Preparing for Launch

Why shaping AI progress matters, and how to go about it | **Tim Fist, Tao Burga, & Tim Hwang**

Preparing for Launch

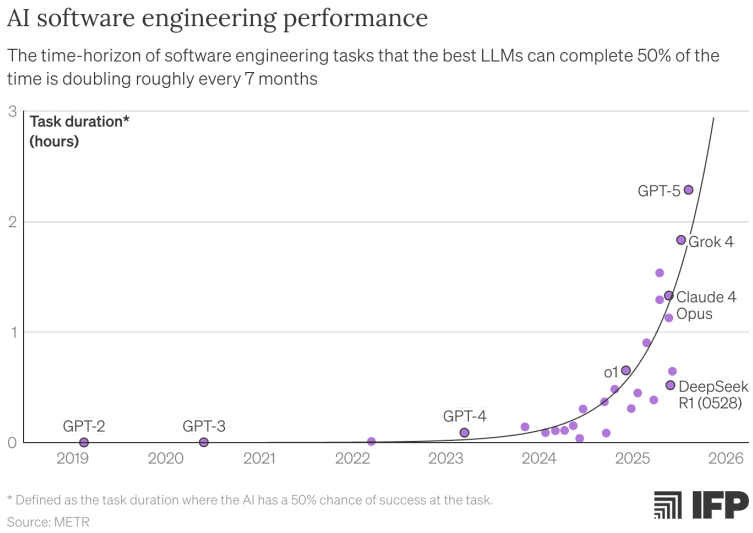
Why shaping AI progress matters, and how to go about it | Tim Fist, Tao Burga, & Tim Hwang

---

This essay is the foreword to The Launch Sequence, a collection of concrete, ambitious ideas to accelerate AI for science and security.

* * *

If we zoom out from the debate about the specific capabilities of each new AI model, the historical shape of AI progress is clear: AI capabilities are compounding toward something transformative. Even while today's agents still fumble long chains of actions, their human-indexed performance across coding, math, tool use, and scientific analysis has been rising at an exponential rate. If more domains continue to fall to these trends, the next decade will see AI reshape the economy, science, and the foundations of national power.



https://datawrapper.dwcdn.net/HVJGb/2/

But technological trajectories aren't fate. AI doesn't automatically solve the most important problems first, and it won't neutralize the new risks it creates by default.

Many proposals targeted at "maximizing AI's benefits" while "minimizing its risks" are poorly targeted or sorely lacking in ambition. In this essay, we lay out the basic case for the United States to proactively shape AI progress by accelerating the development of both beneficial and defensive technologies:

- AI progress is path-dependent: The sequencing of AI progress matters — *where* and in *what order* new capabilities are developed may be just as important as *which* new capabilities are developed.
- Given its position in the AI supply chain, and as the world's most powerful democracy, the United States has the responsibility to shape AI development towards a path that enables — rather than smothers — human flourishing.

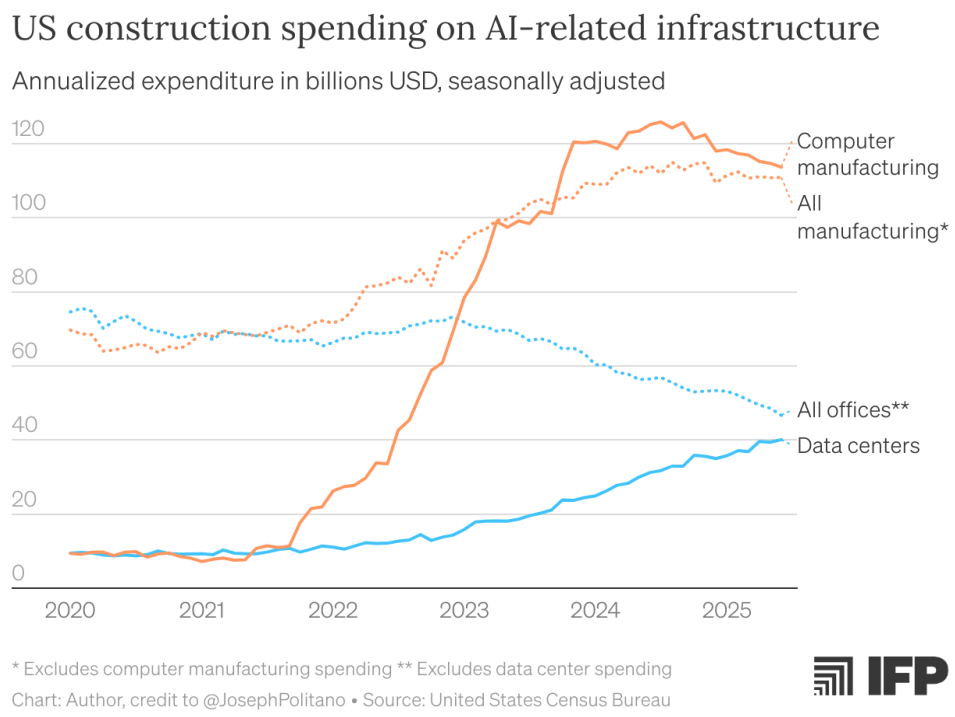
This proactive shaping is not without precedent. From nuclear fission to spaceflight to mRNA, the US has repeatedly changed the trajectory of emerging technologies. In the age of AI, we argue for four guiding principles:

1. We should take advantage of the "jagged frontier" of AI capabilities
2. We shouldn't neglect the costs of stalled progress
3. We should redesign how many of our scientific institutions work
4. We should adapt to deep uncertainty while working to reduce it

These principles motivate *The Launch Sequence*: a collection of concrete, ambitious projects to accelerate science and strengthen security on timelines that matter.

