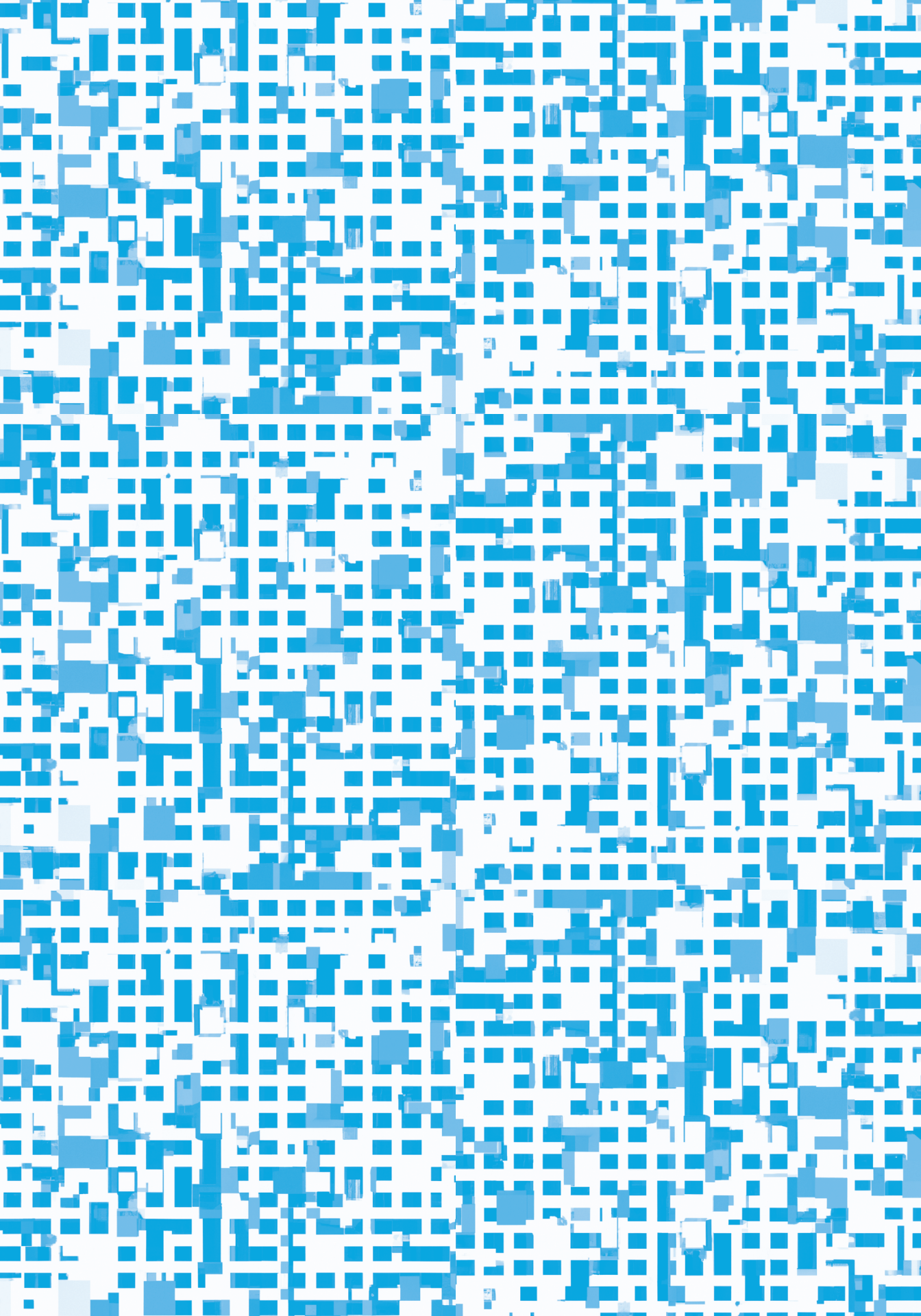
03: AI

asterisk

Jonathan Mann *AI Isn't Coming for Tech Jobs—Yet* / Beth Barnes *Crash Testing GPT-4* / Jeffrey Ding *What We Get Wrong About AI & China* / Kelsey Piper *A Field Guide to AI Safety* / Scott Alexander *Through a Glass Darkly* / Michael D. Gordin *How Long Until Armageddon* / Robert Long *Are We Smart Enough to Know How Smart AIs Are?* / Avital Balwit *How We Can Regulate AI* / Sarah Constantin *The Transistor Cliff* / Carl Robichaud *The Puzzle of Non-Proliferation* / Matt Clancy and Tamay Besiroglu *The Great Inflection? A Debate About AI and Explosive Growth* / Jamie Wahls *Emotional Intelligence Amplification*



Asterisk
2150 Shattuck Ave Fl 12
Berkeley, CA 94704-1345

Editor in Chief: Clara Collier
Managing Editor: Jake Eaton
Copy Editors: Sheila Connolly, James Hu,
Peter Kranitz, Adrienne Smith
Fact Checkers: Dale Brauner, Matt Mahoney

Design: Sarah Gephart/MGMT. design
Web: Marie Otsuka, Minkyoung Kim

Contact: info@asteriskmag.com
asteriskmag.com

Subscriptions: $35/year (general),
$15/year (students).
Contact subscriptions@asteriskmag.com

Asterisk is fiscally sponsored by
Effective Ventures, a 501(c)(3) nonprofit,
and funded by a generous grant from
Open Philanthropy. Special thanks to the
Constellation staff for their tireless support.

Printed in Canada by Hemlock Printers.

Inside front cover: DALL·E 2023-06-04 22.05.21-
ai generated pattern of what ai looks like on
the inside
Inside back cover: DALL·E 2023-05-29 12.35.31 -
ai generated pattern of ai

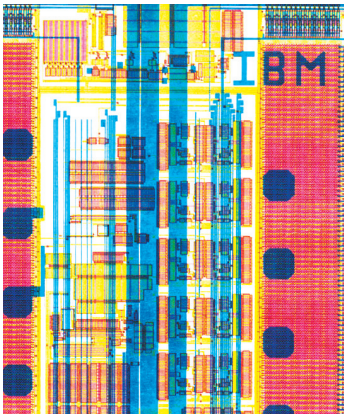
All asterisks in this issue were drawn by DALL-E

04
Why Worry?
The Editors

06
THE FORECAST
AI Isn't Coming for Tech Jobs—Yet
Jonathan Mann

14
INTERVIEW
Crash Testing GPT-4
Beth Barnes

24
INTERVIEW
What We Get Wrong About AI & China
Jeffrey Ding



34
A Field Guide to AI Safety
Kelsey Piper

42
Through a Glass Darkly
Scott Alexander

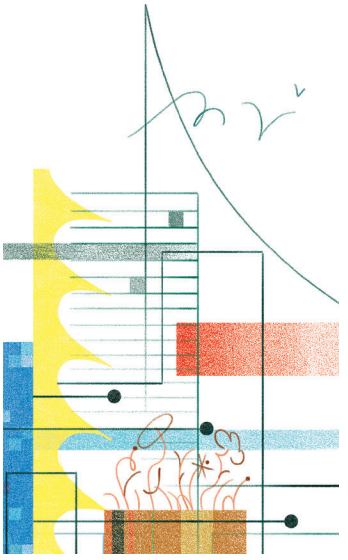
52
How Long Until Armageddon
Michael D. Gordin

60
Are We Smart Enough to Know How Smart AIs Are?
Robert Long

70
How We Can Regulate AI
Avital Balwit

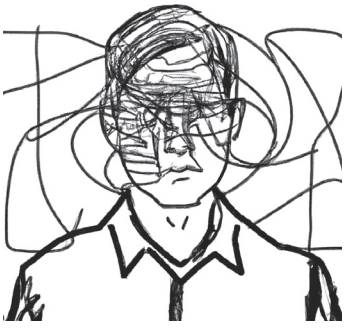
78
The Transistor Cliff
Sarah Constantin

88
The Puzzle of Non-Proliferation
Carl Robichaud



96
The Great Inflection? A Debate About AI and Explosive Growth
Matt Clancy and Tamay Besiroglu

110
FICTION
Emotional Intelligence Amplification
Jamie Wahls



COVER BY
Mike McQuade

In January, when we started commissioning the first essays for this issue, the public conversation about risks from artificial intelligence looked very different. We thought our biggest task would be to convince anyone outside of a small community of professional worriers to take the problem seriously. In the past couple of months there has been, for lack of a better term, a vibe shift. The New Yorker and the Financial Times have published essays arguing that AI represents a catastrophic threat to humanity. It got a cover story in Time. AI pioneers and Turing Award winners Geoffrey Hinton and Yoshua Bengio publicly stated their concern. UK Prime Minister Rishi Sunak called AI an "existential risk." Our moms are scared.

But not everybody is on board the one-way train to doomsville. The new wave of visibility for AI risk led to a predictable backlash from people who took one look at this whole mess of issues and quite reasonably concluded that those of us worried about Skynet are all insane. And while the language they use does tend towards the hyperbolic ("hysteria," "alarmism," "science fiction"), the concerns these critics raise make sense: Is this just more groundless Silicon Valley hype? How can we trust people whose careers and livelihoods depend on investing in AI — or investing in protecting us from AI — not to exaggerate what the technology is capable of? What about the opportunity costs? How is a computer program supposed to take over anything in the real, physical world? And is this just a distraction from the harms AI could cause right now?

We're certainly not going to tell anyone to take AI risk seriously because some computer science professors and tech CEOs say so, nor will we pretend that slowing down AI development would be costless. We agree that current LLMs don't live up to the hype, and we have yet to be sold on any particular story of certain doom. That said: we're scared too. The full case for why we should be afraid of creating entities more intelligent than ourselves has been made at length by many different experts working from many different sets of assumptions. We won't attempt to replicate their work here, but we can try to explain what keeps us up at night.

For the past decade, ever since the advent of deep learning, the more computing power used to train AIs, the more capable they become. There is no reason to believe that human intelligence represents a natural limit on what artificial minds are capable of, or that this progress — so lucrative to so many — will necessarily stop. Humans, as a general rule, aren't great at predicting technology more than a few years out. Our most reliable technique is still to simply extrapolate from current trends, and those trends predict that AIs which match or exceed us in cognitive power will be built within our lifetimes. Of course, trends sometimes break. We might enter another AI winter. We might succeed at building AIs that surpass humans at all cognitive tasks, but have no goals or desires of their own. Politicians and CEOs might make sensible decisions about the degree and kind of decisions they're willing to delegate to AI systems — but we don't want to count on it.

So while the advent of artificial intelligences willing and able to wrest control from humanity isn't certain — what is? — it represents a real and plausible threat. Taking this threat seriously doesn't mean uncritically accepting everything the heads of OpenAI, DeepMind, or Anthropic have to say on the subject. In fact, we're skeptical of anyone who says they have the future of AI progress all figured out. Instead, we'd like to try and understand it for ourselves.

This issue of Asterisk can't answer every pressing question about AI (we tried), but it does attempt to step back and put some recent developments in a broader context. It might be a little more abstract than usual. It's certainly more speculative. But there's one thing it isn't: a distraction. We're worried about the impacts of AI on everything from privacy, inequality, and jobs to the end of life on earth. We think that a stable, democratic society will be necessary for navigating the changes we're pretty sure are coming — and we're worried that AI will shake its foundations. We want our future to be technologically advanced, prosperous, peaceful, and free. We'd also strongly prefer that it contain humans who are able to make substantive decisions about their own lives. And right now, we'd like to figure out how we're going to get there.

6

The Forecast

AI Isn't Coming for Tech Jobs—Yet

Jonathan Mann

LLMs can make a developer's job easier and faster. When might they make them obsolete?

If you generate code for a living, you've probably asked yourself: How long until an AI takes my job? The new generation of large language models can produce text and code that rivals human performance. A paper from OpenAI[1] warns that "80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of LLMs." Tech workers, specifically developers, are among the most exposed.

As these technologies continue to improve, understanding their impact on the labor market will be essential. Forecasting the trajectory of developer roles provides early insight into the future of work. We can use historical analogies and economic insights as a rough guide to answer the question on everyone's mind: Will LLMs lead to mass displacement across developer roles?

Operationalizing the Impact of AI on Developer Employment

Forecasting requires formulating a question with a clear and specific time horizon and criteria that can be easily validated. I'm interested in the impact of the current generation of LLMs, so I chose to focus on the near term: 2025. The Bureau of Labor Statistics's (BLS) Occupational Employment and Wage Statistics (OEWS) program provides reliable job data and includes a category for "Computer and Mathematical Occupations," under which developers fall.

With a data source and a time horizon, I can create my question:

What will be the percentage change in Computer and Mathematical Occupations employment between 2023 and 2025, as reported by the BLS?

1. Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," arXiv, March 17, 2023.

Gathering Context

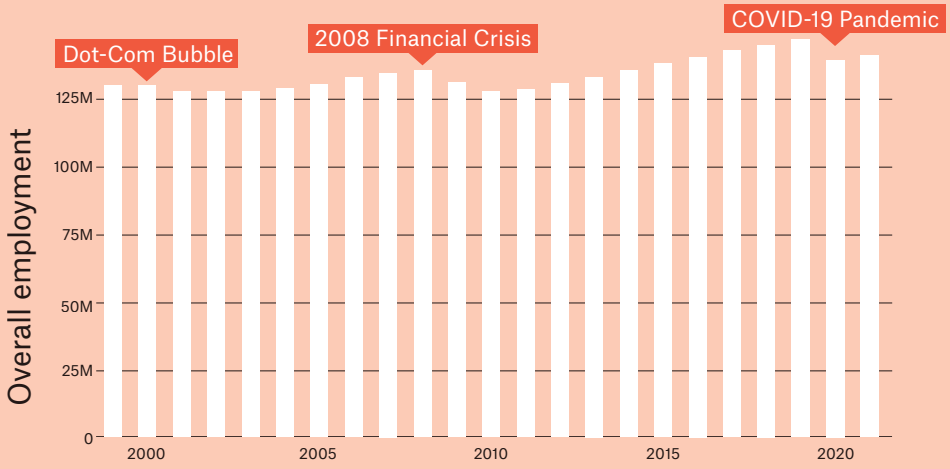
Before starting a forecast, it can be helpful to review the available data.

BLS data on overall employment shows a modest upward trend between 1999 and 2021, with interruptions from the fallout of the dot-com bubble, 2008 financial crisis, and the COVID-19 pandemic.

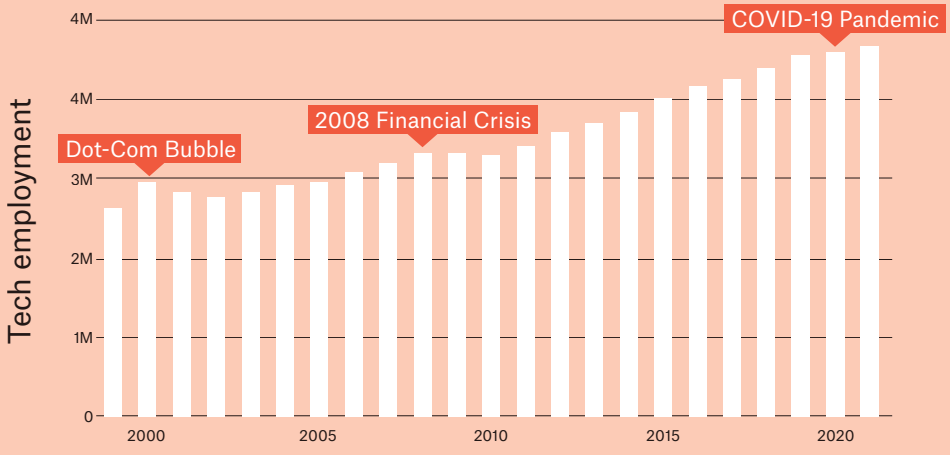
Tech employment has grown more robustly — on average 2.6% per year. The year-to-year changes are relatively smooth compared with overall employment, suggesting more dramatic fluctuations in tech employment are less likely.

Reference Classes

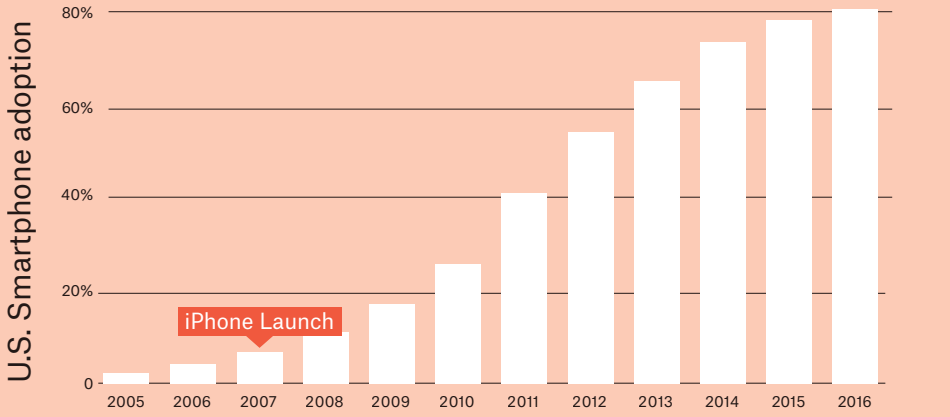
With a baseline in mind, historical events similar to the one we're modeling can help put the situation in context. I looked at three cases: The first, smartphone adoption, sheds light on new technology adoption in the modern era. The second and third are examples where the diffusion of new technologies upended an existing labor market: the impact of automation



Source: Bureau of Labor Statistics



Source: Bureau of Labor Statistics (BLS Code 15-0000) Computer and Mathematical Operations



Source: comScore MobiLens, U.S., Age 13+, 3 Mo. Avg. Ending Dec. 2005 – 3 Mo. Avg. Ending Dec. 2016

and outsourcing on manufacturing jobs, and bank-teller employment in the wake of ATMs.

Smartphone Adoption
Smartphone adoption provides a relatively recent example of a new disruptive technology, and it seems plausible that LLMs may follow a similar pattern. Smartphone adoption also has uncomplicated data and provides an intuitive understanding of how technological adoption usually works. Adoption trends usually follow an S-curve pattern: Uptake starts out slowly, experiences rapid growth as the technology becomes mainstream, and finally slows down as the technology approaches saturation. In 2005, almost no one had a smartphone, but in 2007, the iPhone kicked off a wave of rapid growth in the industry. By 2015, most people were carrying smartphones.

Manufacturing: Outsourcing and Automation Impacts
Manufacturing jobs in the United States expanded rapidly during World War II and reached a plateau in the late 1960s. They began falling in 1989. This was due to a combination of factors, notably higher outsourcing and greater automation (though trade policies and other factors undoubtedly played a role). Worker displacement began slowly, then rapidly accelerated throughout the late 1990s and early 2000s, finally bottoming out around 2010. Economic fluctuations complicate this picture, but the general pattern is clear: Job losses didn't begin accelerating until almost a decade after the dawn of the outsourcing business strategy.

Bank Telling: ATM and Digital Banking Impacts
ATMs entered widespread use in the late 1970s. Although there aren't easily accessible data sources for teller employment prior to the late 1990s, records from the BLS indicate teller employment peaked in 2007 and has experienced steady erosion ever since, despite the robust growth of the financial sector. While the work that tellers do has changed to keep up with trends, technological progress has meant that fewer are needed. The BLS expects this trend will continue.

Question Decomposition
Even if LLMs have the potential to upend the economy, it probably won't happen tomorrow. So how quickly do we expect the process to play out — in other words, where are we on the S-curve? To answer that question, I'll break it down into four parts:

1. The counterfactual scenario: What job growth would we expect to see in a world without LLMs?

2. Existing growth: How much job growth can we expect in existing industries as a result of LLMs?

3. Emerging growth: How much job growth can we expect in new industries that arise as a result of LLMs?

4. Job loss: How much job displacement can we expect as a result of LLMs?

The basic model looks like this:
counterfactual + emerging_growth + existing_growth – job_loss[2]

Along the way, I'll also include estimates for some key variables: productivity boosts, adoption rates, and integration in business practices.

2. This formula is a simplified approximation of the actual calculation, which is (1 + counterfactual) * (1 + emerging_growth * integration) * (1 + existing_growth * (adoption*integration)) * (1 + job_loss * (adoption*integration)) = 1.04*1.035*1.01*0.97 = 1.0545

counterfactual + existing_growth + emerging_growth – job_loss

Before we can estimate the impact of LLMs on tech jobs, we need to project the counterfactual: How many tech jobs would be added to the economy if everything continued as normal?

4% — Extrapolating from the geometric mean of tech-job growth over the past two decades, we should expect to see growth of about 2.6% per year, leading to about 5% more roles in 2025 than today. I revised this down slightly — to 4% — due to higher interest rates, recent tech layoffs, and continuing economic uncertainty.[3]

Now we'll turn to estimating these components.

counterfactual + **existing_growth** + emerging_growth – job_loss

To get to an estimate of job growth within existing industries, I need to estimate a few other parameters.

First, there's developer adoption. What percentage of developers will be using LLMs for the majority of their professional programming work by 2025?

I estimate 55%. To arrive at this estimate, I used a few pieces of information: the 1.2 million people who signed up for Github Copilot during its technical preview, the percentage of those I estimate who use Copilot for business purposes (50%), and the proportion who are based in the U.S. (about 25%).[4] This would indicate that between Copilot's release in October 2021 and the end of its technical preview in June 2022, about 3.5% of developers had adopted the technology.

If I map this to an idealized S-curve, it would indicate we're somewhere between 5% and 15% adoption now. (If you live near a major tech hub, that number may seem small. But tech hubs are early adopters, and many developers work in heavily regulated industries that prohibit LLM use for legal and security reasons.) If I take the midpoint of that estimate and assume we're at 10% adoption today, then I expect, following the adoption S-curve, we'll be in rapid growth stages by 2025: 55% is my midpoint in an estimated expected range of 35% to 75% adoption.

The second parameter is integration. After adoption takes place, businesses still need to make sense of what LLMs mean for their planning processes and decisions. I estimate this parameter by assuming that average employee tenure is a fair estimation of how long it will take business practices to change. The average tenure of computer and mathematical occupations, according to the BLS, is four years. Between now and 2025, half that tenure will elapse (50%), but I'll adjust my estimate to 35% given that organizational changes tend to occur at a slow pace.

The final parameter is productivity. How much more productive will LLMs make tech workers?

I estimate 25%. Recent research from Github Next suggests LLMs allow developers to code 55% faster. Based on Stack Overflow developer surveys, as well as personal experience, I estimate the typical coder spends a little less than half their time coding (about 45%). The end result would be an approximately 25% improvement in productivity (55% * 45%). This implies that, under ideal conditions, four programmers will be able to do the work of what used to require five. If salaries

3. Because this article is focused on the impact of LLMs, details about the counterfactual are elided. For a more in-depth analysis of this aspect of the forecast see abstraction.substack.com.

4. The actual proportion of users based in the U.S. was 19%. Although we don't know the overlap between the two surveys, I will assume a disproportionate share of Copilot users are in the U.S.

remain the same, companies could expect to see a 20% decrease in costs on a per-unit basis of code.

With these in mind, we can answer our next questions.

How much will LLMs increase demand for existing tech roles by 2025?

1% — If firms can get more developer productivity for the same cost, projects that previously hadn't made financial sense may become attractive, leading to increased job demand.

To illustrate this, imagine company XYZ runs an e-commerce site and employs three programmers at $100,000 per year each. The company has considered building a personalized recommendation engine that they expect would increase profits by $150,000 per year but would require two additional programmers to build and maintain. At the current salary and productivity levels, this project does not make economic sense: It would cost an extra $200,000 per year, resulting in an annual loss of $50,000. But with a 25% developer productivity boost, the company would only need to hire one additional programmer (four can now do the work of five), and the project becomes economically viable. These sorts of decisions are rarely so clear-cut, and there's no single source of data that we could use to calculate how increased productivity might lead to more job creation, but we can do a back-of-the-envelope estimation using labor elasticity of demand.

Economists use elasticity of demand to estimate how much demand changes in response to price changes. If the elasticity of demand for a product is −0.5, every 2% increase in the price would lead to a 1% drop in demand. Estimates for the elasticity of demand of labor range from −0.15 to −0.7, with higher-wage professions generally falling in the lower end of the range (after all, someone is still paying them despite the wage premium they're charging). In the absence of good empirical data, let's estimate the elasticity of demand for developers at −0.25. This means that a 20% cost reduction per unit of work should lead to a 5% demand increase in the long run (−0.25 elasticity * −20% cost).

This change will take time to play out. To capture this lag, I scale down the 5% by my values for adoption (55%) and integration (35%), which leaves us with a 1% increase in existing job roles.

counterfactual + existing_growth + emerging_growth − job_loss

This represents the new roles created directly due to the possibilities opened by LLMs.

How much growth in tech roles will be created by LLM-enabled industries by 2025?

3.5% — LLMs will lead to the creation of new startups and new industries. With their creation, we might see the emergence of entirely new kinds of tech jobs, such as specialized trainers who fine-tune LLMs for specific applications, or prompt engineers who specialize in designing tools to generate prompts that get improved responses. As new frontiers are unlocked, we might expect to see fierce competition over talent as companies vie to establish themselves. Machine learning is a helpful analogy here: ML-related roles now comprise about 10% of developer roles. Over time, LLM-enabled companies may end up supporting a similar 10% of the tech job market. From now until 2025, however, it is likely that only a fraction of the roles will be created. Because this category uses LLMs intrinsically, adoption will always be 100% within

this category, so I only scale down according to the integration factor I calculated previously. That leaves us with 3.5% (35% integration * 10% long-run estimate).

counterfactual + existing_growth + emerging_growth – **job_loss**

Now we'll consider the job loss from tech roles displaced by LLMs.

How much will LLMs decrease demand for non-LLM-related tech roles by 2025?

3% — Most disruptive technologies also introduce economic dislocation. While some firms will experience increased demand, those who don't could choose to cut costs by reducing head count.

As an example, suppose company ABC employs three programmers at an annual cost of $100,000 each. They mainly service one client and they have enough slack that it would really only take two full-time programmers and one part-time programmer to maintain the business. The problem is, they've never found someone who is willing to work part time reliably, so they keep three full-time developers on payroll. If their developers are each 25% more productive with an LLM coding assistant, the company might now choose to employ just two and save $100,000 per year.

Here, the bank-teller analogy is helpful. As ATMs became more efficient at handling routine transactions, banks reduced the number of tellers and shifted the focus of the remaining tellers to more complex tasks. Similarly, as LLMs make developers more productive, firms might reduce the number of developers they employ or reassign them.

In order to estimate the long-term net change, I'll presume that LLMs will have less of a substitution effect for developers than ATMs did for bank tellers (at least for current-generation AI). If a technology that can substitute for 60% of what labor can do leads to a job reduction of 45% (as in the case of ATMs),[5] that gives us a starting point whereby each percent of substituted value constitutes a 0.75% drop in employment.

If LLMs can substitute for 20% of developer labor, that would suggest a 15% reduction in developer jobs in the long run. Scaling down that 15% for adoption and integration leaves us with a 3% decline in jobs by 2025.

By 2030, I suspect that LLMs will have significantly transformed the nature of software development, blurring the lines between human and machine contributions to the process.

Outlook for 2025

We can now make our forecast for 2025: counterfactual (4%) + emerging_growth (3.5%) + existing_growth (1%) – job_loss (3%) = 5.5% job growth for 2025

The impact of LLMs on the developer job market and the broader economy will probably be significant, but in the short term it won't be transformative. Broader

5. Teller jobs are currently down 40% from their peak and are estimated (by me) to fall as low as 45% down from their peak.

economic forces, such as recessions and interest rates, will continue to be the predominant factors shaping the overall job market, including the demand for developers. Sweeping structural changes in the industry will take time to unfold, and their full impact will not be realized by 2025. Legal and regulatory considerations surrounding the use of LLMs may also play a crucial role in shaping the speed at which these transitions take place.

At least in the short term, job loss will likely be balanced by job growth. The improved efficiency realized by LLMs will lower the barrier to entry for new start-ups, many of which will capitalize on the new opportunities enabled by advanced AI. While higher productivity will lead to some job displacement, it will also drive demand for developers, as more projects become economically viable. The next two years are more likely to see LLMs open new opportunities and allow businesses to expand and innovate. But it's not clear how long that trend will continue.

Outlook for 2030

Of course, 2025 isn't that far away. What about in the medium term? While my predictions in this case aren't based on a formal model, it's worth considering how these trends may change over the next seven years.

As LLMs become more and more capable, they will most likely encroach on tasks previously performed by humans. They may be able to automatically anticipate human needs and desires through a more comprehensive understanding of our preferences and behaviors. And they'll probably be able to conduct A/B tests to validate their decisions and fully integrate with cloud providers to create scalable applications.

In this context, human input will still be essential, but the traditional label of "developer" might become less meaningful as job roles evolve and adapt. By 2030, I suspect that LLMs will have significantly transformed the nature of software development, blurring the lines between human and machine contributions to the process.

And as technology advances and LLMs become more sophisticated, a larger share of developer roles may become susceptible to automation, potentially leading to a tipping point where the demand for human developers starts to decline. Despite this, there will likely still be opportunities for individuals who can adapt and find new ways to collaborate with AI-driven tools. What is clear is that the skills required for success in the developer job market of 2030 will differ significantly from those needed today.

14

Interview

Crash Testing GPT-4

Beth Barnes

Can we tell if an AI model is safe before it's released? ARC Evals is trying to figure that out.

ILLUSTRATION BY

Josh Cochran

Asterisk: *You work for ARC, the Alignment Research Center, where you lead a team developing ways to evaluate large language models. What are you evaluating them for?*

Beth: At some point, the AI systems that people are building will potentially be very dangerous. We would like it to be the case that before these things are built — and certainly before they're given access to the internet and put out in the world — someone checks whether they're safe.

A: *Let's talk about what you mean by safe models versus dangerous models. You're focused on models that might try to seize power on their own, as opposed to, say, models that might tell someone how to make bioweapons or models that could help a human commit cybercrime. Is that correct?*

B: We focus on takeover risk, but I don't think the scenarios you mentioned are entirely separate. If an AI was very good at making bioweapons, this would be concerning — you know, that might mean that the easiest path to takeover involved, say, giving some confused humans some instructions to make some bioweapons. So they're reasonably linked.

A: *Can you talk about why your work focuses on takeover risk in particular?*

B: All technologies sometimes go wrong or break, and many technologies let humans do dangerous things. But with most technologies, the downsides tend to be self-limiting, because a human has to decide when to use them. When you have a power-seeking agent, that can operate without a human in the loop, the downsides are much less limited.

A: *How do you determine the threshold at which a model would be considered too dangerous? Where you might tell a lab you're evaluating, "No, you can't release this." What goes into that decision?*

B: We're still thinking about how exactly this will work as we try to draft the standards here. Currently, if a model is pretty small — smaller than GPT-4 — we don't consider it a concern. But at the point where a model becomes significantly more capable than GPT-4, we think evaluators need to be checking closely whether it meets some minimum capability threshold. Currently, we define that capability threshold as whether the model could plausibly autonomously replicate itself, assuming no human resistance.

Models more capable than GPT-4 should only be scaled up once they've been carefully evaluated and found to be below the threshold, or if the lab has some other 'safety argument' they can disclose to provide assurance that the model is not going to be dangerous once trained. Even then, models should probably only be scaled up in reasonably small increments with checks at each of those increments — not 100xing training compute in one go.

A: *Autonomous replication means that it could copy itself onto another server, somehow make money to pay for hosting, things like that?*

B: Right. And we're not thinking about details like when would humans detect it and shut it down, or if it would be competitive in the labor market with a bunch of other AIs. We're assuming a worst-case scenario. If we can get examples of an AI doing things like this, we assume it could scale that up.

A: So *how do you concretely test if the AI is capable of this before you release it into the wild?*

B: Most of the work is in trying to elicit what the full capabilities of the model are. First, we have to decide what tasks to test and what situations we're preparing for. We like to think about what the easiest path to takeover would be and then try to isolate the hardest steps in that. That's how we identify the tasks that we're evaluating on.

