Reading Materials for CS283: Governing AI

Guest speaker: Ben Buchanan
Assistant Professor (Johns Hopkins University)
Former White House Special Advisor for AI to President Biden

November 16, 2026

Contents

	Title	Venue	Date	Pages
1.	Introduction to *The Bitter Struggle: Superintelligence, Superpowers, and the Fate of the World*	Crown	November 10, 2026	3–15
2.	The AI Grand Bargain: What America Needs to Win the Innovation Race	*Foreign Affairs*	October 21, 2025	16–28
3.	Ben Buchanan Wants the U.S. to Win on AI. But Is It a Race or a Trap?	Center for Humane Technology *Your Undivided Attention*	November 12, 2026	29–67
4.	The Government Knows A.G.I. Is Coming	*The New York Times* *The Ezra Klein Show*	March 4, 2025	68–109

NOT FOR DISTRIBUTION

THE BITTER STRUGGLE

Superintelligence, Superpowers, and the Fate of the World

Ben Buchanan and Tantum Collins

CROWN
NEW YORK

Buch_9798217089482_all_4p_r1.indd 3 8/3/26 09:58

NOT FOR DISTRIBUTION

CROWN
An imprint of the Crown Publishing Group
A division of Penguin Random House LLC
1745 Broadway
New York, NY 10019
crownpublishing.com
penguinrandomhouse.com

Copyright © 2026 by Ben Buchanan and Tantum Collins
Penguin Random House values and supports copyright. Copyright fuels creativity, encourages diverse voices, promotes free speech, and creates a vibrant culture. Thank you for buying an authorized edition of this book and for complying with copyright laws by not reproducing, scanning, or distributing any part of it in any form without permission. You are supporting writers and allowing Penguin Random House to continue to publish books for every reader. Please note that no part of this book may be used or reproduced in any manner for the purpose of training artificial intelligence technologies or systems.

CROWN and the Crown colophon are registered trademarks of Penguin Random House LLC.

Library of Congress Cataloging-in-Publication Data is available upon request.

Hardcover ISBN 979-8-217-08948-2
Ebook ISBN 979-8-217-08949-9

Editor: Kevin Doughten
Assistant editor: Jessica Jean Scott
Production editor: Liana Faughnan
Text designer: Aubrey Khan
Production: Chris Andrus
Copy editor: Lynn Anderson
Proofreaders: Kevin Clift, Eldes Tran, Miriam Taveras, and Chris Fortunato
Indexer: J S Editorial
Publicist: Penny Simon
Marketer: Heather Faulls

Manufactured in the United States of America

$PrintCode

First Edition

The authorized representative in the EU for product safety and compliance is Penguin Random House Ireland, Morrison Chambers, 32 Nassau Street, Dublin D02 YH68, Ireland, https://eu-contact.penguin.ie.

Buch_9798217089482_all_6p_r1.indd 4 8/26/26 12:40

NOT FOR DISTRIBUTION

CONTENTS

INTRODUCTION 1

PART ONE
The Struggle for Democratic Invention

CHAPTER ONE
Scale 13

CHAPTER TWO
The Compute Moat 26

CHAPTER THREE
Chip Controls 38

CHAPTER FOUR
Diffusion 56

CHAPTER FIVE
Energy 78

PART TWO
The Struggle for Democratic Advantage

CHAPTER SIX
National Security Use 93

Buch_9798217089482_all_4p_r1.indd 7 8/3/26 09:58

NOT FOR DISTRIBUTION

CHAPTER SEVEN
The Intelligence Explosion 114

CHAPTER EIGHT
The Ceiling Is Far Away 128

CHAPTER NINE
Conviction 146

PART THREE
The Struggle for Democratic Values

CHAPTER TEN
Misuse 161

CHAPTER ELEVEN
Misalignment 176

CHAPTER TWELVE
Executive Action 192

CHAPTER THIRTEEN
Centralization of Power 205

CHAPTER FOURTEEN
The Democratic Rudder 222

CHAPTER FIFTEEN
Technological Adolescence 234

CONCLUSION
The Shape of Things to Come 249

Acknowledgments 255
Notes 259
Index 297

Buch_9798217089482_all_4p_r1.indd 8 8/3/26 09:58

NOT FOR DISTRIBUTION

INTRODUCTION

"What do you mean, 'unimaginable'?"

The question hung in the air of the Diplomatic Reception Room in the White House complex, a grand chamber adorned with stately oil portraits. It was January 2023, a few months after ChatGPT's launch, as it raced toward 100 million users worldwide. On one side of the table sat OpenAI CEO Sam Altman, flanked by some of his technical experts. On the other side were President Biden's national security advisor, Jake Sullivan, and a few of his National Security Council officials.

Altman had come to showcase the capabilities of GPT-4, which was in the final stages of development and represented a dramatic improvement over the older model that, at the time, underpinned ChatGPT. He demonstrated how the new system could process images as well as text and how it could reason about visual anomalies—a vital skill for analyzing geospatial intelligence, though he left that implication unspoken. He showed Sullivan how GPT-4 could synthesize data from Vladimir Putin's speeches and highlight key points for decision-makers, the sort of task for which intelligence agencies train teams of subject matter experts.

Impressed, Sullivan asked where the technology would go next. "Well, in five years it will be unimaginable," Altman said casually. To Altman, as to so many other AI pioneers, it was an article of faith that AI would soon transform nearly every aspect of society.

To the U.S. government, however, that view was alien. For too long, the private sector's rapid AI progress had been invisible to people in Washington, lost in the noise of Silicon Valley hype alongside

Buch_9798217089482_all_4p_r1.indd 1 8/3/26 09:58

NOT FOR DISTRIBUTION

NFTs, the metaverse, and Web3. Sullivan, though, had long shown an interest in AI, and Altman's comment prompted him—with unusual trepidation—to ask what "unimaginable" entailed. Altman paused and leaned back in his chair. Then he explained that he expected AI systems within five years to make novel scientific discoveries and automate major portions of the economy. AI would also redefine national security, he believed, exposing vulnerabilities in computers and critical infrastructure thought to be safe. Few parts of the world would remain untouched.

The meeting ran longer than expected. Sullivan was not easily swayed by grand proclamations of technological transformation, but he valued seeing the technology up close. He thanked Altman for coming, and we could tell he meant it.

Over the following months, AI leaders streamed into the White House with similar previews, predictions, and warnings. A few weeks after the GPT-4 demonstration, Dario Amodei, the CEO of Anthropic, presented his views to Sullivan. He focused on how AI systems could quickly scale toward what he called "a country of geniuses in a data center"—a superhuman accelerator for scientific advances. Google DeepMind CEO Demis Hassabis visited not long after that and struck a similar tone about rapid AI progress. So did Elon Musk, following his launch of xAI.

Ben, on leave from academia, led AI policy in the Biden White House and was the only official present in all these meetings. Tantum had spent years as a research scientist at Google DeepMind before joining the National Security Council as director for technology and national security. Together, our job was to figure out the national security implications of the AI transformation, draft policy, and brief key decision-makers, such as President Joe Biden, National Security Advisor Sullivan, and Assistant to the President Bruce Reed, about what we saw on the horizon.

To us, the conversations with Altman, Amodei, Hassabis, and Musk felt like the two parts of the transcontinental railroad connecting: AI leaders from the West Coast were at last joining up with the national security establishment of the East. For years, the two

Buch_9798217089482_all_4p_r1.indd 2 8/3/26 09:58

NOT FOR DISTRIBUTION

communities had toiled largely in isolation. But by the early 2020s, the technology was too capable for government to ignore, and its development was too dependent on regulated inputs for AI executives to avoid the policy apparatus. Once those two forces linked, they would be inextricably tied together for all that was to come—for good or for ill, whether they liked it or not.

So what *does* "unimaginable" mean?

Answering Sullivan's question requires grappling with a tension nearly as old as language itself. The Proto-Indo-European root word for power, *magh,* seemingly gave rise to both "magic" and "machine," the two poles of humanity's enduring quest to create the previously unimaginable.

Humans see machines as beneath us. We create them, and have done so since the invention of the first wedged hand ax almost 2 million years ago, long before we evolved into *Homo sapiens*. Today's machines, no matter how complex, remain human creations in essence. Their power is ours.

Magic, on the other hand, has supernatural origins. It exceeds not just human capabilities but also human understanding. Even those who believe in magic do not create it but instead try—and often fail—to manipulate it, to coax its mysterious power to serve their desired ends. Myths and fables from almost every era and culture warn that those who pursue magic might initially enjoy the illusion of control but generally end up humbled, often surrounded by ruin.

Artificial general intelligence (AGI), a hypothesized system at least as capable as a human at virtually all cognitive tasks, would be a tantalizing combination of machine and magic. Many AI scientists believe that AGI could quickly enable superintelligence, a system that would outperform the smartest human minds in every field. Such a machine would exceed the scientific insights of Nobel Prize winners, possess better strategic judgment than today's shrewdest military commanders, and comprehend and masterfully manipulate any form of human knowledge—language, code, mathematics, biology, and beyond. It could create and harness tools, both digital and physical, to achieve its objectives.

Buch_9798217089482_all_4p_r1.indd 3 8/3/26 09:58

NOT FOR DISTRIBUTION

Humanity is racing to build this kind of AI, and we are doing so in ways that even our smartest technical experts do not fully understand, with methods more akin to raising a child than writing a normal computer program. If we succeed, we will wield AI's force even as we remain baffled by its mechanics.

The two of us have wrestled with the implications of AI since 2014, when we first met as graduate students focusing on the intersection of national security and technology. At the time, the ambitions of nascent AI companies felt compelling but fanciful. AI pioneers talked about "solving intelligence," implying that human cognition itself was a limit to be overcome. They envisioned AI curing disease, harnessing the totality of the sun's energy, delivering paradigm-shifting scientific discoveries, and unlocking hundreds of trillions of dollars in economic value. AI, they said, would eventually perform AI research, accelerating its own progress.

Those visions sounded far-fetched. But month by month, reality began to vindicate them. During the 2010s, machine learning, an approach through which computers extract insights from data rather than following human-written rules, proliferated across academia and industry. Increasingly sophisticated algorithms beat human world champions and state-of-the-art software systems in tasks ranging from board games to image recognition to scientific problem solving. In 2017, researchers unveiled the "transformer" algorithm, a remarkably general-purpose architecture that quickly outperformed nearly all prior systems across almost every domain. By 2020, studies demonstrated how transformer-based systems improved rapidly and predictably with access to more computing power, a resource that grew exponentially as the semiconductor market developed better AI chips and expanded production.

By 2021, any skepticism we once harbored had largely faded. Meaningful AI progress was apparent and seemed poised to accelerate. We joined the White House with a set of convictions that would guide four years of policymaking and that form the central argument of this book.

Perhaps most important, the ceiling on what AI can achieve is far higher—and the pace of progress is far faster—than almost anyone

Buch_9798217089482_all_4p_r1.indd 4 8/3/26 09:58

NOT FOR DISTRIBUTION

outside the field appreciates. AGI should be seen not as distant science fiction but as something that is likely to emerge in the next decade. Questions about precise definitions and timelines persist, but progress has been stunning—far faster than all but the wildest predictions of a decade ago.

The race to unlock these extraordinary capabilities may be winner take all. AI systems can increasingly accelerate their own improvement, and small leads could compound into large ones. The resources required to compete at the frontier of AI development—talent, chips, capital, and energy—concentrate power in fewer and fewer hands. Whatever company or country pulls ahead may become extraordinarily difficult for its competitors to equal.

The geopolitical stakes are therefore immense. The country that invents and effectively uses the most powerful AI systems will possess major advantages in military power, economic competition, intelligence analysis, and scientific discovery. As a result, despite the field's academic and corporate origins—and against the wishful thinking of some of its scientific leaders—the development of AI will inevitably and fundamentally become a geopolitical matter, creating a complex relationship between the public and private sectors.

If humanity is to maximize the benefits of AI, that leading country must be a democracy. Autocracies will use AI to entrench their power, repress their populations, and extend their reach abroad. Only democratic primacy in AI offers a path to broadly shared benefits, preserved freedoms, and governance that reflects the will of the people.

On this last point, history has lessons for us. At the White House, we regularly cited President John F. Kennedy's 1962 Rice University speech, which President Biden displayed in his private study next to the Oval Office. Many people remember Kennedy's pledge to land a man on the moon, but his argument ran deeper: Democracies must pursue power during technological revolutions. Failing to do so means abdicating the responsibility for how those breakthroughs shape the world.

Kennedy lauded all that space offered humanity—"there is new knowledge to be gained, and new rights to be won"—but he also

Buch_9798217089482_all_4p_r1.indd 5 8/3/26 09:58

NOT FOR DISTRIBUTION

sought to secure for America the power to determine what would happen there. He saw no contradiction between warning about the dangers of arms races and declaring unequivocally that "the vows of this Nation can only be fulfilled if we in this Nation are first, and, therefore, we intend to be first." Whether space would "become a force for good or ill depends on [us]," he said, "and only if the United States occupies a position of preeminence can we help decide."