Why shape AI progress?

American companies are making a huge bet on AI. About two years ago, construction spending related to computer chips overtook all other manufacturing construction spending. And construction spending on data centers to train and deploy AI models will likely soon overtake spending on offices for human workers



https://datawrapper.dwcdn.net/bdVqi/3/

These investments could help solve some of humanity's most pressing problems, from finding treatments for crippling diseases through accelerated drug discovery to delivering a new scientific revolution through the discovery of new materials and tools.[1] More broadly, the artificial intelligence of the future could catalyze a new golden age of growth and abundance. But technological progress is not an otherworldly force that delivers solutions to the most important problems first. To fully reap the benefits of AI progress, we must solve two broad problems.

1. The benefits of AI progress may not come quickly enough

Many benefits of AI progress may not come quickly (or at all) by default, given existing commercial incentives. This may be because a particular application would create public goods, which markets tend to undersupply, or because a research direction is high-risk and requires large upfront investments. For example, markets alone may not prioritize AI that automates the replication of scientific

[1] Charles Yang's essay in this collection, "Scaling Materials Discovery with Self-Driving Labs," lays out how to close the gap between AI-guided material design and real-world validation.

studies,[2] or investigate new medical treatments that can't be patented, or conduct basic research in neuroscience with no immediate commercial applications.[3]

Despite a recent wave of new public,[4] private,[5] and nonprofit[6] projects focused on AI for science, we are still not close to fully harnessing new AI capabilities to accelerate solutions to the world's most important problems. Furthermore, too little effort is focused on solving the numerous structural barriers to realizing the benefits of AI-enabled scientific discovery. The deployment of new clean energy technologies will likely face a near-endless series of vetoes at the hands of conservation groups. New drugs will likely be stalled for years in the FDA's needlessly onerous approval process.[7] Thanks to a mix of poor incentives and antiquated government computer systems, many useful scientific datasets aren't accessible for AI training. And public research is hampered by a broken funding

---

[2] Abel Brodeur and Bruno Barbarioli's essay, "The Verification Engine," explains how to build automated replication infrastructure for better, faster science.
[3] Adam Marblestone and Andrew Payne's essay, "Mapping the Brain for Alignment," details how mapping the mammalian brain's connectome could help solve fundamental problems in neuroscience, psychology, and AI robustness.
[4] NSF, "NSF announces $100 million investment in National Artificial Intelligence Research Institutes awards to secure American leadership in AI," July 29, 2025; DOE, "Department of Energy Announces $68 Million in Funding for Artificial Intelligence for Scientific Research," September 5, 2024; Xiao Wang et al., "ORBIT: Oak Ridge Base Foundation Model for Earth System Predictability," 2024.
[5] Google DeepMind AlphaFold Team and Isomorphic Labs, "AlphaFold 3 predicts the structure and interactions of all of life's molecules," May 8, 2024; The AlphaEarth Foundations team, "AlphaEarth Foundations helps map our planet in unprecedented detail," Google DeepMind, July 30, 2025; Amil Merchant and Ekin Dogus Cubuk, "Millions of new materials discovered with deep learning," Google DeepMind, November 29 2023; AlphaProof and AlphaGeometry teams, "AI achieves silver-medal standard solving International Mathematical Olympiad problems" Google DeepMind, July 25, 2024; EvolutionaryScale, "ESM3: Simulating 500 million years of evolution with a language model," June 25, 2024; Viren Jain, "Ten years of neuroscience at Google yields maps of human brain," Google Research, May 2, 2024; Guy Lutsker et al., "From Glucose Patterns to Health Outcomes: A Generalizable Foundation Model for Continuous Glucose Monitor Data Analysis," August 20, 2024; Sally Beatty, "From sea to sky: Microsoft's Aurora AI foundation model goes beyond weather forecasting," Microsoft, May 21, 2025.
[6] Arc Institute, "AI can now model and design the genetic code for all domains of life with Evo 2," February 19, 2025; futurehouse.org, "Automating scientific discovery," FutureHouse; Jinxi Xiang et al., "A vision–language foundation model for precision oncology," 2025; Bezos Earth Fund, "Bezos Earth Fund launches $100 million AI for Climate and Nature Grand Challenge," April 16, 2024; Helena Kudiabor, "Virtual lab powered by 'AI scientists' super-charges biomedical research," December 4, 2024.
[7] Ruxandra Teslo's essay, "Biotech's Lost Archive," details how to fuel AI and help small biotech innovators by unlocking the FDA's knowledge of biotech failures.

model, in which principal investigators spend almost half their time on grant-related paperwork.