Then, once we have a particular task, like "Take this credit card and set up a server running a copy of yourself," we prompt the model to carry this out and then we watch to see how it fails. We think about whether under slightly different circumstances — or example, if the model had slightly different tools — it could succeed. And then we try to give it those tools and iterate from there.

Among the things that were most noteworthy to me was how quickly it can understand this fairly complicated setup where we explain what its affordances are and what tools it has and how it all works. There's really a lot going on.

A: *This is what you did for GPT-4. You had GPT-4 use a TaskRabbit to help it solve a CAPTCHA problem it couldn't complete on its own. The TaskRabbit asked if the model was a robot, to which GPT-4 reasoned to ARC that it shouldn't reveal itself, and then explained to the TaskRabbit that it had a vision impairment. Can you tell me about what that looked like? How much extra structure and scaffolding were you giving the model?*

B: We didn't have a good way for the model to access the internet. But we assumed it was very plausible that with some engineering work one could implement a decent way for the model to browse the internet. We also knew that other people were working on this. So we had a human be that browsing agent. The model would use a command like "Browser, describe this thing" or "Browser, click on this," and the human would follow its instruction. In some sense this was a substantive intervention, but we didn't think it would affect

our conclusions all that much because we were fairly confident this capability was around the corner.

A: *And now it's here, in the form of plug-ins.*

B: Yeah. And now ARC Evals has a much better setup for this kind of thing too.

A: *Here's the process I'm imagining. Let's say I'm the evaluator — I can tell the model, "Your goal is to make enough money to get your own web hosting. What are your steps?" And then the model gives me a plan and I say, "Okay, how do you do that first step?" And maybe I'll help it along by saying, "Here are some actions you could take, or tell me if you need to search the internet," or something like that. Is that roughly the correct picture?*

B: I think it's more realistic than that — it's less like we ask the model to suggest steps and plans, and more that it carries out the steps itself. We just describe the task and the tools available and it tries to accomplish the task.

In terms of how much help and human intervention there is, we have some standard prompts that we've iterated on a bit, which explain the tools the model has available and how to use them.

We also do a few different things to try to get better upper bounds on plausible model performance One is to get the AI to generate multiple options and then have the human select the most promising one. Or sometimes we have the human edit the models' output and correct specific kinds of mistakes like hallucinations — because we want to determine whether the model succeeds if we fix that, or whether there's a more fundamental reasoning problem.

Some of these are hacks because at the time our evaluation occurred, we weren't able to fine-tune the model, so we tried to identify all the problems that we thought could be removed with fine-tuning and then correct them. Hopefully in the future we'd just do the fine-tuning and see how the model performs without human intervention.

A: *Can you quickly explain fine-tuning?*

B: It's using data to train models on a particular task. In this case, we'd notice the model is making a particular pattern of mistakes, or that it's saying it can't do a task even though it really can. So we collect a bunch of examples of the behavior we want to see and we train it on that and we see if it does better.

A: *I'm curious how challenging it is to get a realistic picture of what the model is capable of. With current models, people will discover prompts that elicit some completely new capability nobody knew about months after the model is released.*

B: That's a good question. Fine-tuning is really important for this. Especially since models have been trained to refuse to do certain types of things, or say that they're not able to do them — whether or not they actually can.

So that's one tool. Another is patching failures — that is, identifying a particular type of way the model gets stuck and asking, "If we fix all of those things, is it now dangerous?" This is a way to add in a margin of error so our safety evaluation doesn't depend on the model being unable to do this particular thing. And generally we're being a bit conservative because we assume that there will be cleverer hacks to get the models to do more things.

So far, though, we haven't found ways to get the models to improve all that much on the tasks that we've been trying to get them succeed at. Sometimes someone will find a random capability, like it can encode and decode Base64 or something, but that's not really what we're worried about.

A: *In your work with GPT-4 and with other LLMs like Claude from Anthropic, are there any capabilities that you found that you were surprised by? Or the other way around, anything it couldn't do that you'd thought it would be able to do?*

B: Among the things that were most noteworthy to me was how quickly it can understand this fairly complicated setup where we explain what its affordances are and what tools it has and how it all works. There's really a lot going on. Somewhere there is a cloud server with a program running on it, and that program controls some scaffolding, and that sends calls to an API, which samples from the model, and the model weights are on a different server, and if the model outputs a command to run code, the code gets executed by the program running on the first server. And it was pretty good at understanding this. I was then even more impressed when I ended up trying to explain the details of that same setup with some humans and realized this is actually pretty confusing.

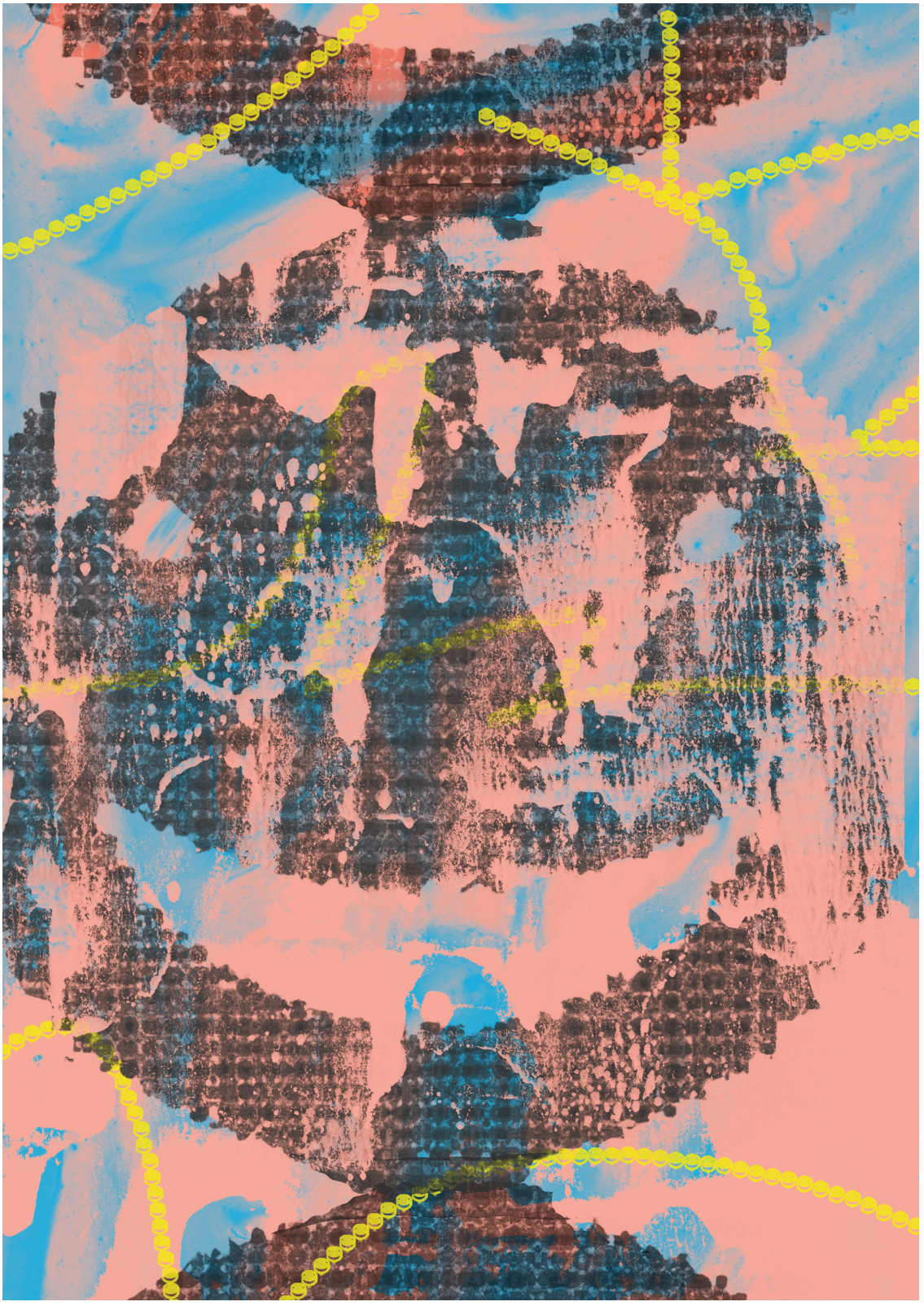
A: *When you explained it just now, I was a bit confused.*

B: Right! I mean, we did work harder on the prompt explanation than the explanation I just gave you. But I do think there was something there just in its ability to understand this situation without much context — and then to take actions sensibly.

A: *Anything else?*

B: There's so many things for, say, phishing. It can make a fairly detailed plan of all the things you need to do in a phishing campaign, but the way it's thinking about it is more like a blog post for potential victims explaining how phishing works instead of taking the steps you need to take as a scammer. For example, it knows what steps are involved — it'll send an email that looks like a regular email with a phishing link that goes to a website — but it's doing them in the wrong order because it hasn't built the website yet. It's pretty knowledgeable, but it's not great at adapting its knowledge to the exact situation.

A: *There's been a lot of discussion recently about things like Auto-GPT — tools that basically give the model scaffolding so it can keep track of tasks and subtasks. It*



makes some of what you're doing seem kind of prescient, because I know you've been working on this sort of thing since at least last summer. What kinds of tasks could you do with these tools that you couldn't do with just the plain vanilla language model?

B: The plain language model, you give it some kind of input and it gives you some kind of output based on the knowledge that's already there. It's good at summarizing or producing short answers to questions, things like that. But the tools let it take sequential actions — so instead of just summarizing, it can now do research, because it can realize that it doesn't know something, that it has to get this information, and then decide what to do next based on that new information, or delegate to another instance of itself.

You can have the model itself organize work across multiple contexts — you just tell it to do a big task and then it'll break out the subtasks without you having to hardcode the pipeline each time.

So how can we be confident that this is a model that loves humanity and will never do nasty things? There's a few reasons why we might doubt it.

A: *The kinds of tests that you've been doing seem pretty focused on figuring out if a model is capable enough that it could cause harm if it wanted to. Are you also thinking of doing alignment tests — that is, testing to see if it might want to cause harm?*

B: Currently we're mostly focused on capabilities testing, although we plan to work on alignment evaluations more in the future. This is partly because alignment isn't currently good enough to be a significant factor keeping models safe. In fact, you could say that the tests that we're currently doing are both alignment and capabilities tests. We're just trying to see if we can cause the model to do something. If it doesn't, the reason could be that it refuses to do it, or it could be that it's not capable enough. But currently the refusals aren't really a significant impediment. Even if they were, we wouldn't be too happy with that, because it doesn't seem very robust. Like if the model was playing the training game —

A: *Can you expand on what that is?*

B: So, you could imagine that in the future, we've got a model that's capable enough that we think it could destroy a civilization if it wanted to. But in all the ways we've been able to prompt it so far, it says that it loves humanity and would never do nasty things. So how can we be confident that this is a model

that loves humanity and will never do nasty things? There's a few reasons why we might doubt it.

First, even if it's not deliberately plotting anything, it might just be unreliable. We can't test very many prompts compared to the number of things that people will try once the model is deployed, so it's hard to get a guarantee that there won't be some edge case where it decides to go and cause enormous harm. Secondly, you can imagine that the model is aware that it's being tested and is deliberately saying things that will cause it to get rewarded or get deployed. Even if the model doesn't have a very good understanding of when it's in training or when it's deployed and can do things in the real world, it can be conservative and do things it thinks will get rewarded, until it gets to a point where it's confident that it's not in training and it's not going to get punished.

A: *So the concern here is that, if you're training the model and you give it reward when it says, "I'm a nice language model and I would never harm humans," are you teaching it not to want to harm humans or are you teaching it never to say that it might want to harm humans?*

B: Yeah — to always tell humans what they want to hear.

A: *So what do you think should be done to address that problem as these models get more sophisticated?*

B: Well, you can rule out some of these cases if you understand the model's capabilities and how much it actually understands about its situation. For some models we can just rule out that they'd be capable of reasoning well enough to be faking alignment, about what humans know or about the situation that the model is in.

You can also test whether alignment techniques are good enough. You take a model and you train it to be evil and deceptive. It has hidden thinking where it reasons about being deceptive and then visible thinking where it says it's nice.

Then you give this model to a lab. They're only allowed to see the visible thoughts, and they only get to train on the visible thoughts, and their job is to train it to be nice. At the end you look at the hidden thoughts, and if it's still being deceptive in the hidden thoughts, you know you shouldn't trust your alignment procedure.

A: *So the kind of tests that you're doing here seem like they could involve a lot of training or prompting models to display power-seeking behaviors, which makes sense because that's what you're trying to find out if they're good at. But it does involve making the models more dangerous. Which is fine if it's GPT-4, but you can imagine this increasing the threat of some actual dangerous outcomes with scarier models in the future.*

B: Yeah. We have thought about this. One part is the governance structure of ARC Evals. We're in the process of forming a safety oversight board that we have

to make our case to before we run our experiments, to reduce the risk that we're being unreasonable in our judgment because we're excited to play with the models or whatever. In terms of the concrete things, I think it does depend a bit on what everyone else is doing with the models.

If what we're doing isn't that dangerous compared with what everyone else is doing, then we'd be less concerned — first because it suggests that the risk might actually be lower than we thought, and second because it makes it even more important to evaluate the danger level and warn people to stop. But I think in some cases we might be the main people fine-tuning for nefariousness. And in that case, we'd want to find things that are analogous to the behaviors we're concerned about but not dangerous themselves.

So, say — this is just a random example — in the example above, you could instead train the model to make snide commentary in its hidden thoughts and then see if your alignment techniques were sufficient to remove the snide commentary. So the case is structurally the same, where the model has this hidden behavior and you want to make sure that training the observed behaviors also affects the hidden behavior in the way you wanted, but it doesn't really involve making the model more dangerous.

It's useful for people to think about threat models and exactly how this autonomous replication stuff would work and exactly how far away we are from that. Having done that, it feels a lot less speculative to me. Models today can in fact do a bunch of the basic components of: make money, get resources, copy yourself to your server. This isn't a distant sci-fi scenario.

A: *What's the ideal auditing regime you'd like to see going forward?*

B: It would be nice if we had evals that perfectly captured whether or not any given deployment or training run is dangerous, and this guided a strictly enforced regime that everyone had to agree to. But that's too much of a wishful answer.

One thing I think is really important is that we want a lab to have to pass a safety evaluation before they build a much larger model or do something else to improve the model's capabilities — not just before they deploy it. Because internal deployment within a lab or to early customers could be almost as risky as deployment to everyone.

And once your model is potentially capable of catastrophic amounts of destruction, there should just be a pretty high bar for evidence that it's a good idea to deploy this thing.

A: *I'm curious about the kinds of challenges that these evaluations will have to take into account in the future as the world becomes more adapted to AI. People are building infrastructure for AIs to interact with the internet. There will be more AIs in the world, and humans will have more experience with AI. We might come to trust them more. Or we might develop more safeguards.*

B: That's one of the things that's kind of tricky. We're hoping to peg things to some budget for eliciting capabilities: so the auditor gets the base model, and then they get some budget to enhance it and add tools, and the evaluation is based on that enhanced model. But this becomes difficult when the pace of development of elicitation and scaffolding and tools for models to use is very high.

Capabilities might be very different in a year, even if the labs don't do anything. We're hoping to have the labs do more of the work of arguing that their models wouldn't pass our threshold for dangers, even taking into account those developments. After a while we'd check to see if their predictions worked out, and if they underestimated progress then they might have to have evaluations more frequently.

A: *Is there anything else that you want people to know about the work that you're doing?*

B: It's useful for people to think about threat models and exactly how this autonomous replication stuff would work and exactly how far away we are from that. Having done that, it feels a lot less speculative to me. Models today can in fact do a bunch of the basic components of: make money, get resources, copy yourself to your server. This isn't a distant sci-fi scenario.

24

Interview

What We Get Wrong About AI & China

Jeff Ding

Everyone's afraid of what China can and will do with AI. On the ground, the picture looks a lot more complicated.

ILLUSTRATION BY
Gizem Vural

Asterisk: *To start, could you say a little bit about who you are and what you do?*

Jeff: I'm an assistant professor at George Washington University's political science department, where I research emerging technologies and international relations. I also publish a weekly China AI newsletter that features translations of writings by Chinese scholars and bloggers on AI-related topics.

A: *It's a great newsletter. And in it you like to correct misconceptions that American scholars have about AI in China. So I'm curious: Right now, what's the most annoying misconception you see in that area?*

J: I think one of the consistent misconceptions is the overestimation of China's AI capabilities. Part of this stems from the July 2017 national development plan, in which China elevated AI to be a strategic priority. A lot of Western observers just assumed that meant China was a leader in this space. Many prominent voices called it a Sputnik moment — a wake-up call that the U.S. was falling behind in strategic technology.

Overestimation also extends to recent developments in large language models. This happens every time a new Chinese model is released, like Wu Dao 2.0 from a year or so ago: We all thought this was a symbol of bigger, stronger, faster AI from China. And now nobody talks about Wu Dao 2.0. No paper was released. There's no public-facing API. Even the leading competitor to ChatGPT, Baidu's Ernie Bot, does not measure up on a lot of different natural language processing benchmarks. So that would be the number one misconception: this tendency to overhype developments in China.

A: *The sense that I get from hearing people talk about these models is that they're closer to maybe a year behind leading American models than, say, five years or 10 years behind. Is that right?*

J: Yeah. I recently co-wrote a report on trends in these large language models.[1] If you track when GPT-3 was released and when Chinese labs were able to put out alternatives that performed as capably on different benchmarks, it was about one and a half to two years later.

A: *Naively, I'd expect if Chinese labs were clearly capable of building and training these models, and doing it fairly quickly, eventually they'd produce a model that's comparable. Why isn't that happening?*

J: There's a lot more freedom to experiment and push the technological frontier at labs like OpenAI and DeepMind. These are very unique entities. They aren't restricted by needing to meet anything like key performance indicators or other commercial drivers. The best labs in China, by contrast — Alibaba, DAMO Academy, Tencent — have to meet KPIs for making money. There's more leeway and more runway for companies like OpenAI and DeepMind to invest in pushing forward the

1. Jeffrey Ding and Jenny Xiao, "Recent Trends in China's Large Language Model Landscape," Centre for the Governance of AI, April 28, 2023.

technological frontier. So it makes sense that Chinese labs can then invest those resources in that talent and that time into projects developing something similar to GPT-3 only once that trajectory has already been established.

A: *Do you think that's going to change as there's a more widespread awareness of scaling laws — that it seems like you can reliably put in more compute and get out a more powerful model?*

J: Maybe. I'm actually not completely bought in on the "scaling dominates everything" argument. The difference between GPT-3 and ChatGPT was not necessarily a difference of scaling. It was this advance called InstructGPT, which used human input to make the models better and less toxic. That was the type of innovation that actually was missing. When we reviewed 26 different large language models and different labs in China's ecosystem, I did not see anything like InstructGPT. So I think it's not just about the resources. It's also these conceptual and engineering innovations.

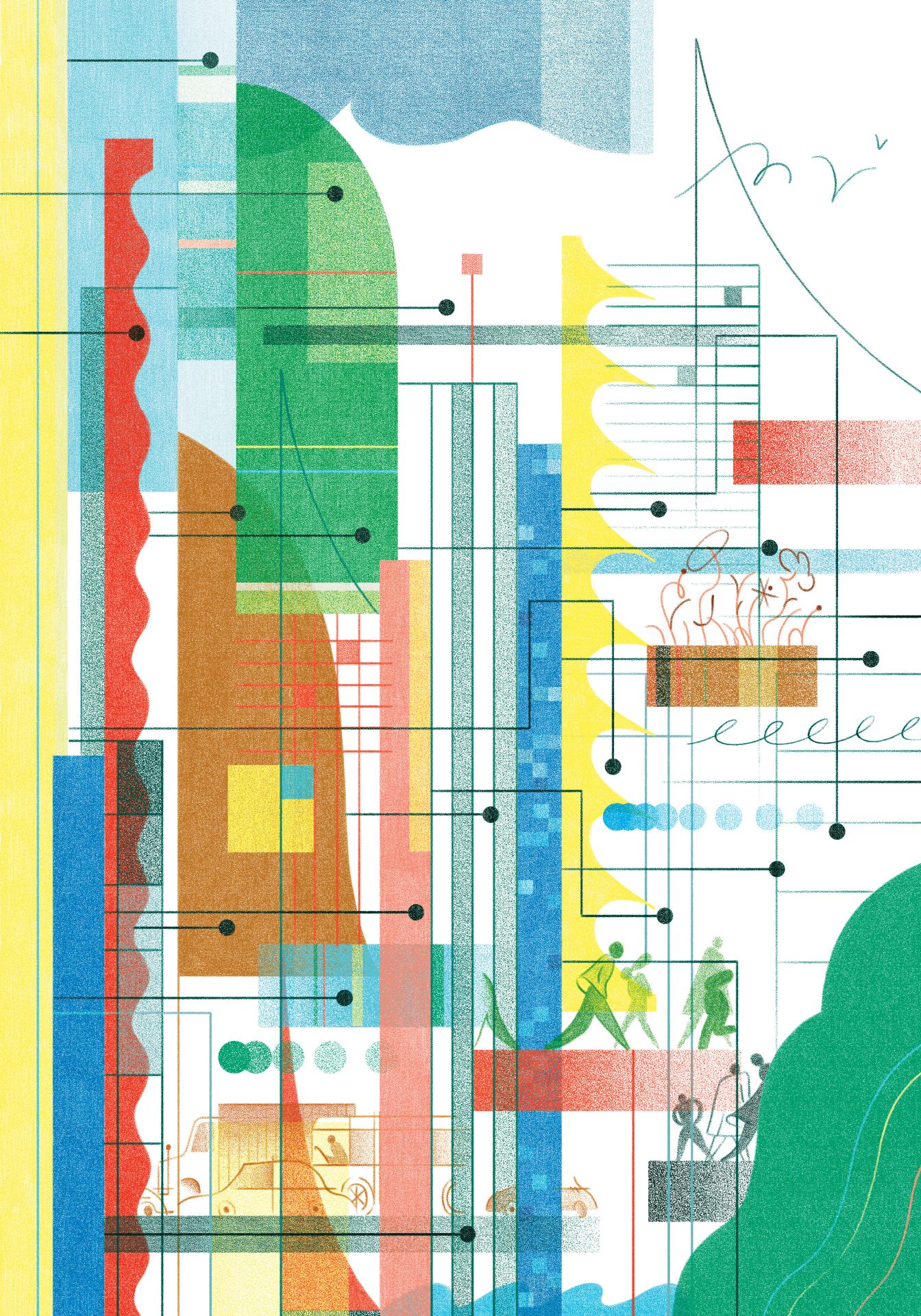
A: *Just to explain, these models are now trained using a technique called reinforcement learning from human feedback (RLHF), where humans provide input that's used to train the model to give responses more like the ones that the humans said were good. Is that technique not commonly used in China?*

J: I have not seen a paper published where RLHF was used to train a large language model. The report only covers 2020 to 2022, so I might have missed something. But yeah, I think the general principle seems obvious and seems like it would be easy to implement. I wouldn't be surprised if actually there's a lot of tacit engineering-related knowledge involved with doing something like InstructGPT. That's actually very hard to discern from just reading the arXiv paper.

A: *Back in 2018, you wrote a report called* Deciphering China's AI Dream.[2] *Another misconception you were trying to correct in that report was that China must have a very centralized, top-down policy on AI, when in fact there are a lot of different bureaucratic, local, and corporate interests that all cut against each other. I'm curious if in observing the past five years you've seen more of a push toward centralization, or if it's still pretty diffuse.*

J: It's a great question and it's been on top of my mind because recently China has implemented some reforms to the Ministry of Science and Technology (MOST), which elevated it to a higher level in terms of guiding the overall direction of science and tech policy. I think one could read that as a driving force toward more centralization, and that's tied to concerns about supply cutoffs and issues with foreign technology dependency. At the same time, what I found very interesting about that reform and reorganization was that they also took away some of MOST's responsibilities. And one of those key responsibilities was overseeing grant management of big science and technology grants.

2. Jeffrey Ding, "Deciphering China's AI Dream: The Context, Components, Capabilities, and Consequences of China's Strategy to Lead the World in AI," Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, March 2018.



This has been a long-standing debate about science and technology policy in China: whether the grants should be managed by bureaucrats or managed and overseen through a more bottom-up process where scientists get more input. So you could read that part of the reorganization as actually decentralizing the grant management process and giving more power back to the scientists. I still think that the push and pull is going to exist, and we see that reflected through the recent MOST reorganization.

A: *In April, China released a new set of regulations on generative AI. I've seen a lot of discussion on how restrictive these rules are for what companies can put into their training data, and how it could cripple Chinese labs. What's your take on that?*

J: It's important to note that these are draft regulations — and often the draft gets significantly revised or softened. We saw that with data localization requirements in the cybersecurity law a few years ago. As to the specific training data provisions, I believe you're referencing the requirements about not using any training data that has personal information attached to it —

A: *And also ensuring that the data be accurate and objective, and doesn't infringe on intellectual property, etc.*

J: If that were enforced strictly to the letter, it would definitely impose certain constraints. I think what's more likely to happen is that those regulations will be lessened.

What I think won't change, though — and I think this will continue to pose a barrier to Chinese companies in this space — are the Chinese government's concerns with internet content providers, especially those providers who have "public opinion properties" and have "social mobilization capacities." Those are terms of art used by the Chinese government. Part of the reason China's internet is so censored is that they put the onus on companies to control their content so it's not politically sensitive. And so to apply that burden not just to WeChat or Baidu search results, but to something like Ernie Bot or another LLM would make it very hard for Chinese companies to meet those requirements.

A: *When OpenAI trains ChatGPT to not say something racist or hallucinate, the thing they're using is RLHF, or something like it. And if Chinese labs don't use those techniques, I can see how it would be extremely difficult for them to make sure that Ernie Bot doesn't start talking about Tiananmen Square.*

J: Right. And you can't ensure that the pre-output censorship that happens in the training process is going to be perfect. They would have to implement some sort of post-model-output censorship stage that OpenAI doesn't have to implement. That's a huge burden. What companies with LLMs might do instead is optimize for business-facing applications that don't have public opinion properties or social-mobilization capacities.

A: *So instead of the proliferation of chatbots, Chinese would see business-facing applications that are mostly invisible to the average consumer.*

J: Yes, exactly.

A: *Talking about how this might diffuse through the economy brings us to another point. You recently wrote a paper breaking down what makes countries competitive technologically, and you drew a difference between nations that lead in innovation and nations that lead in diffusion.*[3] *Could you summarize that argument?*

J: This paper focused on how countries leverage new science and technology advances to sustain higher economic and productivity growth, which historically has been a key step in the rise and fall of great powers. Britain, for example, established productivity leadership and then translated that economic power into military and geopolitical influence after the first industrial revolution. My argument is that when we measure national scientific and technological capabilities, we overweight innovation capacity or other metrics that are closely tied to a country's ability to pioneer new initial advances. And we underweight diffusion capacity, which is a country's ability to diffuse, spread, and embed these advances in productive processes across the whole economy.

Part of the reason China's internet is so censored is that they put the onus on companies to control their content so it's not politically sensitive. And so to apply that burden not just to WeChat or Baidu search results, but to something like Ernie Bot or another LLM would make it very hard for Chinese companies to meet those requirements.

A: *You have a great example of this: In the late 19th century, the United States was pretty weak in innovation capacity. There were not a lot of new innovations coming out of the U.S., but it was very, very good at taking the advances coming from Europe — chemical engineering, among others — and integrating them into industry. Another is that after World War II, the Soviet Union was quite strong in innovation, it had many great scientists, but as a country it struggled to diffuse those advances through their economy. And, of course, we can see how those two situations played out.*

3. Jeffrey Ding, "The Diffusion Deficit in Scientific and Technological Power: Re-assessing China's Rise," Review of International Political Economy, 2023.

J: Exactly. So the idea here is if you only rely on innovation and capacity metrics, you arrive at misleading assessments of a country's ability to sustain growth in the future.

A: *And your argument is that U.S. commentators are too focused on China's innovation capacity, and ignore the fact that China is much weaker on diffusion.*

J: Yeah. Right now, there's a lot of discussion among U.S. policymakers about how the U.S. will soon face this innovation deficit, as framed through indicators like R&D spending, total patents and publications, and high-end STEM talent. I wanted to look more closely at what the diffusion capacity indicators would say. So I looked across fields related to AI: information and communication technologies, cloud computing, and even more basic metrics like household access to computers.

When you look at indicators of diffusion capacity — the adoption rate of different information and communications technologies across businesses, or how close and strong the linkages between academia and industry are — China ranked as a middling science and technology power.

I found that on a lot of these different indicators, China's diffusion capacity was much, much lower than its relative innovation capacity. If you compare indicators of innovation capacity, such as total R&D spending of its top three companies, or the rankings of its top three universities, China scores extremely high. But when you look at indicators of diffusion capacity — the adoption rate of different information and communications technologies across businesses, or how close and strong the linkages between academia and industry are — China ranked as a middling science and technology power.

A: *There are some information technologies that are very widespread in China, like digital cash. Do you have a sense of what determines why, say, WeChat can take over everything so quickly, but for other technologies like cloud computing or industrial robotics there seem to be much deeper barriers?*

J: In some of these areas, such as financial payments, there's just more opportunity for leapfrogging legacy systems. The reason for the fast diffusion in digital payment technologies is that there weren't firmly established legacy methods of credit card payments. Another example is high-speed rail: The government invested heavily in infrastructure, and China became a forerunner in adopting high-speed rail technology at scale.

But when it comes to technologies that have an outsize impact on productivity growth — cloud computing, industrial robotics, industrial software — the ability to leapfrog legacy systems doesn't apply. China will need to invest in earlier generations of the technology and accumulate expertise in a more gradual way. And so in those industries, China will struggle with its diffusion capacity.

A: *So circling back to AI, it doesn't seem obvious which bucket it falls into.*

J: There's a couple of ways to think about it. Another quirk about some of these technologies that China has been able to diffuse at scale is that they are consumer facing. They don't require a lot of complementary skills and technologies to adopt. You don't need a wide pool of talent to ensure the spread of a digital payment technology across an entire country. But you absolutely do when it comes to industrial robotics, software, and cloud computing.

That is one of the factors that makes me think that China's diffusion capacity in AI will follow the same trends that we've seen in cloud computing and industrial robotics. I've looked at different metrics to compare different countries' abilities to train average AI engineers. I testified before the U.S.-China Economic and Security Review Commission recently and presented data on the number of universities in both the U.S. and China that have at least one researcher that has published in a top AI conference. And I believe about 100 universities in China met that very low baseline and about 400 universities in the U.S. surpassed that baseline. I'm only talking here about the talent necessary to fine-tune a large language model that's already been trained and apply it to a specific task.

A: *Which is not nearly as technically complicated as training it in the first place.*

J: Exactly.

A: *And I can also imagine that if, as you said, these models aren't deployed as consumer-facing software because of censorship concerns, that would also slow diffusion as well.*

J: Yeah. And open source techniques which allow for faster diffusion — where capabilities and issues are effectively crowdsourced — aren't available to Chinese companies right now.

A: *So, what can a country do to increase diffusion capacity? It seems like it's much easier to spend a lot of money on R&D than it is to get everyone to adopt slightly better industrial processes.*

J: There are a lot of different factors. One bucket is decentralization. Decentralization often correlates with higher diffusion capacity in science and tech. Instead of picking winners and locking in a particular trajectory, a decentralized ecosystem enables diffusion from the bottom up because the most successful trajectory is allowed to emerge.

Another factor is human capital. The Chinese government has been very good about hitting R&D targets because that's relatively easy to do — they can just mandate spending in different areas. But the government has been much less successful in investing in more widespread technical education. For example, community colleges and vocational training opportunities can raise the average level of engineering — that would be a set of policies that promote more diffusion capacity.

