up to us as individual actors to tilt things in the right direction: if we want AI to favor democracy and individual rights, we are going to have to fight for that outcome. I feel even more strongly about this than I do about international inequality: the triumph of liberal democracy and political stability is *not* guaranteed, perhaps not even likely, and will require great sacrifice and commitment on all of our parts, as it often has in the past.

I think of the issue as having two parts: international conflict, and the internal structure of nations. On the international side, it seems very important that democracies have the upper hand on the world stage when powerful AI is created. AI-powered authoritarianism seems too terrible to contemplate, so democracies need to be able to set the terms by which powerful AI is brought into the world, both to avoid being overpowered by authoritarians and to prevent human rights abuses within authoritarian countries.

My current guess at the best way to do this is via an "entente strategy", in which a coalition of democracies seeks to gain a clear advantage (even just a temporary one) on powerful AI by securing its supply chain, scaling quickly, and blocking or delaying adversaries' access to key resources like chips and semiconductor equipment. This coalition would on one hand use AI to achieve robust military superiority (the stick) while at the same time offering to distribute the benefits of powerful AI (the carrot) to a wider and wider group of countries in exchange for supporting the coalition's strategy to promote democracy (this would be a bit analogous to "Atoms for Peace"). The coalition would aim to gain the support of more and more of the world, isolating our worst adversaries and eventually putting them in a position where they are better off taking the same bargain as the rest of the world: give up competing with democracies in order to receive all the benefits and not fight a superior foe.

If we can do all this, we will have a world in which democracies lead on the world stage and have the economic and military strength to avoid being undermined, conquered, or sabotaged by autocracies, and may be able to parlay their AI superiority into a durable advantage. This could optimistically lead to an "eternal 1991"—a world where democracies have the upper hand and Fukuyama's dreams are realized. Again, this will be very difficult to achieve, and will in particular require close cooperation between private AI companies and democratic governments, as well as extraordinarily wise decisions about the balance between carrot and stick.

Even if all that goes well, it leaves the question of the fight between democracy and autocracy *within* each country. It is obviously hard to predict what will happen here, but I do have some optimism that *given* a global environment in which democracies control the most powerful AI, *then* AI may actually structurally favor democracy everywhere. In particular, in this environment democratic governments can use their superior AI to win the information war: they can counter influence and propaganda operations by autocracies and may even be able to create a globally free information environment by providing channels of information and AI services in a way that autocracies lack the technical ability to block or monitor. It probably isn't necessary to deliver propaganda, only to counter malicious attacks and unblock the free flow of information. Although not immediate, a level playing field like this stands a good chance of gradually tilting global governance towards democracy, for several reasons.

First, the increases in quality of life in Sections 1-3 should, all things equal, promote democracy: historically they have, to at least some extent. In particular I expect improvements in mental health, well-being, and education to increase democracy, as all three are negatively correlated with support for authoritarian leaders. In general people want more self-expression when their other needs are met, and democracy is among other things a form of self-expression. Conversely, authoritarianism thrives on fear and resentment.

Second, there is a good chance free information really does undermine authoritarianism, as long as the authoritarians can't censor it. And uncensored AI can also bring individuals powerful tools for undermining repressive governments. Repressive governments survive by denying people a certain kind of common

knowledge, keeping them from realizing that "the emperor has no clothes". For example Srđa Popović, who helped to topple the Milošević government in Serbia, has written extensively about techniques for psychologically robbing authoritarians of their power, for breaking the spell and rallying support against a dictator. A superhumanly effective AI version of Popović (whose skills seem like they have high returns to intelligence) in everyone's pocket, one that dictators are powerless to block or censor, could create a wind at the backs of dissidents and reformers across the world. To say it again, this will be a long and protracted fight, one where victory is not assured, but if we design and build AI in the right way, it may at least be a fight where the advocates of freedom everywhere have an advantage.

As with neuroscience and biology, we can also ask how things could be "better than normal"—not just how to avoid autocracy, but how to make democracies better than they are today. Even within democracies, injustices happen all the time. Rule-of-law societies make a promise to their citizens that everyone will be equal under the law and everyone is entitled to basic human rights, but obviously people do not always receive those rights in practice. That this promise is even partially fulfilled makes it something to be proud of, but can AI help us do better?

For example, could AI improve our legal and judicial system by making decisions and processes more impartial? Today people mostly worry in legal or judicial contexts that AI systems will be a *cause* of discrimination, and these worries are important and need to be defended against. At the same time, the vitality of democracy depends on harnessing new technologies to improve democratic institutions, not just responding to risks. A truly mature and successful implementation of AI has the potential to *reduce* bias and be fairer for everyone.

For centuries, legal systems have faced the dilemma that the law aims to be impartial, but is inherently subjective and thus must be interpreted by biased humans. Trying to make the law fully mechanical hasn't worked because the real world is messy and can't always be captured in mathematical formulas. Instead legal systems rely on notoriously imprecise criteria like "cruel and unusual punishment" or "utterly without redeeming social importance", which humans then interpret—and often do so in a manner that displays bias, favoritism, or arbitrariness. "Smart contracts" in

cryptocurrencies haven't revolutionized law because ordinary code isn't smart enough to adjudicate all that much of interest. But AI might be smart enough for this: it is the first technology capable of making broad, fuzzy judgements in a repeatable and mechanical way.