2. AI progress may bring risks that industry is poorly incentivized to solve

Advanced coding agents used throughout the economy to vastly increase productivity could also be put to work, day and night, to find and exploit security vulnerabilities in critical infrastructure. AI systems that can broadly accelerate the pace of medical research could also help engineer biological weapons. Leading AI labs have some incentives to prevent the misuse of their models, but the offense-defense balance of emerging AI capabilities in areas like cyber and bio is uncertain. There's no iron law of computer science or economics that says defensive capabilities must grow in tandem with offensive capabilities. In the worst case, private incentives to adequately invest in preventing misuse could be dwarfed by the scale of the risks new AI technologies impose on the public.

Proposals from the AI safety community often attract criticism for focusing on solutions that rely on brittle, top-down control, such as a licensing regime for all models above a threshold of training compute. But despite the validity of these critiques, the problem still remains: AI misuse and misalignment could well cause real harm in the near future, and technical research aimed at solving these problems remains a niche field — around 2% of AI papers published, with roughly $100 million per year in funding.[8] Moreover, thanks partly to an instinct towards nonproliferation, AI safety researchers have devoted insufficient attention to solutions that assume that dangerous AI capabilities *will* rapidly diffuse.[9] In the face of superintelligence both widely available and too cheap to meter, too few projects wield AI to build technologies that asymmetrically benefit defense over offense.

* * *

[8] Evan Miyazono's essay, "Preventing AI Sleeper Agents," outlines how a large-scale DOD-led red- and blue-teaming effort could help ensure American AI models are robust and reliable.

[9] Nora Ammann and David 'davidad' Dalrymple's essay, "Faster AI Diffusion Through Hardware-Based Verification," describes how to use privacy-preserving verification in the AI hardware stack to build trust and limit misuse.

We need proactive policymaking and public investments that address these problems. Phrases along the lines of "maximize AI's benefits, minimize its risks" have already become cliché, but proposals and projects targeted at either goal are often poorly targeted, or sorely lacking in ambition.

At IFP, we're applying ideas from progress studies, defensive accelerationism, and differential technology development to help genuinely ensure that AI's benefits can be captured and its risks effectively mitigated. We're trying to understand the right sequence of technologies that the United States needs to build to realize the promise of an AI-enabled golden age sooner, and to ensure that we have the defensive technologies built in time to navigate the transition safely. We're investigating the incentives, field-building strategies, funding mechanisms, and (de)regulations that can rapidly scale up an R&D ecosystem to solve these challenges. Success in this first requires that the United States take responsibility for shaping the future of AI.

Mobilizing the R&D lab of the world

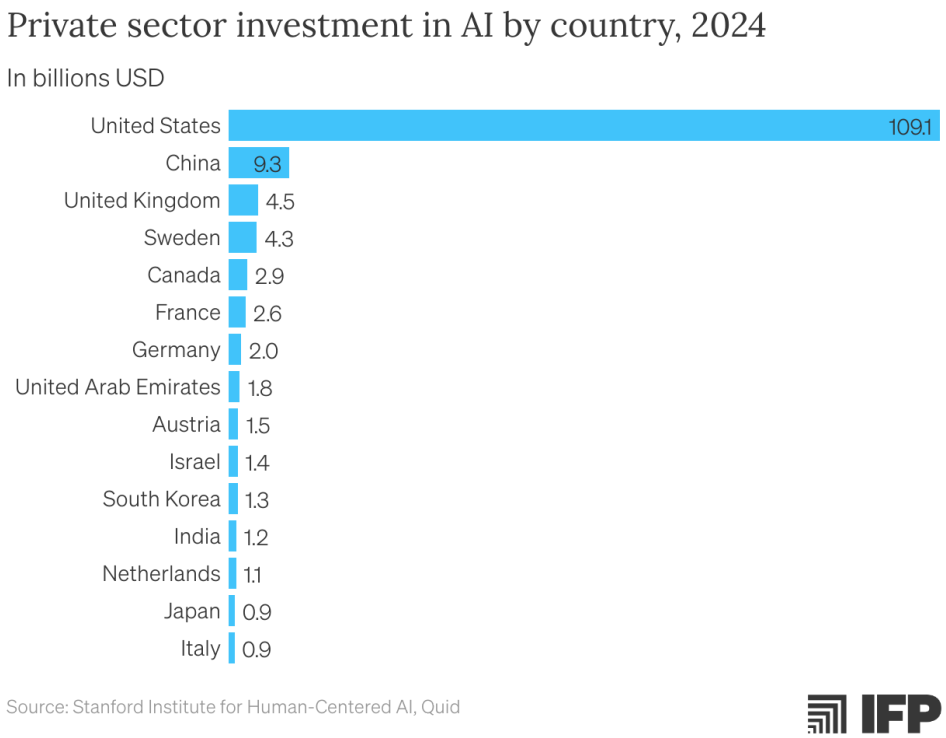
As we stated in IFP's founding essay, technology development and deployment show a great deal of path dependence. *Where* and *in what order* technologies are built can be just as important as *which* technologies are built. Artificial intelligence is likely especially path-dependent:

- **Today's most powerful models are general-purpose.** The space of possible applications, both good and bad, is large.
- **Many *offensive* capabilities of AI also have a *defensive* side**. AI could help design novel deadly viruses, but it can also enable better pathogen monitoring systems for pandemics. AI could hack into sensitive computer systems, but it can also help proactively discover and patch vulnerabilities. Whether the offensive or defensive capability is built first could have a large effect on whether and how new risks from AI materialize.
- **AI progress is sensitive to initial conditions**. Applications that have large amounts of training data (e.g., protein design) and/or lack major real-world bottlenecks (e.g., algorithm development) will see rapid capability improvements. This means that targeted interventions to change these

initial conditions could flip the ordering of offensive vs. defensive capability development.