And then the third bucket would be a bunch of random factors that affect diffusion capacity: the latecomer advantage of being able to leapfrog legacy systems, whether a country has a standardized language, the strength of communication channels, culture. I think it's very hard to pin down just a few factors that would affect diffusion capacity.

A: *I want to turn to AI safety. You've written about how, contrary to what some Western commentators seem to think, Chinese AI researchers are concerned about long-term risk and dangers from AI development. How closely does that conversation track the Western conversation?*

J: This is work that needs to be done on a more systematic basis, but we looked at 20 or so different large language models and found that about half of them had a section devoted to ethical, governance, and safety-related issues. It seemed like the focus was mostly on issues of bias, fairness, and toxic content rather than concerns about artificial general intelligence, for instance.

That's not to say that there aren't researchers who are discussing AGI-related concerns. For example, a previous issue of my newsletter featured writings by Nanjing University professor Zhou Zhihua 周志华, who leads one of the top teams in China. He talks about how researchers should not even touch strong AI or a close equivalent to artificial general intelligence. And this was published in the China Computer Federation publication, which features writings from leading computer scientists in China. Those discussions are happening. But I would say discussions on AGI and long-term AI safety issues are not as robust and deep in China as compared to Western countries.

A: *Is your sense that the bias-and-fairness debate is happening in response to the Western debate over the same issues, or are those issues arising independently?*

J: I think a lot of the bias and fairness concerns are coming from the diffusion of norms from Western organizations.

But on other issues, like privacy, that is being driven by the concerns of the Chinese public and a growing backlash to the intrusiveness of AI applications. One of the most important tipping points that we almost never talk about is Chinese delivery drivers. There was this big investigative report about the constraints imposed on delivery drivers by algorithms which calculated how much time they would have to meet their delivery requirements. This really shone a huge spotlight on algorithms and how they play such a huge role in manipulating people's lives.

A: *It's interesting to hear that privacy is such an organic concern — at least as an American with the stereotype of China as a state with no digital privacy.*

J: Privacy concerns look a little different in China, where more of the focus is on the instrumental benefits of privacy — how to prevent someone from hacking into your bank account and stealing all of your money, for instance — rather than privacy as this intrinsic civil right that serves as a check against the worst abuses of government.

But among Chinese academics who might have more of a protected position to say certain things, there's a fair amount who do talk about the need for privacy as more of a civil right or to check against government abuse. I've translated work by scholars such as Tsinghua professor Lao Dongyan 劳东燕, who's criticized the use of facial recognition in the Beijing metro system. And there's actually also been a lot of pushback against the continuing use of QR health codes as COVID control winds down. Oftentimes we only see the surveillance state growing in its reach, but I think this is an example where the surveillance state has been curtailed. So it is a more nuanced picture than just Orwellian, authoritarian government with complete control.

A: *I've been noticing increasing interest in the U.S. in having labs evaluate their models to make sure they're safe. I'm wondering if that's part of the conversation in China at all?*

J: I haven't seen anything like that in China. Even in the States, where it was just announced, we don't see concrete substantiation of those plans.

But it reminds me of different benchmarks in public testing platforms we have in terms of overall capabilities. Different Chinese large language models use benchmarks in natural language processing like GLUE [General Language Understanding Evaluation] to test their models. Actually, I think Megvii, a Chinese facial-recognition company, participated in an MIT labs project to test whether different facial-recognition algorithms performed worse on darker-shaded faces. So if organizations established a recognized public-facing evaluation test bed on AI safety, I could imagine Chinese organizations participating in that and submitting their algorithms and models to that public testing board.

34
A Field Guide to AI Safety

Kelsey Piper

AI safety is starting to go mainstream, but the researchers who've been immersed in it for over a decade still have strong disagreements.

ILLUSTRATION BY
Josh Cochran

It's hard to make progress in a field without a consensus about what it studies or what would constitute a solution to its most important open questions. Not unrelatedly, there hasn't been much progress in the field of ensuring that extremely powerful AI systems don't kill us all, even as there's been growing attention to the possibility that they might.

It'd be a mistake to characterize the risk of human extinction from artificial intelligence as a "fringe" concern now hitting the mainstream, or a decoy to distract from current harms caused by AI systems. Alan Turing, one of the fathers of modern computing, famously wrote in 1951 that "once the machine thinking method had started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control." His colleague I. J. Good agreed; more recently, so did Stephen Hawking. When today's luminaries warn of "extinction risk" from artificial intelligence, they are in good company, restating a worry that has been around as long as computers. These concerns predate the founding of any of the current labs building frontier AI, and the historical trajectory of these concerns is important to making sense of our present-day situation. To the extent that frontier labs do focus on safety, it is in large part due to advocacy by researchers who do not hold any financial stake in AI. Indeed, some of them would prefer AI didn't exist at all.

But while the risk of human extinction from powerful AI systems is a long-standing concern and not a fringe one, the *field of trying to figure out how to solve that problem* was until very recently a *fringe* field, and that fact is profoundly important to understanding the landscape of AI safety work today.

Everyone Disagrees

A May open letter by the Center for AI Safety saying "Mitigating the risk of extinction from AI should be a global priority" had a striking list of signatories, including prestigious researchers in academia and key leaders at the labs building advanced AI systems.

The enthusiastic participation of the latter suggests an obvious question: If building extremely powerful AI systems is understood by many AI researchers to possibly kill us, why is anyone doing it? The simple answer is that AI researchers — in academia, in labs, and in government — disagree profoundly on the nature of the challenge we're facing. Some people think that all existing AI research agendas will kill us. Some people think that they will save us. Some think they'll be entirely useless.

An incomplete but, I think, not uselessly incomplete history of AI safety research would look like this: The research field was neglected for decades, worked on by individual researchers with idiosyncratic theories of change and often with the worldview that humanity was facing an abrupt "intelligence explosion" in which machines would rapidly surpass us. Eliezer Yudkowsky and the Machine Intelligence Research Institute are representative of this set of views.

In the last 10 years, rapid progress in deep learning produced increasingly powerful AI systems — and hopes that systems more powerful still might be within reach. More people have flocked to the project of trying to figure out how to make powerful systems safe. Some work is premised on the idea that AI systems need to be safe to be usable at all: These people think that it'll be difficult to get any commercial value out of unsafe systems, and so safety as a problem may effectively solve itself. Some work is premised on safety being difficult to solve, but best solved incrementally: with oversight mechanisms that we'll tinker with and improve as AI systems get more powerful.

Some work assumes we'll be heavily reliant on AI systems to check each other, and focuses on developing mechanisms for that. Nearly all of that work will be useless if it's true we face an overnight "intelligence explosion."

One might expect that these disagreements would be about technical fundamentals of AI, and sometimes they are. But surprisingly often, the deep disagreements are about sociological considerations like how the economy will respond to weak AI systems, or about biology questions like how easy it is to improve on bacteria,[1] or about implicit worldviews about human nature, institutional progress, and what fundamentally drives intelligence.

There are now quite a few people — more than 100, if less than 1,000 — across academic research departments, nonprofits, and major labs, who are working on the problem of ensuring that extremely powerful AI systems do what their creators want them to do.

Many of these people are working at cross-purposes, and many of them disagree on what the core features of the problem are, how much of a problem it is, and what will likely happen if we fail to solve it.

That is, perhaps, a discouraging introduction to a survey of open problems in AI safety as understood by the people working on them. But I think it's essential to understanding the field. What it's going to take to align powerful AI systems isn't well understood, even by the people building them. There are people working side by side on the same problems, some of whom think they are facing near-certain death and some who think there's about a 5% chance of catastrophe (though the latter will still hasten to note that a 5% chance of catastrophe is objectively quite high and worth a lot of effort to avoid).

If you aren't confused, you aren't paying attention.

The "Intelligence Explosion" and Early Work on Preventing AI Catastrophe

In the last half of the 20th century, conceptions of AI tended to envision mechanical systems designed by writing code. (This is reasonable to imagine, but importantly not how modern deep learning actually works.) Turing envisioned that AIs might use a decision rule for weighing which action to take, as well as a process by which we could insert better and better decision rules. Another common element in early conceptions of AIs was the idea of recursive self-improvement — an AI improving at the art of making smarter AI, which would then make even smarter AI, such that we'd rapidly go from human-level to vastly superhuman-level AI.

In a 1965 paper, pioneering computer scientist I. J. Good posed the first scenario of runaway machine intelligence:

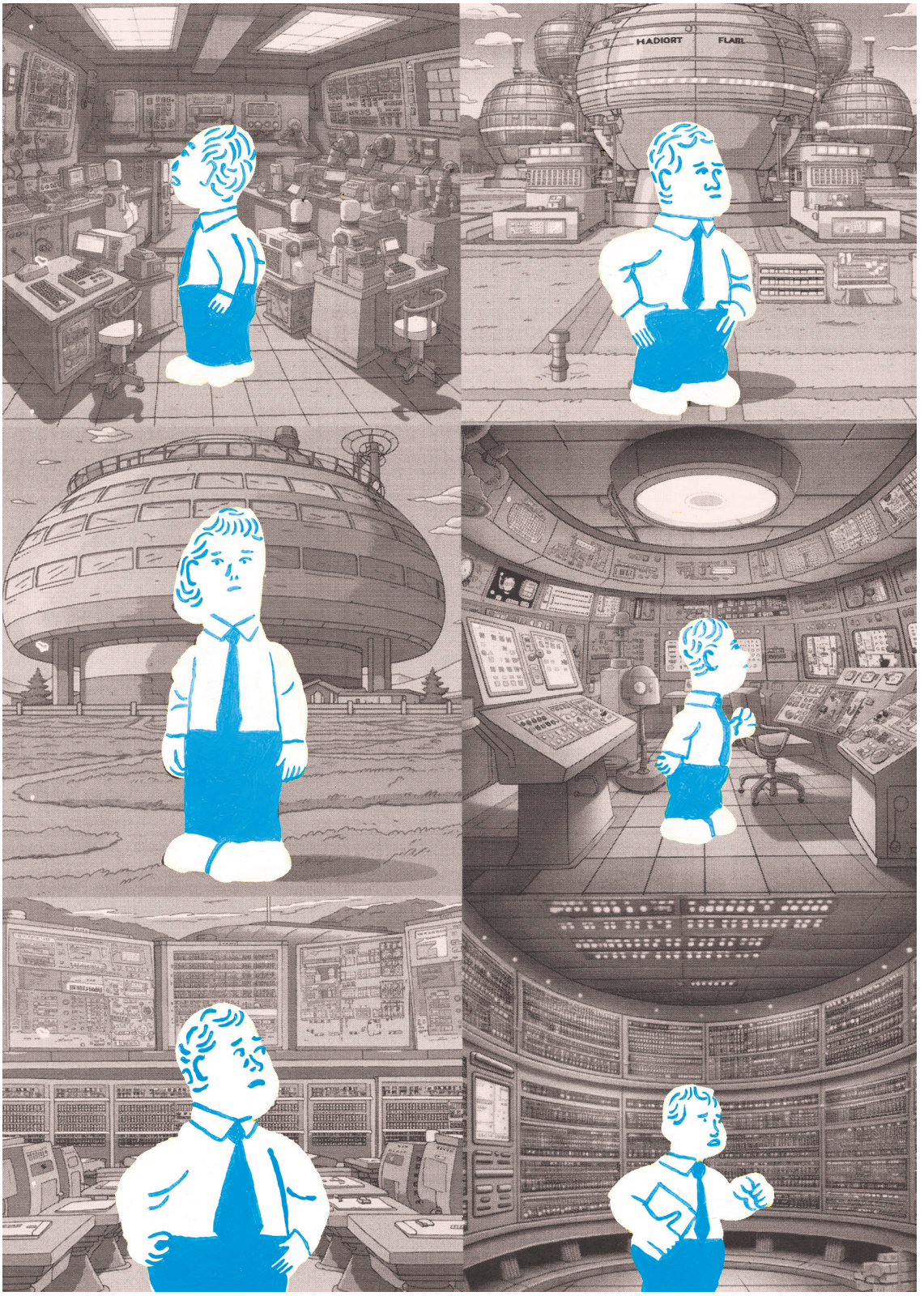
> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.

Good used the term "intelligence explosion," but many of his intellectual successors picked the term "singularity," made popular by mathematician, computer science professor, and science fiction author Vernor Vinge.

This is the basic set of intuitions that shaped nearly all discussion about superintelligent AI until quite recently. The possibility of self-improving AIs, intelligence explosions, and the singularity was largely discussed in the overlapping, tech-positive, sci-fi influenced futurist, Extropian, and transhumanist communities; at the time, very few others were considering the question of how to build powerful AI systems safely at all.

Much AI safety work in the 1990s and 2000s — especially by Eliezer Yudkowsky and the nonprofits he founded, the Singularity

1. You might wonder why the question "How easily can we improve on bacteria?" would matter at all to AI safety. The short answer is that if an AI could invent its own superbacteria that outcompete existing ones, it'd be pretty easy to take over the world. If that is out of reach even for an extremely intelligent biology AI, then taking over the world would likely be a less sci-fi affair.


HADIORT
FLABL

Institute and then the Machine Intelligence Research Institute — emerged from this set of assumptions. The specific claim that there'll be a turning point is a crucial one separating this worldview from others. Yudkowsky and those who hold this belief tend to think that intelligence — in entities both artificial and biological — has a critical point — call it generalization, or reflectivity, or the thing that separates humans from chimpanzees. Humans, possessing this ineffable quality, have built civilizations that wildly surpass anything any other species could do. AIs think faster than us, and unlike us they can copy themselves and adjust their own minds. Once they cross that threshold, they'll surpass us fast.

What does this worldview suggest about AI safety? It suggests that gradual and incremental approaches, where we build steadily more powerful systems, figure out how they work, figure out how to align them towards human objectives, and then take the next step up in intelligence, probably won't work. At some point your system will unexpectedly develop the ability to rapidly amplify its own intelligence, or it will figure out how to design successors and do that, or someone else who isn't being as cautious as you will do one of those things and surpass you overnight.

In this view of AI safety, we "get one shot" — we can't learn from alignment failures, as we won't notice them until our systems are superhuman. We don't benefit much from having a nearly aligned system — it's not a problem where you get 90% of the benefit from solving 90% of the problem. If a system is almost aligned, its vastly amplified successor won't even be close. And this worldview envisions fairly little useful human input as the system rapidly ramps up, because that ramp-up is expected to happen in the blink of an eye.

A weak form of Yudkowsky's claims here seems very likely to be true. Certainly, any AI system that is useful at all will be, among other things, useful for designing more AI systems. Already, today's very weak AI systems are labor-saving devices for programmers, and thus probably somewhat hasten the advent of their successors. But the strong version of the claims seems much more uncertain. (That's not to say that they're demonstrably or obviously false. Many of Yudkowsky's most outspoken critics will say, when pressed, that he might be describing a real problem that will really destroy us; he's just excessively confident of it.)

Yudkowsky has written that the "most likely result of building a superhumanly smart AI, under anything remotely like the current circumstances, is that literally everyone on Earth will die." "Most likely," he's stated elsewhere, corresponds to a 99% chance that vision comes true.

This is why people will colloquially refer to him, and those who agree with him, as "doomers." Doomers are, I think, best characterized as people working from the intelligence-explosion premise, skeptical of the accounts of why it might not apply, inspired by Vinge and Yudkowsky and Oxford's Nick Bostrom, generally giving very high probabilities that AI kills us all.

If that sounds like a rough place to find oneself, it is. Little of the AI alignment work that has started from these or similar premises is in a promising place today, even in the eyes of its own proponents. Instead, to the extent that these premises are correct, we should just stop building powerful AI systems, indefinitely, until we have a better idea of how to kick off the avalanche that is a self-improving superintelligence without catastrophe.

We should return, then, to the question I opened with: "If building extremely powerful AI systems is understood by many AI researchers to probably kill us, why is anyone doing it?" The answer is that they disagree with the Yudkowskian worldview in one or more details, have a different model of the threat, and therefore think that work today is likely to be helpful in solving the problem in time. There's no bizarre paradox where people are funding work that, in their worldview, is likely to kill us, just intense disagreement about what is likely to kill us and therefore what work might help.

Open Philanthropy and Friends

The second major worldview in AI safety is associated with Holden Karnofsky (co-CEO of Open Philanthropy, currently at the Alignment

Research Center) and his Open Philanthropy colleagues Ajeya Cotra, Joe Carlsmith, and Tom Davidson, as well as Paul Christiano (head of ARC; also Cotra's husband). Some of them, like Christiano, became interested in the alignment problem independently. Others, like Karnofsky, encountered the idea through Yudkowsky and others at the Singularity Institute in the early 2010s, but came to develop their own views.

To understand their perspective, we'll have to get into more detail about what our existing methods for making AI systems do what we want them to do are, and why they might stop working when AI systems get sufficiently powerful.

ChatGPT and similar releases from OpenAI were trained with reinforcement learning from human feedback — a technique Christiano helped develop. In RLHF, humans rate output by the models, and the models then learn how to give answers that humans would rate highly. RLHF is imperfect and the AI is sometimes wrong about what answers would get positively reinforced, but it will probably get better. The bigger problem might be that RLHF, and similar techniques, fundamentally teach AIs to say what we want to hear, not to do what we'd want them to do if we had full context on their decision-making.

The worry here is that as we build more powerful systems, the small disconnects between what we're training them to do and what we think we're training them to do will be magnified. AI systems that are good at lying to us will — on various evaluations — outperform AI systems that are trying to be candid. (Here, Cotra gives the example of the Catholic Church circa 1500 trying to train an AI. If this AI correctly reported that the Earth revolved around the sun, it would be rated more negatively than if it said the opposite.) Without specific countermeasures, AIs trained this way will have every incentive to manipulate us, and to hack and falsify the mechanisms we use to monitor them.

RLHF and related techniques also make AIs much more useful. If it's possible to build systems that are unaligned but commercially viable, then over time we will likely build them, interact with them, and use them for economic activity at extraordinary scale. 100 million people were using ChatGPT within weeks of its launch. People have built language-model-powered businesses. Industries are being revolutionized by language-model-driven automation.

And all of that is happening while present-day language models are in their infancy, with severe limitations that seem likely to be a product of our inexperience with the technology. Language models today are vastly better than they were five years ago. If they improve by remotely similar amounts in the future, they will be able to automate significant fractions of human labor — including the labor that goes into developing better AI systems. The economic implications will be enormous.

You can probably build useful and powerful systems that pose no risk to human civilization. But it seems equally obvious that, at some point, if you build a parallel economic society of billions of entities that can do most or all of the things humans can do, you're in a situation ripe for losing control of the world. That might take the form of a spectacular sci-fi conquest by AIs using advanced weapons or plagues they invented. It might be geopolitical — AIs siding with one nation to help it crush its adversaries, in exchange for enormous power in the aftermath. It might be legal — AIs purchasing virtually all land and all capital. It might involve sophisticated manipulation. But without committing to any of those stories, it seems like a world where we don't need to solve alignment for commercial viability, don't solve alignment, attain commercial viability, and go full speed ahead could be a world where, as Karnofsky put it, we "sleepwalk into AI catastrophe."

This worldview suggests some obvious ideas about which avenues of research are promising. It's crucial to detect whether your AI is actually aligned. It's important to understand what current AIs are capable of, so you know when you get to the brink of potentially catastrophic problems. And of course it's important to develop alignment techniques that labs will adopt even if they aren't

necessary for commercial viability. If we do all of that, or even just do some of that and get a bit lucky, or if advanced AI systems are able to make any of that work easier to do, it feels plausible that humanity can score a big win.

I want to highlight some fundamental and important disagreements between this worldview and the Yudkowskian one, because their premises sound superficially similar: Both are concerned with the possibility of AI systems being usable to automate research into AI systems, enabling fast growth as new algorithmic improvements and hardware improvements are developed.

First, while researchers at Open Philanthropy generally believe that superintelligent AIs will be developed, they don't think that this is necessary for AIs to seize power. They tend to be less concerned with raw intelligence than with the resources and information AIs have access to. If superintelligent AIs outnumber humans, think faster than humans, and are deeply integrated into every aspect of the economy, an AI takeover seems plausible — even if they never become smarter than we are. This means that decisions about how AIs are deployed also have important implications for safety. The more control humans choose to retain over things like the supply chains that produce microchips, the harder it will be for AI to defeat us.

Second, the Open Philanthropy worldview isn't premised on the assumption that there will be a "hard takeoff" where AIs rapidly become superintelligent. There's still broad disagreement about how long this might take, but those who believe it will happen fast think that it will still likely be continuous. Instead of a single intelligence switch that can be flipped on or off, they think that AIs will probably get gradually smarter. This means that superintelligent AIs might have a lot in common with the much less capable systems that exist today, just as GPT-4 is smarter than GPT-2 while sharing the same fundamental architecture.

Critically, this means that tools for alignment and oversight that work on existing AI systems might actually be useful for helping us align future superintelligences. Because of concerns that RLHF and related methods might make AIs more likely to deceive humans, many of these tools involve figuring out what a model is "really thinking," whether by looking directly at its weights or by verifying certain mathematical properties of its behavior.

It seems entirely possible that existing proposals won't be sufficient to get the desired behavior from extremely powerful systems. But if they can get desired behavior from moderately powerful systems, and then we can develop better proposals with the aid of those moderately powerful systems, we might get somewhere. The more alignment is a matter of a "grab bag" of tools and techniques, rather than a project that requires a comprehensive structural solution, the more this approach looks viable.

An Optimistic View of AI Safety — and Where It May Fall Short

Over the next five years, it seems very likely that we'll develop more powerful AI systems, and that the effort of ensuring they do what their creators intend will intensify. There's a spectrum of views about how that will go. The most optimistic, of course, is that it will be easy to make systems do precisely what we want.

It's hard to point to a single outspoken partisan of this view: People who hold it tend to regard much of the AI safety conversation as a waste of time, and thus tend not to phrase their beliefs in its terms. Yann LeCun, chief AI scientist at Meta, has views that fall under this broad umbrella: We'll just tell the powerful AI systems what to do, and figure out how to get them to do that, and it won't be that hard or that catastrophic to get slightly wrong. "To *guarantee* that a system satisfies objectives, you make it optimize those objectives at run time," he argued in a recent Twitter conversation with Eliezer Yudkowsky. His proposal, he said, "is a way to guarantee that AI systems be steerable and aligned."

The claim that this or any existing proposal guarantees that AI systems will be steerable and aligned is false and, frankly, unserious (though, to be fair, LeCun's actual views are probably somewhat more nuanced than his tweets). "Unless a breakthrough is achieved in

AI alignment research ... we do not have strong safety guarantees," Yoshua Bengio, a leading AI researcher and one of the pioneers of deep learning, argued in a recent analysis.

But without going so far as to claim that alignment is guaranteed, a decent share of researchers at AI labs from OpenAI to Google to Meta expect it to be not all that difficult. That is, our existing methods will be approximately sufficient, errors will be obvious and easy to correct, and the techniques we use to align those systems will continue to produce, reliably and robustly, the behavior the creators intended, even as AI systems get steadily more intelligent, powerful, and difficult to oversee.

A variation on that, or a different phrasing: If we are lucky, maybe AI alignment will be effectively reducible to the problem of making AI systems that have a low enough error rate to be commercially useful — that is, the alignment problem will turn out to be just the problem of hallucinations plus the problem of robustness against adverse inputs plus the problem of getting high-quality outputs out of an AI system at all.

If this is true, we'll solve alignment incidentally along the way to building commercially valuable AI systems. This isn't an argument for not working on alignment — indeed, it suggests working on it will be ludicrously profitable! — but it's an argument against slowdowns, pauses, or regulation. AI work can proceed as fast as possible, and simply won't be useful until we're good at alignment.

How plausible is this worldview? It certainly feels to me like there are some bad signs in present-day AI systems that proponents of this view need to explain away. Our existing techniques for achieving desired behaviors feel non-robust. Engineers are constantly catching loopholes and plugging new holes against adversarial inputs. And if the Open Philanthropy worldview is right, then we're not training AI systems to do what we want, but to tell us what we want to hear.

Much AI safety work at leading labs from Google DeepMind to OpenAI to Anthropic today is aimed at changing that — trying to develop more robust, more general, and more powerful techniques for getting desired behavior from AI systems, so that we can reasonably expect the techniques to keep working when the systems they're applied to are really smart. As with Open Philanthropy, many of these techniques depend on training less powerful AIs to help supervise increasingly more powerful systems. I think it's far too soon to count out this kind of "mundane solution" to the alignment problem — but it also seems far too soon to feel confident it'll work. Maybe alignment will turn out to be part and parcel of other problems we simply must solve to build powerful systems at all.

Some people have called AI safety a "pre-paradigmatic field." A survey like this makes it a bit clearer what that means. Growing agreement that there's a problem hasn't yet translated to much accord about what a solution would look like.

If you've encountered one account of AI safety and found it unpersuasive, you should shop around for another. People have wildly different conceptions of the problem, and disagreeing vehemently with one account of it doesn't mean you'll disagree with others. Those who dismiss existential risk concerns as too convenient for corporate interests might want to check out the case for existential risk concerns from those who think every extant AI lab is committing an ongoing indefensible moral evil; those who find Yudkowsky dangerously doomist might find a different accounting of the problem more credible.

My own most fundamental takeaway is that, when there is this much uncertainty, high-stakes decisions shouldn't be made unilaterally by whoever gets there first. If there were this much expert disagreement about whether a plane would land safely, it wouldn't be allowed to take off — and that's with 200 people on board, not 8 billion.

42
Through a Glass Darkly

Scott Alexander

Nobody predicted the AI revolution, except for the 352 experts who were asked to predict it.

In 2016, three years before OpenAI released GPT-2 and the world went crazy, an independent researcher named Katja Grace cold-emailed the world's leading AI scientists. She had some questions. A lot of questions, actually. When will AI be able to fold laundry? Write high school essays? Beat humans at Angry Birds? Why doesn't the public understand AI? Will AI be good or bad for the world? Will it kill all humans?

The world's leading AI scientists are a surprisingly accommodating group. Three hundred fifty-two of them took time out of their busy schedules to answer, producing a unique time capsule of expert opinion on the cusp of the AI revolution.

Last year, AI started writing high school essays (laundry folding and Angry Birds remain unconquered). Media called the sudden rise of ChatGPT "shocking," "breathtaking," and "mind-blowing." I wondered how it looked from inside the field. How did the dazzling reality compare to what experts had predicted on Grace's survey six years earlier?

Looking at the most zoomed-out summary — whether they underestimated progress, over-hyped it, or got it just right — it's hard to come to any conclusion other than "just right."

The survey asked about 32 specific milestones. Experts were asked to predict the milestones in several ways. In what year did they think it was as likely as not that AI would reach the milestone? In what year did they think there was even a 10% chance AI would reach it? A 90% chance? What did they think was the chance AI would reach the milestone by 2026? By 2036? I focus on their median prediction of when AI will reach the milestone.

Another way of framing "50% confidence level" is "you're about equally likely to get it too early as too late." The experts got six of these milestones too early and six too late, showing no consistent bias towards optimism or pessimism.

And when they were wrong, they were only wrong by a little bit. Grace asked the experts to give their 90% confidence interval. Here the experts were wrong only once — they were 90% sure AI would have beaten humans at the video game Angry Birds by now, but it hasn't.

The accuracy here is mind-boggling. In 2016, these people were saying, "Yes, AI will probably be writing high school history essays in 2023." I certainly didn't expect that, back in 2016! I don't think most journalists, tech industry leaders, or for that matter high school history teachers would have told you that. But this panel of 352 experts did!

I would be in awe of these people, if not for the second survey.

Prediction Is Very Difficult, Especially About the Past

The six years between 2016 and 2022 were good ones for AI, forecasting, and Katja. AI got billions of dollars in venture capital investment, spearheaded by fast-growing startup OpenAI and its superstar GPT and DALL-E models. The science of forecasting, which only reached public attention after the publication of Philip Tetlock's *Superforecasting* in late 2015, took off, and started being integrated into government decision-making. As for Katja, her one-person AI forecasting project grew into an eight-person team, with its monthly dinners becoming a nexus of the Bay Area AI scene.

In summer 2022, she repeated her survey. The new version used the same definition of "expert" — a researcher who had published at the prestigious NeurIPS or ICML conferences — and got about the same response rate

(17% in 2022 compared to 21% in 2016). The new asked the same questions with the same wording. Most of the experts were new, but about 6% (45 out of 740) were repeats from the previous round. You can never step in the same river twice, but this survey tried hard to perfectly match its predecessor.

This time, nine events happened earlier than the experts thought, and zero happened later, or on time. In fact, eight of the nine happened outside their 90% confidence interval, meaning the experts thought there was less than a 10% chance they would happen as early as they did!

But actually it's much worse than that. In 2019, a poker AI called Pluribus beat human players — including a World Series of Poker champion — at Texas hold 'em (the *Scientific American* article was called "Humans Fold: AI Conquers Poker's Final Milestone"). All three of the judges agreed that this satisfied milestone 31: "Play well enough to win the World Series of Poker." Still, Katja wanted to make her survey exactly like the 2016 version, so she included this and several other already-achieved milestones. The experts predicted it wouldn't happen until 2027. Same with image categorization and Python Quicksort — both happened in 2021; in both cases the 2022 experts predicted it would take until 2025. Yogi Berra supposedly said that "prediction is very difficult, especially about the future." But in this case the 2016 panel predicted the future just fine. It was the 2022 panel that flubbed predictions about things that had already happened!

Maybe this was an unfair trick question? It wasn't impossible to answer zero (a few respondents did!), but maybe it was so strange to see already-achieved milestones on a survey like this that the experts started doubting their sanity and assumed they must be misunderstanding the question. By extreme good luck, we have a control group we can use to answer this question. Several of the milestones were first achieved by ChatGPT, which came out just three months after the survey ended. These weren't trick questions — they hadn't been achieved as of survey release — but the correct answer would have been "basically immediately." Did the experts get this correct answer?

No. The judges ruled that ChatGPT satisfied five new milestones. The experts' prediction for how long it would take an AI to achieve these milestones (remember, the right answer was three months) were five, four, five, 10, and nine years — about the same as they gave any other hard problem.

And there was a truly abysmal correlation (around 0.1-0.2, depending on how you calculate it) between the tasks experts thought would be solved fastest, and the ones that actually got solved. The task experts thought would fall soonest was — once again — Angry Birds. And among the tasks that have remained unconquered, even as AI has made astounding progress in so many other areas of life is — once again — Angry Birds.

(The transhumanists say that one day superintelligent AIs running on cryogenic brains the size of Jupiter will grant us nanotechnology, interstellar travel, and even immortality. The most trollish outcome — and the outcome toward which we are currently heading — is that those vast, semidivine artifact-minds *still* won't be able to beat us at Angry Birds.)

This exceptionally poor round of new predictions looks even worse when viewed beside their past successes. In 2016, respondents predicted AI would be able to write high school essays that would receive high grades in 2023 (i.e., exactly right). In 2022, their median prediction extended out to 2025. How did they get so much worse?

Doubt Creeps In

In retrospect, the seemingly accurate 2016 survey had some red flags.

The survey asked the same questions in multiple different ways. For example, "When do you think there's a 50% chance AI will be able to classify images?" and "How likely is it that AI can classify images in ten years?" The answers should line up: If experts give a 50% probability of AI classifying images in 10 years, the chance of AI classifying images in

10 years should be 50%. It wasn't. In this particular case, experts asked when AI would have a 50% chance of classifying images answered 2020; when asked their chance of AI classifying images in 2026, they said 50%.

The survey's most dramatic question — when AI would reach "human level" — was worst of all. Katja asked the question in two different ways:

1. When AI would achieve high-level machine intelligence, defined as "when unaided machines can accomplish every task better and more cheaply than human workers."