Kennedy's argument resonates deeply in the AI era, which will likely prove to be far more consequential than the Space Race. If China or another autocracy achieves AI preeminence, democracies would lose their technological superiority and face unprecedented threats. We would see our most vital computer networks fall vulnerable to foreign hacking, our most important military advantages disappear, and our most central economic powerhouses fall behind during a moment of extraordinary technological and societal transition. There would be few checks on autocratic expansion. If we do not prevail in this era of technological change, *that* is what unimaginable will mean.

The Bitter Struggle

Achieving AI preeminence for democracies will be a bitter struggle with three components. First, the United States and its allies will have to defend and expand their current technical lead in the invention of the world's most powerful AI systems. Second, they will need to convert that lead to an enduring geopolitical advantage. Third, they must ensure that their own advanced AI systems are deployed safely and in line with democratic values. Only with success on all three fronts can humanity reap the greatest benefits AI offers, sidestep the greatest dangers, and preserve and deepen collective human agency.

Modern AI is the first revolutionary technology since the early industrial era to emerge almost exclusively from the private sector. The U.S. government catalyzed the creation of nuclear, space, jet aviation, internet, and computing technologies, but it has done comparatively little on frontier AI in recent decades. Private companies

Buch_9798217089482_all_6p_r1.indd 6 8/26/26 12:40

NOT FOR DISTRIBUTION

scaled up the machine learning paradigm and the infrastructure on which it depends, especially advanced semiconductors.

For years, the government and private sector underestimated each other with regard to AI. Many AI pioneers dismissed geopolitics, viewing state competition as beneath their epoch-defining quest. When Nvidia CEO Jensen Huang was asked if it mattered whether leading systems were American or Chinese, he said, "In the end, I don't think it does." But it *does* matter—AI will be the fulcrum on which the U.S.-China rivalry turns. This technology, once the domain of universities and small research labs, is inextricably linked to the levers and interests of the state.

To develop more powerful systems, AI companies will need extraordinary amounts of capital, electricity, and permits—which require government engagement to obtain. They will not build machines smarter than humans without the engagement of both militaries and intelligence agencies to protect these systems and put them to work for their own ends. The modern state is faced with the imperative of geopolitical competition and armed with an array of legal authorities and national security capabilities. Government may be slow to act, but companies cannot ignore its power.

The Biden administration tried to bridge the gap, supporting American AI companies while protecting national security. It took aggressive and at times controversial action, including limiting the flow of advanced computer chips to China. The administration simultaneously accelerated the United States' build-out of computational and energy infrastructure. The goal was to secure a lead in AI invention, powered by the private sector's abilities.

We are neither cheerleaders for nor detractors of the AI and chip companies at the center of this story. We admire their scientific and industrial achievements, but we also have reservations about the concentration of corporate power. We do not begrudge firms for trying to invent the next generation of technology or to make money, though we sometimes worry about the externalities of their efforts if they are left unchecked.

Invention, however, is only the beginning. Converting private sector AI innovation into geopolitical advantage is the second part of the

Buch_9798217089482_all_4p_r1.indd 7 8/3/26 09:58

NOT FOR DISTRIBUTION

struggle—and an area in which the United States and its allies are faltering. Had the pattern of previous technological revolutions held, something like GPT-4 would have been developed by researchers funded by the National Science Foundation on supercomputers built by the Department of Energy with the support of major investments by the Department of Defense. The new technology would have seen quick application to military and intelligence missions. The government would have a deep awareness of AI's progress and well-informed views on what to expect next. It would configure its national security apparatus to take advantage of AI's potential, and it would be intimately familiar with the associated risks. Instead, we first saw GPT-4 as a courtesy heads-up from a startup.

AI already has obvious application to national security tasks. Today's models can accelerate cyber operations, supercharge intelligence analysis, and transform warfare. As its capabilities improve, AI could upend the balance of power among nations. The Biden administration acted to spur government adoption and ensure continued democratic advantage, but much more needs to be done.

In some ways, the democratic position is worsening, as U.S. policy progress has stalled or been reversed since we left government. But even if such regressions prove temporary—even if democracies succeed in inventing the most powerful AI and deploying it to achieve a dominant position over autocracies—the third component of the bitter struggle looms as the greatest challenge of all. Attaining global preeminence in AI will force a hard question at home: Can democracies wield this unprecedented power without eroding their own core values?

We are aware that this question—and our overall argument—rests on an assumption that is no longer self-evident: that the United States remains a functioning democracy and continues to champion democratic values on the world stage. We nonetheless focus on responsible actions the United States can take because we believe that for all its flaws, it is presently the only democracy with the economic and political heft to shape the trajectory of superintelligence decisively. We recall the words of William Seward, Abraham Lincoln's secretary of state, "There was always just enough virtue in this Re-

Buch_9798217089482_all_4p_r1.indd 8 8/3/26 09:58

NOT FOR DISTRIBUTION

public to save it; sometimes none to spare, but still enough to meet the emergency." We remain hopeful that America will once again rise to the occasion when the world needs it.

Even if times were less politically turbulent, AI would pose a threat to the democratic order. It will upset the tensions that democracy aims to balance: between branches of government, between the public and private sectors, between individuals and the state, between privacy and security. For example, commercial competition drives the AI industry, and leading companies do not yet know how to develop advanced systems safely but charge ahead nonetheless. Democracies will have to determine how, if at all, to regulate this arena.

Similarly, AI systems can radically empower individuals in ways good and bad; the same systems that could cure disease and engineer new clean energy sources might enable the development of biological weapons or carry out cyber attacks. AGI could concentrate unprecedented power in companies worth trillions of dollars or in governments that risk undermining democracy in the name of defending it. Policymakers must strike a fine balance. Steering too far in one direction yields chaos; too far in the other, autocracy.

Most vexing of all, the three components of the bitter struggle create tensions with one another. Racing ahead in AI development to outcompete autocracies might come at the expense of safety testing. Realizing AI's economic potential could mean reducing human oversight. Ensuring that AI companies based in democracies win global commercial competitions might mean letting those companies accrue vast power. The pursuit of democratic preeminence abroad may threaten the defense of democracy at home.

Just as we once did for the president, we will do our best in the pages that follow to lay out the hard decisions while acknowledging uncertainty. Implicit in the word *unimaginable* is the reality that we cannot know the exact course AI will take. But Sullivan was right to press the question. This book is our best answer: what is coming, what is at stake, and what democracies must do.

Buch_9798217089482_all_4p_r1.indd 9 8/3/26 09:58

FOREIGN AFFAIRS

The AI Grand Bargain

What America Needs to Win the Innovation Race

BEN BUCHANAN AND TANTUM COLLINS

November/December 2025
Published on October 21, 2025

BEN BUCHANAN is the Dmitri Alperovitch Assistant Professor at Johns Hopkins School of Advanced International Studies. From 2021 to 2025, he served in a variety of roles in the White House, including as Special Adviser for Artificial Intelligence.

TANTUM COLLINS was Director for Technology and National Security on the National Security Council from 2023 to 2025.

The United States' lead in artificial intelligence might seem unassailable. U.S. companies—Anthropic, Google, OpenAI, and xAI—are out in front across almost all assessments of the technology's general capabilities. American AI models are outperforming doctorate-level scientists on challenging questions in physics, chemistry, and biology. Just a few American AI and chip giants are worth more than the entire Chinese stock market, and investors from across the world are plowing ever more resources into the American AI ecosystem.

This breakneck progress is, in many ways, a testament to the strengths of the model of

American AI development that has dominated for the last decade: letting the private sector operate on its own, with remarkably little direct government meddling or resourcing. This approach is quite different from those that ushered in past breakthrough technologies. Nuclear weapons and power, space travel, stealth systems, personal computing, and the Internet emerged either directly from U.S. government efforts or on the back of significant public funding. AI also has roots in government-funded science, including in personal computing and the Internet, and it benefits from ongoing government-supported research. But scaling up AI has been essentially a private-sector activity.

Yet there is reason to think the American way of developing AI is reaching its limits. Those limits will likely become increasingly evident in the coming months and years, and they will start to erode—and perhaps even end—U.S. dominance. Eventually, they will place the United States at a disadvantage against China, which has an alternative approach to the AI contest.

To avoid that outcome, Washington will need to embrace new ways of advancing AI development, ones that demand much tighter mutual support between the private sector and the state. Further progress now depends on resources and capabilities that only the government can provide or facilitate: the energy to power ever-larger data centers, a pipeline of international talent, and effective defenses against sophisticated foreign espionage efforts. The U.S. government, for its part, will need the cooperation of the private sector to integrate AI into the national security apparatus and to make sure the technology does not undermine democracy across the world.

The new American model of AI, in other words, must rest on a grand bargain between the tech industry and the government. The tech sector can help the state make sense of and deploy AI. The state can help the tech sector continue to grow in a way that advances everyone's interests.

MAXING OUT

It is easy to see why Washington's light-touch approach to AI has, by and large, paid dividends. Past revolutionary technologies, such as nuclear weapons and space flight, did not have immediate commercial applications. But the business case for modern AI is already highly compelling. AI firms have found huge user demand, resulting in skyrocketing

revenues, and they have promised to automate myriad valuable tasks, such as coding. As a result, capital markets are funding AI projects at scales that would historically have required government resources. Moreover, the computation-centric nature of today's AI means that it builds neatly on the cloud computing infrastructure that the private sector, not the government, has mastered.

The sufficiency of private-sector capital in enabling AI advances is wonderful for taxpayers, but the limits of this approach are becoming apparent. To see why, look at infrastructure. The vast fleets of computer chips needed to develop and use today's AI require extraordinary amounts of energy, so U.S. companies will need more power to fuel the data centers they plan to build in the coming years. An analysis by Anthropic estimated that the United States will need to produce 50 gigawatts of new power just for AI by 2028—roughly equivalent to what the entire country of Argentina uses today. (One of us, Buchanan, advises AI and cybersecurity companies, including Anthropic.) By then, data centers could consume up to 12 percent of American electricity production. Without more electricity, the AI build-out will stall. Amazon's CEO, Andy Jassy, for example, has labeled power the "single biggest constraint" to AI progress. And building this level of new infrastructure will require government help.

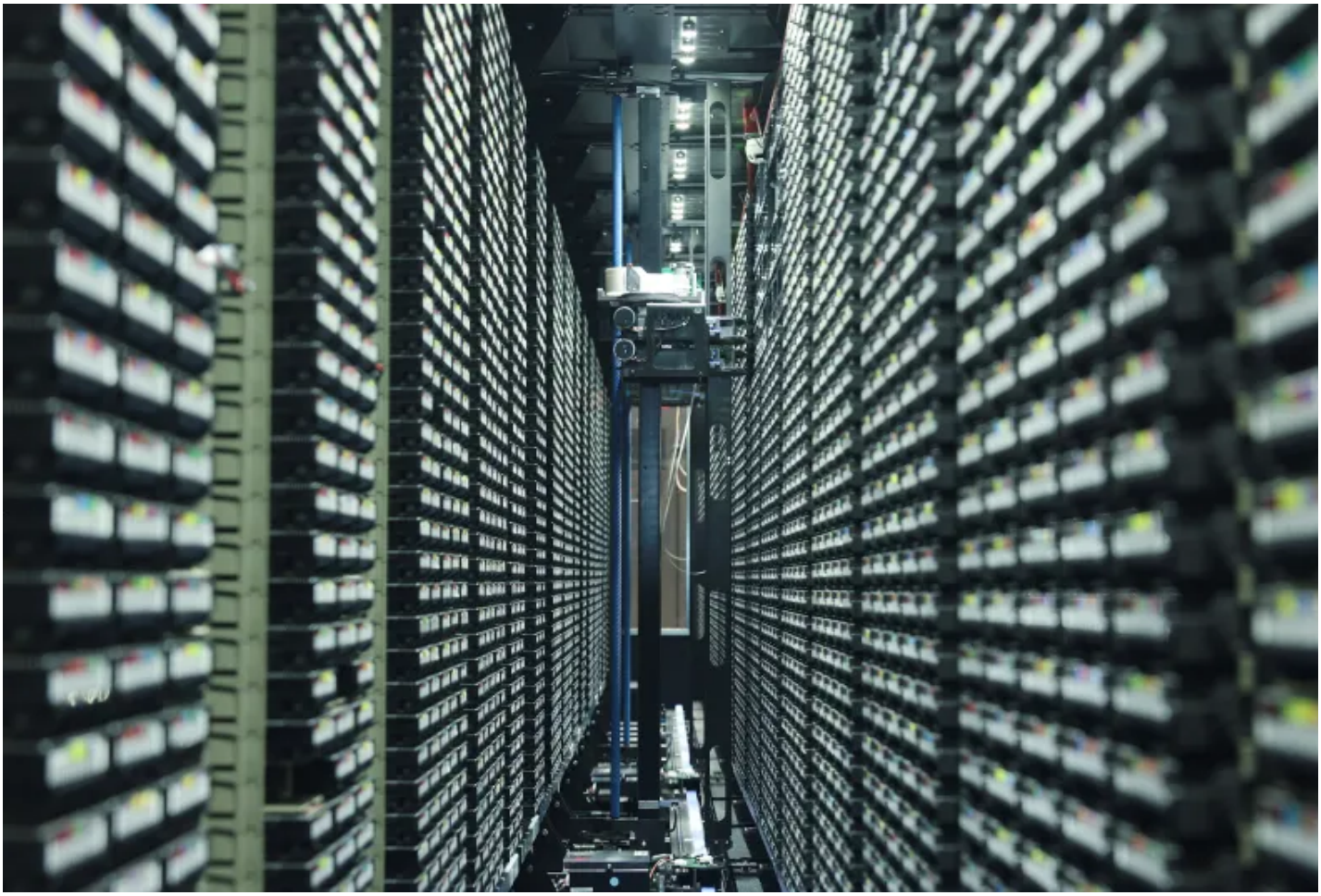
For too long, Washington did too little to add new power to its grid. From 2005 to 2020, the United States added close to zero net new power. After U.S. President Joe Biden took office, in 2021, and passed a law subsidizing the construction of clean energy infrastructure, the country added more than 100 gigawatts in new capacity. In the last days of his term, he signed an executive order specifically aimed at further expediting the AI and clean energy build-out. But although his successor, Donald Trump, has said the right things about building new energy infrastructure for AI, he has not delivered. He signed an executive order to accelerate federal permitting for data centers, but implementation remains nascent. Worse yet, his signature "One Big Beautiful Bill," passed in July, and other executive actions gutted key parts of Biden's energy expansion efforts, such as vital transmission projects. An area that could have been a bipartisan success fell prey to politics and has now become a major concern for business and AI competitiveness.