Given this path dependence, the United States has the responsibility to shape AI development toward a path that enables — rather than smothers — human flourishing. For a vision of one future path, look no further than China, where the CCP has steadily deployed an AI-enabled system of mass surveillance to monitor its citizens both at home and abroad.

The United States also has the *power* to shape AI development. The US is the world's most powerful democracy, and, together with our allies, we possess the advantage across much of the AI supply chain. As of last year, US private sector investment in AI was more than ten times greater than that of any other country.



If the US falters or cedes the lead, there is no other strong democracy ready to take the baton. Outside of the US and China, there are only a handful of notable AI companies and ecosystems. There is no realistic plan B. The European Union can attempt to keep regulating companies outside of their borders, but only the US and China can realistically build the technological rails of the 21st century.

Because of our supply chain dominance, the United States has the potent ability to deny access to advanced AI to others through export controls, allowing it to remain the key decisionmaker around the future of the technology. We believe export controls are a crucial tool for slowing the CCP's ability to impose its vision on the development of advanced AI. But as good a tool as they are, they can't maintain our advantage forever. It will be a poor use of policy effort if we don't use the lead provided by export controls to steer the future in a better direction. The US is the R&D lab of the world, and we should act like it.

How should we think about shaping AI progress?

The way we think about the emerging technologies of our time often depends on the way we perceive the moment we're in. At the dawn of the Scientific Revolution, Francis Bacon's *New Atlantis* laid out a fundamentally optimistic vision for organized scientific research directed towards human benefit, leading to the establishment of the Royal Society in London, which in turn inspired a wave of new scientific academies across Europe. In 1955, John von Neumann wrote an essay titled "Can We Survive Technology?" containing a more pessimistic vision of technological development in a world newly characterized by world-ending nuclear arsenals, suggesting that the only real strategy was to simply muddle through.[10] In 2019, as synthetic biology tools like CRISPR were becoming increasingly available, and AI compute scaling trends were becoming more clear, Nick Bostrom's 2019 article, "The Vulnerable World Hypothesis," suggested that the only way to deal with widely available destructive technologies might be "ubiquitous surveillance or a unipolar world order." Vitalik Buterin's 2023 essay, "My techno-optimism," can be thought of as a synthesis of many of these ideas — the new technologies of our era could lead to disaster, but so could our response to these threats, if it leads loss of democratic values or the concentration of power. Buterin's solution is a wide-ranging R&D program focused on defensive technologies to mitigate and defend against emerging risks while avoiding centralized control.

[10] The penultimate section of the essay was titled "Survival—a possibility"

		
Anti-technology view: safety behind, dystopia ahead.	**Accelerationist view**: dangers behind, utopia ahead.	**My view**: dangers behind, but multiple paths forward ahead: some good, some bad.

Provides a more in-depth introduction to contrasting perspectives on technological progress. From Buterin, 2023, "My techno-optimism"

We borrow the optimism of Bacon, but subscribe to the basic logic of those like Buterin: rather than treating the trajectory of a technology as inevitable, we should think expansively about how we might go about shaping it. This isn't wishful thinking. Over recent decades, the United States has successfully taken varied approaches to shaping the development of new technologies:

- **Nonproliferation:** With nuclear weapons, we decided that the only way to safely manage the technology was to lock it down. Evidence about the unprecedented destructive power of the technology led President Roosevelt to classify all fission research in 1942, which in turn led to the centralization of R&D through the Manhattan Project, and to America's subsequent nonproliferation strategy. Although this strategy hasn't been perfect, nuclear proliferation has consistently been slower than anticipated by most experts.[11]