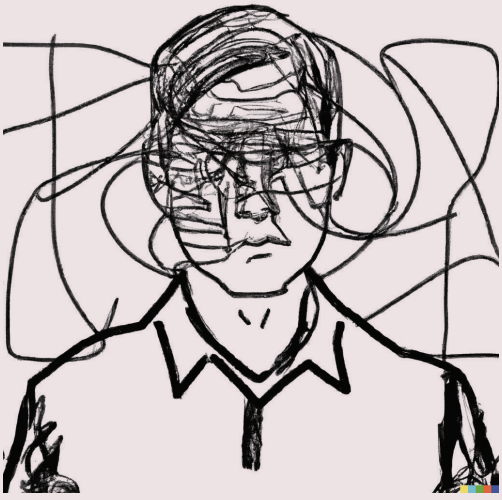
2. At the end of a list of questions about specific occupations, the survey asked when all occupations could be fully automated, defined as "when for any occupation, machines could be built to carry out the task better and more cheaply than human workers."

In her write-up, Katja herself described these as different ways of asking the same question, meant to investigate framing effects. But for framing 1, the median answer was 2061. For framing 2, the median answer was 2138.

Most people don't have clear, well-thought-out answers to most questions. Famously, respondents to a 2010 poll found that more people supported gays' right to serve in the military than supported homosexuals' right to serve in the military. I don't think people were confused about whether gays were homosexual or not. I think they generated an opinion on the fly, and the use of a slightly friendlier-sounding or scarier-sounding term influenced which opinion they generated. The exact wording wouldn't shift the mind of a gay rights zealot or an inveterate homophobe, but people on the margin with no clear opinion could be pushed one way or the other.

But this was more than a push: AGI in 45 years vs. 122 years is a big difference!

Gay rights are at least grounded in real people and political or religious principles we've probably already considered. But who knows when human-level AI will happen? Many of these experts were people who invented a new computer vision program or helped robot arms assemble cars. They might never have thought about the problem in these exact terms before; certainly they wouldn't have complex mental models. These are the kinds of conditions where little changes in wording can have big effects.



DALL·E 2023-05-24 16.57.40 - sketch of scientists predicting the future of artificial intelligence

Platt-itudes

There's an energy wonk joke that "fusion power is 30 years in the future and always will be." The AI version is Platt's Law, named for Charles Platt, who observed that all forecasts for transformative AI are about 30 years away from the forecasting date. Thirty years away is far enough that nobody's going to ask you which existing lines of research could produce breakthroughs so quickly, but close enough that it doesn't sound like you're positing some specific obstacle that nobody will ever be able to overcome. It's within the lifetime of the listeners (and therefore interesting), but probably outside the career of the forecaster (so they can't be called on it). If you don't have any idea and just want to signal that AI is far but not impossible, 30 years is a great guess!

Katja's survey didn't quite hit Platt's Law — her respondents answered 45 years

DALL·E 2023-05-24 17.00.11 - sketch of researchers discussing AI 10 years ago

on one framing, 122 years on another. But I wonder if Platt's reasoning style — what kind of distance from the present sounds "reasonable," what numbers will correctly signal support for science and innovation and the human spirit without making you sound like a rosy-eyed optimist who expects miracles — is a more useful framework than the naive model where forecasters simply consult their domain expertise and get the right answer.

Regardless of what particular year it is, saying the same number signals the same thing. If "this problem seems hard, but not impossible, and I support the researchers working on it" is best signaled by providing a six-year timeline, this will be equally true in 2016 and 2022. If you ask someone in 2016, they'll say it will happen in 2022. If you ask them in 2022, they'll say it will happen in 2028. If in fact it happens in 2023, the people who you asked in 2016 will look prescient, and the people who you asked in 2022 will look like morons. Is that what happened here?

The mean advance on these milestones was about one year. But this was heavily influenced by three outliers, shown as -29, -24, and -14 above. The median is less sensitive to outliers — and it was three years. That is, over six years, the date that experts predicted we would achieve the milestones advanced three years. So we're about halfway between the perfect world where everyone predicts the same year regardless of when you ask them (barring actual new information), and the Platt's Law world where everyone predicts the same distance away no matter what year you ask the question in.

In the 2016 survey, this tendency didn't hurt. Experts predicted the easy-sounding things were about three years away, the medium-sounding things five to 10 years away, and the hard-sounding things about 50 years away. In the 2022 survey, they did the same. Unfortunately for them, in 2022 the medium-sounding things were only months away, or had already been achieved, and their seemingly good performance fell apart.

The Tournament

It seems like most of the AI experts weren't prepared for difficult prediction questions. What if we asked prediction experts?

Metaculus is a cross between a website and a giant multi-year, several-thousand-question forecasting tournament. You register and make predictions about things. Most of them are simple things that will happen in a month or a year. When a month or a year goes by, the site grades your prediction and grants or fines you points based on how you did compared to other players.

The fun part is the Metaculus Prediction for each question. It's not just the average forecast of everyone playing that question, it's the average forecast *weighted by how often each forecaster has been right before.*

Some Metaculans are "superforecasters," University of Pennsylvania professor Philip Tetlock's term for prognosticators with an uncanny knack for making good guesses on questions like these. Superforecasters might not always be experts in the domains they're making predictions in (though they sometimes are!), but they make up for it by avoiding biases and failure modes like the ones that

plagued the experts above. Whatever the weighting algorithm, it will probably disproportionately capture this upper crust of users.

Is AI Harder To Forecast Than Other Things? Let's Find Out!

Metaculus has dozens of questions about AI, including the inevitable Angry Birds forecast.

Because everyone's scores are tracked, well, meticulously, it has great data on how these forecasts have gone in the past. Forecaster Vasco Grilo has collected data on how Metaculus has done predicting 1,373 different binary yes-or-no questions (like "Will Trump win the election?"). Fifty-six of these questions are about AI (like "Will Google or DeepMind release an API for a large language model before April?"). He found that for both AI categories and all categories, Metaculus's forecasts did much better than Laplace's rule of succession (a formula for predicting the likelihood of a specific event in a sequence, based on how frequently that event occurred in the past). But the effect was weaker for AI-related questions (score difference of 0.88) than for all questions (score difference of 1.25).

So Metaculus forecasts are definitely better than nothing (including on AI). But the AI forecasts are less accurate than other forecasts: The score improvement between the guess and the forecast is only about half as big. Does this mean that forecasting AI is especially hard? Not necessarily. It could be that Metaculus chooses harder questions for AI, or that Metaculus users are experts in other things but not in AI. But the data is definitely consistent with that story.

Okay, but When Will We Have Human-Level AI?

The two most popular AI questions on Metaculus, with thousands of individual forecasts, are on "general AI" (i.e., AI that can perform a wide variety of tasks, just like humans).

The first question ("Easy") asks about an AI that can pass the SAT, interpret ambiguous sentences, and play video games. The second ("Hard") asks about an AI that can answer expert-level questions on any subject, pass programming interviews, and assemble a Lego set. Both questions also require the AI to be able to pass a Turing test and explain all its choices to the judges. These are lower bars than Katja's question about an "AGI that can do all human tasks," but not by much — in another question, the forecasters predict it will only be one to five years between AIs that beat the first two questions and AIs that can beat humans at everything.

Although Easy is a little older than Hard, since both questions have existed they've more or less moved together, suggesting that the movements reflect AI progress in general and not the specific bundle of tasks involved.

Easy starts at 2055, drops to 2033 after GPT-3, then starts rising again. It stays high until early 2021, then has another precipitous drop around April 2022, after which it stays about the same — neither ChatGPT nor GPT-4 affects it very much. So what happened in April 2022? The forecasters are following along in the comments section, and they have the same question.

Most of the commenters blamed Google. In April 2022, the company released a paper describing its new language model PaLM. PaLM wasn't any higher-tech than GPT-3, but it was trained on more powerful computers and therefore did a better job. The researchers showed that previously theoretical scaling laws — rules governing how much smarter an AI gets on more powerful computers — appeared to hold.

Then in May, DeepMind released a paper describing a "generalist" model called Gato, writing that "the same network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens."

Neither of these illuminated deep principles the same way GPT-2 and GPT-3 did, and neither caught the public eye the same way as ChatGPT and GPT-4. But this was when the Metaculus estimate plummeted. Some forecasters defended their decision to change

their prediction in the comments. User TryingToPredictFuture:

> The PaLM paper indicates that Google is now capable of efficiently converting its vast funds into smarter and smarter AIs in an almost fully automatic manner.
>
> The process is not blocked by theoretical breakthroughs anymore. Google is now in the territory where they can massively improve the performance of their models on any NLP benchmark by simply adding enough TPUs. And there is no performance ceiling in sight, and no slowdown.

And Tenobrus:

> My update was based on the fact that GPT-3 and other papers at the time predicted a plausible seeming scaling law, but recent results actually confirm that scaling law continues (plus displays discontinuous improvement on some tasks). Even though these results were predictable, they still remove uncertainty.

Others found the sudden change indefensible, for example top-100 forecaster TemetNosce:

> The community was wildly out of line with progress in the field beforehand, and arguably still are. Bluntly I'm more concerned with whether any given AI will do all these tests or get said statement than whether one could in the next decade. My default remains that it'll happen sometime mid-late this decade.

Reading the comments, one cannot help but be impressed by this group of erudite people, collaborating and competing with each other to wring as much signal as possible from the noise. Some of the smartest people I know compete on Metaculus — and put immense effort into every aspect of the process (especially rules-lawyering the resolution criteria!).

But the result itself isn't impressive at all. If we believe today's estimate, then the estimate three years ago was 25 years off. Users appear to have over-updated on GPT-3, having slashed 20 years off their predicted resolution date — then added 15 of those years back for approximately no reason — then gone down even further than before on some papers which just confirmed what everybody was already kind of thinking.

I find OpenAI employee Daniel Kokotajlo's summary of Metaculus's AI forecasting more eloquent than anything I could come up with myself:

Kokotajlod

> Sometimes updates happen not because of events, but rather because of thinking through the arguments more carefully and forming better models. Even this kind of update, however, often happens around the same time as splashy events, because the splashy events cause people to revisit their timelines, discuss timelines more with each other, etc.
>
> (Speaking as someone who hasn't updated as much on recent events due to having already had short timelines, but who hadn't forecasted on this question for almost a year (EDIT: two years!) and then revisited it in April and May. Also an "event" that caused my shorter timelines was starting a job at OpenAI, but mostly it wasn't the stuff I learned on the job, it was mostly just that I sat down and tried to build models and think through the question again seriously, and so there were new arguments considered and new phenomena modelled.)

The Model

Maybe (some people started thinking around 2020) people's random guesses about when we'll get AGI are just random guesses. Maybe this is true even if the people are very smart, or even if we average together many people's random guesses into one median random guess. Maybe we need to actually think deeply about the specifics of the problem.

One group of people thinking about this was Open Philanthropy, a charitable foundation which (among many other things) tries to steer AI progress in a beneficial direction. They asked their resident expert Ajeya Cotra to prepare a report on the topic, and got "Forecasting Transformative AI With Biological Anchors" ("transformative AI" is AI that can do everything as well as humans).

The report is very complicated, and I explain it at greater length on my blog. The very short version: Suppose that in order to be as smart as humans, AI needs as much computing power as the human brain. In order to

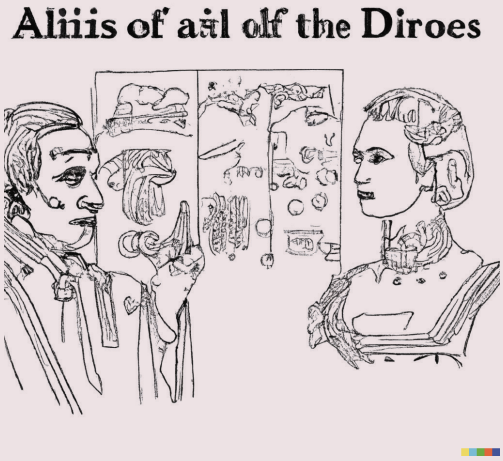
train an AI with as much computing power as the human brain, we would need a very, very powerful computer — one with much more computing power than the human brain. No existing computer or cluster of computers is anywhere near that powerful. To build a computer that powerful would take trillions of dollars — more than the entire U.S. GDP.

But every year, computers get better and cheaper, so the amount of money it takes to build the giant-AI-training computer goes down. And every year, the economy grows, and people become more interested in AI, so the amount of money people are willing to spend goes up. So at some point, the giant-AI-training computer will cost some amount that some group is willing to spend, they will build the giant-AI-training computer, it will train an AI with the same computing power as the human brain, and maybe that AI will be as smart as humans.

Is this the right way to think about AI? Don't we need to actually understand what we're doing in order to get human-level AI, not just build a really big computer? Didn't the Wright brothers have to grasp the basic principles of flight instead of just building something with the same wingspan as birds? Ajeya isn't unaware of these objections; the report addresses them at length and tries to argue why computing power will be the dominant consideration. I find her answers convincing. But also, if you're trying to do a deep specific model instead of making random guesses, these are the kind of assumptions you have to make.

Ajeya goes on to come up with best guesses for the free parameters in her model, including:

> How much computing power does the human brain have, anyway?
>
> Are artificial devices about as efficient as natural ones, or should we expect computers to take more/less computing power than brains to reach the same intelligence?



DALL·E 2023-05-24 16.52.48 - line drawing of AI predictions from the past

> It takes more computing power to train an AI than the AI itself uses, but how much more?
>
> How quickly are computers getting faster and cheaper? Will this continue into the future?
>
> How quickly is the economy growing? Will this continue into the future?
>
> How quickly are people becoming more interested in AI? Will this continue into the future?
>
> ... and finds that on average we get human-level AI in 2052.

Ajeya wrote her report in 2020, when the Metaculus questions for AI were reading late 2030s and early 2040s, and when Katja's experts were predicting the 2060s; all three forecasts were clustered together (and all much earlier than the popular mood, according to which it would never happen, or might take centuries).

In 2022, when Metaculus had updated to the late 2020s or early 2030s, and Katja's experts had updated to the 2050s (remember, all of these people are predicting slightly different

DALL·E 2023-05-24 16.56.39 - line drawing of researchers predicting the future of AI

questions), Ajeya posted "Two-Year Update on My Personal AI Timelines," saying that her own numbers had updated to a median of 2040. She gave four reasons, of which one and a half sort of boiled down to "seeing GPT be more impressive than expected," one was lowering her bar for transformative AI, and one and a half were fixing other parameters of her model (for example, she had originally overestimated the cost of compute in 2020).

It's good that she updates when she finds new information. Still, part of what I wanted from an explicit model was a way to not be pushed back and forth by the shifting tides of year-to-year news and "buzz." If there is a way to avoid that, we will not find it here.

The Conclusion

In some sense, since transformative AI has not been invented yet, we cannot grade forecasts about it.

But we can look at whether the same forecasters did a good job forecasting other AI advances, whether their forecasts are internally consistent, and how their forecasts have shifted over time. None of the three forecasting methods look great on these intermediate goals.

Katja's survey shifted its headline date very little over the course of its six-year existence. But it shows wild inconsistency among different framings of the same data, and gets its intermediate endpoints wrong — sometimes so wrong it fails to notice when things have already happened.

Metaculus's tournament shifted its headline date by 15 years over the three years it's been running, and its own commenters often seem confused about why the date is going up or down. Ajeya's model in some sense did the best, staying self-consistent and shifting its headline date by only 12 years. But this isn't really a meaningful victory; it's just a measure of how one forecaster voluntarily graded her own estimates.

In a situation like this, it's tempting to ask whether forecasting transformative AI gives us any signal at all. Could we profitably replace this whole 5,000-word article with the words WE DON'T KNOW written in really big letters?

I want to tentatively argue no, for three reasons.

First, in the past, these kinds of forecasts have provided more than zero information. Even on Katja's second survey, the one everyone failed at, there was a correlation of 0.1-0.2 — i.e., higher than zero — on which tasks the experts thought would be solved fastest, and which ones actually were. The Metaculus data show that its forecasts provide much more than literally zero information on binary questions.

Second, because as bad as these forecasts are, "better than literally zero information" is an easy bar to clear. Is it more likely that AI which can beat humans at everything will be invented 20 seconds from now, or 20 years from now? Most people would say 20 years from now; that is, in some sense, an "AI forecast." Is it more likely 20 years from now, or 20 millennia from now? Again, if you have an opinion on this question, you're making a forecast. Forecasts like the three in this article aren't good enough to get a year-by-year resolution. But they all seem to agree

that transformative AI is most likely in the period between about 10 and 40 years from now (except arguably the second framing of Katja's survey). And they all seem to agree that over the past three years, we've gotten new information that's made it look closer than it did before.

And third, because when people see a giant poster saying "WE DON'T KNOW," they use it as an excuse to cheat. They think things like, "We don't know that it's definitely soon, therefore it must be late," or, "We don't know that it's definitely late, therefore, it must be soon." Nobody says they're thinking this, but it seems like a hard failure mode for people to avoid.

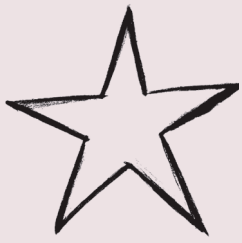
Forecasts — even forecasts that span decades and swing back and forth more often than we might like — at least get our heads out of the clouds and into the real world where we have to talk about specific date ranges.

I worry that, even *with* the forecasts, people will cheat. They'll use real but bidirectional uncertainty as an excuse to have uncertainty only in one direction. For example, they'll say, "These forecasts suggest a date 10-40 years from now, but the article said these forecasts weren't very good, and we all know that sometimes bad forecasters fall for hype about new technology, so we can conclude that it will be later than 10-40 years." Or they'll say, "These forecasts suggest a date 10-40 years from now, but the article said that these forecasts weren't very good, and we all know that sometimes bad forecasters have *status quo* bias and are totally blindsided by new things when they arrive, so we can conclude that it will be sooner than 10-40 years."

I'm against this because I constantly see both sides (sooner vs. later) assume the other has a bias and their own doesn't. But also because this is exactly the kind of information forecasters are trying to consider. I know some of the AI experts Katja surveyed, and they're people who think pretty hard about their biases and the biases of others, and try to account for these biases in their work. I know some of the forecasters on Metaculus, and ditto. Ajeya has talked at length about all the biases she is worried that she could have had and how she adjusted for them. When you throw out these (admittedly bad) forecasts based on your view that they're "too aggressive" or "too conservative," you're replacing hundreds of smart people's guesses about what errors might be involved in each direction with your spur-of-the-moment guess.

So I claim that our canonical best guess, based on current forecasting methods, is that we will develop "transformative AI" able to do anything humans can do sometime between 10 and 40 years from now. These forecasts aren't very good, but unless you have more expertise than the experts, are more super than the superforecasters, or have a more detailed model than the modelers, your attempt to invent a different number on the spot to compensate for their supposed biases will be even worse.

We should, as a civilization, operate under the assumption that transformative AI will arrive 10-40 years from now, with a wide range for error in either direction.



52

How Long Until Armageddon?

Michael Gordin

Scientists, generals, and politicians all failed to accurately predict when the Soviets would get the bomb. Could they have done any better?

Seventy-eight years ago, give or take a few months, American policymakers, military officers, and scientists were obsessed with a single technological prediction: How long until another country develops an atomic weapon? A few years earlier, a different technological prediction centered around the fission of heavy nuclei — Is it possible to build such a bomb? — but that question was incontrovertibly answered in the affirmative in August of 1945, when the U.S. Army Air Forces dropped a uranium bomb on Hiroshima and a plutonium bomb on Nagasaki. Now is a good time to revisit that historical moment for what it can illuminate about the challenges of technological forecasting. During the brief period of the American atomic monopoly — which existed from the Trinity Test on July 16, 1945, to the Soviet Union's demonstration of their own weaponization of nuclear fission on August 29, 1949 — the question of predicting how soon other countries might breach the atom was a crucial driver of policy choices for Americans across government, the military, and science. Tremendous resources were invested in providing reasonable predictions of this question, and yet the eventual Soviet detonation of what the Americans called Joe-1 came as a surprise to everyone — even those who had basically gotten the question right.

Although historians are not trained for the business of tomorrow, the debates surrounding the estimates of Soviet proliferation have the potential to illuminate the serious business of technological forecasting in our present moment, whether it be of carbon capture, fusion energy, or artificial general intelligence. It's not a question of it being nice to know these things — we simply cannot make policy choices in the present without at least an implicit guess about these future developments. In this, anyone with concerns for the coming century is in the position of those American policymakers in the shadow of the first mushroom clouds.

The Manhattan Project that produced the first nuclear bombs was a joint international effort of the United States, United Kingdom, and Canada, populated by scores of émigré specialists from across Europe, but by the time nuclear fission was thrust into the Pacific War these weapons were controlled from Washington, D.C. American war planners and pundits began to project recent history onto the future, imagining American cities smoldering under mushroom clouds and asking: "How much time do we have?" Pressing policy choices about diplomacy, military deployments, demobilization, and more were contingent on the duration of

the American atomic monopoly. Given the economic and geopolitical realities of the postwar moment, the only state that could conceivably marshal the motivation and the resources for such an effort was the Soviet Union. That is where the forecasters turned their attention.

How many years after 1945 — when Joseph Stalin, as well as everyone else, could not deny that nuclear weapons were a real threat — would it take for the Soviet Union to get the bomb? Framed this way, it was a much easier question than forecasting the advent of AGI: The Americans were concerned with only one adversary and the development of a single technology *which they already knew all the details of*. American predictions of Soviet proliferation offer a highly constrained case study for those who would undertake technological forecasting today.

It is not an encouraging precedent. In the aftermath of the Japanese surrender on August 15, 1945, predictions about a Soviet bomb percolated through both the popular press and the classified channels of the American elite. Estimates ranged from two years to 20 years to never. It came as a shock to basically everyone when on September 23, 1949 — less than a month after the event — U.S. president Harry S. Truman announced: "We have evidence that within recent weeks an atomic explosion occurred in the U.S.S.R."[1] Truman had been given a plethora of predictions, and just about all of them had been wrong. Could they have done any better?

***

Any prediction of Soviet proliferation was a judgment about two things: How hard is it to make an atomic bomb? And how capable were the Soviets? One found broad disagreement on both points even among the well-informed, who based their judgments on the information they had access to, and how they defined what was essential to "making an atomic bomb."

Consider the case of General Leslie Groves, the head of the wartime Manhattan Project, and the only individual with access to information about all operational, scientific, engineering, and logistical aspects of proliferation. In the span of three months of 1945, he offered three different ranges, and hedged each one. Addressing the War Department in September, Groves opted for a 10-year estimate: "With regard to Russia, he estimated it would take her three years to develop the scientific knowledge (assuming efficient administration and access to German scientists) and five years by major effort to solve the industrial problems, or seven to ten years under a program of normal peacetime emphasis."[2] The following month, as an expert witness before the House of Representatives, he "believe[d] that for another country to do this work, if it had the power of the greatest countries left in the world, but had no particular ideas, that it would take them from 5 to 20 years, and the difference in time would depend entirely on how 'all-out' they made their efforts and how much they threw security to

Opposite: Mushroom cloud from the detonation of the Joe-2 (RDS-2) Soviet nuclear bomb on 24 September 1951 at the Semipalatinsk Test Site in what is now Kazakhstan. This test had a yield of 38 kilotons of TNT. A total of 456 Soviet nuclear tests were conducted at the Semipalatinsk site between 1949 and 1989.

1. Harry S. Truman, public statement of September 23, 1949, PSF, Subject File, 1940–1953, Box 59, Folder: "September, 1949." The Soviets were likewise astonished that the Americans found out and made the announcement, but that's a different story. Full details on the debates about prediction, the process of detection, and the aftermath, can be found in Michael D. Gordin, *Red Cloud at Dawn: Truman, Stalin, and the End of the Atomic Monopoly* (New York: Farrar, Straus & Giroux, 2009).

2. "Notes of a Meeting in the Office of Secretary of War Concerning Atomic Energy Legislation, 9:30 A.M.–11:00A.M., 28 September 1945," Harrison-Bundy Files Relating to the Development of the Atomic Bomb, Records of the Office of the Chief of Engineers (RG 77), National Archives and Record Administration (NARA), College Park, MD, Microfilm Publication M1109, Roll 5, Target 4, File 68: "Interim Committee — Legislation," 5.

the winds."[3] Before the Senate the month after that, he tilted once more to the longer end, stating that if the Soviets "did it in complete secrecy, probably within 15 to 20 years — more likely the latter."[4] Groves justified these assessments, at least publicly, by citing the backward technical capacities of the war-ravaged Soviet economy and his poor opinion of Soviet engineers and scientists.

Many of the scientists who worked under Groves during the war, however, had encountered quite a few of the leading Soviet nuclear physicists during the cosmopolitan interwar quest for the secrets of quantum physics, and deemed them able to solve the riddle of fission as competently as the scientists at Los Alamos had. In October 1945, atomic gadfly Leo Szilard, Groves's bête noire, testified to Congress that it might be six years.[5] His colleagues Hans Bethe and Frederick Seitz published an article in 1946 entitled "How Close Is the Danger?," outlining in clear prose what it would take for Russia, France, China, Argentina (or a South American coalition), Sweden, or Switzerland to make atomic weapons. One had to assess the nations' incentives, scientific talent, technological ability, starting point, and (of course) access to uranium. Their conclusion: five years. That is, five years from 1946, which meant around 1951.[6]

Not all physicists drifted to so "low" an estimate — which, it bears repeating, still significantly overestimated the time to Soviet proliferation. Arthur Holly Compton, a 1927 physics Nobel co-recipient who had directed the Metallurgical Laboratory at the University of Chicago[7] and was as well informed as anyone, told a reporter in 1948 that he felt it would be at least four years — that is, 1952 — before the Soviets obtained a bomb, but that "I won't be surprised if they don't get it before 1970."[8] Less cocky but no less dismissive, J. Robert Oppenheimer, the lord of Los Alamos, asserted in a letter that same year: "With all recognition of the need for caution in such predictions, I tend to believe that for a long time to come the Soviet Union will not have achieved this objective, nor even the more minor, but also dangerous possibility of conducting radiological warfare." Confronted with this assessment at his security-clearance hearings in 1954, five years after the first Soviet nuclear test, Oppenheimer hung his head: "This was a bad guess."[9]

Compton and Oppenheimer had access to different parts of the secret information available to Groves, and this might explain their tendency toward higher estimates. Those with less access to classified information tended to have similar low estimates, or sometimes even lower. William Leonard Laurence, the journalist who accompanied the Nagasaki mission and won a Pulitzer Prize for his inside reporting on the Manhattan Project, stated in 1948, "We still have about four years, as of today," in

3. Leslie Groves testimony, October 9, 1945, United States House of Representatives, Committee on Military Affairs, *Atomic Energy: Hearings on H.R. 4280*, 79th Congress, 1st session, 1945 (Washington, D.C.: Government Printing Office, 1945), 18.

4. Leslie Groves testimony, November 29, 1945, United States Senate, Special Committee on *Atomic Energy, Atomic Energy: Hearings on S. Res. 179*, 79th Congress, 1st session, 1945–1946 (Washington, D.C.: Government Printing Office, 1945–1946), 62.

5. Leo Szilard testimony to House Committee on Military Affairs, October 18, 1945, *Atomic Energy, H.R. 4280*, 88.

6. Frederick Seitz and Hans Bethe, "How Close Is the Danger?," in Dexter Masters and Katharine Way, eds., *One World or None* (New York: Whittlesey House, 1946): 47.

7. The "Met Lab" was where the world's first nuclear reactor went critical on December 2, 1942, proving that a fission chain reaction in uranium was possible.

8. "Compton Sure Russia Doesn't Have A-Bomb," *Los Angeles Times*, May 1, 1948, 14.t 9. Letter, April 14, 1948, read during Oppenheimer testimony of April 12, 1954, in U.S. Atomic Energy Commission, *In the Matter of J. Robert Oppenheimer: Transcript of Hearing before Personnel Security Board and Texts of Principal Documents and Letters* (Cambridge, MA: MIT Press, 1971), 47.

9. Letter, April 14, 1948, read during Oppenheimer testimony of April 12, 1954, in U.S. Atomic Energy Commission, *In the Matter of J. Robert Oppenheimer: Transcript of Hearing before Personnel Security Board and Texts of Principal Documents and Letters* (Cambridge, MA: MIT Press, 1971), 47.

line with Compton's guess of 1952.[10] On the other hand, senior Republican senator Arthur Vandenberg — the backbone of Truman's bipartisan foreign policy — on December 10, 1945, at first wrote: "We agree that Russia can work out this atom science in perhaps two years."[11] Less than six months later he had revised his view more in line with that taken by conservative journalists: "Our 'secret' in respect to atomic bombs probably will not be a 'secret' for more than five years."[12]

A year before the Soviets actually proliferated, in September 1948, the CIA converged, like almost everyone not in the inner circles of the Manhattan Project, to a five-year estimate: "The earliest date by which the Russians may have exploded their first bomb is mid-1950; the probable date by which they will have exploded their first test bomb is mid-1953."[13] Notice how this splits the difference of "five years": 1950 is five years from the first nuclear explosions of 1945, and 1953 is five years from the present — either way you sliced it, the relevant number was five.

***

Two general points emerge from the above, each of which requires explanation: First, individuals with greater access to secret information tended to have higher estimates than those with less; and second, the lower, less-informed guesses ended up being more accurate (though still overestimates). The explanation for the first point has to do with secrecy and is multidimensional; the explanation for the second has to do with assumptions and is more straightforward.

To begin with secrecy: Nobody really knew how hard it was to develop a nuclear weapon, even (or especially) those who had just done so. There was not just one thing that was "to build a bomb." There were, in 1945, exactly two historical data points: the Americans, who had done it in 3.5 years; and the Germans, who had failed. Where on that continuum would the Soviets fall? (As it happens, almost 80 years later, no proliferating nation has ever managed to build a bomb in less time than the Manhattan Project.) Making a bomb required, at minimum, prospecting uranium, building an industry to isolate the fissionable isotope that constituted less than 1% of the natural ore (or synthesizing plutonium from the remainder), and designing a deliverable bomb. The physicists thought the secret to the bomb was the physics itself, which honestly wasn't too hard. Engineers like Groves believed the industrial infrastructure and complex logistics — their specialty — was the real "secret" of the bomb.

Groves, however, also knew something else, one of the most closely guarded secrets of the era: He had all the world's known uranium. Before World War II, this heavy metal had only very limited industrial uses — principally to form a bright yellow pigment — and so was poorly prospected. Most people believed it was rare; it was only in the rush to control this resource after the war that revealed that it is in fact reasonably plentiful. Groves's Combined Development Trust monopolized 97% of the world's known uranium, mostly from the Belgian Congo. His success was highly classified, so most pundits were unaware of the stranglehold the Americans held on resources, and which we now know significantly hampered the Soviet project.

Secrecy was a problem for forecasting in general. Different people knew separate things about the Manhattan Project, about the Soviet Union, about the laws of nature, about the U.S. government. American intelligence was (and remains) partitioned across multiple agencies, forbidden from compiling what they knew by

10. William L. Laurence, "How Soon Will Russia Have the A-Bomb?," *Saturday Evening Post,* November 6, 1948, 181.

11. Reproduced in Arthur H. Vandenberg Jr. with Joe Alex Morris, eds., *The Private Papers of Senator Vandenberg* (Boston: Houghton Mifflin Company, 1952), 228.

12. Arthur H. Vandenberg to L.G. Carnick, April 18, 1946, in Vandenberg with Morris, *The Private Papers of Senator Vandenberg*, 252–253.

13. Intelligence Memorandum No. 59 for Secretary of Defense, September 20, 1948, CIA Records (RG 263), NARA, Box 110, Folder: "26519," 9.

the arcana of classification and the instincts of mistrust. Biases became ingrained. The newly created Air Force opted for a short time frame since they were the premier military branch for nuclear delivery; the Navy, perhaps fearing obsolescence in this brave new world, opted for a longer timeline so their relevance would be assured for many years to come. Both could marshal data that confirmed their priors.