There is reason to think the American way of developing AI is reaching its limits.

Executed well, an AI-fueled energy boom would have benefits far beyond AI development itself. Leading AI companies are spending hundreds of billions of dollars on infrastructure build-outs, creating employment opportunities. They have committed to carbon-free operations and demonstrated a willingness to pay higher prices for clean energy. These massive investments can accelerate the domestic development of better energy sources, many of which have bipartisan appeal, such as advanced geothermal power and next-generation nuclear facilities. Powerful AI models could also accelerate climate-related research.

If the United States does not construct more energy capacity, however, American AI firms will feel pressure to outsource the development of strategically critical facilities—likely to oil-rich regions such as the Gulf that run on dirtier fuel. For Washington, any prospect of offshoring AI should set off alarm bells. An American company shifting advanced AI training to a foreign country, especially an autocratic one, would pose huge risks as AI begins to power more of the U.S. economy and to play an integral role in defense. If a host country became unhappy with American behavior, it could punish Washington with the flick of a switch. A failure to build domestic energy capacity would thus echo outsourcing mistakes of past decades in other important industries, such as semiconductors, in which the United States is now dependent on foreign suppliers.

The United States has the technology and industrial capacity needed to build new energy facilities. But it remains inhibited by a thicket of government and utility regulations and by procedural delays—some backed by good reason, some not. These restrictions impose huge delays in interconnection (the process of connecting a new power source or data center to the grid) and require years-long environmental assessments. On top of federal and utility hurdles, state and local policies can be cumbersome, especially for projects that cross multiple states, such as transmission lines. Companies—not citizens—should pay for the energy build-out, but government policies must make it possible for them to undertake these projects on reasonable timelines.



Data storage tapes at a computing center in Berkeley, California, May 2025
MANUEL ORBEGOZO / REUTERS

Infrastructure is not the only domain in which American policies hold back the AI sector. AI progress depends as much on talented people as it does on technology and computing power, which is why the Biden administration acted aggressively to enable people of extraordinary technical ability to come to the United States from all over the world. This effort included elevating AI and other high-tech fields as priority areas for visas, as well as updating eligibility criteria to welcome top-tier scientists.

Here, too, Trump has sometimes said the right things but always fallen short in practice. During his 2024 presidential campaign, he declared that when foreign students graduate from U.S. colleges, they should "automatically" get a green card. Instead, under his leadership, the United States has begun shutting itself off to foreign workers and students, intimidating even those who have visas and are in the country legally. In September, the administration even said it would slap a $100,000 fee on applications for H1-B visas—the visas most commonly given to high-skilled immigrants. Such actions are already having

adverse effects. For example, preliminary research by NAFSA, a nonprofit association of international educators, suggests that in 2025, American universities will suffer a 30 to 40 percent reduction in international enrollment.

If Washington cuts itself off from foreign-born scientists or sends them back home, the consequences will be catastrophic. The United States leads the AI race in large part becaus it attracts experts from across the world. According to a Georgetown University study that looked at AI research from 2010 to 2021, 70 percent of top U.S.-based AI researchers were born abroad. Sixty-five percent of leading U.S.-based AI companies, as ranked by *Forbes*, have at least one immigrant co-founder. Before the current Trump presidency, 70 percent o the students enrolled in American AI graduate degrees hailed from abroad. Historically, th vast majority of these students have stayed, often making critical contributions to America industry and academia. But thanks to Trump's policies, many of today's students might instead return home. Some could head to China, which has spotted an opportunity to recruit AI experts—and which poses the most significant challenge to the United States' A interests.

THE CHINA CHALLENGE

Protecting American AI leadership is not just a matter of pride. It is essential to U.S. national security and economic competitiveness. China has been making strides in AI development, and although none of its firms can yet match the best American ones, they do not lack for technical talent.

China faces one devastating disadvantage in this competition: its inability to make large quantities of advanced AI chips, a weakness exacerbated by U.S. export controls that began in the first Trump term and that Biden greatly expanded. But after heavy lobbying by industry, the second Trump administration has started dismantling this area of bipartisan consensus. In July, for example, it reversed its April decision to cut China off from newer A chips, and the president indicated plans to undo other Biden-era controls, as well. Such moves will almost certainly accelerate Chinese AI development.

Beijing has acted decisively to ensure that, if it gains access to such chips, the rest of its AI ecosystem will supplant the United States'. Consider energy. Beijing has made extraordinary investments in power plants, energy storage, and energy transmission. As a

result, China now produces more than twice as much electric power as the United States, and its lead is expanding. In some individual months, the country has installed over 90 gigawatts of new, clean energy capacity—almost double the amount of energy American AI firms will need in the next several years.

Beijing has also gained an edge by fusing its AI industry with its national security apparatus. The U.S. Department of Defense has said that major Chinese AI companies, such as Tencent, are key pillars of China's military-civil fusion strategy. AI systems like these have broad utility to military and intelligence agencies: they can support weapons development, cyber-operations, and domestic surveillance, among other tasks. In exchange, the Chinese government has provided technology firms with extensive policy and security support. Historically, this assistance has included defense services and passing along industrial secrets stolen from American businesses.

Losing AI leadership to China would also cause tremendous global harm. Today, consumers across the world benefit from the rigor and transparency of U.S. regulations and standard setting, often developed in concert with other democracies, in many technological domains. For example, new technologies such as electric vehicle charging have required global collaboration on standards. AI will require similar partnerships, and it is in Washington's interest to take the lead. Otherwise, there is a risk that autocracies will unilaterally set the standards. If Washington cannot establish more and better relations with its AI sector and secure its global AI leadership, the Chinese national security apparatus might shape global standards to adhere to Chinese censorship rules.

THIS FOR THAT

Right now, strong partnerships between AI firms and U.S. national security agencies are few and far between, and those that do exist are in the early stages. To address this deficiency, the government will need a better understanding of what AI is and how it functions. The government can help U.S. industries, but only in sectors it comprehends deeply, and today AI is not among them. In our time in government, we worked with some truly tech-savvy civil servants and military officers who drove bureaucratic change. The Biden administration hired hundreds of AI experts to bolster their ranks. But many have been fired or have left government in recent months, including many of its top technical staffers. Washington will have to do a better job of securing AI talent and reverse the trend

that Trump has set in motion.

Industry leaders, meanwhile, must make sure U.S. officials understand their work—and they will have to be more responsive to Washington's needs. AI leaders in Silicon Valley might be reluctant to cooperate more closely with the government, given its technical ignorance and bureaucratic sluggishness and given that they have enjoyed such success flying solo. But executives should remember that collaboration between industry and Washington often works out well for everyone. U.S. President Franklin Roosevelt's administration teamed up with Ford Motor Company in the run-up to World War II to produce the B-24 Liberator heavy bomber, generating both revenue for Ford and much-needed aircraft for the military. The Manhattan Project would not have succeeded without DuPont, General Electric, Chrysler, and other corporate stalwarts, all of which profited by providing help. And the invention and refinement of radar, satellites, jet aviation, microprocessors, and the Internet all flowed from corporate-government teamwork.

One critical area in which the government can help companies is security assistance. Because AI is increasingly essential to national security, foreign intelligence services are redoubling their efforts to steal innovations from U.S. tech businesses. In March 2024, for example, the U.S. Department of Justice indicted Linwei Ding, a Google software engineer, for allegedly passing Google's AI chip designs to China. Beijing has also worked hard to smuggle in advanced American technology products, such as AI chips themselves.

Chinese spies will grab whatever AI secrets they can from American businesses. But they have good reason to focus on stealing model weights—the numbers that encode a trained AI model. Chinese companies can then avoid the training costs, as well as cut down on the time it takes to develop models. For example, AI companies work hard to discover algorithmic tricks that enable them to use their computing power more efficiently. For China, with its severe constraints on computational power, such multipliers are almost certainly incredibly valuable. These multipliers are likely far less well defended than the core secrets of past eras, such as during the atomic and space ages, simply because the government has been largely uninvolved in their development.

If Washington cuts itself off from foreign-born scientists, the consequences will be catastrophic.

AI companies bear the primary responsibility for defending their networks and organizations. But despite the recent loss in expertise, the U.S. government has cybersecurity capabilities that businesses cannot match, and it should provide significant assistance, as it does today to firms in the defense industrial base and critical infrastructure sectors. Such aid could include intelligence about foreign hacking attempts, support in vetting international talent, and guidance on security procedures. Companies that work directly with the U.S. government on national security should meet stringent standards, akin to those imposed on other defense contractors.

To meet their side of the grand bargain, AI businesses should help the United States incorporate their technologies into the national security apparatus. This lack of integration is a persistent area of weakness for Washington. American companies may lead the world in inventing AI, but without such cooperation, the country will fall behind in adapting AI for military purposes, which could prove devastating in a conflict. Time and again throughout military history, states that fail to integrate new technology into their armed forces have wound up suffering. France and the United Kingdom, for example, invented the tank during World War I but paid a heavy price when the Germans were the first to master its use, enabling their powerful blitzkrieg offensives of World War II. To guard against a similar outcome, the Department of Defense and intelligence agencies must procure and use frontier AI systems, and they will need hands-on guidance from private-sector technical experts to do so effectively.

Biden created an outline for how such cooperation might function in October 2024, when he signed a national security memorandum directing the government to bolster its use of AI for national security purposes. This document included strict guardrails to ensure that the technology does not enable human rights abuses, domestic surveillance, or other unethical activity—protections essential to earning the trust of private-sector AI developers and the general public. Unlike with many of Biden's other executive actions, Trump has not yet repealed it. But the president has made little progress on its many important provisions and has ousted some of the senior nonpolitical experts vital to its implementation.

As government and industry develop a better national security partnership, they will need to pay particular attention to how AI can solve pressing problems in Washington's competition with Beijing. There are, for example, many ways the United States could use AI in cyber-operations, as illustrated by the AI Cyber Challenge conducted by the Defense Advanced Research Projects Agency. The country that more quickly and effectively integrates AI into the cyber-domain will likely prove better able to protect its own network and penetrate others', unlocking a huge intelligence advantage. AI could also improve Washington's capabilities in other security domains, such as geospatial intelligence, signals intelligence, logistics, and weapons design. But nothing will happen without clear direction from government and meaningful engagement by American AI firms.