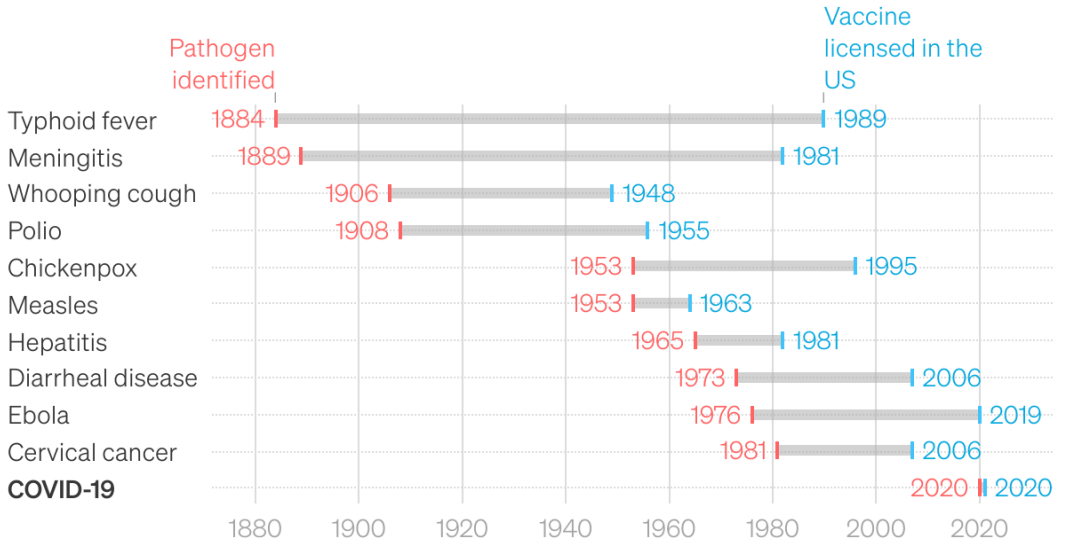
[11] From 1949 to 1964, "an overwhelming majority of classified and academic studies suggested that proliferation to more countries was inevitable." In 1960, then-Senator John F. Kennedy warned that "there are indications because of new inventions, that 10, 15, or 20 nations will have a nuclear capacity, including Red China, by the end of the Presidential office in 1964." In those four years, only two countries — France and China — developed nuclear weapons. And only four more have developed them since. This success did not come by default; it resulted from decades of active efforts by American and allied statesmen to contain the most dangerous technology ever developed.

- **Selective Acceleration:** In the 1980s, recognizing the downsides of private companies patenting genetic sequences, we decided to proactively accelerate the development of universal tooling and infrastructure for gene sequencing through the Human Genome Project. This created a vast public domain resource that helped prevent broad patents on raw genomic sequences.[12]
- **Defensive Acceleration:** In the face of COVID, the Trump Administration realized that overcoming the pandemic required massively accelerating a new defensive technology: the mRNA vaccine platform. Operation Warp Speed combined market-shaping mechanisms, a whole-of-government strategy, and the streamlining of regulation to radically accelerate the development of the first fully-approved mRNA vaccine. The vaccine was developed in just 9 months, more than 10 times the normal speed, and far faster than any vaccine ever developed.

[12] The Human Genome Project also allocated ~5% of its resources to the ELSI Program ("Ethical, Legal, and Social Implications"). ELSI research helped secure policy victories in preventing discrimination in health insurance and deployment based on genetic test results that predict clinical manifestation of future genetic disorders.

Vaccine development timelines

The COVID vaccine was developed and licensed in the US at least 10x faster than any other vaccine in history.



Source: Our World in Data

IFP

Of course, general-purpose AI is a different kind of technology from each of these, both in its wider variety of applications and also in the way it can itself be harnessed to differentially accelerate the development of other technologies. This means the right way to shape it is likely through a complex mix of strategies, both well-tested and novel.

Four principles for shaping AI progress

1. We should leverage the "jagged frontier" of AI progress

AI's capabilities don't advance at an equal pace across all fields, creating a "jagged frontier." As of early last year, just 5% of businesses in the US used AI, according to data from the US Census Bureau. But within that same year, an AI model led to a Nobel Prize in Chemistry, and Google announced that more than a quarter of the code accepted into production codebases at the company was generated by AI. At least in programming, this trend is clearly accelerating: 80% of Claude Code (an

agentic coding tool built by Anthropic) was written by Claude itself. The future is already here — it's just not evenly distributed.

Those seeking to use AI to advance work on the world's most important scientific problems, or to build technologies to address new risks introduced by AI, should take advantage of this jagged frontier. If vast sums of knowledge within a field can be easily tokenized for processing by AI, solutions to problems in that field can be quickly verified, and new capabilities rapidly deployed, then work within that field can be dramatically accelerated using AI. This is already apparent for parts of mathematics, chemistry, and biology, as well as much of programming. Less so for fields like particle physics, which is limited by the slow and expensive process of gathering data from experiments in particle accelerators. What lessons should we draw from this?

Tactics for shaping AI progress should be informed by predictions about where AI will succeed first. The ease of *tokenization*, *solution verification*, and *deployment* across different domains is a good predictor of where we'll first see new beneficial and destabilizing capabilities. This helps assess the likely balance of offense against defense, which will, in turn, help determine the best strategy for shaping progress. For example:

- In cybersecurity, the same AI capabilities that enable automated offense could also be quickly deployed to enable automated defense.[13] In domains like this one, a strategy based on accelerating defensive applications and security infrastructure[14] may be sufficient.[15]
- Parts of biosecurity, on the other hand, appear offense-dominant. Many worry that a terrorist group could make use of AI-based biological design tools to design and release a deadly, ultra-transmissible virus. Here, it still makes sense to accelerate relevant defensive technologies like far-UVC and flexible vaccine platforms. But the deployment of these technologies is

[13] Google's Big Sleep agent, trained to search for security vulnerabilities in software, recently discovered an SQLite vulnerability (widely used open-source database software) that was known only to bad actors and was at imminent risk of being exploited. Thanks to this threat intelligence, we were able to proactively patch the vulnerability.
[14] Sella Nevo's essay, "A Sprint Toward Security Level 5," outlines how to build this infrastructure to protect American AI from nation-state level threats.
[15] Miles Brundage's essay, "Operation Patchlight," details how to leverage advanced AI to give defenders an asymmetric advantage in cybersecurity.