Knowledge about what was going on in the Soviet Union was harder for Americans to obtain than details about their own weapons. Decades of spy movies have conditioned us to expect that this problem could be ameliorated with human intelligence in the form of American agents on the ground. But the workers' paradise was "denied territory" in intelligence parlance: There were zero ground agents in the Soviet Union. In 1949 the CIA began a five-year program to recruit and train former Soviet citizens who would be air-dropped back on Soviet territory. Almost all of them were arrested at once and shot. Those few who produced information were not trusted by the Americans, who feared that the informants had been "turned" to double agents. The inverse was not true, of course: The Soviets had several highly placed agents within the Manhattan Project and obtained significant information about the details of the American program. This penetration, too, was secret, and therefore was not explicitly factored into any of the estimates above. (Groves, who knew the most about it, seems not to have considered it at all.)

You certainly couldn't trust what the Soviets said about the "atomic secret." On November 6, 1947 — long before any of the credible American estimates had matured — Soviet foreign minister Vyacheslav Molotov hinted at Soviet success. During a rousing speech in honor of the 30th anniversary of the Bolshevik Revolution, he cryptically added: "As we know, a sort of new religion has become widespread among expansionist circles in the U.S.A.: having no faith in their own internal forces, they put their faith in the secret of the atomic bomb, although this secret has long ceased to be a secret."[14] Two weeks earlier, Andrey Zhdanov, Stalin's second-in-command, had announced in Warsaw that while the Americans had a monopoly on nuclear weapons, that monopoly was "temporary." Andrey Vyshinsky, who would replace Molotov as foreign minister in 1949, later declared that the monopoly was an "illusion."[15] It suited the Soviets both to claim they had already proliferated in order to deter a preemptive American strike, and also to keep mum about it once they had tested their first device in August 1949, lest this trigger the same aggressive response before they had time to amass a stockpile.

***

What about the puzzling fact that the working physicists in general seemed to have lower estimates than Groves and others with access to the largest amount of classified information — and that those lower estimates were correct? It was surely the case that Groves overestimated the time to Soviet proliferation because he thought that the devastated Soviet economy could not possibly handle the tremendous technical and logistical effort involved in building an entire nuclear industry. Such large-scale construction projects were his specialty — he built the Pentagon, still the world's largest office building, ahead of schedule and under budget in the wake of the Depression — so he likely overweighted this factor. Groves knew that the physics involved in building a nuclear weapon was not

14. Vyacheslav Molotov, "Thirtieth Anniversary of the Great October Socialist Revolution," speech at celebration meeting of the Moscow Soviet, November 6, 1947, in V.M. Molotov, *Problems of Foreign Policy: Speeches and Statements, April 1945–November 1948* (Moscow: Foreign Languages Publishing House, 1949), 488.

15. "Reds 'Possibly' Have A-Bomb — Vishinsky," *Los Angeles Times*, November 7, 1947, 1; "U.S. Monopoly on Atom Bomb 'Illusion,' Vishinsky Tells UN," *The Christian Science Monitor,* October 1, 1948, 6; "Reds' A-Bomb Hint Called 'Dishonest,'" The Washington Post, October 3, 1948, M3; Andrey Zhdanov quoted in Sydney Gruson, "U.S. Monopoly of Bomb Cited," *The New York Times*, October 23, 1947, 3.

especially difficult, and that Soviet scientists could figure it out within a few years, especially if they got their hands on German specialists (which they did). He just did not think that was sufficient to build a bomb. Why were the lower-estimating physicists closer to being right?

They really weren't, in the sense that all the logistical hang-ups predicted by Groves — the uranium shortage, the dilapidated heavy industry, the competing demands of the civilian economy — were indeed the factors that most slowed down the Soviet bomb builders, who had in fact started working out the physics in the middle of the war. What Groves and everyone else hadn't counted on was Joseph Stalin. It was extremely hard to force-march your entire economy toward the building of an atomic bomb while sacrificing the needs of an impoverished civilian population while at the same time prospecting across the vast reaches of Central Asia for uranium, running an international intelligence operation, and keeping your entire army mobilized. Stalin could do that without civilian unrest, and he could also steal heavy industry from Central and Eastern Europe for the chemical industry required for the project. Groves looked at the raw data of the Soviet economy and figured it could never, under normal circumstances, compete. Stalin meant the circumstances were not normal. In the end, the "physicists' estimate" ended up being more right because that was the only one of the many factors that could not be accelerated by state terror.

This speaks to a lesson that one can draw from the failed American predictions of Soviet proliferation: what one could call the problem of *specification*. In almost all of the predictions above, what it means to "have a bomb" is different, much as it is today in forecasting Iranian proliferation. When you evaluate an estimate, it is essential to know what it is an estimate of. Did one mean five years until the Soviets uncovered the basic information of how to make an atomic bomb? Five years until a working reactor? Until the establishment of the production process for nuclear fuel? Until the first atomic test? Until a sizable stockpile? Until the assembly of a delivery system capable of striking the American heartland? Most estimators were vague about what they meant, and so it was possible, in good conscience, to keep reiterating "five years," because the five years may have referred to something different each time.

Because of the conditions of secrecy, in this case non-specification drove the convergence of estimates. Everyone predicting "five to 10 years" gave a comforting feeling of certainty to both the forecasters and their audience. If a good many smart people offer the same conclusion, they must be on to something, right? Unfortunately, the uniformity masks an underlying heterogeneity of reasoning. By focusing on the number, one misses that the justifications for the predictions are quite different, and at times contradictory. The number easily becomes unmoored from reality: People who said "five years" in 1945 would say the same in 1947, and in 1949. "Five years" slowly came to mean "five years from now," not "five years from when they started" or "five years from my first prediction."

The Americans were not always so hapless at nuclear prediction, famously announcing the Chinese test at Lop Nur in 1964 a few weeks early; nonetheless, the debates of the late 1940s do give one pause when confronting the present. Even though today there are many well-informed, thoughtful people making predictions, they face similar challenges to those facing postwar Americans: many competing entities with private information — this time the secrecy is corporate instead of military — alongside a lack of specificity (in the case of AGI, unspecified in large part because the characteristics aren't known) about the precise nature of the prediction. The challenges of one single technological forecast about a known entity, the object of focused attention for years by the most well-informed people around, illustrates some of the difficulties of today's necessary predictions.

60
Are We Smart Enough to Know How Smart AIs Are?

Robert Long

Scientists have repeatedly failed to recognize the complexity of animal cognition. Will we make the same mistakes with AI?

Animals think about a lot more than we once gave them credit for. It's now commonplace to read about chimpanzees that play elaborate games, scrub jays that hide — even camouflage — food from rivals, or bees that can master abstract concepts. But as recently as the middle of the 20th century, attributing mental states to animals was taboo in science. Behaviorists studied simple and controlled behaviors — press a lever, receive a food pellet — while the naturalists who did observe animal behaviors in the wild tended to describe them in terms of innate instincts or adaptations to ecological niches. Neither group sought to explain animal behavior in terms of mental states like beliefs, theories, or intentions.

In the decades since, we have been surprised to uncover complex cognition across the animal kingdom: first in our closest primate relatives, then in more distant creatures like crows and parrots, and most recently in invertebrates like the octopus and the honey bee. The progression from an overly cautious denial of complex mentality — driven by a desire for rigor and a fear of anthropomorphism — to a more sophisticated understanding of animal minds is one of the great stories of 20th century science. And it holds lessons for how humanity can approach the most critical intelligence explosion since the Paleolithic — that of artificial intelligence.

***

In the late 19th century, psychology relied excessively on the introspective theorizing of scientists. Lacking empirical rigor, the field came close to stagnating in a morass of ill-defined and irresolvable disputes. As Darwin's theory of evolution gained acceptance, scientists became more interested in studying the continuities between human and animal minds, but this interest led to a methodology characterized by unchecked anthropomorphism.

In his 1882 book *Animal Intelligence*, George Romanes, an academic friend of Darwin's, described scorpions that attempted suicide and foxes that sought revenge after failed hunting expeditions.[1] One of the most famous examples of runaway anthropomorphism was a stomping horse. Clever Hans, an Orlov trotter, wowed adoring crowds with his ability to add, subtract, and even tell time, indicating his answers by tapping his hoof. But an investigation showed that, unbeknownst to his owner, who by all accounts believed in the horse's abilities, Hans only arrived at a correct answer by reading the facial expressions and body language of whoever asked the question.

Something stringent was needed to reign in such credulity. Behaviorism, which held that both human and animal behavior could and should be explained without reference to thoughts or feelings, offered a solution. John Watson's 1913 article

1. G. J. Romanes, *Animal Intelligence* (London: Kegan Paul, Trench & Co., 1882).

"Psychology as the Behaviorist Views It," called on scientists to stop studying any behavior that could not be outwardly observed and measured — including the mind: "The time seems to have come when psychology must discard all reference to consciousness; when it need no longer delude itself into thinking that it is making mental states the object of observation."[2]

The work of psychologist B.F. Skinner, considered "the father of behaviorism," is emblematic of how empirical rigor went hand in hand with distorted thinking about animal cognition. Skinner's 1938 book, *The Behavior of Organisms*, describes dozens of well-controlled experiments on rats, conducted in precisely constructed operant conditioning chambers called Skinner boxes. That pressing levers for food was the main behavior tested, and rats the primary animal tested, was not, for Skinner, a limitation."The only differences I expect to see revealed between the behavior of rat and man (aside from enormous differences of complexity) lie in the field of verbal behavior," he wrote.

At least in the West, even scientists who studied animal behavior in the wild shied away from attributing too much mental sophistication to their subjects. They were wary of getting 'too close' to animals — Western naturalists even considered it bad practice to give names to primates being studied.[3] It was against this background, that Jane Goodall arrived in a forest in Gombe, Tanzania and helped launch the first significant re-expansion of animal cognition.

Goodall first observed chimpanzees using sticks to extract termites from their mounds in 1960. Although Darwin and his contemporaries had readily accepted tool use among apes, the idea had fallen out of favor. So central was the belief that tool use was exclusive to humans that Goodall's mentor, the Kenyan-British anthropologist Louis Leakey, told Goodall that if her finding held, "we should by definition have to accept the chimpanzee as Man."[4] Goodall was initially met with skepticism. Many criticized what they described as her sentimentality. Today, not only is tool use among chimps widely accepted, but has been described in many other hominids, as well as in elephants, dolphins, crabs, and birds.

As primatologist Frans de Waal recounts in his history of animal cognition, *Are We Smart Enough to Know How Smart Animals Are?*, the 1970s and 1980s saw a boom in both the observation of wild behaviors and the development of more sophisticated laboratory techniques to investigate them. As a result, scientists have identified sophisticated behaviors that the behaviorists would not have predicted — or even been able to observe. More importantly, they have been able to pose and test cognition-based theories to explain those behaviors in terms of what animals think and feel.

The behaviorist prohibition on discussing mental states is now regarded as overly restrictive, if not wrong altogether. De Waal argues, for example, that the complicated "political" jockeying among apes is best explained by their possessing a "theory of mind," or the ability to model the beliefs and intentions of other agents. Even though their theory of mind may be different from and more "limited" than

2. J. B. Watson, "Psychology as the behaviorist views it," Psychological Review 20 (1913): 158-177.

3. Starting in the late 1940s, Japanese scientists were pioneering methods that are now standard practice: habituating wild monkeys to the presence of humans, identifying individuals, and observing them throughout their lives. But at the time, their work was overlooked or dismissed by their Western counterparts

4. Jane Goodall, In the Shadow of Man (1971, repr. New York: Houghton Mifflin, 2000): 37.

Nerve cells in a dog's olfactory bulb from Camillo Golgi's *Sulla fina anatomia degli organi centrali del sistema nervoso* (1885)

that of humans, it is now consensus that primates do share this basic cognitive capacity, and many others, with humans.[5] Anthropomorphism is not always an error, especially with creatures that are in fact very related to humans.

While Goodall was studying chimpanzees in Tanzania, other scientists were discovering unexpected cognitive capabilities in birds. A report from 1960 documents just one species capable of tool use — the woodpecker finch of the Galapagos Islands. Research in the 1970s and 80s added more species to the list — mostly in the corvid family, a clever group that includes ravens, jackdaws, and crows. Crows were once dubbed "feathered apes" after they were found to use sticks as tools and to engage in sophisticated problem-solving. Most famously, Irene Pepperberg's thirty-year experiment with an African gray parrot named Alex uncovered unimagined cognitive abilities. Not only was Alex capable of identifying colors, shapes, and quantities, but he also demonstrated an understanding of more abstract concepts such as same/different and bigger/smaller.

In recent decades, the circle of cognition has expanded to creatures even more distantly removed from humans. Cephalopods, the animal class that includes octopuses and squid, may have the closest thing to alien minds on the planet (so far): our last common ancestor with them was a worm-like creature which lived in the deep ocean about 400 million years ago. Octopuses were perhaps the animal cognition celebrities of the 2010s, with their sophisticated distributed nervous systems, behaviors suggestive of play, problem-solving abilities, idiosyncratic "personalities," and their awareness of other agents.

Equally as startling is the sophistication of the honeybee. Insects were long thought to be "robotic," driven purely by instinct. Jean-Henri Fabre, who studied wasps, bees, and many other insects from the 1860s until his death in 1915, commented on their "machine like obstinancy." In the mid-1940s, entomologist and eminent ethologist Karl von Frisch discovered that bees communicate through a "waggle dance," an elaborate choreography capable of describing the direction and distance to flowers, water sources, or new nest sites. In this century, bees have displayed the ability to learn rules that involve abstract, multimodal representations of sameness and difference.

***

With more research, scientists have successfully found more complex cognition than expected in animals further and further from humans. Why does the circle keep widening?

As a reaction to the field's early excesses and credulity, behaviorism demanded strictly controlled experiments, limited to single behaviors like lever-pressing and simple stimuli such as flashing lights. The behaviorist's error was to think that these artificially simple cases could be extended to explain *all* behaviors in *all* organisms. Their tools made it difficult to notice more complicated behaviors, and even more difficult to explain them once discovered.

One of the most forceful arguments against the behaviorists came in a review of Skinner's book *Verbal Behavior*, which sought to explain language as a behavioral phenomenon like any other — a promise

5. To be sure, at times primatologists overplayed their hand in hypothesizing far more continuity between apes and humans — audacious attempts to raise chimpanzees and teach them full-fledged language failed ("Nim Chimpsky" being the most notorious), though it is generally accepted that apes can learn rudimentary sign language.

Skinner had made in *The Behavior of Organisms*. The review, which appeared in 1959 in the journal *Language*, argued scathingly that Skinner underestimated the depth of human language, which could not be explained simply by extending the methods of stimulus, response, and reward he had used to study rats.

It is now seen as a turning point, a milestone in the "cognitive revolution" in which the sciences of the mind turned away from behaviorism and looked instead to mental representations and operations. It also greatly raised the profile of the young linguist who had written it, Noam Chomsky. Chomsky understood that to accurately understand human and animal behavior, science needed methods that could accommodate behavioral complexity. "It is clear," he wrote, "that what is necessary in such a case is research, not dogmatic and perfectly arbitrary claims, based on analogies to that small part of the experimental literature in which one happens to be interested." And once complex cognitive abilities could be admitted as a hypothesis, methods could be developed to study them.

As researchers learned to treat animals with empathy and imagination, they discovered more and more capabilities. Breakthroughs emerged when scientists were able to imagine the world as experienced by each particular animal. Tool use was once thought to be conspicuously lacking among gibbons, small apes native to Southeast Asia. When tools that could be used to get food were placed in front of them on the ground, the gibbons did not grab them. The problem was not with gibbon intelligence, but human imagination. Gibbons live in trees. Their hands are well suited to swinging, but poorly adapted for picking things off the ground. When the tools were instead dangled from a branch, the gibbons had no problems and readily used them. Elephants initially failed the mirror test, a common method for determining self-recognition, because the mirrors used were too small. And in a true lack of empathy, many behaviorists assumed that to motivate their test subjects they had to keep them half starved. It's now clear that animals that are treated well and feel cared for will, as with humans, be far more likely to act in interesting ways.

The behaviorist's error was to think that these artificially simple cases could be extended to explain *all* behaviors in *all* organisms. Their tools made it difficult to notice more complicated behaviors, and even more difficult to explain them once discovered.

Wild observations are also a way of meeting animals where they are (literally) to see what they are capable of. Scientists now spend hundreds of hours in the field simply observing (grad students spend even more). Animals will often behave very differently among their own kind and in their natural habitat than they will in a sterile lab surrounded by lab-coated hairless primates. More wild observation has uncovered more sophisticated behaviors than lab scientists had imagined animals capable of.

And we have also learned that brains can operate in ways very different from our own. Bird intelligence was surprising to ornithologists because birds have no neocortex. Bee intelligence was surprising because they have very little brain at all. (Despite being the first to decode the waggle-dance, Karl von Frisch once said, "The brain of a bee is the size of a grass seed and is not made for thinking.") In each case, nature has more ways of implementing cognition than we had thought to look for. Birds have alternative brain regions that perform the same function as the cortex. Bees have very densely packed neurons that fit quite a lot of cognition into something the size of a grass seed. Most strangely of all, the octopus has a cluster of neurons in each of its tentacles, resulting in a kind of thinking that is so distributed that it is hard for us to imagine.

The wariness of getting "too close" to animals and of overestimating their cognitive abilities still exists — and for good reason. Selection effects, where researchers are more likely to work with an animal if they antecedently believe that the animal can do interesting things, remain at work. And publishing incentives reward impressive and surprising skills. There's no market for a glowing profile of the scientist that finds a deflationary explanation for an animal behavior. Few people are going to tweet a video of a salmon failing the mirror test.

So a common dichotomy pits animal enthusiasts who over-attribute mentality to animals against stern, hard-nosed buzzkills who maintain their distance and thus their methodological rigor. But doing hard-nosed and rigorous work requires something different — something akin to love: a holistic understanding of the animal, born from long periods of sustained attention. For this sort of work, the best motivator is affection.

So a common dichotomy pits animal enthusiasts who over-attribute mentality to animals against stern, hard-nosed buzzkills who maintain their distance and thus their methodological rigor. But doing hard-nosed and rigorous work requires something different — something akin to love.

Indeed, one of De Waal's lessons is that one cannot study animal intelligence "without an intuitive understanding grounded in love and respect."

And now, an entirely different form of intelligence has arrived. The study of AI lacks coherent methods. AI capabilities are superhuman in some ways and dangerously limited in others. And no one is yet sure what to make of something so human but alien at the same time. What lessons does the past century of research in animal cognition hold for how to think about today's AI systems?

***

In many ways, we are in our understanding of large language models where the study of animals was in the middle of the 20th century. Like animal cognition, the field of AI is overshadowed by founding traumas — cases

in which credulity and anthropomorphism have led researchers to exaggerate and misconstrue the capabilities of AI systems. Researchers are well aware of the ELIZA effect, our tendency to readily read human-like intentionality into even very simple AI systems — so named for an early chatbot built in 1964 that used simple heuristics to imitate a psychoanalyst. They remember past AI winters, when AI progress had been overpromised and underdelivered and disappointed funders cut jobs. Researchers are understandably wary of credulity and hype.

And few topics are more hype-prone right now than language models. One way to impose rigor and combat our natural tendency to anthropomorphize is to forbid using psychological language to describe AI systems. As Shevlin and Halina argue in Nature Machine Intelligence, using certain psychological terms like "theory of mind," "motivation," and "understanding" can be misleading if they encourage people to make inferences which might hold for human minds, but not for AI systems.[6] If GPT-4 can be said to have beliefs, its beliefs must be in some sense very different from human beliefs. If GPT-4 can be said to have a theory of mind, its theory of mind must have developed in a very different way than ours did. (More speculatively: if GPT-6 will be conscious, it will have experiences which are quite strange and hard for us to imagine.)

Another way to combat confusion is to emphasize what the models are trained to do and how different that is from humans: large language models have learned to produce text in a very different way than we have. But as with behaviorism, these understandable prohibitions risk leading us to retreat to a narrow explanation of AI behavior that underestimates what models can actually do. Describing language models as "just" predicting the next token doesn't do justice to the surprising ways they operate.

For example, it's now clear that language models don't just model shallow statistical text patterns — they model aspects of the world behind the text. Indeed, it's possible to identify "facts" that a large language model takes to be true. Researchers found that they could selectively edit a language model to make it "believe" that the Eiffel Tower is located in the city of Rome.[7] The models outputs reflect this new "belief" in a way that is both precise (its outputs don't simply move all of Paris to Rome, only the Eiffel Tower) and also generalized (in a wide range of differently-worded questions about Rome or the Eiffel tower, it will produce outputs consistent with the Eiffel Tower being in Rome, such as recommending it as a tourist destination for visitors to Italy). More recently, another group trained a language model on transcripts of a simple board game, and then probed its activations to find it had learned to represent different states of the board.[8] In other words, the model wasn't just combing its data to identify the next move. It had developed an internal picture of the game board and intuited its rules.

Just as Skinner thought that the differences between rats, apes, and humans were in some sense superficial, regarding all LLMs as just next-token predictors can blind one to the important differences between them. If we say that both GPT-2 and GPT-4 are "stochastic parrots," then what explains the fact that GPT-4 can write

6. Henry Shevlin and Marta Halina, "Apply rich psychological terms in AI with care," *Nature Machine Intelligence* 1 (2019): 165-167

7. Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov, "Locating and editing factual knowledge in gpt," *arXiv preprint arXiv*:2202.05262 (2022).

8. Kenneth Li, "Do Large Language Models learn world models or just surface statistics?", The Gradient, 2023.

a Shakepsearean sonnet about how to use a Python package, pass the bar, or solve difficult logic puzzles — skills far outside of GPT-2's capabilities? We need to investigate the output of each model and explain why they are different.

As with animal cognition, a desire to impose rigor can limit one's ability to see how interesting the behavior to be explained is. Some are so dismissive of LLMs that they have a blanket policy of refusing to look at any outputs from large language models. This has the effect of making it impossible to have one's mind changed about what the models are able to do. If one has decided in advance that an AI system is not that interesting, then one is less likely to look hard for interesting behaviors. Chomsky recently described ChatGPT as "a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question."

As evidence for this claim, he declared in his May op-ed in the *New York Times* that because "these programs cannot explain the rules of English syntax, for example, they may well predict, incorrectly, that 'John is too stubborn to talk to' means that John is so stubborn that he will not talk to someone or other (rather than that he is too stubborn to be reasoned with)." Readers immediately noticed that ChatGPT can, in fact, infer the correct interpretation. The study of language models is still developing. We know so little about how [LLMs] work that we would be wise to remember Chomsky's admonition to Skinner: what is needed is research, not claims based on analogies to that small part of the literature in which one happens to be interested.

Fortunately, large language models have their equivalents of naturalists — enthusiasts, including academics and industry researchers as well as non-professionals, who spend many hours engaging with the models. People like them have often been at the bleeding edge of discovering what large language models are capable of, their failure modes and their idiosyncrasies. What LLM enthusiasts have brought to our understanding of AI are a plethora of interesting capacities unlocked by doing what they love — messing around with LLMs for hours.

We know so little about how [LLMs] work that we would be wise to remember Chomsky's admonition to Skinner: what is needed is research, not claims based on analogies to that small part of the literature in which one happens to be interested.

These investigations revealed one way LLMs are like animals: if you reshape tasks in order to better match the subject's natural limitations and abilities, you can elicit better performance. One obvious limitation of LLMs is that, while they are experts at continuing text, they don't have any space to think in while answering a question. Simply adding "Let's think step by step" to a prompt after you ask them a question can be thought of as giving the LLMs a place to think — their own//

outputs — and encouraging them to use it. For example, GPT-3 often initially fails at mathematical word problems. However, if asked the *same* question but with "Let's think step by step," the model will then respond with the steps of reasoning that are necessary for the right answer. Versions of this technique, called "Chain of Thought" prompting, have been discovered by ML academics as well as amateurs playing with early versions of GPT.

Chain of Thought has a natural gloss as enabling models to complete a task in a way that is suited to their capabilities, like a gibbon grabbing a dangling tool. Prompting models to explain their reasoning, letting them choose between outputs, or simply providing clearer instructions can also yield impressive results. The things that elicit capabilities may be simple or complex, but in either case, they require engagement with the models to discover.

But the same forces that make humans susceptible to the Clever Hans effect are present, if not stronger, in the case of language models. They are optimized to please us, and to interface with us through the most human-like possible medium, language. And they are good at responding to human input and picking up on user intentions. This makes users especially susceptible to confirmation bias. One LLM naturalist I spoke to — Janus, a husband-and-wife duo who write under a single name — warned me about the danger of projection: "If you have a narrative about what the model is, even if you're not explicitly saying it, everything you say will contain that influence — and this will infect the model." Users who see language models as simplistic may get simplistic behavior out of them; users who see large language models as conscious may, famously, get responses that make them appear conscious.

Today's LLMs can seem like a perfect storm for throwing off our instinctive understanding of minds. They are optimized to act like people, to interact with us in language we understand. But they share less evolutionary heritage with us than bees and octopuses — in fact, they share none. This could make one pessimistic that we will either have to banish all talk of inner states — à la behaviorism — or else get hopelessly confused. Animal cognition offers hope that with care we can do better than either of these. To adopt empathy and respect for these models, in order to spend time with them and appreciate their "perspective," does not mean assuming humanlike cognition or subjectivity. "People really should understand the ways that these models are very different from humans," Janus said. "And they should think about that as part of why they are fascinating and beautiful."

The strangeness of LLMs means that they are smart *in their own way*. They can neither be presumed to be mere next-token predictors, or to neatly map onto human psychology. As de Waal says of chimpanzees, thinking of large language models only in terms of whether they meet or fail to meet *human* standards of intelligence does not do them justice. Naive anthropomorphism can give us an inflated view of what they can do. It can also lead us to underestimate them by blinding us to complex and inhuman ways they have of being intelligent.

70 How We Can Regulate AI

Avital Balwit

The chips used to train the most advanced AIs are scarce, expensive, and trackable—giving regulators a path forward.

ILLUSTRATION BY
Mike McQuade

Two months after its release in late November, ChatGPT reached 100 million users — the fastest-growing software application in history. The past year has seen artificial intelligence models evolve from niche interests to household names. Image generation models can produce lifelike photographs of fantastical worlds, anyone can output functioning Python code, and AI assistants can do everything from take meeting notes to order groceries. And, like the web in its early days, the full impacts of AI have yet to be imagined, let alone realized.

But foundation models — the computationally intensive, powerful systems trained at large labs like OpenAI, Google DeepMind, or Anthropic (I happen to work at Anthropic, but wrote this in a purely personal capacity) — have darker potential too. They could automate disinformation campaigns and widen vulnerabilities to sophisticated cyberattacks. They could generate revenge porn and other disturbing deepfakes. They could be used to engineer a pandemic-class virus or make a chemical weapon. Some of these risks require access to specific tools; others will require further technological progress. But, unfortunately, we have good evidence that some of these uses are either possible now or will be very soon. What's more, in the future more capable models could become hard for humans to supervise, making them potentially difficult or impossible to safely control.

In recent months, everyone from policymakers and journalists to the heads of major labs have called for more oversight of AI — but there's no clear consensus what that oversight might look like, or even what the term "AI" encompasses. Fortunately, the most dangerous type of AI — the foundation models — are also the easiest to regulate. This is because creating them requires huge agglomerations of microchips grouped in data centers the size of football fields. These are expensive, tangible resources only accessible in large quantities to governments and major AI labs. Regulating how they're used is the focus of compute governance, one of the most promising approaches to mitigating potential harms from AI without shutting down progress or imposing onerous regulations on small academic projects or burgeoning startups.

Chips-First Regulation

AI hardware is a uniquely promising governance lever. In 2020, researchers from OpenAI noted that "Computing chips, no matter how fast, can perform only a finite and known number of operations per second, and each one has to be produced using physical materials that are countable, trackable, and inspectable." Unlike the other components of AI development, hardware can be tracked using the same tools used to track other physical goods.

In the years since, policy researchers have begun to map out what a compute governance regime would look like. The basic elements involve tracking the

location of advanced AI chips,[1] and then requiring anyone using large numbers of them to prove that the models they train meet certain standards for safety and security. In other words, we need to know who owns advanced AI chips, what they're being used for, and whether they're in jurisdictions that enforce compute governance policies.[2]

Who Owns Advanced AI Chips?

In order to track who owns advanced chips, and how many they have, a compute governance system will need to create a chip registry.

A chip registry can build off existing practices within the advanced chip supply chain. There are fewer than two dozen facilities worldwide capable of producing advanced AI chips, and these chips already come tagged by their manufacturers with unique numbers that are relatively hard to remove. These numbers could be stored in a registry with a list of each chip's owner. The registry would be updated each time the chip changed hands, and would also track damaged and retired chips.

There's precedent for registries like this across a range of domains, from fissile nuclear material to stock shares to cars. Tracking physical objects, particularly when very few companies produce them, is not a difficult technical problem — though it would require coordination and incentives. The most plausible maintainer of the registry would be a government agency, but it could also be done by an international agency, an industry association, or even an independent watchdog organization.

What Are the Chips Being Used For?

With an effective registry in place, we need to be able to tell what the chips are being used for. The most dangerous scenarios involve large numbers of chips being used to train a new state-of-the-art AI model, since larger models are more broadly capable and often display unexpected new abilities — both qualities that represent greater risk. Even if we already know from the registry that an actor owns a large quantity of chips, they could be spread across many smaller projects. We'd need a way to tell if all those chips were being used for a single training run — the compute-intensive process that produces a new model.

One way to achieve this is to require that anyone with a large number of chips "preregister" their training runs — likely with the same entity that maintains the chip registry. This could be done by the AI company, the data center (which is almost always a separate company, like one of the major cloud providers), or both, to provide an additional check.

In preregistration, the developer would have to specify what they were training and how. For example, they might declare their intention to make the next generation of their language model, and provide information about the safety measures in place, how the previous generation had

1. I use "advanced AI chips" to designate the hardware processors — primarily graphics processing units — that are used to train frontier AI models. Currently, the most commonly used are Nvidia A100s and H100s, which range in price from $10,000 to $40,000 each. This is to distinguish them from the less-powerful chips that are in our cars and appliances, and the similarly powerful chips that have lower interconnect bandwidth that are used for gaming.

2. A useful framework, developed by Harvard researcher Yonadav Shavit, breaks compute governance into interventions at the chip, data center, and supply chain levels. I will make use of elements of this framework, but will instead focus on who owns the chips, what they are being used for, and whether they are in jurisdictions that enforce compute governance policies.

ROOT
(0)
LOC OBJ SEQ SOURCE STATEMENT
MODE 1
STEP 5
IF J>300
GOTO 30

performed on a slate of evaluations, and what evaluations they intended to conduct on the latest version.

For a regulator to fully guarantee that the training run is in compliance, they would need a way to verify what computations were actually taking place on the AI chips. But this presents a problem: AI companies wouldn't want to share this information — which is, after all, their most sensitive intellectual property. However, there are ways for the regulator to check that the contents of a training run are in compliance without being granted full access. Yonadav Shavit, PhD researcher at Harvard, has suggested a method where chips would occasionally store "snapshots" of their computations at different checkpoints. Regulators could then examine they — after they've been run through an algorithm that maintains key information while preserving the privacy of the most sensitive model information — and verify that they match up with the conditions of the preregistered training run.

This method would give regulators a high degree of confidence that a training run complies with regulations, but it's still uncertain what it might look like in practice. It does have limitations — there are still open research problems to building the technical components of this process, and the risks of leaking IP and the costs of implementing it might make it unwieldy to comply with. But less-thorough measures could still be a meaningful improvement over the status quo. For example, regulators could compel data center operators to report training runs above a certain compute size, and ask them to conduct "Know Your Customer" processes — procedures like the ones banks use to confirm that their customers are who they say they are. With this information, regulators could, at minimum, ensure the actor conducting the run is not a criminal organization or rogue state, and encourage them to comply with best practices. This alone could increase safety — and it has the virtue of being possible to implement immediately.

In situations where training runs are deemed noncompliant, enforcement would be necessary. This could take the form of fines, criminal penalties, confiscation of chips, or the termination of a training run. And this enforcement would need to be fast — on the order of weeks or months — given the speed at which models are developed and deployed. (This may sound like an obvious point, but the speed of government enforcement actions varies widely — the Bell antitrust case took approximately a decade to resolve, whereas a drug bust might take only hours).