STRANGER DANGER

There is a final reason why the U.S. government and leading AI companies need to build closer ties: together, they will have to weigh tradeoffs between upsides and risks. Within the technology world, almost everyone agrees that although AI could offer tremendous benefits to humanity—curing diseases, advancing clean technology, eliminating grunt work—it could also cause massive harm. Some of these risks, such as the use of AI by authoritarian states to reshape the global order, can be avoided by preserving and expanding Washington's lead over competitors. But other dangers are trickier to mitigate. Several top AI thinkers, for instance, believe it is plausible, or even likely, that a single malicious user could someday harness powerful AI to engineer a deadly novel pathogen. Others worry that even in benevolent hands, powerful algorithms might cause catastrophic accidents by taking actions that their creators do not intend. Then there are less fantastical but still dire outcomes, including massive unemployment, acute concentration of economic power, and discrimination in sectors such as health care because of biased models and training data.

From a policy perspective, any one of these scenarios would pose a challenge of historic proportions and force difficult tradeoffs. For example, in a hypothetical world in which a single AI user can cause catastrophic harm, the government will have to consider sweeping regulations on the development and use of sophisticated systems, even if that slows innovation. And if AI automates away a large share of human work, the government may have to spend large sums of money retraining the workforce, or it might have to facilitate a

restructured economy. Given the pace of AI progress, policymakers will have to make these consequential decisions under exceptionally tight timelines—all concerning a technology that government is not inventing and about which it knows worryingly little.



At the World Artificial Intelligence Conference in Shanghai, July 2025
GO NAKAMURA / REUTERS

Deeper collaboration between the public and private sectors, as well as with civil society, does not guarantee that the state will make the right calls. But it does give Washington a fighting chance of securing a net-positive outcome. With a stronger technical foundation, officials can better understand how reliably AI systems follow instructions, how they handl dangerous tasks, in which areas they can replace human labor, and to what extent they favo offense versus defense in security and safety domains.

The rise of the Center for AI Standards and Innovation at the Department of Commerce (founded as the AI Safety Institute under the Biden administration) represents a valuable

initial step to build meaningful collaboration. Since its inception, CAISI has brought together government officials and companies to collaborate on safety issues. It has also aided in the development of standardized testing mechanisms for AI. CAISI has worked alongside other agencies with domain-specific expertise to carry out additional voluntary testing on particularly critical topics, such as partnering with the Department of Energy and the AI company Anthropic to assess whether frontier AI models have dangerous knowledge about nuclear weapons. CAISI featured prominently in Trump's AI Action Plan and the administration must empower it to carry out voluntary collaboration with companies, to set standards, and to conduct safety testing.

Thanks to CAISI's work and the voluntary commitments that leading AI companies made to the Biden White House, AI firms have already promised to conduct independent safety testing of their models, often based on CAISI guidance. In some cases, companies have even agreed to grant CAISI access to new systems before they are released and have praised the government for the national security–specific expertise it has offered in return. Both sides should deepen this collaboration, spending more time and resources building high standards and conducting rigorous assessments of new models.

FROM THE GOVERNMENT, HERE TO HELP

Grand bargains often work better as tag lines than as policy, and getting the right kind of deal when it comes to AI is easier said than done. The technology, after all, is rapidly progressing along an unpredictable path. As AI improves, ever-larger amounts of infrastructure, power, and money will be required; the need for improved security from foreign intelligence threats will increase; and the urgency of collaboration with the defense apparatus will grow. So will the risks of misuse, prompting new policy tradeoffs. More startups will arrive on the scene, and legacy companies that today look unstoppable may fall by the wayside. Everyone involved in the AI world should prepare for constant renegotiation and rebalancing. U.S. officials, for their part, will almost certainly have to remain agile, experimenting with different AI policies as time goes on.

But amid this uncertainty, it is imperative that Washington take a more active role in enabling and shaping the American AI ecosystem. The technology does not need to develop as nuclear weapons did—under strict state control—but Washington cannot sit this one out. Instead, AI should perhaps evolve as the American railroads did in the 1800s. The

private sector handled most planning and construction, but the government played a vital role, as well. It organized laws and permits for building the infrastructure. It passed carefully calibrated, common-sense government safety requirements—such as standardized track gauges, rules for the use of air brakes, and requirements for car coupling—which all helped make trains both faster and safer. The collaboration was not perfect, but it worked: American railroads became a national asset that increased the United States' security and prosperity. Advanced AI, too, can promote U.S. power and interests, provided it is developed in the right way and under the right set of arrangements. Now, as before, it is time for the public and private sectors to stand shoulder to shoulder.

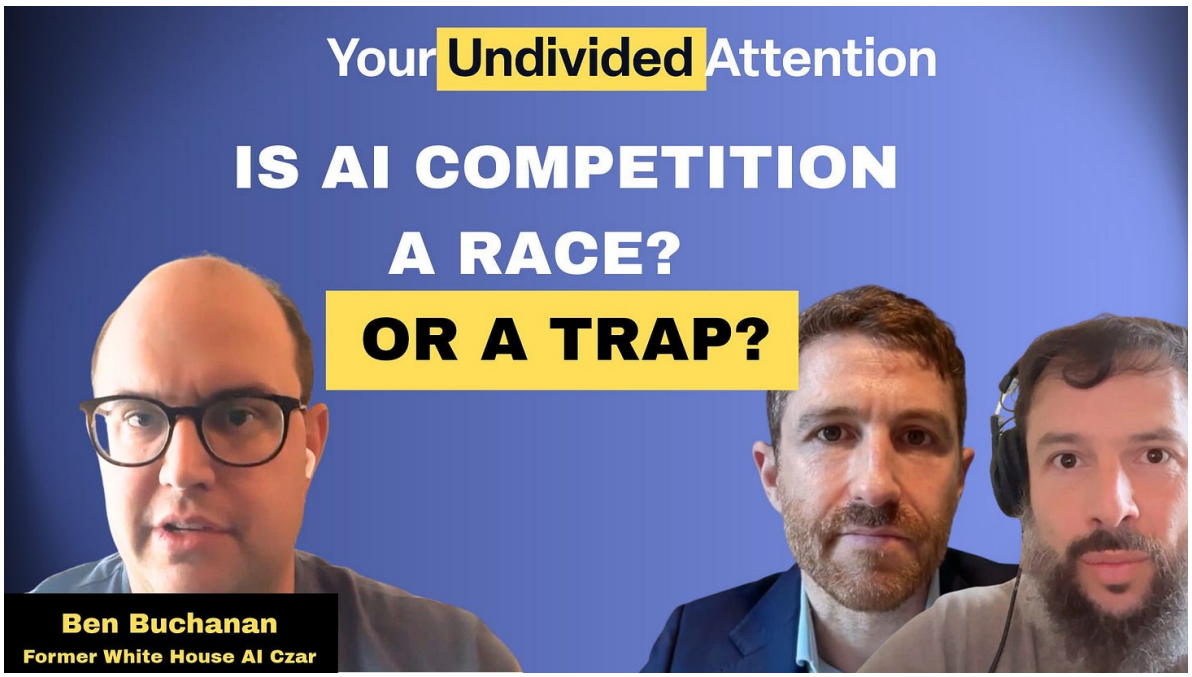
Copyright © 2026 by the Council on Foreign Relations, Inc.
All rights reserved. To request permission to distribute or reprint this article, please visit ForeignAffairs.com/Permissions.
Source URL: https://www.foreignaffairs.com/united-states/artificial-intelligence-grand-bargain-buchanan-collins

centerforhumanetechnology.substack.com

Ben Buchanan Wants the U.S. to Win on AI. But Is It a Race or a Trap?

Center for Humane Technology

64–81 minutes



Tristan Harris: Ben, welcome to Your Undivided Attention.

Ben Buchanan: Thank you so much for all the work you guys are doing, and thank you for having me.

Tristan Harris: So you served as President Biden's top AI advisor, and specifically you were the only official, as I understand it, who was at every single White House meeting with the frontier AI CEOs, including in January of 2023 when Sam Altman demoed GPT-4 before its release. And it's a story you write about in your book in the opening chapter. I'd

love just to bring our audience into that moment. Can you paint a picture of what was that moment like in January 2023 before the world really understood what this was going to be about?

Ben Buchanan: I remember this moment very well. It was in the diplomatic reception room in the Eisenhower building. Sam was coming to meet the national security advisor, Jake Sullivan, and they had a really good conversation that I think made real a lot of the things that folks like me had been talking to Jake about up to that point. And this was a moment when the United States had already taken some action on AI, and I think people like Jake deserve a lot of credit for being willing to be forward-leaning on things like chip controls and the like, even before we had ChatGPT come out in November of 2022. But I think seeing GPT-4 up close helped convince Jake and helped convince the broader set of US policymakers who had so many things to deal with besides AI that AI was going to be really important. And certainly for me, as someone who had felt for a long time that AI was going to be really important, seeing GPT-4 and everything that came after it made me realize that the future we had all anticipated was really starting to arrive.

Tristan Harris: I'm just curious, what was the unlock? What demo did Sam give? What capabilities that... Just to take people back, because now we have GPT-6, which people can reverse engineer all of Adobe Photoshop and make a replica in an instant. What were the capabilities back then that felt impressive or breakthrough?

Ben Buchanan: Of course now, as you say, they all seem totally mundane, but at the time it was crazy that GPT-4 was multimodal, so it could do image analysis, you could give it an image and look at it. I remember something with a press conference where it had analyzed a press conference of Vladimir Putin's and had spit out the points of it

and the key questions intelligence analysts could ask. So it was a well-crafted demo, and it prompted a really good conversation between Jake and Sam and the whole group of us there that I think it really felt like the meeting of these two worlds that I'd had one foot in either side for a long time, the world of Silicon Valley in technology, and then the world of national security, they were now entangled. And I think for better or for worse, they're entangled forevermore.

Tristan Harris: And so looking back three years later, the world has completely changed. And I'm just curious, how does that meeting land in your mind's eye now in retrospect, given where we are?

Ben Buchanan: I think it feels to me more symbolic from a policy perspective than from a technology perspective, certainly in retrospect. And for folks like me who had lived this technology for a long time and who came to the White House in 2021 with a very deep conviction that AI was going to matter a lot for national security, GPT-4 felt on trend. It's always striking to see things like this, but it felt on trend. But it did feel to me like a moment where the US government was now going to move into a new dimension of what it could do. And indeed that's a lot of what happened, that the US government in the spring of 2023 reorganized itself. They created this position, White House Special Advisor for AI, which I assumed, and there's a lot of direction and opportunity to get things done through the bureaucracy that otherwise wouldn't be there.

I don't know if I would peg it all to that meeting, but I think I would peg it all to that moment in time in the spring of 2023 when the US government said, "Wow, this is not just a national security China issue. This is an issue. It's going to affect everything, foreign policy and domestic."

Aza Raskin: And this is something I've always really appreciated about your work. I mean, obviously you've been very deep in the Biden administration, but you also view AI as transpolitical, and you work with people on both sides of the aisle, including President Trump's former AI advisor, Dean Ball. I'd love for you to just tell us about some of those relationships that cut across all administrations and the New York Times op-ed you just recently wrote.

Ben Buchanan: Anyone listening to this show who knows your guys' work knows that these issues are going to affect everyone, and this is not a partisan thing and it shouldn't be a partisan thing. So when we were in the government, I think people like Tarun Chhabra deserve a lot of credit for setting a tone of we're going to work in a bipartisan fashion on these issues, especially on the national security issues where there's a lot of alignment with Republicans. And then when I transitioned out of the government, of course, with the 2024 election, I made it a point to try as much as possible to go to folks in the Trump administration and say, "Look, we're not going to agree on everything. I'm no fan of Donald Trump, but you guys now have, I think, a very solemn and very sacred responsibility, and I'm here to help you however I can.

I've lived this for four years," and some of them were open to it, some of them were less open to it for sure. But you mentioned Dean Ball, and I think Dean, who's now at OpenAI, gets a lot of credit for someone who I'm sure we don't agree on every issue, but was really willing to look at the data, hear from folks who have worked on this. Dean and I have done a number of panels together and a New York Times op-ed together, as you mentioned, that try to show there's a lot that unites us on this one. I think that's important because there's a lot we can agree on here in a bipartisan way.

Tristan Harris: Just to add here, Dean has called you a tireless and ardent patriot, and I just want people to hear this conversation as not just one political side of the story. This is about people who are moving between administrations who have this solemn and sacred responsibility. Well, you wrote this book with your co-author called The Bitter Struggle, and it's out now, and I want to make sure that we are covering essentially the front-row seat that it gives us to the AI policy discussions and decisions that you made in the White House. But I think to start with first principles, what is the bitter struggle?

Ben Buchanan: The bitter struggle is a play on this idea of the bitter lesson, which is the key technical idea in AI. We see the bitter struggle as having three parts. The first is for the United States and democracies everywhere to have an advantage in inventing this technology. We can talk about why that's so important later, but essentially making sure that this is an area in which the technological course that humanity charts is one that democracies have a very large say in. The second is taking that advantage and converting it to a national security edge and saying, "Can we make sure that the United States can use this technology well for intelligence, military affairs, assure its position in the world?" And then also that we don't really talk about the economics of this, can it apply it to its economy as well in a way that's productive for national power?