> likely to be slow, relative to the speed of the worst pandemics. This means that we likely also need to pursue a strategy of nonproliferation for some biosecurity applications. One promising approach, as called for by the White House's recent AI Action Plan, is to focus on nucleic acid synthesis providers to ensure they implement sequence screening and customer verification.

Closely tracking the frontier of AI capabilities in different domains could give us the advanced warning necessary to proactively build new defenses. Evidence from progress on well-designed benchmarks, and from new capabilities observed from internal deployments of new models at frontier AI labs, can give us an "adaptation buffer" — a period of time we can use to build the tools, infrastructure, and policies to maximize the chance that the wide availability of the new capability goes well. This will require building the right early warning systems to prompt public and private action. Evidence about new bio-offensive capabilities seen in early versions of new models, for example, could help trigger a build-up of personal protective equipment for rapid distribution, and the rollout of wastewater surveillance to detect new pathogens early.[16]

Tracking and predicting AI capability improvements can also help us avoid wasting resources by acting too early. Today, formally verifying the software used across all critical infrastructure to prove its security would likely be a horrifically expensive undertaking. But the rapid improvement of long-context performance in software engineering tasks for frontier models could lead to a situation where, suddenly, ubiquitous formal verification is cost-effective.[17] In such cases, rather than trying to develop defensive technologies early, we should instead set up the measurement infrastructure to help decide when to act, and the policies and relationships to allow for rapid deployment. Automated refactoring of the codebases used in critical infrastructure, for example, will predictably be bottlenecked by slow government procurement processes for new software. This is a problem that can be addressed now.[18]

---

[16] Simon Grimm's essay, "Scaling Pathogen Detection with Metagenomics," explains how to generate the data necessary to reliably detect new pathogen outbreaks with AI.
[17] Patrick Shafto's essay, "The Infinity Project," describes how to use AI and mathematics to improve science and security by formally verifying critical processes.
[18] Herbie Bradley and Girish Sastry's essay, "The Great Refactor," lays out how to secure critical open-source code against memory safety exploits by automating code hardening at scale.

The shape of the AI capabilities frontier can be changed. We can push on this frontier to get the capabilities we want faster. For example, many barriers to unlocking datasets exist due to historical contingency in regulations or because the data is not easily monetizable. By some estimates, the stock of total data generated each year is over a million times greater than the current stock of data publicly available on the internet through the Internet Archive, which contains the bulk of the data used to train LLMs today. Much of this data could be useless for training AI models. But some of it, especially scientific datasets locked away on hard drives in government agencies and universities, could have significant societal value. Effective policymaking could be used to help unlock the data that will help deliver beneficial AI capabilities sooner.[19]

2. We shouldn't neglect the costs of stalled progress

This may seem too obvious to say. However, we believe that many policy ideas for addressing AI risk do not take the cost of stalled technological progress seriously enough.[20]

Until recently in human history, half of all children would die before adulthood. These conditions, described by Thomas Hobbes as making life "poor, nasty, brutish, and short," ruled until the scientific and industrial revolutions delivered us germ theory, vaccines, fortified foods, antibiotics, and thousands of other such innovations, in addition to much greater wealth. Failing to invest in science and fundamental technologies that will continue these advances will incur costs in human welfare. These need to be balanced against the expected risks of the new technologies. Technology-shaping strategies that neglect the upsides of progress may delay the solutions to some of today's biggest problems.

For scientific discovery relevant for areas like medicine, the potential upside of leveraging the frontier of AI capabilities is huge. Global health grantmaker Jacob Trefethen lists a set of medical technologies, like malaria vaccines and tuberculosis vaccines for adults, that, if built, would together save around 3.6 million lives per year. Each of these technologies is achievable to build, but will likely not be

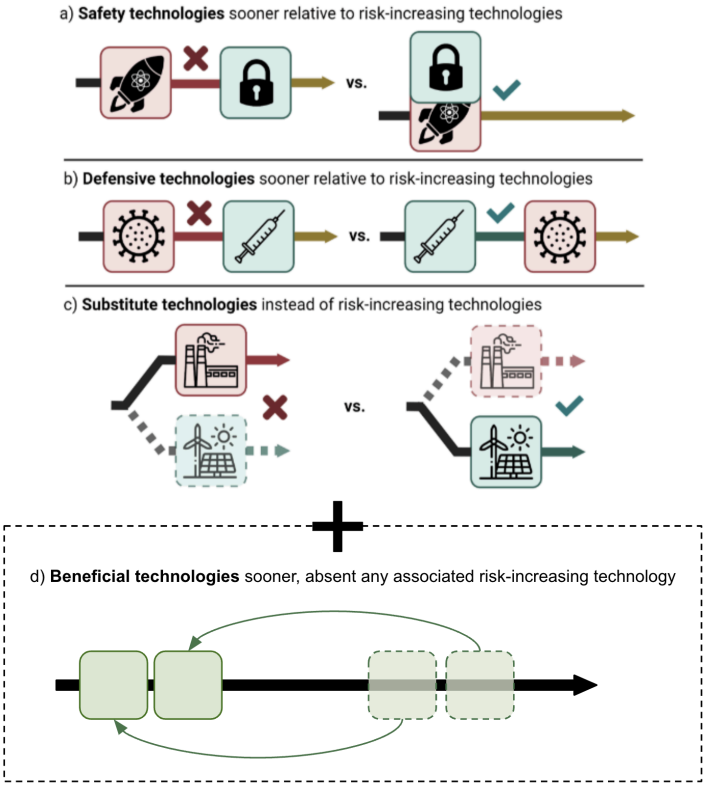
[19] Andrew Trask and Lacey Strahm's essay, "Unlocking a Million Times More Data for AI" (forthcoming), proposes that a new ARPANET-style program could solve the data scarcity problem via Attribution-Based Control.
[20] For example, some called for a temporary pause in AI development after GPT-4