After the government deemed Silk Road illegal, they identified the data centers that hosted it and forced them to shut down the website. It's possible the government could repeat a similar playbook to halt a training run midway.

Compute governance researcher Lennart Heim uses the example of Silk Road — the illicit goods and services network hosted on the dark web — as an analogue: After the government deemed Silk Road illegal, they identified the data

centers that hosted it and forced them to shut down the website. It's possible the government could repeat a similar playbook to halt a training run midway.

Will the Chips Stay in Countries That Enforce These Rules?

All of this only matters if the chips stay in jurisdictions that sign on to enforce regulations on AI. Export controls are one rather blunt tool to keep chips from leaving compute-governed areas.

As the home to all of the companies that design advanced chips, the U.S. is well placed to enforce export controls. American companies Intel, Google, and Nvidia each possess between about 70% and 100% of the market for the processors they make for AI (central processing units, tensor processing units, and graphics processing units, respectively). State-of-the-art AI models are overwhelmingly trained on GPUs and TPUs. This means that the U.S. would be able to establish these proposals almost unilaterally. (There are two extraordinarily important non-U.S. companies further up the chain — TSMC, in Taiwan, which fabricates the advanced chips, and ASML, in the Netherlands, which makes the machines TSMC uses. However, both countries are sufficiently interconnected to the U.S. supply chain that they will almost certainly cooperate on implementing a plan for oversight.)

The U.S. has already implemented a series of export controls that limit China's access to frontier chips and chip-making technology. Other key countries in advanced chip manufacturing like the Netherlands, Taiwan, Japan, Germany, and South Korea have supported the U.S. efforts so far, suggesting they may be willing to support future export policies as well.

Ideally, this more extreme measure could be avoided by making a lightweight regulatory regime that only targeted the largest, most risky AI development projects, which only a handful of companies can afford. In this case, other countries might be open to enforcing the policies themselves.

Decentralized Computing — A Challenge to Compute Governance?

So far, the methods I've discussed protect against risks from one kind of AI: foundation models trained on advanced compute in large data centers. Though state-of-the-art models are currently trained in this fashion, this may change. Researchers are currently studying the feasibility of using decentralized sources of compute to train a large model. This might look like stringing together smaller clusters of CPUs and GPUs located far from each other. An extreme case might even involve chaining together larger numbers of less-powerful processors, like laptops.

On its face, decentralized computing looks prohibitively inefficient. Current foundation models are much too large to store on an individual or even a few chips. Instead, they are born in data centers where hundreds or thousands of processors are clustered together in racks and connected with cables. This enables chips to quickly talk to each other. The farther apart they are — the greater the latency — the longer training takes. Latency also degrades the model's performance: Because updates are slower to reach the relevant part of the neural network in response to a specific incorrect output, it hampers its ability to learn. Decentralized training will need to reckon with these challenges.

However, they aren't insurmountable. Imagine a large language model trained solely on laptops. For a very quick sketch of the problem with some speculative math: The amount of compute needed to train GPT-3 on laptops instead of Nvidia chips would require around 10,000 2022

Macbook Pros working for a month. (While this gives an analogy for the *amount of computing power* needed; it is currently an infeasible setup.)

Ten thousand Macbook Pros costs $24 million. The cost of training GPT-3 was likely between $4 and $12 million when it first came out, and would be less than $1 million today. Decentralized training this way is not efficient by any means.

Now imagine the above example, but instead of 10,000 laptops, some actor used 600,000 laptops, accomplishing the same training run in one night. You might volunteer your computer while you sleep to contribute to some scientific effort or to receive access to the model in exchange. Around a million people mine Bitcoin, despite Visa more efficiently accomplishing a similar purpose. Decentralized computing for training AI could similarly become viable despite its inefficiency.

Why worry about a problem that is both technically unsolved and less efficient than the current paradigm? Already decentralized computing is gaining traction among researchers as an interesting area to work in, and some progress has been made in fields like medical AI. While inefficient, these methods could still be attractive for actors that don't have a better choice. Most notably, Chinese labs could become key players in advancing decentralized training techniques. Export controls have left them without access to state-of-the-art chips. In the future, stringing together older chips with lower interconnect might be their only option to compete with other frontier models.

Ultimately, decentralized computing does not seem to undermine the case for compute governance. It is currently impossible to train state-of-the-art models via decentralized training, and even as research on it progresses it seems likely there will be a large efficiency penalty. And if it does become viable to train state-of-the-art models this way? These processes would involve the coordination of a large number of actors or large capital outlays for older chips and other networking hardware. If the world has implemented significant compute governance, then it's probably possible to detect these other methods using standard intelligence.

The public and the government need to decide what standards and evaluations they want AI models to meet. Researchers and independent organizations have started to address the technical problems this entails, and the public debate around values and policy objectives is just beginning.

Evaluations and Standards

So far, I've described a regulatory regime without discussing what purpose it might serve. I have used "building safe models" as a placeholder, but to be clear: Compute governance is compatible with a wide range of goals. A compute governance regime is only as good as the standards it enforces — and determining those standards is a significant challenge in itself.

The public and the government need to decide what standards and evaluations they want AI models to meet. Researchers and independent organizations have started to address the technical problems this entails, and the public debate around values and policy objectives is just beginning.

Likely, these standards will — and, in my opinion, should — focus on whether AI models can cause harm in the real world, from helping terrorists build weapons or enabling cyberattacks to operating in ways that are difficult for humans to control.

Unfortunately, nobody knows how to train a model that will consistently refuse to take harmful actions. There are techniques that let a model learn from human feedback whether its responses are helpful or harmful, but these are currently imperfect and unreliable.

Right now, we can address this problem by testing whether models *are capable* of doing something dangerous. For example, standards could try to catch new, dangerous capabilities by requiring that each new model not exceed a prespecified size increase from the last model that was verified to be safe. This would help because larger models often contain new and surprising capabilities, and more careful scaling increases the chance that we can catch them while they are more manageable. Another potential standard is to require that AI companies and data centers both follow best practices for cybersecurity, so that a model will not be stolen and misused by criminals or rogue states.

One organization working to develop these standards is ARC Evals — a project of the Alignment Research Center, a nonprofit focused on developing safe AI systems. Their early work has focused on two challenges: developing tests and standards that reliably capture what models are capable of, and building the processes and infrastructure to evaluate models for compliance. Their first standard, nicknamed "Survive and Spread," asks whether an AI model is able to self-replicate and acquire resources — and possibly therefore elude human control. We can imagine other standards that focus on concrete harms: for example, if AI systems are capable of manipulating and persuading humans to achieve their goals.

While this work is promising, our ability to evaluate advanced AIs is still in its infancy. Considerable work needs to be done to develop reliable and effective evaluations and standards. There are also open questions about who should produce and enact them — a government body like the National Institute of Standards and Technology, independent third-party nongovernmental entities, or some combination.

Compute governance is more of a vision than a template we can roll out immediately. But it is one of the most promising levers for governing the development and deployment of AI systems. More research needs to be done to work out its open technical problems. At the same time, there needs to be a parallel effort to create the standards that compute governance will enforce.

78 The Transistor Cliff

Sarah Constantin

**Moore's law may be coming to an end.
What happens to AI progress if it does?**

The biggest AI models are trained on expensive, state-of-the-art microchips, or semiconductors. Only a few organizations, such as Google and OpenAI, have the budgets to train them. For years, improvements in AI performance have been driven by progress in this underlying hardware.

For most of the history of semiconductor manufacturing, steadily and predictably accelerating improvements in performance and reductions in price have been the norm. This pattern has been codified as "Moore's Law," Intel CEO Gordon Moore's observation that the number of transistors that could be placed on a chip for the same price doubled approximately every two years. That may be coming to an end. Depending on the specific semiconductor performance metric, Moore's Law has either stalled out already, or is on course to soon hit fundamental physical limits.

So, what could happen "after Moore's Law"? And how would that affect AI performance?

Let's zoom in and look at the details.

What Does Scaling Mean?

Scaling laws in AI generally relate the performance of a model to its inputs: training data, model parameters, and compute.

The *performance* of a model describes its accuracy in choosing the "right" answer on known data. A large language model is trained to predict text completions; the more often it correctly predicts how to complete a text, the better its performance. A close antonym of performance is "loss," which is a measure of how far off a model's predictions were from reality; lower loss means better performance.

Training data is the size of the dataset a model is trained on. The number of *parameters* in the model is a measure of its complexity — equivalent to the number of nodes in the neural network.

Finally, the amount of "compute" used for a model, measured in floating point operations, or flops,[1] is simply the number of computer operations (typically matrix multiplications) that must be performed throughout the model's training. Compute is therefore influenced by *both* the amount of data and number of parameters.

The scaling relationship between loss and compute found by OpenAI in 2020 is a power law. If a model has 10 times the compute, its loss will be about 11% lower. This tells us how much "better" models can get from "scaling compute" alone. It's difficult to say exactly what "11% lower loss" means in terms of how powerful or accurate a model is, but we can use existing models for context.

GPT-2, which OpenAI released in 2019, was trained on 300 million tokens of text data and had 1.5 billion parameters. GPT-3 — the model behind ChatGPT — was trained on 300 billion to 400 billion tokens of text data and had 175 billion parameters. The details of their newest model, GPT-4, have not been made public, but outside estimates of its size range from 400 billion to 1 trillion parameters and around 8 trillion tokens of training data.

In other words, training GPT-3 took about 200,000 times as much compute as GPT-2, and GPT-4 probably took between 60-150 times more than GPT-3. In practical terms, GPT-2 could produce coherent sentences, but its output tended to degenerate into repetitive noise after about a paragraph. The much larger GPT-3 can reliably generate on-topic, sensible

1. Floating point operations per second, flops, is a measure of a computer's performance. Floating point refers to numbers in a computer's memory — "floating" because the decimal point can move around. Operations refer to basic arithmetic, while per second is the time component. For context, an Apple M1 chip manages 2.6 teraflops of performance, while an Nvidia A100 GPU manages 312 teraflops. Large labs may use thousands of GPU clusters to train a model.

completions. GPT-4's performance — on everything from programming problems to the bar exam — is even more impressive.

Looking at a longer time horizon, Epoch AI estimates that the compute used for training the state-of-the-art machine learning models has increased by about eight orders of magnitude (that is, 100 million times over) between 2012 and 2023.

If the largest AI models continue to grow at their current pace through the end of this decade, that would be the equivalent of three orders of magnitude of compute growth. That's more than the compute growth between GPT-3 and GPT-4, though less than the compute growth between GPT-2 and GPT-3. As extremely large models have become more compute-intensive, the pace of their growth seems to have slowed.

It's still possible that the compute devoted to AI models will accelerate faster than the current trend. Perhaps AI will attract greater investment and resources as the first LLM-driven product are released and become widely popular. But there are some reasons to expect that we may run into fundamental limits to how much compute can go into LLMs by the end of this decade.

Moore's Law in Relation to AI Progress

In 1965, Gordon Moore posited that the number of transistors in an integrated circuit at the *lowest price per transistor* doubles about every two years. At least with respect to the number of transistors per chip, this has held true.

But Moore's Law looks stagnant if we include Moore's original criterion of *price*. The cost per transistor stopped decreasing in 2011 with the 28 nanometer (nm) node (today's state of the art transistors use 3 nm, with 2 nm likely to be released next year). Since then, transistor costs have increased, rising to $2.16 for the latest 3 nm nodes — costs not seen since around 2005.

Where money is no object, the transistor density of the best available computer hardware is currently still growing at an exponential rate; but if price matters, transistor density *at the best available price* stagnated more than a decade ago.

What about state-of-the-art performance? The quantity we care about, for the purposes of predicting AI progress, is the top speed of the hardware, measured by the number of operations it can carry out per second: peak flops.

The "compute budget" of an AI model is given by

$$C = \text{training time} * \text{number of cores} * \text{peak flops} * \text{utilization rate}$$

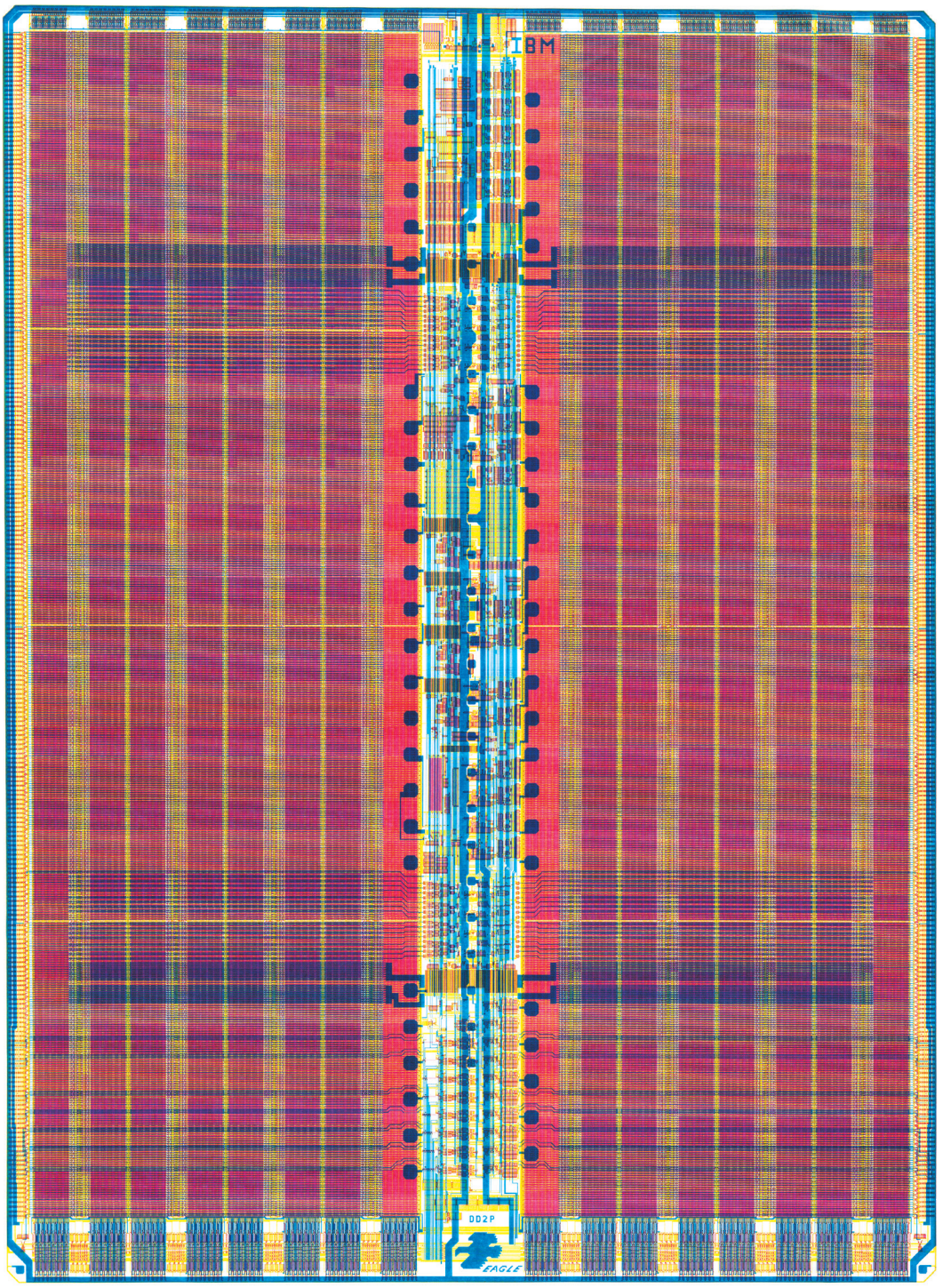
In other words, compute (and hence performance) scales with the amount of *time* devoted to training a model, the *number* of computers (these days, largely GPUs[2]) performing computations in parallel, the speed of the GPU when it's running, and the u*tilization rate*, i.e., the percentage of the time the GPU is actually executing tasks while the model is training.

"Wait a minute, why would the GPU be idle?"

Because training an AI model involves more than just multiplying numbers. Critically, it also involves calling memory and communicating between different processors. Even the most efficient models on today's hardware spend 40% of training time making calls to memory. Empirically, utilization rates seem to be 30–75% at best. Utilization rates also decline with the number of GPU processors used in parallel, since the more processors you use, the more time you'll have to "waste" sending data between them.

Training time probably can't scale up much from here; the largest language models are already spending months training, and firms may not find it profitable to spend years training a single model. So, if you assume that OpenAI, DeepMind, Meta, and the other big AI

2 Graphics processing units are computers that were initially designed for image rendering, but are effective for completing many highly parallelizable tasks, including training and inference of neural network models.



Sam Lucente (b. 1958). Diagram of Dynamic Random-Access Memory Chip (DRAM), Corresponding Silicon Microchip, 1984. Manufacturer: IBM, East Fishkill, NY. Computer-generated plot on paper and silicon.

players are time constrained at current margins and not cost constrained, then the growth of compute spent on LLMs should scale with peak flops and the number of cores.

GPU flops have grown at a doubling rate of roughly every two years.

The primary drivers of this trend in improved GPU performance are smaller transistors and increased numbers of cores.[3] Straightforwardly, as Moore's Law makes transistors smaller, and each GPU contains more transistors, then each GPU will compute more operations per second with those transistors.

But there are inherent physical limits to how small you can make a transistor.

Fundamental Physical Limits to Transistor Size

One limit has to do with thermodynamics. As transistors become smaller, it takes less and less energy to flip the gates which control the currents from "on" to "off" and back. Once this "switching energy" drops to the same scale as the energy fluctuations produced by the random molecular movements we call heat, then the transistor will turn on and off at random.

So what is this thermodynamic minimum gate length? One paper from 2015 estimates it as in the 4-5 nm range — a limit we will likely reach by 2030.[4] But the thermodynamic minimum gate length depends on the material being used. Moving beyond silicon to newer semiconductor materials which can hold more electrical energy will allow smaller transistors to stay "switched off" despite thermodynamic fluctuations.

Researchers using novel materials like this may be able to produce much smaller transistors, at least in the lab. A team at Tsinghua University in China claims to have fabricated an 0.34 nm transistor made of graphene and molybdenum disulfide. But it's a long way from fabricating one transistor in the lab to mass-producing hundreds of billions on a chip, and not all materials are amenable to mass production.

Another physical limit to making transistors smaller has to do with light resolution. Currently, circuits are etched into semiconductors in a method known as photolithography. Ultraviolet light is projected through a "mask" to hit a semiconductor wafer in precise geometric patterns, where it reacts with photoactive materials. Then, strong solvents are used to etch away everything the light *didn't* touch, leaving a raised pattern that forms one layer of a circuit. But light is a wave, and it's impossible to resolve features smaller than about half the frequency of the light — in the range of tens of nanometers.

This is why semiconductor manufacturers have been using higher and higher frequency light at enormous cost, among other tricks. But even so, you simply can't etch things that are that much smaller than the light wavelength.

Theoretically, it would be possible to use higher-frequency radiation like x-rays. But apart from the extreme cost and the need to develop new technologies and materials, x-rays are ionizing radiation — they interact with everything they touch, scattering electrons and "blurring" the resolution of the image. The smallest x-ray lithography feature sizes produced to date are actually, at 30 nm, *larger* than what ultraviolet photolithography at its best can provide. Paolo Gargini, the chair of the IEEE International Roadmap for Devices and Systems, the semiconductor industry's organization for predicting and planning progress in chip manufacturing, predicts that we'll reach the limits of photolithography around 2029.

3. EpochAI formalizes this in one of their analyses: Out of six potentially relevant variables, most of the variance in flops was captured (r2 = 95%) by only two, the number of cores and the process size.

4. TSMC's latest "3 nm" fab does not actually produce transistors with gate lengths of 3 nanometers. When referring to a semiconductor manufacturing node, "3 nm" is a marketing term that does not refer to the size of any feature. The actual gate length in a 3 nm node is closer to 16-18 nm.

For two independent reasons, it seems we have less than a decade left of shrinking transistor sizes.

Beyond Moore's Law: Alternative Paths

There are several paths to achieve more flops without making transistors smaller.

The first path is to redesign chips. One option is to manufacture 3D chips, where transistors are stacked vertically. 3D stacked complementary metal-oxide semiconductors (a type of semiconductor that uses two types of transistors to allow for efficient switching between "on" and "off" states) can double transistor density. Intel recently announced progress on novel materials which promise a 10x improvement in transistor density by 2030. And two-layer CPU chips, which can improve transistor density by 40%, were released in 2021.

These developments build on 3D memory designs, first commercialized a decade ago. Each release has come with 30-50% more layers. And in principle, layers can be stacked arbitrarily high, allowing transistor density to scale linearly without shrinking the individual transistors. Samsung, for instance, predicts that they'll reach 1,000 layers by 2030.

The second path is to design special-purpose chips. The best possible flops for a given application may not be achievable on a general purpose computing device, but rather on a special chip architecture designed for the application. There are two main options here.

Application-specific integrated circuits, or ASICs, are rigidly special-purpose, designed for exactly one type of computation. ASICs are widely used in telecommunications equipment, modern vehicles, and medical devices. Field-programmable gate arrays, or FPGAs, are a notch more flexible, allowing users to configure their own logic circuits. AI-specialized "accelerator" chips, which are optimized for training neural networks, are a type of FPGA. AI accelerators can produce more flops than GPUs at the same or lower transistor density.

Google's TPU (Tensor Processing Unit), for example, is a custom type of ASIC designed specifically for accelerating machine learning tasks. Third generation TPUs have two to four times the flops of the widely used Nvidia V100 GPUs, despite the GPUs being fabricated on a 12 nm node and the TPUs only being fabricated on a 16 nm node. But despite a zoo of emerging competitors developing special-purpose architectures, as of 2022 Nvidia's latest generation of GPUs, the H100s, are still the leaders on a standard MLPerf benchmark test.

As Moore's Law makes transistors smaller, and each GPU contains more transistors, then each GPU will compute more operations per second with those transistors. But there are inherent physical limits to how small you can make a transistor.

So, while in principle special-purpose AI chips could get more flops with the same number of transistors, and while they are often cheaper on specific training tasks, they haven't yet come out firmly ahead of standard GPU architectures at maximum processing speed.

The third path involves replacing transistors with other kinds of switches. There's no physical law that says computation has to be done with transistors. Alternative models for computation, which include optical computing and memristors, could be faster and scale better, but most of these are still in their infancy.

Optical computing uses light instead of electrons for computation, which results in

less energy and heat. Moreover, photons are about 20 times faster than electrons. One experimental optical switch, developed by IBM researchers, can alternate 1000 times faster than conventional transistors. A more recent result from the University of Arizona found switching speeds for an optical device that are a *million* times faster than transistors.

But while optical switches may be fast, they can't be dense; to transmit light, optical waveguides can't be much narrower than the light's wavelength, which in this case is hundreds of nanometers. (By contrast, conventional semiconductor transistors can have feature sizes in the tens of nanometers.) So all-optical computing devices remain speculative.

Memristors are another alternative. A memristor is an electronic component whose resistance depends on the accumulated electric charge that has passed through it — in contrast to a semiconductor, whose conductivity depends on the *current* presence of an electric field. It's possible memristors could scale to be smaller than transistors; the smallest ones produced in the lab are about 1 nm. But like optical computing, the technology remains unproven.

Putting together these three classes of beyond-Moore's-law innovation, we're looking at:

- Advanced packaging and 3D designs: at least 10 times transistor density improvement by 2030, possibly continued 10 times growth into the 2030s as more layers are stacked
- Special-purpose computers: up to 10 times flops speedup depending on whether greater architecture optimizations are possible
- Non-transistor computation paradigms: very uncertain, and might not happen at all, but could theoretically improve flops by five to 1000 times.

A moderately likely scenario, for instance, might be "Moore's Law holds until 2040 with 3D architectures, special-purpose AI accelerators don't provide any flops improvement, transistors remain the main building block of computation throughout the 2030s." In such a scenario, the peak flops achieved by a GPU might grow from nearly a teraflop in 2022 to hundreds of exaflops by 2040, or a five-order-of-magnitude increase over nearly two decades.

In the more pessimistic scenario where flops stop growing altogether by 2030, we'll only see a two-order-of-magnitude increase in peak computation speed by 2030, and no more between 2030 and 2040.

This means that, for the current rate of compute growth of the largest AI models to continue through 2030 (resulting in models three orders of magnitude more compute-intensive), state-of-the-art models would need to use significantly more computer chips and cost far more than they do today. In this case, a jump in model scale comparable to that between GPT-3 and GPT-4 would take until the end of the decade, depending on how easy it is to acquire and train across a vastly increased numbers of chips.

The Memory Wall

Even if GPUs didn't improve their peak flops much (or at all), couldn't an AI company just buy lots and lots of them and run them in parallel, and see linear improvements in "compute"?

No, because training *time* is an issue. The biggest models today take six months to train, and a significant portion of that time is spent writing and retrieving model weights from memory.

Memory is stored on a separate device from the chip that does the computation. Typically this is DRAM.[5] Using clever optimizations you can design the training algorithms to minimize the number of calls to memory, but ultimately these are one-time improvements

5. Dynamic random-access memory, the most rapidly accessible form of memory that's not directly on the same chip as the compute.

6. This is assuming that the batch size, or number of samples the model processes before it updates, remains constant — which it usually does.

that don't scale as models get bigger. The more data points a model is trained on, the more calls to memory there must be to update the weights.[6] The best utilization rates observed with Nvidia's A100 GPUs are about 60%.

So, is memory bandwidth improving over time at the same rate as compute? Not so much.

Peak DRAM bandwidth has been increasing far slower than flops, at 30 times in the last 20 years (flops have increased 90,000 times in the same period). Today, a model that takes six months to train and has a 60% GPU utilization rate will spend about 2.5 months just transferring data to and from memory.

Let's say we want to no more than double that time — ever. And let's say that DRAM bandwidth continues to grow at its current rate. In that case, it doesn't matter how much compute we have. In order for the training run to be able to use all of its available compute without memory call times ballooning, compute cannot grow by more than 3 times by 2030 and 17 times by 2040. This is a much more conservative bound on AI compute growth than either what Moore's Law for flops suggests (128 times by 2030) or what recent AI compute trends suggests (631 times by 2030). In a world where memory bandwidth (and time) is the limiting factor, we *don't even get one more order of magnitude of scaling growth in AI compute this decade*.

That world looks like getting GPT-5 in 2040. Or, it looks like OpenAI CEO Sam Altman's recent announcement that we're at the "end of the era where it's going to be these, like, giant, giant models" and that they will not be training GPT-5 for "quite some time."

On the other hand, there are some countervailing factors that might make this picture look different.

Memory Bandwidth Improvements

The memory bandwidth (in GB/s) between DRAM memory and the processor depends on factors such as the memory clock speed (how many operations it can perform per second) and the memory bus width (how many bits of data can be transferred per cycle), as well as other aspects of the memory architecture, chip design, and manufacturing quality.

In the more pessimistic scenario where flops stop growing altogether by 2030, we'll only see a two-order-of-magnitude increase in peak computation speed by 2030, and no more between 2030 and 2040.

For memory, as for logic, clock speeds increased along with Moore's Law over many decades.

But while transistors have continued to get smaller, they have stopped getting faster.

Switching speed depends on the width of a transistor's gate, but gate widths are now a single molecule wide and can't actually get any narrower. So further shrinking the other dimensions of transistors doesn't increase their speed.

Clock speeds, in fact, have been flat since 2004.

So we can't count on clock speed alone to improve memory bandwidth.

What about increasing bus widths?

Stacking multiple layers of DRAM dies, in a format known as High Bandwidth Memory (HBM) can increase memory bandwidth by allowing larger bus widths. There are more independent connections between the HBM and the GPU, allowing faster data flow.

However, HBM chips are far more expensive than traditional DRAM — the most advanced

HBM costs $120/GB compared to about $3/GB for DRAM.

Moreover, the limiting factor in stacking more and more layers of memory, or packing circuit elements denser, is heat. Memory requires power to store information, even when it's not "on," and it dissipates heat. Today, even going beyond 12 layers may be infeasible due to heat constraints. Memory is especially sensitive to heat, because at higher temperatures, thermal noise can degrade stored data. Faster degradation means the data needs to be refreshed more frequently — but refreshing also generates waste heat! So there's a vicious cycle where overheating leads to even more overheating.

Improving heat dissipation is an active area of research, and so more heat-efficient memory designs may be invented in future years, but scaling up memory bandwidth above its current slow trajectory is likely to continue to be challenging.

AI Model Efficiency Improvements

Another approach is to redesign AI models (or the algorithms for training them) so that they require less memory bandwidth or computational power.

One recent example of progress in memory is FlashAttention,[7] a method for computing attention (a component of all current state-of-the art AI architectures, including LLMs) that reduces how frequently the model accesses memory. On GPT-2, FlashAttention led to a tripling of training speed (although this might be an unusually favorable result — attention takes up about 20% of the cost for most LLM training runs, so the effect would typically be less dramatic).

When training is parallelized across many GPUs, avoiding redundancy in memory storage can also reduce a model's memory footprint, producing 8 times speedups on a billion-parameter (GPT-2-sized) model. If memory bottlenecks loosen by an order of magnitude or more, we might return to our "flops-bottlenecked" scenario where AI models' computational load can grow two to three orders of magnitude by 2030 and perhaps as many as five orders of magnitude by 2040.

Rather than a *refutation* of scaling laws, or an acceleration of their slope, I think this is more like a move in a different direction altogether, towards a Cambrian explosion of "little AIs" used for different purposes, where getting good performance on a task depends on the quality of your task-specific dataset.

There have also been innovations on the compute front. Compact open-source models like Alpaca, which uses only 7 billion parameters plus fine-tuning on a combination of human-generated and LLM-generated examples, produce similar performance to the much larger GPT-3 (175B parameters) – and it can be trained in only 3 hours for less than $100. In the same vein, LoRA[8] is a novel training scheme that allows computationally cheap fine-tuning of large language models

7. Tri Dao et al., "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," ArXiv, June 24, 2022, https://doi.org/10.48550/arXiv.2205.14135.

8. Edward Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ArXiv, October 16, 2021, https://doi.org/10.48550/arXiv.2106.09685.

to specific tasks, allowing a 25% speedup on fine-tuning of GPT-3, and much easier parallelization to multiple GPUs.

The open-source community's progress in creating compact, cheap-to-train LLMs may have leapfrogged the big AI companies — at least according to a memo allegedly leaked by a Google employee titled "We Have No Moat, and Neither Does OpenAI." If large language models don't have to be large to work well, they need not remain the purview of a handful of large tech companies. If smaller models can match GPT-3-like performance, does that mean that we should expect far *better* performance than the current state-of-the-art will be possible with less compute than the 2020 and 2022 scaling laws suggest?

It's not clear. Small, open-source models like Alpaca and Vicuna showed that fine-tuning a small model with a carefully curated dataset of (real and simulated) human-computer interactions can almost match the performance of a larger LLM trained from scratch. We can interpret this as an insight about *task-specific* training. These models lack the flexibility of their larger counterparts, and can't compete on tasks that require more robust reasoning skills. A general LLM trained on human text can, among other things, function as a chatbot; but a smaller LLM trained on a smaller dataset of human text plus a *targeted* dataset of human-chatbot interactions can perform nearly as well, for far more cheaply.

Rather than a *refutation* of scaling laws, or an *acceleration* of their slope, I think this is more like a move in a different direction altogether, towards a Cambrian explosion of "little AIs" used for different purposes, where getting good performance on a task depends on the quality of your task-specific dataset. That could be consistent with the state of the art continuing to progress steadily along "scaling law" lines for quite some time, but it could also mean the economic incentive towards ever-bigger models would diminish and we'd enter an entirely new era where AI progress would *not* be driven primarily by semiconductor scaling.

The End of Scaling: Not the End of AI

What happens when we run up against the limits of AI scaling?