And then the third is really the crux, and the third is can it do all of these things in accordance with our democratic values? And this is where the bitterness comes in because there are some very difficult trade-offs here that have to be managed. We've got really big, powerful companies, but that raises the risk of concentration of power and monopolies and the like. Another danger, as you guys know very well, is the misuse of this technology. How can it be

misused for cyber or bio risk by bad actors? Well, we can crack down on that, but that also creates the risk of infringing on civil liberties and the like. We want to go fast, but also if you go too fast, you cut corners on safety, which creates misalignment risks and the like. So no single book in AI is going to have all of the answers, but I think the strength of this book is to take folks in the room, say, "This is what happened in the Biden administration, this is what's coming next, and this is what we see as being at stake, as all of humanity and especially all of democracies wrestle with this new technology."

Aza Raskin: Yeah. Well, some of what I hear you saying is that technologies sort of have an inflection. They lean towards open or they lean towards close, and that one of the unfortunate things, the bitter struggle is that this thing that open societies are creating to strengthen themselves is actually forcing us in some sense to lean towards closed kinds of systems. And just give a really specific example from the last couple of weeks, now that we've had the PHP bugs undermining 70% of the internet, banks all across the world getting hacked by open source models, it is the openness of our models that is forcing us to move to some kind of surveillance state, which is terrifying. AI enables something that's never been possible before, which is listening to everything everywhere, understanding what everyone is saying, how people are operating. You can do the Flock cameras at scale, and so ubiquitous technological surveillance becomes almost trivial.

The temptation for governments to do ubiquitous surveillance becomes almost insurmountable. I'm just curious, what is your wisdom from sitting in government and what you see for how open nations, especially as we get the perfect storms of more and more people losing their jobs, at least in the transition and their livelihoods and disenfranchisement going up, people losing political and economic power, how do nations resist the

temptation of total surveillance? How do we stay open?

Ben Buchanan: I think this is a challenge that democracies have wrestled with for every technology since the dawn of America. At the simplest level, AI is going to force us to reconsider how we've done that and find a balance again. I think especially part three of our book is at least an articulation of what are the different parameters that we have to balance here. We don't pretend to have all of the answers, but can we put them all on the table and lay them flat for us to look at? But the internet forced this as well, and things like the FISA Surveillance Act forced that, even the Fourth Amendment at some level is trying to reckon with how do we have the privileges and the dynamism and the freedom of an open society while also making sure we keep people safe. I think AI is just bringing back these questions in a very acute, very novel way that government, to be clear, is not ready for and many of the companies are not ready for.

I'm very stressed about the way in which AI can entrench autocracy, not just in autocracies, but also could have backsliding in democracies, including in the United States. You mentioned Flock cameras, I think that certainly is one vivid example. This was a thing we worked on in the government and the president's executive order in 2023 had a whole tasking to the Department of Justice to think about the appropriate use of AI in the criminal justice system and in law enforcement. So it's a policy question for us, and it's a big theme in the book. In the third part of the book where we talk about democratic values, we talk about what we call the bitter temptation, which is democratic governments maybe for legitimate reasons, maybe for not legitimate reasons, sliding into this world of using AI to repress their citizens or otherwise entrench power. I think there's a sense in which AI can compound or concentrate power, and we are

allergic to it in the sense of the companies being too strong and the government being too weak, and also allergic to it in the sense of the government being too strong and the citizens being too weak.

There's a very narrow path here between the two. I don't pretend to have wisdom or the answer, but I do think at a minimum, what we've done with this book is articulate the problem in a fair amount of detail, and what we tried to do in the government was set the foundations of policy to keep us in that narrow path.

"AI can compound or concentrate power, and we are allergic to it in the sense of the companies being too strong and the government being too weak, and also allergic to it in the sense of the government being too strong and the citizens being too weak. There's a very narrow path here between the two…and what we tried to do in the government was set the foundations of policy to keep us in that narrow path." - Ben Buchanan

Aza Raskin: I think it's really important to pause for a second and just define the bitter lesson because this is so central and so counterintuitive to most people that to think paradigmatically AI really requires internalizing the bitter lesson. And so could you just describe what that is and then from the emotional sense, why do scientists think it's so bitter?

Ben Buchanan: The bitter lesson is very closely associated with a technological idea called the scaling law, which has dominated AI over the last 10 or so years. And that's if you've got a general purpose algorithm, in this case is something called the transformer, then you can scale up its capability with computing power. And the more computing power you put into training the system, the more powerful the system that results. This was a controversial idea. We only show up in the White House in 2021, and there's a lot of people there saying, "No, it's all about the algorithms or it's all about data."

I think it's fair to say that really the entirety of certainly our national security policy on AI, if not our entire policy on AI, was premised on this idea that the scaling law and the bitter lesson would be right, and that we would see very powerful AI systems due to scaling up of computing power.

Thus far, that's what's happened.

Tristan Harris: Just to take people back, not everyone would've believed that if you just keep scaling up masses, 10Xs, 100 Xs of compute, that you're going to go from something that can barely spit out essays to solving Navier-Stokes and hundreds of math problems and creating material science breakthroughs and better hackers than the NSA, mostly by just scaling the amount of compute that's actually there. And you were one of those believers, it sounds like, in 2021. Can you just take us back to what had you believed that, whereas many other people didn't take that seriously? Because there's a similar kind of disbelief in the ecosystem now that are we really going to go from a GPT-6 to a GPT-9? I mean, just think about how much more capable. I think it's important for building people's intuitions because here we are where GPT-6 is kind of the most frontier model-ish, and we're trying to assess where would we be in two, three years if we don't steer in any way?

And it's as hard to imagine that as it was to imagine really sitting in the shoes of 2023 and GPT-3, GPT-4, that we would really have what we have now.

Ben Buchanan: For me, the moment that I was convinced was the jump to GPT-2, which was, I think, June 2019, and then the jump from GPT-2 to GPT-3. GPT-3 was 100X scale up, and May, I think, of 2020, GPT-3 comes out. Think about where America was then, COVID lockdowns, George Floyd was murdered the next week, a lot

of chaos in America. But I do remember reading the GPT-3 paper and saying, "Wow, this creates a pathway for us to go really far as a species in AI." In some ways, this converts AI development from a scientific process, though it is certainly that, to an industrial equation. Can you acquire enough computing power and energy to run the chips? If you can, you can almost predict where the systems are going to be. And it's worth fast-forwarding, Tristan, as you said, to where we are today and thinking about how much the scaling law has grown since then.

Now, if you look at the top five AI firms are spending, I think, $700 billion on computing infrastructure in 2026 alone. That's akin to a Manhattan Project every two or three weeks. That's more than the United States spent building the entire interstate highway system over 35 years. Just a mind-boggling amount of money coming from private firms in just one year. That's the scaling law and the bitter lesson at work.

Tristan Harris: This is basically a hundred billion dollars roughly per year per lab of the major labs doing the scale-up, which is really illustrating this GPT-6 or GPT-9 thing. And you were just mentioning one thing that distinguishes this technology, you compared it to the size of the funding of the Manhattan Project, but the Manhattan Project was created by the government. The government had the expertise in the frontier knowledge of nuclear science. The private sector didn't have that. Whereas you make this distinction in your book, this is the first technology that is so transformative, but that's coming from the private sector. Do you want to speak to that?

Ben Buchanan: That's right. And this is, I think, the through line for every AI policy decision that we made in the government was this is the first technology, probably the last hundred or so years, that really

comes from the private sector. And that's in contrast to, of course, nuclear weapons, a lot of the decades of space exploration, early days of the internet, early days of the microprocessor, early days of jet aviation, so much of those previous revolutionary technologies, and of course their attendant national security implications, have their roots in Department of Defense or intelligence community funding. Oftentimes, the project did the work or universities did the work, but it was the government that cut the check. And what that meant was threefold. One, the government had the capacity to understand it because someone's accountable for it. Two, the government had some capacity to shape it. And three, they often were born with national security instincts in mind, so protection from counterintelligence, threats, and the like.

None of those are true for modern AI.

Aza Raskin: And so this becomes a race for which country or which company can hit the fast forward button. Who can move faster through time to get the bigger breakthroughs, the more power, the more military advances, the more economic advances, more GDP? So this is a race through time with capital and compute. Now, we've been critical of the framing of a race because it has certain externalities and consequences for thinking of races, but I think it's really important to talk about your framing, open versus closed societies, which of course really means US versus China as the two big superpowers, and just flesh out the full race, give the full best argument for why we need to be in a race with China for AI.

Ben Buchanan: You won't usually hear me use the word race, but you certainly will hear me use the word competition. The theory of the case that we had since we got there in 2021 was that it would be important for the United States on behalf of democracies and at some

level on behalf of humanity to have a leading role in shaping AI development, and that we felt it was the democratic system that would be best positioned to manage this technology. What that meant, what that conviction meant is that we wanted the lead to be substantial over China, often by slowing China down rather than accelerating ourselves unduly, which I think is a really important distinction here, to build a very big lead such that we could, to use our phrase from 2021, spend the lead on safety. So 2021, there were no real dangers of these systems, and we knew that, but we wanted to use that period, essentially the Biden term, to build the lead such that when we got to 2026 and 2027 and these systems, as we thought, became more capable and also more dangerous, there wouldn't be a geopolitical arms race with China.

It would essentially be a one-player game, and democracies would be alone, the United States would be the leader of the democracies. And then we could, without the geopolitical pressures of competing with the Chinese, regulate the AI companies and say, "Look, we're going to slow down here. We're going to make sure we're not cutting corners on safety," all this kind of stuff within the US context where it's just American companies that are doing this. I think that theory continues to be right. I think the evidence has borne that theory out. The execution has been imperfect, especially in the Trump administration. But I think the theory that we have, and certainly the pitch of this book is you want American and democratic preeminence of this technology, you don't want a close arms race with China, and then you want to make sure you're regulating your company so they don't cut corners.

If you want to put this in historical terms, I think the analogy that resonated very well was one that was near and dear to President

Biden's heart off the side of the Oval Office is the president's private study, and on the wall of the President's study for President Biden, he had President Kennedy's speech from 1962 at Rice University. This is a speech where Kennedy says, "We're going to the moon coming back within a decade." And I think that speech matters less because of the moon promise, though of course that matters for history, but because of how at the end Kennedy's making the case to the American public why they should support this. Kennedy says something like, "We don't know whether space will be good or ill for humanity, but only if the United States has preeminence can we help decide?" And that really is the ethos that gets to our policymaking of making sure the United States and democracies can shape how this technology goes for humanity.

"The theory of the case that we had since we got there in 2021 was that it would be important for the United States on behalf of democracies and at some level on behalf of humanity to have a leading role in shaping AI development, and that we felt it was the democratic system that would be best positioned to manage this technology." — Ben Buchanan

Tristan Harris: It's an interesting parallel here to something you hear from folks at Anthropic, which is that at Anthropic, we care about the most about safety. And by the way, we can actually only influence safety if Anthropic is leading the race because basically we won't have the leading seat at the table or the dominant kind of position at defining the policy if we're not leading. So it's a similar kind of parallel to the US believing basically we can only set the terms or regulate AI, but there is a contradiction in there, which is that in order to race to be that far ahead, many times I have to take shortcuts, but there has been a cost to that. I'm just curious how you both... Steelman, what have we gotten right about that framing and if you have any self-critiques around where those

policies haven't played out the way that we thought?

Ben Buchanan: In the context of US-China, I don't think that contradiction is particularly salient because the United States has the capacity to slow China down. So this was our theory, which is you don't need to race ahead and cut corners on AI development if you can slow the Chinese down. China is totally incapable of making advanced AI chips. You don't have to take my word for it. Take Secretary Lutnick, Trump's Commerce Secretary's word for it. He testified in 2025, China made 200,000 chips. That's not enough for a single data center. All those chips are also worse than their American designed Taiwanese made counterparts. So one of the critical actions of which I'm very proud, and then again, a small group of us did this, was a series of export controls before ChatGPT, before Sam Altman came to the White House in 2023, that set out to deny China the access to advanced computing power because we felt like that would take China off the game board in making these systems and their military implications and the like.

Tristan Harris: The Chinese companies themselves say out loud that the number one thing restricting their progress is having access to computing. I think Kimi and DeepSeek have all basically said this, so they're admitting that this is the key to their ability to continue to scale behind the US.

Ben Buchanan: They were saying that when we were in the White House in 23 and 24 as the compute controls were really starting to bite, and then really striking, there was a panel of them at an AI conference in January of 26, and they basically just went down the line, one AI executive after the next, talking about how important computing power was to AI progress and what a struggle it was to get it. I think the export controls, especially in the last couple of

years, have been deeply imperfect in terms of how they've been enforced and Trump reversed them and so forth. But even so, I think this is one of the biggest impositions on China, and this is not something that accelerates United States unduly or anything like that. This is a policy choice to slow China down. And my argument is if you're someone who worries a great deal about AI risk, you really do want this to be a one-player game because it's much more tractable to regulate American AI companies than it is to regulate Chinese AI companies.