I am not suggesting that we literally replace judges with AI systems, but the combination of impartiality with the ability to understand and process messy, real world situations *feels* like it should have some serious positive applications to law and justice. At the very least, such systems could work alongside humans as an aid to decision-making. Transparency would be important in any such system, and a mature science of AI could conceivably provide it: the training process for such systems could be extensively studied, and advanced interpretability techniques could be used to see inside the final model and assess it for hidden biases, in a way that is simply not possible with humans. Such AI tools could also be used to monitor for violations of fundamental rights in a judicial or police context, making constitutions more self-enforcing.

In a similar vein, AI could be used to both aggregate opinions and drive consensus among citizens, resolving conflict, finding common ground, and seeking compromise. Some early ideas in this direction have been undertaken by the computational democracy project, including collaborations with Anthropic. A more informed and thoughtful citizenry would obviously strengthen democratic institutions.

There is also a clear opportunity for AI to be used to help provision government services—such as health benefits or social services—that are in principle available to everyone but in practice often severely lacking, and worse in some places than others. This includes health services, the DMV, taxes, social security, building code enforcement, and so on. Having a very thoughtful and informed AI whose job is to give you everything you're legally entitled to by the government in a way you can understand—and who also helps you comply with often confusing government rules—would be a big deal. Increasing state capacity both helps to deliver on the promise of equality under the law, and strengthens respect for democratic governance. Poorly implemented services are currently a major driver of cynicism about government.

All of these are somewhat vague ideas, and as I said at the beginning of this section, I am not nearly as confident in their feasibility as I am in the advances in biology, neuroscience, and poverty alleviation. They may be unrealistically utopian. But the important thing is to have an ambitious vision, to be willing to dream big and try things out. The vision of AI as a guarantor of liberty, individual rights, and equality under the law is too powerful a vision not to fight for. A 21st century, AI-enabled polity could be both a stronger protector of individual freedom, and a beacon of hope that helps make liberal democracy the form of government that the whole world wants to adopt.

5. Work and meaning

Even if everything in the preceding four sections goes well—not only do we alleviate disease, poverty, and inequality, but liberal democracy becomes the dominant form of government, and existing liberal democracies become better versions of themselves—at least one important question still remains. “It’s great we live in such a technologically advanced world as well as a fair and decent one”, someone might object, “but with AIs doing everything, how will humans have meaning? For that matter, how will they survive economically?”.

I think this question is more difficult than the others. I don’t mean that I am necessarily more pessimistic about it than I am about the other questions (although I do see challenges). I mean that it is fuzzier and harder to predict in advance, because it relates to macroscopic questions about how society is organized that tend to resolve themselves only over time and in a decentralized manner. For example, historical hunter-gatherer societies might have imagined that life is meaningless without hunting and various kinds of hunting-related religious rituals, and would have imagined that our well-fed technological society is devoid of purpose. They might also have not understood how our economy can provide for everyone, or what function people can usefully serve in a mechanized society.

Nevertheless, it’s worth saying at least a few words, while keeping in mind that the brevity of this section is not at all to be taken as a sign that I don’t take these issues seriously—on the contrary, it is a sign of a lack of clear answers.

On the question of meaning, I think it is very likely a mistake to believe that tasks you undertake are meaningless simply because an AI could do them better. Most people are not the best in the world at anything, and it doesn't seem to bother them particularly much. Of course today they can still contribute through comparative advantage, and may derive meaning from the economic value they produce, but people also greatly enjoy activities that produce no economic value. I spend plenty of time playing video games, swimming, walking around outside, and talking to friends, all of which generates zero economic value. I might spend a day trying to get better at a video game, or faster at biking up a mountain, and it doesn't really matter to me that someone somewhere is much better at those things. In any case I think meaning comes mostly from human relationships and connection, not from economic labor. People do want a sense of accomplishment, even a sense of competition, and in a post-AI world it will be perfectly possible to spend years attempting some very difficult task with a complex strategy, similar to what people do today when they embark on research projects, try to become Hollywood actors, or found companies. The facts that (a) an AI somewhere could in principle do this task better, and (b) this task is no longer an economically rewarded element of a global economy, don't seem to me to matter very much.

The economic piece actually seems more difficult to me than the meaning piece. By "economic" in this section I mean the possible problem that *most or all* humans may not be able to contribute meaningfully to a sufficiently advanced AI-driven economy. This is a more macro problem than the separate problem of inequality, especially inequality in access to the new technologies, which I discussed in Section 3.