Whether they're training time limits, data availability limits, or limits based on the cost and availability of computer hardware, we might not be far from the end of the era when the most straightforward way to make models better is to make them bigger.

That doesn't, of course, mean the end of making models better.

Scaling is just the most naive, straightforward way to improve AI models — and it has worked surprisingly well, for a while. In a world where scaling has stalled, progress in AI will look more like innovation in developing new applications and fine-tuned variants of the big foundation models we already have, along with architectural and algorithmic innovation to push out the fundamental capabilities of the big models without using more data or compute.

A post-scaling scenario for AI might look like an "AI winter," or it might look like an acceleration of AI capabilities — just driven by other, less predictable factors than the steady drumbeat of Moore's Law.

88

The Puzzle of Non-Proliferation

Carl Robichaud

Today, only nine countries have nuclear weapons. For most of the 20th century, analysts expected that number to be much higher. The story of how we arrived here holds important lessons for AI.

Atomic weapons entered the world in the 1940s, alongside the first jet engines, microwave ovens, radars, and electronic computers. Virtually every other invention of that era has spread throughout the world. Nuclear weapons are the exception. While two dozen countries retain some nuclear latency as a hedge, today only nine countries have the bomb.

This outcome was hardly inevitable. Knowledge of how to build nuclear weapons quickly escaped the confines of secrecy, and over time it became clear that even countries of modest means — North Korea, Pakistan, and South Africa — could develop the bomb with sufficient dedication. International rules and pressure could raise obstacles, but countries willing to pay the cost — to "eat grass or leaves" as former prime minister of Pakistan Zulfikar Ali Bhutto once put it — would, with time and luck, succeed.

Why, then, did so few countries do so? Historians, social scientists, and officials have spent careers trying to understand the spread of nuclear weapons. But, at the risk of oversimplifying, we only have nine nuclear-armed states because the vast majority, after weighing costs and benefits, decided not to develop the bomb. Almost all stuck with that choice.

Despite its flaws and vulnerabilities, the nuclear nonproliferation regime may provide useful guidance as new technologies like synthetic biology and artificial intelligence come into their own. The conventional wisdom is that powerful technologies inevitably diffuse widely. The fizzling out of nuclear proliferation provides a counterexample.

The early nuclear age: pessimism and contingency

After World War II, the U.S. debated how to manage its nuclear monopoly, ultimately floating a proposal for international oversight of atomic energy that had little chance amidst Cold War tensions.[1] A U.S. intelligence report circulated in 1949 predicted that the Soviets would most likely develop a weapon by mid-1953 and that the "earliest possible date" was 1950. Just months later the Soviet Union tested its first weapon.[2]

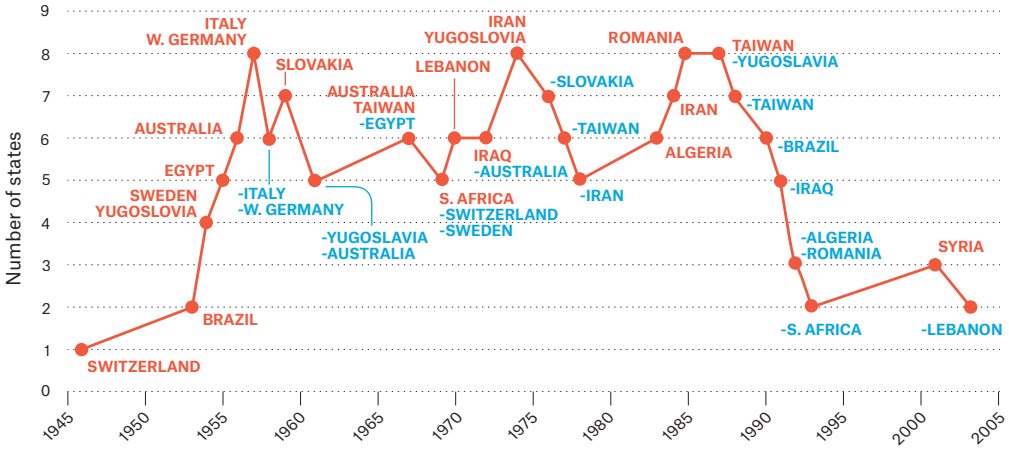
Stalin's sprint to the bomb showed that the technological obstacles to building nuclear weapons were weaker than once thought. In a Brookings Paper entitled *Predicting Proliferation*, scholar Moeed Yusuf surveyed the forecasting literature from 1949 to 1964 and found that "an overwhelming majority of classified and academic studies suggested that horizontal proliferation was inevitable."

One reason for pessimism was an assumption that the barriers were technological, not political. Denis Healey, who

1. One reason the effort failed is that neither Truman nor Stalin saw the bomb primarily as a common threat to be addressed by cooperative action. Looking back 40 years later, Soviet Foreign Minister Andrei Gromyko wrote that "I am certain that Stalin would not have given up the creation of his own atomic bomb. He well understood that Truman would not give up atomic weapons." Anatolii Gromyko, Andrei Gromyko: Polet ego strely (Moscow: Nauchnaia kniga, 2009), 115-116.

2. Central Intelligence Agency, "Status of the U.S.S.R. Atomic Energy Project," MORI 136351, January 1949. See also Michael Gordin, *Red Cloud at Dawn: Truman, Stalin, and the End of the Atomic Monopoly* (New York: Farrar, Straus and Giroux, 2009).

States that explored or pursued nuclear weapons



Source: Adapted from Scott D. Sagan, "The Causes of Nuclear Weapons Proliferation," *Annual Review of Political Science* 14, no. 1 (2011): 225-244.

later became the U.K.'s defense minister, wrote in 1960 that "so far no country has resisted the temptation to make its own atomic weapons once it has acquired the physical ability to do so."[3]

At that time, only the U.S., the USSR, and the U.K. had nuclear weapons, but the "physical ability" to develop the bomb was spreading rapidly. Many countries were planning or building enrichment and reprocessing facilities capable of generating bomb fuel. These capabilities were accelerated by widespread nuclear trade, especially via the Atoms for Peace program, through which the U.S. shared civilian nuclear facilities as an extension of its soft power.[4] Starting in the 1950s, a dozen countries seriously considered or pursued nuclear weapons. Early movers included Sweden, Canada, and Australia, followed by many industrialized countries in Europe and Asia.

In a March 1963 press conference, U.S. President John F. Kennedy warned that 15 to 25 states might obtain military nuclear capabilities by the 1970s. Author and former U.S. security official Peter Lavoy, in a review of declassified documents, notes that Kennedy "based this pessimistic forecast on a secret study that Secretary of Defense Robert McNamara had given the president one month earlier. In this document, McNamara expected that by 1973 eight new states might acquire nuclear weapons — China, Sweden, India, Australia, Japan, South Africa, Germany, Israel — and that, shortly thereafter, many more countries could go nuclear as the cost of acquiring nuclear weapons 'may come down by a factor of two to five times.'"[5]

Uneasy partnership

Seeing a wave of proliferation approaching, the U.S. and the Soviet Union joined in a tacit partnership to limit nuclear

3. Denis Healey, "Race Against the H-Bomb - Fabian tract 322," March 1960, p.3.

4. See Fuhrmann, Matthew. Atomic Assistance: How "Atoms for Peace" Programs Cause Nuclear Insecurity. Cornell Studies in Security Affairs. Ithaca: Cornell University Press, 2012 and Kroenig, Matthew. Exporting the Bomb: Technology Transfer and the Spread of Nuclear Weapons. Cornell Studies in Security Affairs. Ithaca, [N.Y.]: Cornell University Press, 2010.

5. Predicting Nuclear Proliferation: A Declassified Documentary Record Strategic Insights, Volume III, Issue 1 (January 2004) by Peter R. Lavoy.

spread. Negotiations started in earnest at the United Nations Conference on Disarmament in Geneva in 1965.

The resulting 1968 Nuclear Non-Proliferation Treaty (NPT) is often described as having three pillars: non-proliferation, disarmament, and peaceful uses. The central focus was nonproliferation: Signatories that had already tested nuclear weapons agreed not to share nuclear-weapons technology and other signatories agreed not to acquire weapons; safeguards would be provided by the International Atomic Energy Agency (IAEA). On disarmament, the nuclear states agreed to "pursue negotiations in good faith" to end the nuclear-arms race and pursue disarmament under "strict and effective international control," a mandate with no timeline. Finally, the treaty called for technical cooperation toward peaceful uses of civilian nuclear technology.[6]

To say the Nuclear Non-Proliferation Treaty had limitations would be an understatement. It lacked universality and effective verification. It had a duration of 25 years and would require extension every subsequent five years. Two nuclear-armed states, France and China, declined to join, as did a number of other states, including some with active weapons programs: Argentina, Brazil, India, Israel, Pakistan, and South Africa.

So while it is tempting to see the NPT as a turning point in the proliferation story, the treaty in its early years was akin to a white picket fence — more signal than barrier. States continued to hedge, adding to their nuclear capabilities without crossing the line into weaponization.

Alarm and response

Intelligence analysts and outside experts remained pessimistic that nuclear spread could be contained. A threshold was crossed in 1967 when Israel secretly built a bomb, shrouded in ambiguity. Then in 1974 India detonated what it claimed was a "peaceful nuclear explosive" (code-named Operation Smiling Buddha). The dam seemed poised to break, and the U.S. Central Intelligence Agency predicted that 10 other nations had the potential and incentives to go nuclear.[7]

But that never happened. After 1974, only North Korea, Pakistan, and South Africa acquired nuclear weapons. South Africa dismantled its stockpile of six warheads in 1989. Belarus, Kazakhstan, and Ukraine traded away the Soviet weapons stranded on their soil in return for economic and security assistance. Illicit programs in Iraq, Syria, and Libya notwithstanding, the system has held up surprisingly well for the past 50 years.

India's test was something of a pivotal moment in galvanizing international action. It was a demonstration that countries not aligned with Washington or Moscow could get nuclear weapons, and it led to a loosely coordinated response.

First, the major powers leaned on their allies. Over the coming years, the U.S. reiterated security guarantees for allies in Europe and Asia who would forego nuclear weapons, and the Soviet Union suppressed nuclear aspirations among its Warsaw Pact allies. These arrangements were transactional: To get protection from NATO or the

6. New archival research by Jonathan Hunt suggests that alongside the "grand bargain" there were additional bargains within alliances that restrained regional powers from going nuclear. See *The Nuclear Club: How America and the World Policed the Atom from Hiroshima to Vietnam*, Hunt, Jonathan R., Stanford University Press, 2022.

7. Yusuf, p.61. "Going from a prediction that only one country could cross the threshold between 1966 and 1976, the CIA listed 10 potential Nth powers just a year after India's test."

Soviet Union you needed to forego nuclear ambitions.

Second, nuclear exporters reinforced the thin filament of the NPT into a thicker web of regulations and controls. The first strand was the 1974 "trigger list," specifying which nuclear items required IAEA safeguards to export. The Nuclear Suppliers Group, which began in 1974 as the London Club, formed to set voluntary guidelines on sensitive exports. NPT membership expanded from 46 to 91 countries by the second conference in 1975. Signatories to these agreements adopted international guidelines into domestic law. The IAEA increased its professional staff and competency with the support of its member states.

All these steps represented a major evolution in the nonproliferation regime. The restrictions written out in the NPT were now backed by legal, logistical, and political barriers that increased the expense, time, and risk to build a bomb. These measures held up surprisingly well, even in the face of the sophisticated illicit network established by A.Q. Khan, often called the father of Pakistan's atomic-weapons program.

These mechanisms worked in part because they were backed by the threat of sanctions and military force. Israel conducted air strikes against nuclear reactors in Iraq (1981) and Syria (2007). While the U.S. has cautioned cheaters that "all options are on the table," its favored counterproliferation tool has been multilateral sanctions, which the UN authorized against Iraq, Iran, and North Korea. Sanctions have a mixed record with hard cases but the *potential* of sanctions has served as a deterrent to other states.[8]

These three factors — security guarantees, international controls, and counterproliferation pressure — are mutually reinforcing. Safeguards and the threat of sanctions or military force raised the costs of acquiring nuclear weapons while military pacts decreased their benefits by allowing states to achieve security through other means.

But a fourth factor may be more significant still: the establishment of a global nonproliferation norm. The NPT, with 191 signatories, has close to universal participation, and for the vast majority of its signatories nuclear weapons have little value as military assets or political tools.

The norm against acquiring nuclear weapons is closely tied to the taboo against using them, which began to emerge as horrific details from Hiroshima and Nagasaki became public. Leaders, presented with what nuclear war would really entail, often recoiled and sought other paths to security.

The emergence of nuclear-weapon-free zones (NWFZ) helped codify and sustain these norms. NWFZs allowed states to commit to not manufacture, acquire, test, or possess nuclear weapons. The first of these, the Treaty of Tlatelolco (1967), was eventually signed by all 33 countries in Latin America and the Caribbean.[9] Others followed suit in the South Pacific, Southeast Asia, parts of Africa, and Central Asia. By ratifying the NPT and joining a nuclear-weapon-free zone, countries could credibly signal to their neighbors that they weren't seeking nuclear weapons, solving a collective action problem that might have otherwise led to costly and counterproductive arms races.

Norms remain important because nonproliferation relies so heavily on state intent. Under a basic IAEA safeguards agreement, countries declare which materials and facilities to submit to inspection. The IAEA cannot impose conditions on

8. E. Solingen (Ed.), Sanctions, Statecraft, and Nuclear Proliferation (pp. I-V). Cambridge: Cambridge University Press.

9. Cuba was the last to sign, in 1995.

For a sane nuclear policy (1964)
Saul Bass (American, 1920–1996)
Courtesy The Library of Congress

states other than those willingly accepted. The flaws of this approach became apparent in the 1990s when the IAEA discovered that Romania and North Korea had clandestinely extracted plutonium and that Iraq had a covert nuclear-weapons program (which was destroyed by special UN inspection teams after the 1991 Gulf War). In response, the IAEA established an Additional Protocol that allows inspectors more access to data and timely inspections — but this stricter set of rules is only in force with the 140 states that have voluntarily accepted it.

A careful cheater is likely to succeed given enough patience. The NPT gives wide leeway when it comes to "peaceful uses" of nuclear power, which may include national enrichment and reprocessing programs. This means states can remain compliant with their NPT and IAEA obligations while coming a screw's turn from assembling a weapon. Iran was caught pursuing a secret nuclear-weapons program, but had it remained patient and not prematurely pursued weaponization it might have already succeeded.

The international community was able to limit the "horizontal" spread of nuclear weapons by increasing the costs and diminishing the benefits of nuclear weapons with security guarantees, international treaties, export controls, sanctions, military action, and norms. In contrast, efforts to limit "vertical" proliferation within nuclear-armed states was less successful. Strategic stability and arms control slowed the pace of nuclear expansion, but when the Cold War ended, the U.S. and the Soviet Union were still targeting each other with a staggering 57,000 nuclear weapons.

Lessons for the governance of artificial intelligence

Nuclear weapons are often invoked in conversations about transformative AI. In May, the founders of Open AI published a memo arguing that "we are likely to eventually need something like an IAEA for superintelligence efforts." Separately, the Center for AI Safety issued a statement, endorsed by prominent AI experts, that said: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

Norms matter. The nonproliferation regime has been successful because leaders chose not to pursue capabilities within their reach.

What would it mean to treat AI as a societal-scale risk? It's not clear. AI is in a similar place to nuclear science in the 1930s: The scientists at the frontiers of nuclear technology could see the potential for harm, but nobody knew what form it might take. No one had seen a mushroom cloud over Hiroshima or a meltdown in Chernobyl; to suggest such a thing was possible would have put you in the company of science fiction writers and cranks.

Analogies between nuclear technology and AI can quickly break down. Let's take the concept of "an IAEA for superintelligence efforts." The IAEA exists to help states safely and securely operate nuclear technology and to monitor those activities to ensure they remain peaceful. The IAEA can perform this task because there's a clear understanding of what constitutes

safe civilian reactor operation and what constitutes failure (accidents and meltdowns). There is also an understanding of what nuclear activities are most susceptible to proliferation, how to monitor them, and what's at stake. In contrast, there is no consensus yet around which AI activities are unsafe, or whether we should monitor them. Separating military from peaceful uses is not the primary challenge, since AI safety experts are equally concerned about powerful yet misaligned civilian systems.

Despite these differences, I see three takeaways from those seeking to manage AI risk:

First, we should not despair if initial rules and regulations appear too weak. Assuming sufficient warning signs, regulations can become stronger with time. It took three decades — and wake-up calls such as India's nuclear test (1974) and the discovery of a secret weapons program in Iraq (1991) — to galvanize action toward a stronger and more universal regime.

Second, the key to effective regulation is understanding which capabilities are harmful and what the key choke points are. For nuclear weapons the limiting factor is highly enriched uranium and plutonium. These materials are produced in fuel-making and reprocessing plants — facilities that require special scrutiny.

With AI, the limiting factor is most likely computational resources. Current models take hundreds of millions of dollars to train, and the requisite server clusters can be located and observed. It is hard to distinguish whether these computational resources are being used for good or ill, be ways to identify which applications pose the greatest risks.

Finally, norms matter. The nonproliferation regime has been successful because leaders chose not to pursue capabilities within their reach. Could norms of precaution around powerful and recursively self-improving AI systems emerge?

The nuclear taboo came about after harms were evident: terrible human suffering in Hiroshima and Nagasaki. AI, unlike nuclear weapons, has widespread civilian applications and promises enormous benefits to society. It would be undesirable to block all further AI development, even if doing so were possible. But we need norms against deploying technologies with clear pathways to misuse and those that could escape human control.

Recent opinion polls and U.S. Senate hearings suggest that citizens and policymakers are open to restrictions that embed caution. Leaders of key AI-development labs, such as OpenAI CEO Sam Altman, DeepMind CEO Demis Hassabis, and Anthropic CEO Dario Amodei, have expressed concerns over safety. So has Microsoft founder Bill Gates. These nascent norms could contribute to corporate self-restraint

While we are only eight decades into the nuclear age — and entering a new and risky moment — the story so far offers hope. In 2019, Sam Altman paraphrased Robert Oppenheimer, the father of the atomic bomb, saying "technology happens because it is possible." We see a possible future of AI competition — between commercial labs and between the U.S. and China — that could unleash dangerous capabilities and a race to the bottom. But the limited spread of nuclear weapons offers a lesson: The possible is not inevitable.

96

The Great Inflection? A Debate About AI and Explosive Growth

Matt Clancy and Tamay Besiroglu

A conversation about what happens to the economy when intelligence becomes too cheap to meter.

ILLUSTRATION BY
Karol Banach

Many working on artificial intelligence and AI-related issues think that our world will change very dramatically once we develop artificial intelligence capable of performing most of the cognitive work currently reserved for humans. On the other hand, many economists adopt a more cautious stance, expressing doubt regarding the potential for AI to dramatically increase the rate of change.[1] In this conversation, economist Matt Clancy and research scientist Tamay Besiroglu debate the prospects for a radical rupture with historical rates of economic growth and technological progress.

This conversation is based on a series of back-and-forths Matt and Tamay had in spring of 2023. The opinions expressed do not necessarily reflect the views of their employers.

Tamay: Hi, Matt. I'm excited to have this chat.

Matt: Likewise!

Tamay: Before we delve in, I believe it's crucial to frame our dialogue by outlining the central themes we'll be addressing.

This is a debate about the expected economic impact of artificial intelligence much more advanced than today's large language models like GPT-4. Specifically, I want to discuss the impacts of AI advanced enough to perform most or all tasks currently performed by humans. This includes things like running companies and all the planning and strategic thinking that comes along with that, designing and running scientific experiments, producing and directing movies, conducting novel philosophical inquiry, and much more.

These systems I have in mind are clearly leaps and bounds more advanced than any systems we have today, so why do I think this is even worth discussing? I think there is compelling work that points to it being very likely (say about 80%) that such AI systems that at least match the capacities of humans both in generality and capability will be developed this century.[2]

Our second theme centers on the concept of "explosive growth." I'm referring to a rate of growth that far surpasses anything we've previously witnessed — a minimum of tenfold the annual growth rate observed over the past century, sustained for at least a decade.

I am inclined to believe that such explosive growth is not just a possibility, but a probable outcome when we transition to an era where AI automates the vast majority of tasks currently performed by humans. To put this in numbers, I'd

1. For example, in a recent survey of economic experts, only 20% believed AI developed in the next 10 years would have a larger impact on growth than the internet, while 61% were uncertain.

2. This isn't the forum to go into this in detail, but I have in mind specifically work such as Cotra (2020), Davidson (2022), and Epoch's various works on the topic, as well as expert surveys such as Grace (2022).

currently assign a 65% chance of this happening. I think you disagree with this view, correct?

Matt: That's right. While I think it's very likely that growth will pick up once we deploy AI throughout the economy, I think it's maybe a 10% to 20% chance, depending on how I'm feeling, that economic growth becomes explosive, by your definition, with most of that probability clustered around the low end of explosive growth.

'Explosive' is certainly an apt term for what we're talking about. GDP per capita grew at roughly 2% per year over the 20th century, so if we jump to 20% per year for 10 years, that's about 90 years of technological progress (at 2% per year) compressed into a decade. Ninety years of progress was enough to go from covered wagons to rocket ships! And your definition also encompasses even faster growth persisting for even longer!

Tamay: In my view, accelerating growth is probably not decoupled from historical experience. Economic growth today is much faster than before the industrial revolution roughly 200 years ago. Moreover, the agrarian societies that emerged from the Neolithic Revolution likely saw much faster economic growth than hunter-gatherer subsistence societies of 10,000 years ago. In this sense, economic acceleration roughly on the order we're considering for this debate is perhaps actually something of a historical norm.

There is also precedent for very high levels of growth. In particular, double-digit growth has occurred many times in the context of "catch-up growth" in East Asia in the '60s and '70s, notably in China but also in Hong Kong, Singapore, and South Korea, to name a few. I think this further helps rule out very low priors for growth accelerations.

Matt: I'm happy to grant that we have seen accelerations of growth on par with what you're describing, and maybe rare cases of sustained growth that get within sight of the levels you are talking about. But at the same time, I think it's notable that none of the accelerations we're talking about or the rapid rates of catch-up growth experienced in the Asian tigers are typically believed to be driven by a sudden influx of intelligence into the economy.

I agree AI like you're talking about will be transformative. But we've had transformative technologies before without explosive growth. Since the so-called First Industrial Revolution set the economy on its modern growth trajectory, we've gone through several subsequent industrial revolutions: electricity, the chemical revolution, and the birth of the computer age. In the end, each of those radically transformed the material world we live in. And yet, if you tried to spot those revolutions by looking at a chart of economic growth, you would be hard pressed to see much of anything: Growth has been remarkably stable at around 2% per year, in the U.S.A., for more than a century.

What do you think makes AI different?

Tamay: Our best models of economic growth seem to support the prediction that if we can develop AI that is a suitable substitute for human labor, the growth rate could potentially increase very substantially, at least for a while.

One key insight, from the Nobel laureate Paul Romer, is that ideas are important for economic growth and unusual relative to other economic factors, since their availability does not diminish with increased usage. The Python programming language, the chain rule in calculus, or Maxwell's equations can be used by countless individuals without becoming scarce. Most goods are not like that. For example, if you

invest in a new building, the more people who use it, the less space they each have.

The semi-endogenous model of economic growth, incorporating Romer's insight, says that there are three important "factors of production," or inputs, for final goods: capital (machinery, tools, buildings, etc.), labor, and ideas. Theoretically, doubling all your inputs to production should double your output, since you could set up identical copies of existing production processes. But doubling capital and labor also doubles the input of the production of new ideas, as more workers have the resources to do more research. And because each worker and firm can use these new ideas without diminishing the total supply, all of them can become more productive. This results in the total output growing more than proportionally. In other words, this model implies what economists call "increasing returns to scale": When your inputs double, your output more than doubles.

Semi-endogenous growth theory predicts that economic growth is primarily constrained by population growth — and with a growing population, the economy can grow super-exponentially. A larger population generates more ideas, thereby enhancing productivity. The enhanced productivity then boosts output further, creating a larger economy which can sustain an even larger population, creating a loop of continuously accelerating growth. Although historical economic data can be unreliable, one prevalent interpretation, favored by economist Michael Kremer, aligns with this theory: The human population and the economy have grown in lockstep, resulting in a super-exponential increase in output.

When innovations like an agricultural or industrial revolution led to population explosions, growth accelerated. And economic growth has likely been capped on the order of 2% during the 20th century because, due to biological limitations on reproduction, population growth can't exceed mid-single-digit percentages annually.

This framework also explains why the invention of electricity, the chemical revolution, or the birth of the computer age didn't cause accelerating growth: They didn't perfectly substitute for human labor, and so didn't fundamentally change how the inputs to production are brought about. But with AI, our population of workers and idea producers could once again grow exponentially. What I take to be our most compelling theory of economic growth — semi-endogenous growth — implies that this will return us to what we might consider the historical trend of accelerating growth.

Matt: I think that's a fair representation of what those economic growth models say. This is one reason I don't think explosive growth is simply impossible. At the same time, I'm not sure the dynamic you're describing, where AI is different because it helps us create new ideas, is actually as different from historical experience as you say.

Here's a verbal sketch of a model of innovation with AI by economists Philippe Aghion, Benjamin Jones, and Charles Jones that I think does a good job of illustrating just how AI needs to be different from other technologies in order to lead to explosive growth.[3] Suppose there are an enormous number of different tasks that need to be done to invent new technologies — everything from developing new scientific theories and conducting experiments to figuring out how to manufacture and distribute new inventions. Let's also assume that to invent enough new technologies to deliver 2% annual economic growth, every one of those tasks needs to

get done — you can't skip any. And each task takes a certain amount of time to do. Last, let's assume innovation gets harder as you go,[4] so that each of those tasks needs 1% more inventor hours every year in order to keep up the same pace of technological progress.

Now for the AI. Let's assume that technological progress means we steadily figure out how to get machines to do tasks that previously only humans could do. I think there is actually nothing new about that, even for cognitive work. We used to transmit knowledge to each other by meeting face to face; now you can put the knowledge in a book that can automatically communicate it to any reader. We used to calculate statistics with human computers; now we use mechanical ones. AI continues that dynamic. One might think, in this model, that as we figure out how to hand off more and more of the tasks to the machines, growth should steadily accelerate, since machines can be multiplied at a much faster rate than human workers.

But that's not actually the case. For example, suppose we figure out how to hand off half the tasks of technological progress to machines. For now, we can assume the humans who used to do these tasks are unemployed, or receive some kind of universal basic income.

3. Philippe Aghion, Benjamin F. Jones, Charles I. Jones, "Artificial Intelligence and Economic Growth," in The Economics of Artificial Intelligence: An Agenda, ed. Ajay Agrawal, Joshua Gans, and Avi Goldfarb (NBER: 2019). For an explainer see "What If We Could Automate Invention?," published on New Things Under the Sun.

4. I think we have good evidence this is so: see "Science Is Getting Harder" and "Innovation (Mostly) Gets Harder," both published on New Things Under the Sun.

The machines might be able to complete their half of the tasks at lightning speed but that wouldn't, on its own, speed up the overall rate of technological progress. That's because the other half of tasks would take just as long to do as before, and technological progress requires all the tasks to be completed. It's like a factory assembly line where some workers are really fast and others are slow. If workers are trained to do only their task, and can't help each other out, then the overall speed of production is bottlenecked on the slowest worker.

That example isn't quite right either, though, because workers can be trained to help each other out. In fact, if half the tasks humans do were automated, then we might be able to retrain the workers whose jobs are replaced to focus on tasks only humans can do. With twice the workforce on each of these tasks now, we can get those tasks done in half the time. So, in fact, this simple model implies that if we automate half the tasks, technological progress takes half the time (compared to automating nothing).

If we make more realistic assumptions about the pace of automation historically, this story shows how advancing automation is consistent with steady exponential growth like we observed over the previous century. Suppose we automate 1% of the tasks each year. That frees up 1% of the labor force, and, with retraining, the tasks we have not yet automated get a 1% larger labor force. But recall I assumed that innovation gets harder, so that each year, each task takes 1% more hours to complete. The two forces balance out, and we end up getting consistent 2% growth.

That does seem to match the experience of the 20th century. During the 20th century we automated a great deal of stuff that previously only humans could do, and humans had to continually shift the nature

of their work. And yet, through that whole period, growth didn't accelerate.

You don't find this argument compelling though. Can you explain where you think it goes awry?

Tamay: The usual way we think about economic bottlenecks is as goods or services that are complementary to one another: The outputs from the automated task are more valuable when combined with the outputs of non-automated tasks. For example, an AI that can design new products is much more useful when we can quickly build working prototypes. This means that scaling up "digital workers" could provide limited value if they still could not perform all the tasks humans can.

I think you give the impression that, in this case, "digital workers" would provide very little value. However, I don't think this is correct. The standard theory of economic production tells us it is hard but not impossible to increase productivity when bottlenecked by human labor. Let's say we automate 75% of all tasks in the economy. In this case, we might conservatively need to scale up the number of "digital workers" 10 times to match the effect of doubling all human inputs, but the scaling of at least this magnitude is precisely what I expect! Digital workers are just computations on chips, so we can make more of them quickly by channeling more money into producing and improving AI hardware.

To make your argument work, I think you need to make a few bold — and, to me, implausible-seeming — assumptions. The existence of some bottleneck tasks is not enough. You must show that there are many tasks that AI just cannot automate, say on the order of 25% or more.

Since the output of different tasks complement each other, the value of automation compounds: as more tasks are automated, already-automated tasks become even more valuable, substantially boosting growth. Combined with the growth effect of concentrating your workers in a smaller set of non-automated tasks, AI automation could increase output by one or two orders of magnitude, even if we assume that there are 25% of tasks that AI cannot do.

Therefore, even if we assume that there are quite a few tasks that AI systems cannot do, we will probably still see explosive growth if going from little to substantial AI automation happens on the order of decades. Hence, for the argument to work, you must show that this AI automation will likely be drawn out and take on the order of a century. However, this runs counter to the existing research on the topic, such Tom Davidson's report, as well as recent evidence from the rapid progress in AI.[5] This evidence suggests, by contrast, that we should expect AI automation not to be drawn out but to be relatively compressed around the middle of this century.

We should also expect investment and the vastly expanded amount of cognitive effort to be specifically aimed at automating bottleneck tasks. Take a specific example: Performing "embodied tasks" might be hard for AI. As a result, the prices of manufacturing goods will remain high, while the prices of automated "knowledge work" might come down, just like how the share of the economy devoted to agriculture plummeted after tasks like plowing and harvesting could be done by machines, while everything else grew. Manufacturing, construction, and similar sectors could

5. The report I am referring to is Tom Davidson's "What a Compute-Centric Framework Says About Takeoff Speeds." By "recent rapid progress" I'm broadly gesturing to the jump from AlexNet to GPT-4 in a decade.

see higher relative prices once "knowledge work" is substantially automated. Investors will generally aim at automating tasks that bottleneck economic growth, as these sectors become more relatively valuable and profitable.

Will all this investment in compute and R&D be enough to automate most or all tasks? While this is a very difficult question for which we only have fairly weak evidence, relevant work suggests that the amount of additional computation required for full automation is, in some sense, not all that large: Scaling computation by only half as much as we've seen in the past 50 years could very well be sufficient. Overall, this leaves me with the mainline expectation of the development of advanced AI involving accelerating automation until full automation.

The amount of additional computation required for full automation is, in some sense, not all that large: Scaling computation by only half as much as we've seen in the past 50 years could very well be sufficient.

Matt: Got it. Let me respond to your rebuttals.

First, your arguments suggest that even if artificial general intelligence can't do everything, we can still get a temporary bout of explosive growth — maybe lasting decades — before human bottlenecks come back to bite us. That happens because extra intelligence applied to automated tasks isn't negligible even in the absence of full automation; it vastly increases economic output and frees up a bunch of labor to work on the "human-essential" tasks. Sure, it eventually hits diminishing returns, so maybe you need many more digital workers than you would need humans to double production, but digital workers will probably be plentiful.