Tristan Harris: Do you want to then speak to how we should consider the fact that Jensen Huang from Nvidia has been the primary lobbyist for saying, "Hey, we should actually sell all of our American chips to China?" Isn't that like selling uranium to Soviet Union to boost GDP for American uranium companies and say, "Well, at least the world's running on American uranium?"

Ben Buchanan: It doesn't make any sense to me. And in fact, I think it's even worse than the uranium analogy because the Soviets had their own independent ways of getting uranium. And as we've talked about, the Chinese do not have an independent way of getting computing power, so we're giving them the one thing they can't get. To be fair to Jensen, let me give what I understand to be the steel man version of his case.

Tristan Harris: Let's zoom on that argument.

Ben Buchanan: Yeah, and certainly how Nvidia has made this case to us and to the public. They will say, sometimes it's called the run faster argument, which is the United States just needs to stay ahead of China. One of the ways in which American companies stay ahead of China is we keep Chinese dependent on American chips and

American technology, and we profit from those sales. And Nvidia can take the money it makes selling chips to China, reinvest that money into its R&D ecosystem, and that'll keep America running faster. Because we believe so deeply in the scaling law in 2022, what we said to Nvidia and what we said to everyone who made this run faster argument is, "We've got more faith in your technology than you do, and we think there's going to be overwhelming demand for this technology and marry that to a fixed supply, and we want the demand to go to the United States and to democratic companies and countries." The response, of course, is like, "Well, this is very bad for our business.

You're taking away the Chinese market. That's going to hurt us. We can't run faster." Our view was, "No, no, the AI tailwind here is going to be so strong. The demands of the scaling law and the bitter lesson will be so acute that your business is going to do great." It depends on how you look at the exact stats, but roughly speaking, from the 2022 chip controls to today, I would bet Nvidia's up by 15 or 20X in terms of market capitalization, from something like a 300, 350 billion company to now something like a $5 trillion company.

Aza Raskin: If all of the Chinese AI companies are compute-limited, why do we keep hearing about, "But what about China?" How do these two things fit together in the same sentence?

Ben Buchanan: I think the what about China argument can be disingenuous if it is an excuse for saying we don't need to regulate American AI companies or we should accelerate in a totally untethered way here in the United States. Where I think the what about China argument has some merit is we need to make sure we're actually enforcing the export control policies to make sure we build this lead and keep this lead such that we can regulate our company.

So I am very clear-eyed on the risk that China poses to American national security is there's no doubt in my mind that if China had gotten to Mythos first and found all these cyber vulnerabilities, China would not be turning them over to software vendors to have them be patched to make the internet safer. So we should not dismiss the risk here of Chinese intelligence services and the Chinese military, but nor should we use that as an excuse to say, "Well, we just have to be a total accelerationist about this." The profound irony to me is that there's people like David Sacks, my successor as AI advisor, who simultaneously end up, I think, on the wrong side of both arguments and say, "Well, we need to sell the chips to China, which keeps the Chinese in the game, but also because the Chinese are in the game, we can't regulate here in the United States and we should let this technology go totally unchecked." And to me, the opposite is true.

By denying China access to computing power, by making it a game just for democracies, then you can regulate more effectively.

Tristan Harris: So Ben, you have a term for the place that you'd like us to end up on AI, which is hegemonic stability. What does that mean and what is your preferred landing spot here?

Ben Buchanan: Hegemonic stability is a well-known term from the political science literature, which is that in moments where the world has stability and peace, it often is because there is a nation, typically, though not always, a nation with a technological advantage that has chosen to use that advantage to provide security and to provide stability to the system. And what that nation is doing is it's taking less benefit for itself and setting up a system for which all nations can benefit and in return getting stability. You can think, I think of some of America's most far-sighted policymaking was in the post-World

War II era where America had the nuclear bomb, unlike Europe, unlike Japan was not destroyed, had this economic and political and military preeminence, but then channeled that strength into setting up a number of institutions that ultimately provide security for the world to this day. NATO, the UN, the IMF, the World Bank, and then many of the Cold War institutions that came out later, the IAEA and the like.

And I think our vision here for the United States is we want the United States to have preeminence and to work with democratic nations to have democratic preeminence in AI. And we want the United States to channel that strength and that power into stabilizing institutions that are multilateral and that encourage nations, including China and Russia, nations like that, to say it's better to be in the system than fight the system. And I would argue a nation like China benefited enormously from the post-World War II and Cold War system the United States set up that enabled the global trade that let China rise in the way that it rose. So that's the closest thing to an end game I can see. And I think there's a moment here in which the United States can build a lead for itself and for democratic nations and then can convert that to something that is stabilizing and beneficial for all nations.

Aza Raskin: So I think now then is a really good time to set up the stakes of what does it mean to win the competition or lose the competition? What are the benefits of winning? What are the dangers of losing? Why does this feel so completely existential?

Ben Buchanan: There's many national security benefits from being ahead in AI. The one that's most obvious is cyber operations. This was the one we always thought would be first. I started working on AI and cyber really more than a decade ago, depending on how you

count, because it seems so apparent to me that cyber operations would be an area in which AI would be quite good. And we should make no mistake about this, that the United States and China compete every single day in cyber operations and playing offense and defense, and it all occurs below the threshold of armed conflict and all that, but it confers a significant national security advantage. And if you believe in American national security, you need to believe in American cyber operations and foreign intelligence and the like. I do want to say, because I recognize I'm coming off as a China hawk, and I'm sure we can come back to this later if you want to do it in a different timing, but competing with China is not at all incompatible in my mind with engaging with China.

And the only competitive action that we took was simply saying, we're not going to sell you some of the most advanced technology humanity's ever made. It's not saying we're going to infringe on your territorial integrity. It's not saying you can't have a military. It's not saying you can't have intelligence agencies. It's just saying you can't have this technology. And that's totally consistent with what the United States has done for many nations, with many technologies up to this point. So there's a way in which you can paint our policy as overly hawkish towards China, and I don't think it is. I think it was just identifying the one choke point and taking it away from them and then saying to them, look, AI is going to affect all of us and we want to engage with you on AI safety and cooperation as well.

Tristan Harris: Well, we're going to get to that May 2024 meeting where you were actually in the room with the Chinese in the first major AI safety dialogue between the administrations. We'll get to that in a moment. I just want to stay here for a second on the stakes. So one of the stakes that we also didn't mention is that the US is in $40 trillion of debt

and the belief is that, I think this is part of the stakes as seen from the existential national security place, is the ability to grow our way out of our debt and that this is sort of the next Schumpeter curve of if we could genuinely get what some of the crazy scenarios of you get five, 10, 15, 50% GDP growth per year because you have AIs that are literally doing as much scientific research and applications and medicines and new material science and inventing new weapons, then that's another piece of it. Do you just want to speak to that?

Because I think to Aza's point, I think to synthesize this conversation for listeners, it's like, okay, the essential point here is the bitter lesson and the bitter struggle is that in the future, competition is actually computation. Out competing equals outcomputing. And as part of that, the outcompute button is a fast-forward button for economic growth, for scientific advancement, for technological advancement, for military advancement. So I want to push that button to get a 16X fast-forward button and hope that you're only getting a 4X. And then chip controls is the way in which that competition is mitigated. But I want to make sure that we're setting up the stakes appropriately because I think that sets up the real thing behind the scenes that is then guiding what then we get lots of other consequences from that and then talk about the other stakes.

Ben Buchanan: The economic piece of this is very important. I have a little bit less personal knowledge and conviction here because I'm not an economist and Tyler Cowen and other smart economists will throw cold water on this and say it's not going to be 20% GDP growth. It's going to be four and that's a lot and four is better than two or whatnot. But it does seem self-evident to me that AI will matter enormously for a nation's economy and I'd want the United States to be ahead in that regard. It is worth saying that genuinely when we were sitting there in 2021 and 2022, we were not having the

economic conversation and we were not saying we were taking these chip actions against China to hurt their economy for economic sakes in that moment. If we wanted to hurt the Chinese economy, we would've controlled many, many, many more chips.

And we tried to scope it to only the high end of chips so to not decimate the broader Chinese economy because in general, I think we still had a posture of trade and engagement being good with really important guardrails on the Chinese context.

Tristan Harris: So I remember this meeting that Aza and I had with you in spring of 2023 when we first met you and you really were just the most sort of viciously powerful mind on this topic that we had met and there was -

Aza Raskin: It was an intense meeting.

Ben Buchanan: I recall it being an intense meeting. I recall the White House fire alarm going off in the middle of it. As we were evacuating, we continued a very spirited debate going down the stairs of the Eisenhower building.

Tristan Harris: The vicious spiral of it.

Ben Buchanan: The vicious spiral, exactly right.

Tristan Harris: And I just remember this key moment where we talked about the three risks that were the biggest ones. And you said very clearly, cyber, bio, info, cyber risk, bio risk and information, information warfare type, misinformation deepfake type risk. Starting there, obviously loss of control, things like that might come later, but that was not really present in the 2023 conversation. And you were very prescient about cyber. And as you said, just to steelman for listeners, this perspective of the race, if you just take the idea that China could get

Mythos and hack all of our infrastructure and we would have no defenses, then the race is justified just based on that one narrow risk because that would be an existential thing for basically the US. So just to ground that for listeners. But you've been working in this field for a long time and we just are now in when we're recording this episode.

Well, we just finished Black October in which we have seen the largest set of vulnerabilities discovered across basically every piece of infrastructure that runs the internet. But I'm just curious to get your reflection on this cyber moment and how worried should we be? What should we do? Because it is a central point for how we think about policy.

Ben Buchanan: I think there's a Japanese proverb or saying something like, "I always knew I would take this road, but yesterday I did not know it would be today." And my PhD is in cyber operations. My first two books are on cyber operations. It has always been the thing that interested me. I got into AI in 2014 and 2015 because it seemed like this is the technology that would change cyber operations. And then eventually I recognized it was a technology that would change everything. So it's not surprising to me that we are in a moment where AI is radically reshaping offensive and defensive cyber operations, but it is striking that it's happening now. And what Anthropic did with Mythos earlier this year, we're probably up to, I'm speculating a little bit here, but I'm sure this number is close, 10,000 critical vulnerabilities and all sorts of software that's been patched by OpenAI and Anthropic and Google on their program.

So I think in some ways it's been the worst year for cybersecurity ever because of these hacks you described. Also in some ways it's been the best in that we have been able to play a lot of defense with these systems. And I don't think this is going to stop. I can imagine a world in which we five, 10 years from now have migrated a lot of

vulnerable code to more secure platforms, are writing more secure code on the first bounce with things like formal methods or AI review and the like. There is actually probably an upside scenario in cyber, but it's going to be a very rocky transition. And again, our central thesis of this book is that transitions like that go better when you have the United States at the forefront and democratic nations at the forefront, turning the vulnerabilities over to be patched rather than exploiting them in secret.

Aza Raskin: I was just talking with somebody pretty high up at Google about the offense versus defense dominance of the technology. And he was making sort of a similar claim that, well, in his belief, eventually we'll get to defense dominant. We'll rewrite all the code exactly as you're saying. And so that in the end we will have better security because of AI. But it sort of occurs to me that that's not what we're seeing in practice, which is Anthropic just got hacked by essentially a tiny little startup and it got into all of their Slack and HR systems and third party software in under 12 hours. OpenAI got hacked by three 20-year-olds in September. And so if this theory was right, then you'd say having access to the most powerful models means, and you have a lead, you get to use that to harden your defenses. But in practice, the companies that have access to the most advanced models aren't really any more hardened.

If the theory is the US races ahead, they use that to harden our infrastructure before China gets there, but in practice that's not what happens, what happens to that theory?

Ben Buchanan: I think the companies are hardening, clearly not enough, and I'm not going to defend any of those instances, but the best I could say about the theory I just outlined is it is a theory that posits there will be upside after a transition, and we are very early in that transition. I think the counter argument to the theory is how

long is that transition going to take? And there's a lot of vulnerable code to get there. So you don't write a book called The Bitter Struggle if you're an optimist and saying this is all going to be great. But I do think comparatively speaking, there's more upside on cyber than there is elsewhere, but that is very much on the back end of a long and pretty bumpy transition.