First of all, in the short term I agree with arguments that comparative advantage will continue to keep humans relevant and in fact increase their productivity, and may even in some ways level the playing field between humans. As long as AI is only better at 90% of a given job, the other 10% will cause humans to become highly leveraged, increasing compensation and in fact creating a bunch of new human jobs complementing and amplifying what AI is good at, such that the "10%" expands to continue to employ almost everyone. In fact, even if AI can do 100% of things better than humans, but it remains inefficient or expensive at some tasks, or if the resource *inputs* to humans and AIs are meaningfully different, then the logic of comparative

advantage continues to apply. One area humans are likely to maintain a relative (or even absolute) advantage for a significant time is the physical world. Thus, I think that the human economy may continue to make sense even a little past the point where we reach "a country of geniuses in a datacenter".

However, I do think in the long run AI will become so broadly effective and so cheap that this will no longer apply. At that point our current economic setup will no longer make sense, and there will be a need for a broader societal conversation about how the economy should be organized.

While that might sound crazy, the fact is that civilization has successfully navigated major economic shifts in the past: from hunting and gathering to farming, farming to feudalism, and feudalism to industrialism. I suspect that some new and stranger thing will be needed, and that it's something no one today has done a good job of envisioning. It could be as simple as a large universal basic income for everyone, although I suspect that will only be a small part of a solution. It could be a capitalist economy of AI systems, which then give out resources (huge amounts of them, since the overall economic pie will be gigantic) to humans based on some secondary economy of what the AI systems think makes sense to reward in humans (based on some judgment ultimately derived from human values). Perhaps the economy runs on Whuffie points. Or perhaps humans will continue to be economically valuable after all, in some way not anticipated by the usual economic models. All of these solutions have tons of possible problems, and it's not possible to know whether they will make sense without lots of iteration and experimentation. And as with some of the other challenges, we will likely have to fight to get a good outcome here: exploitative or dystopian directions are clearly also possible and have to be prevented. Much more could be written about these questions and I hope to do so at some later time.

Taking stock

Through the varied topics above, I've tried to lay out a vision of a world that is both plausible *if* everything goes right with AI, and much better than the world today. I don't know if this world is realistic, and even if it is, it will not be achieved without a huge amount of effort and struggle by many brave and dedicated people. Everyone

(including AI companies!) will need to do their part both to prevent risks and to fully realize the benefits.

But it is a world worth fighting for. If all of this really does happen over 5 to 10 years—the defeat of most diseases, the growth in biological and cognitive freedom, the lifting of billions of people out of poverty to share in the new technologies, a renaissance of liberal democracy and human rights—I suspect everyone watching it will be surprised by the effect it has on them. I don't mean the experience of personally benefiting from all the new technologies, although that will certainly be amazing. I mean the experience of watching a long-held set of ideals materialize in front of us all at once. I think many will be literally moved to tears by it.

Throughout writing this essay I noticed an interesting tension. In one sense the vision laid out here is extremely radical: it is not what almost anyone expects to happen in the next decade, and will likely strike many as an absurd fantasy. Some may not even consider it desirable; it embodies values and political choices that not everyone will agree with. But at the same time there is something blindingly obvious—something overdetermined—about it, as if many different attempts to envision a good world inevitably lead roughly here.

In Iain M. Banks' *The Player of Games*, the protagonist—a member of a society called the Culture, which is based on principles not unlike those I've laid out here—travels to a repressive, militaristic empire in which leadership is determined by competition in an intricate battle game. The game, however, is complex enough that a player's strategy within it tends to reflect their own political and philosophical outlook. The protagonist manages to defeat the emperor in the game, showing that his values (the Culture's values) represent a winning strategy even in a game designed by a society based on ruthless competition and survival of the fittest. A well-known post by Scott Alexander has the same thesis—that competition is self-defeating and tends to lead to a society based on compassion and cooperation. The "arc of the moral universe" is another similar concept.

I think the Culture's values are a winning strategy because they're the sum of a million small decisions that have clear moral force and that tend to pull everyone together onto the same side. Basic human intuitions of fairness, cooperation, curiosity, and

autonomy are hard to argue with, and are cumulative in a way that our more destructive impulses often aren't. It is easy to argue that children shouldn't die of disease if we can prevent it, and easy from there to argue that *everyone's* children deserve that right equally. From there it is not hard to argue that we should all band together and apply our intellects to achieve this outcome. Few disagree that people should be punished for attacking or hurting others unnecessarily, and from there it's not much of a leap to the idea that punishments should be consistent and systematic across people. It is similarly intuitive that people should have autonomy and responsibility over their own lives and choices. These simple intuitions, if taken to their logical conclusion, lead eventually to rule of law, democracy, and Enlightenment values. If not inevitably, then at least as a statistical tendency, this is where humanity was already headed. AI simply offers an opportunity to get us there more quickly—to make the logic starker and the destination clearer.

Nevertheless, it is a thing of transcendent beauty. We have the opportunity to play some small role in making it real.

Acknowledgements

Thanks to Kevin Esvelt, Parag Mallick, Stuart Ritchie, Matt Yglesias, Erik Brynjolfsson, Jim McClave, Allan Dafoe, and many people at Anthropic for reviewing drafts of this essay.

To the winners of the 2024 Nobel prize in Chemistry, for showing us all the way.