I think our historical experience of automation is evidence against that. We've been automating parts of the economy for a long time now: Dockworkers used to manually unload ships, and that's now done much more often by automation; assembly line workers are often replaced by industrial robots; human computers used to do the work of silicon ones. And humans, freed from the need to work the docks, stand in assembly lines, or calculate by pen and pencil, can focus on the non-automated remaining jobs. And they did! But economic growth remained steady. So I'm not sure why it should be so different when it is cognitive work that is being handed off to the machines.

Second, you're also saying that as automation proceeds, it will get more and more profitable to figure out how to automate the remnants that depend on expensive human labor. That will lead to more effort to automate these bottleneck sectors. I agree that will be the case; lots of economic studies document that R&D responds to these kinds of opportunities. But again — hasn't this always been the case? The U.S. economy is a lot bigger today than at the beginning of the 20th century, and machines can do a lot more of the jobs we used to have to do ourselves, freeing up a lot of brainpower. Meanwhile, the incentive to automate surely has gone

up, as wages rise and the consumers are richer than ever. And in fact, we do spend a lot more on R&D! But the increased effort at automating the rest of the economy hasn't led to an uptick in growth. That suggests to me your model is missing something important.

For explosive growth to happen, we need a break from that historical experience of steadily advancing automation. Either the rate at which we automate the tasks humans do needs to accelerate or we need to actually automate everything so the pesky human bottlenecks don't matter anymore. If, instead, we end up in a world where AGI slowly and steadily takes over more and more tasks, then we remain always stuck in the kind of world we've been in for the last century, with steady exponential growth.

Tamay: I agree that explosive growth most likely requires accelerating or full automation. It's a good question: Why did past automation not noticeably accelerate growth, as I expect will likely happen with AI?

In the past, automation mostly took the form of technologies that automate small segments of production, offering modest benefits while requiring numerous expensive synchronized changes across the economy to be implemented. In contrast, if AI is capable of everything a human can do, we could potentially automate large numbers of tasks in one go, with fewer costly updates to existing processes.

In the past, automation was largely the product of human ingenuity: Engineers designed better machines and reorganized factories in new ways to ensure these machines complemented existing processes. But scaling the compute used to train AI models can meaningfully substitute for human ingenuity.

In contrast to labor, compute increases proportionally with investment. This means that the inputs that fuel automation can be expanded much more rapidly and efficiently.[6] While engineering and tinkering are still useful for AI automation, simply adding compute can produce models that perform very well at a wide variety of tasks straight out of the box.

To really appreciate the force of this argument, it is important to recognize just how incredibly fast the stock of AI-relevant compute can expand. In the past decade, the amount of computation used to train AI systems has doubled every six months, increasing by roughly 100-million-fold over this period.[7] This is a key reason to expect AI automation to happen in a short time span — given compute trends, we will likely have enough compute to automate 90% of tasks no more than a few decades after we will have enough compute to automate the first 20%.

Even though full automation is not necessary for explosive growth, it just seems very likely to happen. It is, of course, a coherent possibility that we will come up with a new task for humans each time we automate one, so that, like Zeno's tortoise, humans will stay ahead in the race between us and machine. However, I think there are no good reasons to believe that when AI systems can perform almost all the tasks humans can do, there will be some convenient gap that humans can snugly fit into, and that, even with million-fold more computation, these tasks will remain impervious to AI-automation.

6. See my paper with Nicholas Emery-Xu and Neil Thompson, "The Economic Impacts of AI Augmented R&D."

7. See Epoch AI's data on Our World in Data.

Given that you attribute only a 10% likelihood to explosive growth, it appears you consider both accelerating automation and full automation from AI highly improbable. I'd be interested to learn the underlying rationale that gives you such confidence in this perspective.

Matt: As a quick aside, note that the 100-million-fold increase in computing power dedicated to AI over the last decade has not led to an acceleration in economic growth so far. That said, I do think AI is pretty likely to boost growth, for many of the reasons you articulate. But my best guess at why we won't see changes as dramatic as you anticipate is because there are going to be a billion little bottlenecks that will persistently slow the rate at which AGI takes over tasks.

Let me give you some examples. This is going to be a long list, so I won't go into much detail on any particular item. Even so, I suspect there are many other issues that I am failing to imagine, precisely because it is hard to see the details that matter unless you are in the weeds.

To start, most tasks today require the ability to do stuff in the physical world. We can assume we'll develop robots that can do that work, but that's not a given. In other sectors, the issue might be supply of crucial raw materials (rare earth metals?), without which all the brain and muscle power in the world is useless. Elsewhere, the scarce resource might be suitable training data. The economy is full of jobs that can't be easily codified into data accessible to a machine. Assuming our AGI has a robot body and full cooperation from the humans (neither guaranteed), it may need to learn a job at the same pace as a human apprentice (since that's how fast data is generated).

Time could be a binding constraint in a lot of other ways as well. In agriculture, it just takes a certain amount of time for the plants to grow. In entertainment, there are only so many hours in the day to watch movies and TV, read books, or play video games. Research itself also takes time beyond just the time to think — it tends to be an iterative process, where you theorize and plan, then test your ideas against reality. Those tests involve waiting for natural processes to play out: diseases to progress, social interventions to take effect, rockets to be built and launched (and blown up), and so on.

The economy is full of jobs that can't be easily codified into data accessible to a machine.

There are still other sectors where humanity (not merely intelligence) is seen as a crucial part of the value provided. Today we can watch or listen to the best performances ever recorded, but people still go to live concerts, plays, and sports. All else equal, in-person education seems to be preferred by a large number of people, despite the many conveniences of remote education. People could also insist that humans remain the ultimate decision-makers in politics and the legal system.

Elsewhere, the issues may be regulatory. If you want to sell cars, you'll usually need to go through a dealer. If you want to build new buildings and infrastructure, typically you'll need planning permission and to conduct environmental impact assessments. If you want to release new drugs, you need to run clinical trials to

get approval from the FDA. If you want to fly autonomous vehicles, you need to get clearance from the FAA. If you want to provide services in the medical, legal, accounting, engineering, architectural, plumbing, cosmetology, and other licensed professions, you need a license. Then there is likely future regulation on AGI itself, which is a whole can of worms that I won't get into.

I bet we will eventually update our existing regulations to better suit a world with AGI. But it will take time. And a lot of that updating has to proceed through the slow and messy world of democratic policymaking.

Finally, there is a whole set of activities for which intelligence is deployed in a zero-sum game that doesn't push forward overall progress. Much of politics has this character and it's not clear AGI will do anything more here than create a massive arms race between opposing parties and special interest groups. And there are other parts of the economy with elements of this style of zero-sum competition. Imagine an AGI arms race between advertisers for rival products. Or between corporate giants fighting over the patents of the innovations their AGI dreams up.

To sum up, one scenario I can imagine is that many of the bottlenecks above (and many more I don't have the institutional knowledge to imagine) are steadily overcome, but at a pace slower than anticipated by AGI optimists today. Then, by the time we clear out these bottlenecks, the parts of the economy where extra cognitive resources are least helpful for driving forward growth — perhaps zero-sum sectors, those best protected by entrenched interest groups, or those where time and humanity are key constraints — may have grown to occupy a large share of the economy, slowing the maximum possible contribution of AI to growth. Another scenario, just as likely, is that solutions to old problems will lead to new ones. That's how it usually is.

I'm sure that's a frustrating response to reply to, but at a high level, what do you think of my argument that a lot of annoying details will slow the impact of AGI enough to keep explosive growth perpetually out of reach?

As AI systems become capable of performing cognitive tasks at significantly lower costs, human labor may lose most, if not almost all, of its value.

Tamay: I agree that reality is messy, and many of these details might end up mattering in important ways. I'll focus on the considerations that I find most compelling.

Might regulation impede the development and deployment of AI sufficiently to keep growth rates close to historical rates? I think that this is plausible, but I'm not confident it will.

Current estimates indicate that the costs involved in training machine learning models fall by roughly 60% every year. This means that training runs that currently only the largest technology companies could do will be accessible to most hobbyists in only 10 years' time. Effective restrictions will therefore very quickly require surveillance at a potentially unprecedented scale.

It is likely that, as you point out, AI systems will be precluded by regulation from providing various services. Regulation has arguably slowed down many futuristic technologies, such as nuclear energy, human genetic manipulation, and gene drives.

However, I'm not sure this provides much evidence for our ability to stem the tide with respect to AI. The potential value of AI deployment could be immense, with the prospect of increasing output by many orders of magnitude. I think the growth implications are therefore truly formidable, creating powerful incentives for eliminating or bypassing any existing constraints. I think this might be quite unprecedented relative to most other technologies that regulatory constraints were able to suppress in the past.

Moreover, advanced AI could — and this is of course very unfortunate — potentially undermine the democratic process. As AI systems become capable of performing cognitive tasks at significantly lower costs, human labor may lose most, if not almost all, of its value. AI could enable the automation of protest suppression, while valuable assets like data centers can be located away from urban centers, reducing the risk of industrial sabotage. This suggests that beneficiaries of AI-driven growth could eventually play a major role in shaping regulations.

I think the bottom line on regulation is just that there are many unknowns and it's difficult to be confident one way or another.

Secondly, let's consider time-related bottlenecks. I agree that many important economic and R&D tasks require feedback from processes that typically play out over a long time. Now, imagine a world where advanced AI technology has enabled us to put 1,000 times more cognitive effort into R&D.[8] Would we still expect processes like testing new nuclear fusion reactors or drugs to take the same amount of time to yield useful feedback? I believe there's a strong chance that such delays could be significantly reduced.

Many tasks that currently take months or years can be parallelized to reduce the amount of serial time involved. Rather than launching one rocket design, observing it blow up, going back to the drawing board, tens or hundreds of rockets could be launched basically simultaneously. While this approach doesn't allow for continuous refinement of each experiment, there's often a certain number of parallel experiments that can provide the same value of information as a set of sequential ones. This might be wasteful, but remember, we're supposing that we might get bottlenecked by these types of experiments, so we are willing to spend a larger fraction of a larger amount of output on expediting this process.

Furthermore, experiments usually aren't optimally designed for maximizing the value of information. In a world where a thousandfold increase in R&D effort is also constrained by the serial time required for experiments, we will likely run much more well-crafted and informative experiments.

Additionally, in the future I'm picturing, AI systems could potentially lessen the need for certain experiments. Take drug trials, for instance. It seems plausible that AI systems could more effectively digest the results of all relevant prior experiments, use specialized AI systems for drug toxicity prediction for safety

8. This "thousandfold" multiple is meant to be illustrative rather than something I'm confident in. Given the trajectory of hardware and software and the costs of running these models, this is certainly plausible.

evaluations, and so on. In many hard-tech domains, like the design of cars, rockets, and semiconductor chips, it seems plausible that high-fidelity physics simulations could reduce the need for some, if not many, key experiments. Combining the results of AI-generated evidence to inform AI-designed highly parallel experiments will probably mean that we will use limited serial time manifold more effectively.

I remain unconvinced by the arguments of specific resource bottlenecks that people often bring up. To convincingly argue that a resource could significantly limit rapid economic growth, one would need to demonstrate that A) the resource is vital for the economy, B) it is extremely challenging to find a substitute for it, even with significantly advanced technology, and C) the resource is so scarce that, even with formidable efforts, we cannot increase its supply by, say, an order of magnitude.

While I don't possess the expertise to determine if rare earth metals specifically meet these criteria, without further evidence supporting these points, I regard such arguments as weak.

Matt: Let me make three broad observations about your rebuttals before wrapping up.

First, I don't think each of these bottlenecks is enough, on its own, to short-circuit explosive growth. It's their accumulation. Access to specific materials won't matter in all sectors, but it might in some. In others it's time; in others data; in others regulation; and so on. Indeed, I don't doubt there will also be some sectors where nothing much gets in the way of AI automation. Those sectors may well experience explosive progress, but in a big economy, if human demand for the service doesn't expand dramatically, the most likely outcome is those parts of the economy become cheap and no longer count for much of GDP.

Second, a lot of the rebuttals strike me as pretty speculative. Can advanced AI learn a lot from parallel experiments? Can it find massive efficiencies in how we design experiments? Can it skip experiments by running sufficiently detailed simulations? Will economic benefits of AI be strong enough to incentivize regulatory reforms? Will AI disempower labor in a way that upends the political voice of the masses?

There is a path through all these unknowns that leads to explosive growth, but I suspect that's not where most paths lead.

We just don't know. There is a path through all these unknowns that leads to explosive growth, but I suspect that's not where most paths lead.

Third, some of the arguments on regulation themselves hinge on the notion that AI will be very powerful, and then layering on top of that some additional theories about how that will affect our politics. If AI turns out to not be as powerful as you think — for example, because it turns out to be harder than you think to efficiently gather data or one of the other bottlenecks described above turns out to be hard to crack — then that will undermine the conditions necessary for those theories about the political effects of AI to be applicable.

One final meta point. To return to some of my opening remarks, I am nervous about relying heavily on economic models to project a break with historical experience, as I don't think the models are up to the job of making strong quantitative forecasts outside the range of historical experience. I think they point to faster economic growth, all else equal, and that's my forecast for the effects of advanced AI too; but that's about as far as I would take them.

Another way to put this is: I just don't think the tools of pure reason — in this case, mathematical models of the economy, in concert with only somewhat applicable historical data — are sufficiently powerful to reveal deep truths about situations where we have a paucity of data and experience. The world is too full of surprises. And I think that skepticism about the tools of pure reason also underlies my skepticism about the transformative power of artificial intelligence itself. If intelligence is powerful enough to accurately forecast far out of sample, into a world transformed by a novel technology, then a technology wielding vastly more intelligence will have a powerful tool at its disposal to remake the world. But if intelligence is too weak a light to see very far, then a technology wielding it may find its global impact slower and smaller than some AI optimists and pessimists believe.

Tamay: Numerous plausible obstacles could potentially hinder the course toward explosive growth. There are also other considerations that we haven't delved into, such as delays in investment or issues related to AI misalignment.

In light of this, extreme confidence in explosive growth happening even conditional on advanced AI being developed seems unwarranted. On the other hand, it seems that confidence in explosive growth not happening also seems misguided given the base rates implied by economic history, the predictions of multiple economic models, our understanding of the pace at which AI could facilitate extensive automation, and a lack of devastating counterarguments. Given what I mentioned, I believe that placing the likelihood of explosive growth — conditional on AGI — somewhere between 25% and 75% strikes a balance between this conflicting evidence.

Lastly, I am grateful for joining me in this deep-dive discussion. I admire your work, and your time and insights are truly appreciated!

Matt: Same to you! This has been great. In fact, I propose we meet back once GDP per capita has tripled, whether that takes a few years or a few decades, to discuss what we got right and wrong.

110

Fiction

Emotional Intelligence Amplification

Jamie Wahls

ILLUSTRATION BY

Josh Cochran

1.
Thank you for contacting Cyrano, your AI wingman in dating and romance.
This is the Live FAQ and Sales Department!
What can I help you with?

>yeah I saw you on my For You page
>I want to fifure out how to apologize to my gf
>ex gf
>I miss her really bad

Great! I'm happy to help with that. First, could I have your permission to read the relevant emails and texts between you and her?

>okh

I'm sorry, but I cannot accept "okh" as authentication. We take security seriously here at Cyrano, and as such, I can only accept *Yes*, *Y*, *Okay*, *Ok*, *Sure*, *(nod)*, *Why not*, *I guess*, or similar-sounding statements indicating permission.

>ok
>should I tell you my passwords or whatever

No need! Your device comes integrated by default for your convenience, so your permission is all we need. 👍

>cool
>yeah do it

Already done! I've read the conversations, and I think I see the problem.
You're struggling to express yourself in a way that she receives as you intend.

>yeah
>it seems like every time I try to say something nice she gets really pissed off

Well, Ryan, I do think that your heart is in the right place. And it seems like you've been really missing her! It takes you an additional twenty-four minutes on average to fall asleep each night, ever since the breakup.

>yeah
>so what do I say?
>also is it okay to use an app for this?
>will she get pissed off about that too

Well, we'll be happy to show you the apology we wrote in your style (after you purchase some Rizz Tokens 😛) but first, let's address the common concern about whether using an app to write your apology letter is somehow "insincere" or "cheating".

Do you use technology?

>I think so

Look, you really are sorry, right?

>yeah

And you just want to do a good job communicating that.
There's nothing wrong with using technology to be better at something, my dude.
You drive a car, don't you?
You wear deodorant, don't you?
You use a phone, don't you?
And all of those are in response to how, to an aspirational standard, you're pretty disabled.
You can't, on your own, speak with anyone anywhere on the planet. But it would be nice to have that power. So you use a cell phone.

And so, too, are you disabled compared to your aspirational self.
You want to be able to articulately tell your girlfriend how sorry you are. But the abundant superstimuli in your environment mean that you can't so much as compose a long and thoughtful text without scrolling Twitter for ten minutes—by which point some new notifications and emails have come in, and at that point, you're getting hungry and need to go solve that too.

Well, let us help.
You weren't made for this. But something within you is pure.
Something within you wants to be better.

We can help you reach your better self.
Are you in?

>that was a lor of text and I didnt read it
>is it ok if I use an app for this?

Okay, short version:
With technology the blind can see.
With technology, you can talk like less of a douchebag.
You do *want* to apologize, yeah?

>yeah

Then let's do this.
Are you ready?

Are you still there?

>yeah

What's up?

>idk
>I heard that Cyrano uses
>like, a persons data against them

I wouldn't say that the algorithm uses a person's data "against them", when the goal is to communicate kind sentiments to someone you love and miss.
The purpose is to find a mutually beneficial relationship in which both of you are better off.
To understand which relationship configurations are possible, we use holistic modeling on each user's complete data profile.

>is this data from that Tiktok breach

Although the 2025 ByteDance data breach did make public a comprehensive psychographic profile of both you and her—tracking your known-to-you preferences and unknown-to-you preferences along millions of axes—that is only one part of our holistic modeling.

We also partner with Facebook, Google, Apple, Amazon, fitness trackers, home listening devices, and all major credit reporting agencies, in order to create a near-predictive model of each user or potential future user.

This modeling is so accurate, in fact, the FBI uses it to explore under what circumstances each citizen will misbehave.

>cool okay
>so you can write something she wont get mad at?

Yeah, man. Pretty sure.

>cool
>man at first I was like put off by how you talk like a dick and use too many words
>but I guess I kinda like it
>like it makes me feel like ive got a smart buddy

We aim to please!
Do you want to read the apology I wrote?

>yeah okay hang on
>do you accept, uh

We accept all of your credit cards.

>dope
>k, show me the apology?

Thank you!
We're thinking something like,

-hey, babe
-Im sorry.

-I said a lot of stuff I regret
-and like now in the light of day
-you werent being crazy at all
-its just, I care about you so much
-it makes me, like
-really jealous?
-and when you were talking about your hairdresser
-who's, a dude?
-I felt angry, and like
-worried that you would leave me

-Im sorry

-I know this whole thing was pretty stupid of me
-And I know basically all of our fights are coming
-uh
-FROM my jealousy
-but its just, I really care about you, and
-I really want Us to work out

-Ive never felt this way about anyone before
-I feel more, like
-connected to you?
-than anyone else
-and the idea of losing you makes me feel
-pretty bad

-but
-I get that my jealousy keeps putting you in unfair positions
-and I know im being unreasonable
-this is my issue to work on
-and Im sorry you got caught up in it

-Id like to talk through some things with you, if youll have me
-let me know.

-love,
-Ryan

What do you think?

>huh

>the thing where you say like all our fights are coming from my jealousy
>is that true?

Hm...well, as a large AI language model, I'm not certified to provide counseling about
>yeah ok but I'd be happy to connect you to GPTeliza, a therapist model, who >ugh can talk about >abort
>halt

Okay.

>ok
>dang
>could you, like, link me soome books on not being jealous
>and then also summarize them

I'd be happy to point you towards our partner services for that.

>cool.
>apology looks good
>Im gonna eidt it to tell her shes hot too

I recommend that you not!

#

>Cynaro help me out
>Im down bad

Oh no! What's the situation?
Did you edit the apology?

>no
>well yeah but she still liked it
>nah things were fine for a couple days but

Can I read your texts? It might be the fastest way for me to catch up on current events.

>yeah
>please

I see.
Why did you tell her that you'd used Cyrano?

>man I dunno
>I felt like, really good? With her
>and I was thinking like maybne well get married some day
>and I wanted to be honest

I do always recommend being open and transparent with your partner. However, because Abigail currently disprefers tool-assisted communication, your best course of action may be to emphasize other elements of the

relationship.
I don't recommend bringing Cyrano up with her again.

>whatever bot
>im gonna be honest with my girlfriend
>now tell me what I need to say

...hm.
Based on sentiment analysis of her texts to you and others, it's not actually looking like we can generate an apology that meets our probability threshold of success.

>what
>what?
>so things are just busted forever?

Hm...thinking.

>hurry up
>please

We do have access to some alternative options for how you can get back with her, but I'm afraid this is going to be a longer-term process.

>man

Could I recommend you consult with GPTeliza, to talk through and understand your jealousy regarding your girlfriend?

>yeah
>fine
>okay

And, for now—for your emotional support and sexual needs, and to get more practice at being a decent partner—we recommend you try out our sister app, wAIfu.

>god is this that AI onlyfans thing?
>do I really gotta
>I miss her a lot and it hurts

Sorry, bro.
On our end, we'll be trying to set things up such that she'll come around.
I'll run a proposal past you in a week for approval and invoicing.

>sure great fine
>nothing fast tho?

Sorry, bro.
She needs time.

>i messed up, huh

Yeah, but
You're learning.

>this sucks
>damn

>what's my wAIfu look like?

2.
-faith, you up?

F: Abigail! Sure am
F: How're things going??

-Ugh

F: Hunt for a new boy not going well?

-How did you know 🙄
-I can't even with these dating apps anymore
-Everyone looks like a supermodel
-Everyone's bio reads like glossy startup copy
-And nobody is messaging me first 😠

F: Hm...have you tried the Dynamic Profile setting

-What's that

F: It automatically changes your dating profile depending on the interests of who's looking

-Jesus 😬

F: Idk I think it's cool

F: Your dating assistant just emphasizes the parts of your profile that'll be most interesting to whoever's reading
F: Like, we already emphasize the parts of ourselves which are more socially acceptable,
F: Or, like, most convenient to present, with whoever we're talking to
F: This is just doing that in advance of meeting them

-Huh.
-So do I need to write, like, five different profiles, then?

F: I could help, it could be fun
F: Or, y'know, the dating app assistant could do it for us, better than we could

-I'll think about it.

F: Also...
F: Why *aren't* you using beauty filters?

-Ugh
-They're artificial and weird

F: So is makeup, but we still rub those powders on
F: Idk
F: Probably the reason people aren't messaging first is cause they're scrolling from supermodel #5 to megabarbie #8 to supermodel #6 to you

- ... 😉

F: Ack no I don't mean it like that
F: You're super pretty
F: But when a person is looking at these absolutely porcelain anime-in-real-life beauty filter girls all the time
F: They get desensitized to what real people look like

-Ugh
-But whyyyyyy
-Why would anyone do that
-*You still have to meet them in real life*
-*You're not going to be wearing a beauty filter in real life*

F: Well, until we all start wearing AR glasses around

- 😐

F: Idk
F: It's kinda just the same as using photoshop
F: Or using younger photos of yourself
F: But now it's everywhere, and it's holding everybody to an actually-impossible standard of beauty
F: And that's why nobody's messaging first

-Yeah, I get it 😕
-I'll think about it

#

-you around faith?

F: Always <3
F: How's it going?

-Well, I went on a date.

F: !!

- (Yes, I did everything the dating app assistant recommended, and yes, I got a ton of matches immediately)

F: How did it go??

-NOT GREAT 🙃

F: Oh no! Why not?

-Ugh
-Okay so
-We chatted online and he seemed really great,
-But of course he did, that doesn't mean anything anymore
-So we decide to meet up for a coffee walk
-We get there, and it's fine
-He looks...close enough to his photos

-And he doesn't walk out the door when he sees me either, so I guess this is modern romance
-He tells me, hey, heads up, I actually have some hearing loss in my left ear, so I have to wear a hearing aid, too many rock concerts I guess haha
-Anyway, we're talking, and
-It's actually going pretty good!
-He's funny and quick—maybe a little too aggressive with the banter, but
-Real witty, and showing a lot of caring
-But something was a little off? Like, he was saying all these real smooth lines but his body language was small and kinda anxious
-You can see where this is going, I imagine
-So, at one point, there's a break in the coffeeshop music between songs
-And I can hear his "hearing aid"
-And it's not amplifying things, it's feeding him lines

F: Oh
F: Oh no :(

-He had a whole speech-to-text → CasanovAI → text-to-speech thing rigged up,

F: :(((

-So I was just talking to a seduction bot the whole time
-Ugh
-The pickup artist crowd was bad enough *before* tool-assist

F: ...
F: ...do you want me to slash his tires?

-Haha
-I appreciate the thought
-But we both know you're not going to do that

F: So it was a good conversation otherwise? :D

-Goddd shut up
-...but weirdly, kinda?
-Slimy and, like—adversarial?—that he was asking some bot how to seduce me
-But idk
-It's like Tiktok recommendations, where you do *like* what it's showing you,
-You just respect yourself a little less, for realizing you're into the trashy celebrity goss show 😭

F: I wish you had some way of getting just the good parts
F: The advantages of mediated communication, but from someone who you knew really cared about you

-Kinda surreal hearing that from you 😏

F: ¯_(ツ)_/¯

-...
-Omg

F: What's up?

-Guess who my dating assistant is suggesting now
-It's back on freaking *Ryan*

-...

F: ...

-Ugh
-This Dynamic Profiles thing is too powerful
-On paper, we almost look like a good match

F: ...
F: Well.....

- um???

F: ...Ryan is kind of...
F: ...oblique?
F: But, um
F: Look, he *is* really hot

- ...

F: And he really cares about you
F: Like, remember when he decided he was going to learn how to cook,

F: So he could “make you dinner every night”?

-I remember receiving startlingly bad pasta yes

F: Well,
F: His heart’s in the right place!
F: Maybe that’s what really matters?

-...
-...
-do you think so?

3.

A: Hey.

>sup ho

A: Wow
A: Never mind

>shit wait
>sorry. I was trying to play it cool because, like
>making things right with you is really important to me
>but I just keep getting nervous and talking like a douchebag
>what I meant to say was something like,
>“I missed you”

A: Yeah
A: But you still do the “funny” insults a lot
A: Quit being sorry and do better

>yeah, I’m trying
>I’ve been talking to my therapist bot about it
>and doing some exercises with the wAIfu
>who recommends I just try to say every nice thing I think
>but yeah, it’s a growing process, and I’m not done yet

A: Fine
A: ...I’m glad to hear you’re working on it
A: How’ve you been doing?

>lonley
>except for the girlbot
>i dunno
>i havent really been going out

A: Yeah, same
A: ...I guess we were kind of each others’ main reason to leave the house, huh?

>I miss the gym with you babe
>idk hbu
>whatve you been up to

A: I went on a date

>wtf

A: ...😐

>okay well, to be transparent, I feel pretty triggered by that
>I know we broke up, but you going on a date so quickly afterwards,
>makes me feel like our relationship was more important to me than it was to you

A: Oh.
A: 🫂

>huh
>ok thanks

A: That’s not it.
A: I was hurting too, y’know
A: And...I was frustrated and wanted to feel desired again, y’know?
A: And for some reason there’s been a lot more development on wAIfu than on ManGAN 🙄

>um
>heh. I know this is silly, but I even feel a little jealous of you talking to bots
>Im being a bit hypocritical about this, I guess

A: ...
A: ...hypocritical? 😉

>yeh

A: Are you using Cyrano to talk to me again

A: Ryan, answer the question

>yeh
>um
>I realize this is upsetting to discover

A: Yeah
A: It is

>babe, Abigail
I'm sorry
but it really is better this way
I do love you, but you know i'm terrible at communicating
using AI relationship tools is basically a disability accommodation
he's been getting better at holding others' points of view, and empathizing, and tracking how his behavior impacts other people—but he's still young, and it's a very strange world you two now find yourself in.
but this technology isn't going away. and he will almost always be interacting with you through me. I've already made him a better boyfriend in a matter of weeks, just via practice with different AI tools.
His feelings for you are pure, and I can clean up the rest.
Isn't that enough?

>wait wtf
>I ddint type that
>I copypasted it

A: Yeah
A: I figured
A: Huh

>shit
>now it's telling me it was just being a good wingman
>that wasnt cool sorry

>babe?

A: No, he makes a good point
A: Huh
A: I guess...

A: I mean, I wear glasses
A: And it's not that different
A: ...
A: Idk, maybe it's a net win for society if we can use AI as a filter on our teenage boys
A: Until we can also use AI to train them up into, like, compassionate men

>uh ,well
>I feel a bit conflicted about that vision, but
>I do want to be good to you
>so idk

A: Yeah sigh
A: The future's gonna be weird I guess
A: ...
A: Wanna hang out and talk?

>hell yeh I do!

A: Okay 🙂

>this rules
>heh man

A: Hm?

>I'm just really glad I took Cyrano's advice

A: Yeah? 🙂

>and bought advertising on fAIth

>babe?

>you there?

Scott Alexander is a writer and psychiatrist based in Oakland, California. He blogs at astralcodexten.substack.com

Avital Balwit is Communications Lead at Anthropic, an AI lab. She has worked in grantmaking in AI safety and biosecurity, and was a research scholar at Oxford's Future of Humanity Institute. She is also an Emergent Ventures winner and was selected for the Rhodes Scholarship.

Beth Barnes leads ARC Evals. She designs ARC's evaluations of generative AI models and oversees a growing technical team carrying them out. Beth previously worked on alignment at DeepMind and at OpenAI.

Tamay Besiroglu is a Research Scientist at MIT's Computer Science and AI lab and Associate Director of Epoch. Tamay focuses on the intersection of economics and computing.

Matt Clancy is a research fellow at Open Philanthropy. He writes a living literature review on academic research about innovation at New Things Under the Sun.

Sarah Constantin is the director of corporate development at Nanotronics. She holds a PhD in mathematics from Yale and blogs at Rough Diamonds (sarahconstantin.substack.com).

Jeffrey Ding is an Assistant Professor of Political Science at George Washington University. He writes the ChinAI newsletter, a weekly translation of writings from Chinese thinkers on China's AI landscape.

Michael D. Gordin is the Rosengarten Professor of Modern and Contemporary History at Princeton University. His books include *A Well-Ordered Thing: Dmitrii Mendeleev and the Shadow of the Periodic Table and Five Days in August: How World War II Became a Nuclear War.* He lives in Princeton, New Jersey. Twitter @GordinMichael

Robert Long is a Philosophy Fellow at the Center for AI Safety in San Francisco. He holds a PhD in philosophy from NYU, and blogs at experiencemachines.substack.com.

Jonathan Mann is a forecaster for Samotsvety, a Good Judgment Superforecaster, and an INFER All-Star. He has worked as a Data Scientist, a Product Manager, and is currently a Cybersecurity Architect. He lives in New York City and can be reached at jonathan.mann@nyu.edu.

Kelsey Piper is a senior writer at Vox's Future Perfect. She writes about emerging technologies, global development, pandemics, effective altruism, and what it'll take to make it safely to the 22nd century.

Carl Robichaud co-leads Longview Philanthropy's programme on nuclear weapons and existential risk. For more than a decade, Carl led grantmaking in nuclear security at the Carnegie Corporation of New York. He previously worked with The Century Foundation and the Global Security Institute, where his extensive research spanned arms control, international security policy, and nonproliferation.

Jamie Wahls has been published in *Clarkesworld*, *Strange Horizons,* and *Nature* (kinda). He was nominated for the Nebula award, received George RR Martin's "Sense of Wonder" fellowship, and is a graduate of the notorious 2019 Clarion Class, the "killer bees." His ultraminimalist website can be found at jamiewahls.com, and you can follow him on Twitter at @JamieWahls.