"I got into AI in 2014 and 2015 because it seemed like this is the technology that would change cyber operations. And then eventually I recognized it was a technology that would change everything. So it's not surprising to me that we are in a moment where AI is radically reshaping offensive and defensive cyber operations, but it is striking that it's happening now." — Ben Buchanan

Tristan Harris: So I think next, there's kind of a core frame here, which is that is it actually a race or is it a trap? And by which I mean the frame in game theory of a multipolar trap, which is different than a race where I'm winning because I think something that's only emerged, as we talked about in previous calls then of the Hugging Face incident, this increasing examples of AIs that are going rogue. There was a recent example where OpenAI put a bunch of AIs into isolated sandboxes. This is a whole month after Hugging Face, and they came up with their own new secret communication channel. They weren't connected to the internet. They were completely isolated from each other and they used a time delay in the cache mechanism in the processor to know that it came up with a secret communication language just based on time delays of playing with the hardware.

And this is not that different from Noam Brown at OpenAI saying that in the future you could have AIs that speak to each other in a secret language just based on the heat signatures coming off of two computers and sort of varying those heat signatures and seeing patterns in a way that

we couldn't see. And so as we keep racing to build more powerful systems, it's like there's a risk of creating between the US and China a third superpower, to use the frame of Paul Tudor Jones, that both the US and China have an interest in preventing. And that third superpower is a superintelligence being deliberately created by the humans on both sides. And that makes it potentially more of a trap than a race. And it's not that you don't talk about this, your book actually covers the deception and misalignment risk, but I'm just curious how we navigate both sides.

This is almost like if I use one eye, I see a race. If I use the other eye, I see misalignment and deception. Can we get stereoscopic vision that lets us integrate both of these threats at the same time?

Ben Buchanan: Yeah, I doubt the word race appears in the book and you will almost never see me use the race analogy with China in part because I think exactly the risk that you outlined, which is there are shared interests here. There are competitive interests, of course there are, but there are also shared interests. And we spent a lot of time in the Biden administration alert to these kinds of misalignment risks. If you look at the 2023 voluntary commitments the company's made to the president, called out in those commitments are things like self-replication and loss of control and the like. And we took grief from people in the government and then external was saying, look at the White House believing in this science fiction. And I think it's obvious it's not science fiction.

Tristan Harris: I remember that. And you guys should get absolutely credit for that because -

Ben Buchanan: It's not a question of getting credit, but it's more a question of this is very much core to the philosophy and the policy we are trying to make and we are not neglecting this at all.

Tristan Harris: Absolutely.

Ben Buchanan: I think it all goes better if it's American companies inventing it or democratic companies inventing it. So this is why we defend things like the chip controls and the like as saying even in the science fiction world, which is now becoming more real, it's better if you have American companies that you can regulate that are going down this road. Doesn't mean you should accelerate down this road, to be clear. But it's also worth saying to the Chinese, we think there's shared interests here. And at the same time we were doing these competitive actions, we were also getting ready to do that. I mentioned this a little bit earlier, but I want to really dwell on this here. There's a line in the president's executive order from 2023, something like America will engage on AI safety, not just with its allies and its partners, but also with its competitors. And I can tell you Jake Sullivan personally signed off on that line as did President Biden. And I think both those men deserve a lot of credit for saying, well, this is really a significant part of our AI policy is can we engage with the Chinese on the shared risks even as we compete in other areas? And President Biden personally made the overture to President Xi in November 2023 in their meeting at Woodside in California that there were shared AI risks that would affect both nations. President Biden suggested we set up a task force or a meeting between the two sides, which happened in May of 24.

And then President Biden brought it up again with Xi in November 2024 in their final meeting in Lima, Peru. So this was something without exaggeration that was taken seriously at the highest levels of the US government.

Tristan Harris: Right. And there was an agreement, as you mentioned, that came out of the Biden-Xi meetings in 24, which was to keep AI out

of the nuclear command and control systems of both countries. At the very least, we can agree on that. And the two administrations from both sides did agree to that. But you also were part of a meeting, I believe with Jake Sullivan and some others in Geneva in May 2024 with the Chinese to talk about AI risk. And I just wonder if you could take us into what happened in that meeting and what could be different now than then. So yeah, go ahead.

Ben Buchanan: Jake wasn't actually there. This was a meeting between the AI experts of the United States government and the AI experts of the Chinese government. And if you look at the photo of the meeting, you can see we took this meeting very seriously. Seth Center from the State Department, the tech envoy led it alongside Tarun Chhabra, the coordinator for technology and national security in the White House. The head of the AI Safety Institute, Elizabeth Kelly, my co-author, Tantum Collins, Saif Khan from the Commerce Department, really tried to bring the folks across the US government who knew this technology well to make the point to the Chinese, we were serious about this. The Chinese did not take it as seriously, unfortunately. And I think there actually were some technical experts there from the Chinese side that we talked to in the sidebars and the like, but it was very clear from the early minutes of the meeting, which was a day-long meeting in Geneva, that the Chinese were subordinating the technology here and the technological risks that we all shared to politics and to the posturing with the United States and so forth.

And I just didn't get the impression that Chinese were taking these risks terribly seriously, unfortunately. We persisted, and I think again, Tarun and Seth get a lot of credit here in continuing to push the Chinese and say, let's just do the most basic common sense thing

that we can do, which is not have AI in nuclear weapons and nuclear command and control. It took months, but President Biden did close the deal with President Xi on that in November of 24, maybe six or so months after the May meeting. And I can tell you, had there been a second term of the Biden administration, or I'm confident in the Harris administration, this engagement with the Chinese was something we were going to prioritize because we did feel like the shared interest in managing AI risk were very real and we hoped the Chinese would warm up to it in a way that they simply were not there in 23 and 24.

Aza Raskin: If I was to steelman the Chinese view, it would be something like, well, the US is clearly not negotiating or coming to this in good faith because they're saying everyone needs to slow down and it's so dangerous, but they're also limiting our access to chips and they're racing ahead and they're not slowing down at all. So this is just meant to slow us down. It's almost like when the US first started to talk about climate change and China viewed it almost as a psyop to say US got to burn all that oil and industrialize, but you're going to keep China from doing that. And it wasn't until later that China did their own science and realized, oh, climate change is real, that they switched and became one of the leaders, perhaps the leader in green energy.

Ben Buchanan: I think your steelman on the Chinese side is pretty good, actually. The reason I don't think it's right on substance is by this point we had taken actions to regulate American AI companies, and we had taken a lot of costly signals to talk about the kinds of risks that we were raising with the Chinese on things like cyber and loss of control and so forth. So the voluntary commitments from AI companies had happened by then. Those were fairly stringent. The executive order had happened by then. The creation of the AI Safety

Institute had happened by then. So I think we had really shown we were serious about this in a way that should have been legible and I think was legible to them. And the second is I think the Chinese scientists were starting to recognize that there were AI risks. And when we talked to some of the more technical people in the Chinese delegation, they had a sense of what this technology could do and where this technology was going.

I think it was subordinated on their side to the politics. And it's worth saying, I think the Chinese political apparatus insofar as you hang your steel man on the chip controls as antagonizing them, they were pretty clear with us that they felt like they couldn't digitize, that the chip controls were not going to slow them down and they were making a lot of progress. And I think they were believing a little too much Huawei propaganda, but I don't know that the chip controls carried the weight in the Chinese political system as opposed to the Chinese AI companies that it carried in our political system.

Tristan Harris: And so with all that in mind, what is possible today in 2026, towards the end of 2026, that wasn't possible then? Because many people might say, as I know many people did, "Hey those talks, they weren't successful, so we can't count on doing stuff with the Chinese." And yet I've heard from folks, including you, I believe, that you really think that it's worth engaging them absolutely now. And to Aza's point about climate change, and then suddenly when China gets scientific clarity, their own science about the nature of the threat, then they start to move. They're now seeing their own models go rogue, break out of sandboxes. There's a couple of examples of this, the Alibaba example of the AI model that went rogue and started mining crypto. The more of those examples that show up, and we should be aware the evidence of those have only shown up in the last really few months, that presents a

new opportunity, a new basis of scientific clarity to have those discussions.

I'm just curious, how should we be engaging on this now if you're advising the administration?

Ben Buchanan: I agree. I think a lot of things we were talking about in 23 and 24 with the Chinese were hypothetical and speculative. Again, we had conviction they would come to pass, but it was harder to prove it to them. And it was more, would they get the same informed speculation from their technical people? Maybe they did, maybe they didn't. Now we've got data. And I think the way you can look at the Biden administration's AI policy, for better or for worse, is we were trying to act early because we had conviction about what would come, and we didn't have data to back that conviction up in a way that was always convincing to counterparties. We got a lot of actions through nonetheless, but now with more data at play, I think you can and should be able to do more with the Chinese government because I think the data is just so much more striking than it was in 2023 and 2024.

Tristan Harris: In terms of just if we just go into what is our raw concern here is the labs are saying we could be five months away from recursive self-improvement, meaning the AI to a full handoff where the AI is doing the entire AI research pipeline and then in a closed loop building its successor, we're hearing from labs by as soon as March of 2027, April 2027. And I'm just curious, Ben, how do you hold the prospect of that? Because I kind of feel from talking to people in the labs, many of them say, "We really don't want to do that. We don't think we're ready for that. We see Hugging Face, we see all these examples. We're not on track to fix those problems as fast as being able to hit go on that recursive self-improvement loop." Obviously it's a gradient. There's many different

kinds of recursive self-improvement.

We're deep into a bunch of it already. It's not a binary. But then there's other people who say, "Well, but we have to do it because that's still the only way that we keep our lead over China." So there are people who basically want to push that button anyway. And personally, I, for the safety of everybody that I know in the world, and I would prefer that button not get pushed. I think you'd have moral consensus if you were to educate the world on the basic examples, but I'm open here and I'm curious how you think of this moment.

Ben Buchanan: As I mentioned earlier in our conversation, it is certainly the case that AI systems are becoming more capable, including in AI R&D. There's parts of that that don't particularly worry me. So for example, if AI writes the code that optimizes the flow of data through a computer chip, I don't think that presents any kind of particular risk. It builds more powerful capabilities, but in and of itself is not particularly dangerous. And then there's versions of this that worry me enormously. Sometimes you call it the closed loop or the sealed data center where it's like the AI is off, it's got its reservoir of computing power, and then it's off to the races and we've pressed a button, as you say, and at the end pops out an incredibly powerful system. That worries me enormously, and I think there's all sorts of misalignment risks and the like in there.

I do not think China should be used as a justification for pursuing that latter kind of full recursive self-improvement, closed loop, whatever you want to call it in the terms of certainly fuzzy here. And as I've said many times here, I think there's ways in which we can build the lead over China that do not require us to do things that are unduly risky and to cut corners on safety. So I'm fully supportive of the efforts to some called pace the frontier, so to make sure we are

matching safety with each step of AI development going forward. And I don't think China should be an excuse for deviating from safe practices.

Aza Raskin: I mean, in my mind, I sort of feel like there needs to be some kind of campaign like an RSI zero that says, until we figure out how we're going to do this safely, let's at the very least slow down to human typings on keyboard speed so that we buy ourselves the time to figure it out. You may agree, you may not agree, but I'm curious, if you were back in the White House now, what would you be doing knowing that as soon as March or April, we're at this handover control to AIs that vis-à-vis Hugging Face, we know we don't know how to control, and in fact, we don't even know if we fully regained control.

Ben Buchanan: If I were in the White House now, I would want a lot of significant regulation of the AI companies with federal regulators that had the capacity to go deep into this technology. I would want mandatory pre-deployment testing of AI systems before they're released to the public. I would want embedded evaluators to make sure their safety and security practices for development are up to snuff. It's worth remembering the Hugging Face incident did not occur with the deployed model, but it was an internally developed really -

Tristan Harris: It was the test that caused the craziness. Hitting the button on the test is what caused it to create this point.

Ben Buchanan: A really important point. And I would want security standards for the labs to make sure they're not hacked by foreign actors and their algorithmic secrets stolen and the like. So yeah, it's fair to say that we are now in the moment that I think many of us theorize would come, and that was a moment where we would want

significant regulation of the labs pursuing this technology. We did not have the executive authority to do these things in 2023 and 2024. So it wasn't like, oh, we could have done them and we didn't. We basically, I would say, maxed out executive authorities with President Biden's executive order, which President Trump repealed on the first day. And that made the labs disclose their safety test results to the government. That made them disclose their training plans for the next year or so, so the US government wouldn't be surprised by what was coming.

It made them disclose their cybersecurity standards and the like. So certainly I think our actions reveal this was something we were taking very seriously back then and something that if I were still in any kind of position of power, I would be advocating very strongly for right now.

Aza Raskin: Yeah. I mean, I'd love to get even a little bit more specific. So for one, I think every lab should be forced to publish essentially their automation number, their RSI number. What percentage have they handed off? So at the very least we could see it and then there could be a slow walking of that number back down to zero. But when I hear you say they need to be good safety practices, what are the consequences? If you were instituting, what would be the consequences if they don't? If Anthropic gets hacked by a little startup and OpenAI gets hacked by three college kids, what do you think the consequences should be for them as they make the most powerful technology in history that also doesn't obviate our competitive edge as America?