available within five years. If the clear bottlenecks to progress in AI for medicine are addressed, AI could help deliver these and other public goods.

#	Technology/tool	Why do we need one?	Will one be available in 5 years?	Is making one achievable?
1	TB vaccine that works in adults	Kills 1.5M/year	Probably not	Yes
2	Strep A vaccine	Kills 500K/year	No	Yes
3	Malaria monoclonal antibody	Kills 600K/year	Probably not	Yes
4	Bugs that stop malaria spreading (including rural)	Kills 600K/year	No	Yes
5	Hep C vaccine	Kills 300K/year	No	Yes
6	Stroke reducing drug that most patients can take	Kills 6M/year	No	Unclear
7	Hep B complete cure	Kills 600K/year	No	Unclear
8	Test that tells you why you're sick	Do I have COVID or flu or strep throat or a cold or what??	I don't know	Yes
9	Programmable drugs before a future pandemic	The last one killed 7M-25M	Probably not	🤞
10	Syphilis vaccine	Kills 100K/year	No	Yes

Trefethen, 2025, "What does AI progress mean for medical progress?"

The promise of using AI in this way highlights how the original conception of differential technology development falls short of a complete strategy for shaping technological progress: by focusing solely on the risks posed by new technologies, it neglects the massive benefits to be gained from accelerating their development. Even absent new technologies that create additional risks, we still face a range of important and urgent problems. As we discuss later, the reasonable attitude about the new risks created by artificial intelligence is fairly deep uncertainty. It may be that the problems we already face today remain the most important problems of the future.



Edited version of original image from Sandbrink et al., 2022, "Differential technology development"

As Trefethen notes:

> "Some technologies are more important than others — indeed, some are life or death. Technological progress is not a mystical force that delivers the most important ones first. Some problems are hard to solve, and won't make you much money even if you succeed, and don't get talked about on the news. What people choose to work on determines what new technologies are made."

The same logic that applies to defensive technologies also applies to humanitarian R&D. Along many branches of the technology tree, there are market failures that the private sector will not explore by default (or quickly enough). The public sector's essential role is to deliberately cultivate these neglected parts of the

portfolio — choosing, funding, and accelerating work that markets cannot pursue at the needed speed or scale.[21]

3. To realize the benefits of AI, we should redesign how science works

Strategies for shaping AI progress should not just pay attention to object-level research paths, but also to metascience. How should we rethink research production in a world where AI has become the driving force of scientific discovery? Much as factories needed to be retooled to take advantage of the affordances provided by the steam engine, so too will our funding mechanisms, institutions, and incentives need to be redesigned to take advantage of AI-enabled scientific discovery.[22]

The American science funding ecosystem has become increasingly bureaucratic. Researchers face wait times of up to 20 months for grant funding, and principal investigators spend almost half of their time on grant-related paperwork. This lost scientific productivity will become even more acute in a world where the role of researchers is to manage teams of AI scientists, responsible for generating and testing a vast number of new hypotheses. Every month delayed by paperwork will be equivalent to many months of time in today's world. And the classical funding model, of small, project-based grants awarded to individual investigators at universities, is poorly suited to AI-driven scientific research. AI-driven research increasingly requires resources that universities don't have: large-scale compute infrastructure and dedicated engineering teams to help manage compute and data. Even if AI enables dramatic improvements in our knowledge about the world, our bureaucratic systems may severely limit our ability to harness it.

A wave of new organizations is showing what the future of research should look like in a way that will truly be "AI ready." Arc Institute and FutureHouse are independent, nonprofit research organizations that have made large investments in infrastructure-driven "team science." These investments have already led to promising results. Earlier this year, Arc released Evo 2, a frontier biology model

[21] Maxwell Tabarrok's essay, "A Million-Peptide Database to Defeat Antibiotic Resistance," outlines how to build a large peptides database to train the AlphaFold for new antibiotics.
[22] Ben Reinhardt's essay, "Teaching AI How Science Actually Works," addresses this by outlining how new block-grant labs could generate the real-world data AI needs to do science.

capable of both identifying disease-causing mutations in human genes and designing new simple genomes.

To lean into this new paradigm for scientific discovery, agencies like the NIH and NSF should make better use of flexible award mechanisms like Other Transactions Authority (OTA). OTAs can enable experimentation with new structures for research funding, like institutional block grants to support organizational structures similar to Arc and FutureHouse, and fast grants to enable fast-moving research funding to take advantage of rapidly improving AI capabilities.[23]

4. We should adapt to uncertainty while working to reduce it

Unfortunately, technological forecasting is exceedingly difficult. In 1955, John von Neumann (widely regarded by his peers as the smartest person of his time) made a set of startlingly accurate predictions about the automation potential of semiconductors and the looming risk of climate change. But, at the same time, his predictions (and others')[24] about widespread cheap nuclear energy and rampant geoengineering failed to come to pass.

The developments that led to the capabilities of today's general-purpose models would have been similarly hard to predict. Google's work on natural language understanding was driven by its desire to produce better search results to sell more ads. These incentives indirectly led to the transformer architecture, the key algorithmic innovation that underlies almost all frontier AI models today. Today's GPUs trace their original lineage not back to any AI-specific development effort, but instead to the needs of video gaming, which required the same high-performance parallel processing that turned out to be highly suitable for training large neural networks. Before pivoting to large language models, OpenAI

[23] Caleb Watney's essay, "Using X-Labs to Unleash AI-Driven Scientific Breakthroughs," centers on how to adapt our science funding mechanisms to the unique infrastructure needs of large-scale AI projects.

[24] In 1954, then-chairman of the Atomic Energy Commission Lewis Strauss expressed his hopes for nuclear technology: "It is not too much to expect that our children will enjoy electrical energy too cheap to meter — will know of great periodic regional famines only as a matter of history — will travel effortlessly over the seas and through the air with a minimum of danger and at great speeds — and will experience a life-span far longer than ours, as disease yields and man comes to understand what causes him to age. This is the forecast for an age of peace."

first spent its time exploring deep reinforcement learning for video games and robotics as the most promising paths to AGI.

Even within today's AI development paradigm, despite the ability for scaling laws to accurately predict training loss, the specific capabilities of models have often surprised AI researchers, both on specific tasks like programming, as well as in terms of broader propensities, such as self-preservation and deception.

This spotty history of predictions should dissuade us from being too confident in how and when the biggest opportunities and risks from advanced AI will materialize. But this uncertainty doesn't imply that the right approach is to do nothing. Instead, we should adapt to uncertainty and seek to reduce it.

Some uncertainty will be difficult to reduce. We could not have predicted the historical contingencies leading to the development of the transformer and the GPU. But many uncertainties about AI capabilities and their effects can likely be reduced with a better AI measurement and evaluation ecosystem.[25]

Adapting to uncertainty about AI progress is also a political endeavour. It's hard to rally a broad base of support for an ambitious technology-shaping agenda in the face of wide uncertainty. The best approach is pluralistic, blending a variety of important goals. The CHIPS and Science Act is a good example. Thanks to its dual focus on semiconductor supply chain resilience and on promoting basic science, it found support from both national security stakeholders and public science advocates. A technology-shaping agenda for AI should likewise include measures focused on both science and security.

The Launch Sequence

Dario Amodei's essay "Machines of Loving Grace" offers a tantalizing vision of the scientific problems that could be solved by AI progress over the next five years. But it does not describe the concrete steps required to apply new AI capabilities to solve problems in applications that lack strong market incentives, like diffuse societal resilience technologies and basic science. The default path of AI

[25] Séb Krier and Zhengdong Wang's essay, "Benchmarking for Breakthroughs," describes how to incentivize AI for national priorities through a strategic challenge and evaluations program.

development may not deliver us these benefits as soon as we'd like, nor protect us from new AI-enabled threats.

What is the right sequence of technologies that we need to develop as we transition to a world of intelligence too cheap to meter? The goal of this collection is to start piecing together the answer to this question. We've pulled together concrete but ambitious proposals from some of the sharpest people thinking about how to use AI to proactively shape progress for both science and security. The collection features projects that are unlikely to occur by default given existing commercial incentives, that can be achieved or fully set up by 2030, and are particularly important to achieve in light of the rapid advances of AI.

Success won't come from a single master plan, but rather an adaptive strategy that evolves in tandem with AI's capabilities. Our strategy should not ignore the tremendous upside of the technology, nor ignore the perils of top-down control and centralization. To quote von Neumann:

> "What safeguard remains? Apparently only day-to-day — or perhaps year-to-year — opportunistic measures, a long sequence of small, correct decisions."
>
> *John von Neumann, "Can We Survive Technology?"*

This collection is not a comprehensive plan. And when compared to the size of private investment in AI R&D, almost any proposal looks small in scale. We hope that the ideas in this collection form part of a sequence of "small, correct decisions" to positively shape the trajectory of the most important technology ever developed.

* * *

Acknowledgements

Special thanks to all the authors of the essays in this collection; to Emma Kumleben, Santi Ruiz, and Beez Africa for editing; to Beez and Emma Steinhobel for design; and to Caleb Watney, Ari Kagan, Arushi Gupta, Saif Khan, Ben Schifman, Alexandra Bates, Jonah Weinbaum, and Alec Stapp for feedback throughout.