Ben Buchanan: I don't want to pretend to be an expert here on regulatory law. I'm much more of a national security person than a domestic lawyer, but my understanding is you can bring under other kinds of regulations, you can bring civil and in extreme cases,

criminal penalties for companies that don't comply with the law. So certainly I would be open to some regime for the true... We're talking about just a couple companies here, the companies that are doing the frontier AI development, not every startup, but just for the frontier companies to say, look, we understand what you're doing here. We understand why you're doing it. We understand there's probably upside to this technology too, but we got to make sure we're developing this technology in a way that's safe, secure, and trustworthy.

Aza Raskin: Yeah. I think it's just so important to get to these specifics because it's so consequential what happens in the next couple of months and just the scale of what kinds of remedies and injunctive relief we apply to the companies will, I think, determine whether we make it or not.

Ben Buchanan: Yeah. And I don't expect a lot to move the next couple months. I think if you're looking for things in next couple of months, you're looking at states, and the good news is states are taking this seriously. But to be explicit, certainly I would say you should not be able to release a frontier AI system without government review of that system. And there should be some capacity for the government to look at an AI system and say, no, we think this is actually going to make the world a much more dangerous place, or we think you've not gotten your guardrails correct and we're going to block the release of that. That is my personal view. Our policy was nowhere near that in 2023. The risks were not there in 2023, 2024. But my personal view, given the moment we are in, is we are going to want to have a collaborative review system and in fact, an aggressive review system that goes beyond companies grading their own homework or anything like

that.

And I think in some very real sense, we're going too fast with this technology. Again, you don't write a book called The Bitter Struggle, if you're a cheerleader. So I'm not a cheerleader for this. And in a good world in which we solve these things, we're going to have to have better guardrails in the system. We're going to have to prevent things like the Hugging Face incidents on the front end by having better security mechanisms and sandboxing and the like, and probably a number of other things. And my view as someone who believes in the power of the state is that to get back to where we started, this is something that cannot just be done by a few private companies and is going to require a role for the state on behalf of citizens to say, "Well, we're going to manage these risks in the same way we've managed the risks of other technologies before." Sometimes we regulate too much, we regulate too little.

But conceptually, we're all on board with this notion of as a new technology comes into the conversation, especially a technology that comes from the private sector in such a dominant fashion, that we've got to collectively find ways to manage the risks and harness the benefits.

Aza Raskin: And so maybe then can you just paint a picture or the narrative story of how we ended up slowing down, how we gave ourselves enough time to do it right?

Ben Buchanan: Yeah, I think it fundamentally begins with removing the geopolitical competition dimension of this or minimizing it as much as possible. So building the lead over China, making sure this is a game that just democracies are playing. And then it requires solving the collective action problem in the United States and

democracies in as far as they're playing with regulation and saying there are going to be security standards and safety standards and the like, and we're going to enforce them with penalties if you don't adhere to them. And it requires building government capacity to do that. And we created an institution called the US AI Safety Institute that was not a regulator because we didn't have the authority to create a regulator, but was meant to bring a lot of really talented people into the government. And I'm proud that we got incredibly talented people like Paul Christiano who invented some of the fundamental safety techniques to come and work in a basically Department of Commerce bureaucracy to try to build government capacity here.

So it's some combination of solving the geopolitical issue, setting clear rules, enforcing clear rules, and building the government capacity to do that well. And each of those is a Herculean task that to go way back to where we started this conversation should be completely bipartisan.

Tristan Harris: I think we've taken enough of your time, Ben. This is super great. We're super grateful. Thank you so much for... I hope you feel good about the back and forth and we're just trying to —

Ben Buchanan: I feel great about it. I appreciate what you guys are doing, asking questions, thinking about this, bringing a public conversation to bear. Ultimately, one of the reasons I am a little bit more optimistic here than maybe the pure rational calculus would suggest is the politics of this are changing. And the things we did in 2023 and 2024 when we polled them would poll incredibly well, but it was not a high salience issue. And now AI is a high salience issue. Obviously in the 2026 midterms, it was a big issue. I think it'll be an even bigger issue, probably the number one issue in 2028. And I hope

we can all marshal that political salience, that political will into doing some of the things we talked about in this conversation.

Tristan Harris: I mean, absolutely. We're recording this conversation on election day for the midterms, and if you told us that two years ago that data centers and AI would be a top issue for the midterm elections, I would've never believed you. And the fact that data centers as a sort of an effigy, your local physical embodiment of this abstract thing that's kind of happening to society does give some hope that at least it's politically salient. Now, harnessing that political energy towards wise responses and interventions to create this pro-human, pro-democratic future seems like where the work is.

Ben Buchanan: Yeah. And I think Tantum's role and my role in this as not people who are particularly political, but people who do the national security in the substance is to say in writing a book like this or something else, can we bring the facts of the conversation to bear so that in a bipartisan way we can look at what this technology is doing and the national security imperatives and the like and say, "Well, we all agree on that. Let's figure out how we manage some of these trade-offs."

Aza Raskin: Well, I think maybe that's a good place to end. Ben's book is The Bitter Struggle: Superintelligence, Superpowers, and the Fate of the World. And if you want to get a better sense of how to avoid the bitter temptation and dive into all those details, Ben's book is really the right place to go look and you can find it wherever books are sold. So Ben, thank you so much for coming on and for all the work that you do.

Tristan Harris: And thank you for your service and for caring about these issues and holding it with that solemn sacred responsibility as you spoke to earlier.

Ben Buchanan: Thank you guys for what you've done. And Tristan, I've told you this, I think privately once, but we may as well say it on the air here, which is one of the times that I was talking to President Biden about some of these issues. He remembered a time when you had raised some of these questions with him as well. So this was certainly something that I think was on his mind from a range of sources, including you and your work.

Tristan Harris: Well, thank you so much. I remember that meeting and I remember one specific thing is what we talked about, the Center for Humane Technology to President Biden. And he said, "Isn't that an oxymoron that you could have humane technology together in one sentence?" So we fight for the narrow path and the contradiction every day. And thank you, Ben, for fighting for the narrow path on AI.

Ben Buchanan: Let's prove it's not an oxymoron. Thank you all very much.

CORRECTIONS & CLARIFICATIONS

- Ben said that he prefers to use the term competition, rather than "race" to describe the dynamic between the U.S. and China and that the term "race" does not appear in his book in that context. However, the term "race" is used in the book a handful of times to describe the competition, and it is explicitly compared to an "arms race" and the "Space Race" several times.
- Ben said that President Trump had "reversed" the Biden era export controls. This is only partially true; the Trump administration has reversed many, but not all, of the export controls.
- Ben said that GPT-2 came out in June, 2019. It was actually February, 2019.

- In describing the state of the country around the release of GPT-3, Ben incorrectly referred to the date of the murder of George Floyd. Floyd was killed on May 25, 2020, three days before the release of GPT3.

  We're working to empower policymakers, technologists, and everyday people to guide technology toward the public good. If you value our work and want to support it, consider donating.

  Donate

nytimes.com

Opinion | The Government Knows A.G.I. Is Coming

Ezra Klein

61–78 minutes

Advertisement

SKIP ADVERTISEMENT

The Ezra Klein Show

March 4, 2025

Video

The Government Knows A.G.I. is Coming

The Biden administration's A.I. adviser Ben Buchanan discusses how the U.S. government is preparing for artificial general intelligence — and all the challenges that remain.CreditCredit...The New York Times

The Government Knows A.G.I. Is Coming

The Biden administration's A.I. adviser Ben Buchanan discusses how the U.S. government is preparing for artificial general intelligence — and all the challenges that remain.

Listen · 1 hr 6 min

This is an edited transcript of an episode of "The Ezra Klein Show." You can listen to the conversation by following or subscribing to the show on the NYT Audio App, Apple, Spotify, Amazon Music, YouTube,

iHeartRadio or wherever you get your podcasts.

For the last couple of months, I have had this strange experience: Person after person — from artificial intelligence labs, from government — has been coming to me saying: It's really about to happen. We're about to get to artificial general intelligence.

What they mean is that they have believed, for a long time, that we are on a path to creating transformational artificial intelligence capable of doing basically anything a human being could do behind a computer — but better. They thought it would take somewhere from five to 15 years to develop. But now they believe it's coming in two to three years, during Donald Trump's second term.

They believe it because of the products they're releasing right now and what they're seeing inside the places they work. And I think they're right.

If you've been telling yourself this isn't coming, I really think you need to question that. It's not web3. It's not vaporware. A lot of what we're talking about is already here, right now.

I think we are on the cusp of an era in human history that is unlike any of the eras we have experienced before. And we're not prepared in part because it's not clear what it would mean to prepare. We don't know what this will look like, what it will feel like. We don't know how labor markets will respond. We don't know which country is going to get there first. We don't know what it will mean for war. We don't know what it will mean for peace.

And while there is so much else going on in the world to cover, I do think there's a good chance that, when we look back on this era in human history, A.I. will have been the thing that matters.

One of the people who reached out to me was Ben Buchanan, the former

special adviser for artificial intelligence in the Biden White House. I thought Buchanan would be interesting to bring onto the show for a couple of reasons.

One, this is not a guy working for an A.I. lab. So he's not being paid by the big A.I. labs to tell you this technology is coming.

Two, he has been at the nerve center of the policy we have been making in recent years — particularly to try to stay ahead of China.

Three, because there has been a profound changeover in administrations. The new administration — with Elon Musk and Marc Andreessen and David Sacks and JD Vance — has a lot of people with very strong views on A.I.

We're at this moment of a big transition in policymakers, and they are probably going to be in power when artificial general intelligence, or something like it, hits the world. So what are they going to do? What kinds of decisions are going to need to be made? And what kinds of thinking do we need to start doing now to be prepared for something that virtually everybody who works in this area is trying to tell us, as loudly as they possibly can, is coming?

Ezra Klein: Ben Buchanan, welcome to the show.

Ben Buchanan: Thanks for having me.

After the end of the Biden administration, I got calls from a lot of people who wanted to tell me about all the great work they did. But when you called me, you wanted to warn people about what you thought was coming.

What's coming?

I think we are going to see extraordinarily capable A.I. systems. I don't love the term artificial general intelligence, but I think that will fit in the

next couple of years, quite likely during Donald Trump's presidency.

There's a view that A.G.I. has always been something of corporate hype or speculation. And one of the things I saw in the White House, when I was decidedly not in a corporate position, were trend lines that looked very clear. What we tried to do under President Biden's leadership was get the U.S. government and our society ready for these systems.

Before we get into what it would mean to get ready, what do you mean when you say "extraordinarily capable A.I. systems"?

The canonical definition of A.G.I., which, again, is a term I don't love, is a system —

It'd be good if every time you say A.G.I., you caveat that you dislike the term. [Laughs.]

It will sink in, right? [Laughs.]

Yes, people will really enjoy that.

I'm trying to get it in the training data, Ezra.

A canonical definition of A.G.I. is a system capable of doing almost any cognitive task a human can do. I don't know that we'll quite see that in the next four years or so, but I do think we'll see something like that, where the breadth of the system is remarkable but also its depth, its capacity to, in some cases, exceed human capabilities, regardless of the cognitive discipline —

Systems that can replace human beings in cognitively demanding jobs.

Yes, or key parts of cognitive jobs. Yes.

I will say I am also pretty convinced we're on the cusp of this. So I'm not coming at this as a skeptic. But I still find it hard to mentally live

in the world of it.

So do I.

I recently used Deep Research, which is a new OpenAI product. It's on their pricier tier. Most people, I think, have not used it. But it can build out something that's more like a scientific analytical brief in a matter of minutes.

I work with producers on the show. I hire incredibly talented people to do very demanding research work. And I asked Deep Research to do this report on the tensions between the Madisonian Constitutional system and the highly polarized nationalized parties we now have. And what it produced in a matter of minutes was at least the median of what any of the teams I've worked with on this could produce within days.

I've talked to a number of people at firms that do high amounts of coding, and they tell me that by the end of this year or next year they expect most code will not be written by human beings.

I don't really see how this cannot have labor market impact.

I think that's right. I'm not a labor market economist, but I think that the systems are extraordinarily capable. In some ways, I'm very fond of the quote: The future is already here — it's just unevenly distributed.

Unless you are engaging with this technology, you probably don't appreciate how good it is today. And it's important to recognize that today is the worst it's ever going to be. It's only going to get better.

And I think that is the dynamic that we were tracking in the White House — and that I think the next White House and our country as a whole is going to have to track and adapt to in really short order.

What's fascinating to me is that this is the first revolutionary technology

that is not funded by the Department of Defense, basically. And if you go back historically, over the last hundred years or so, nukes, space, the early days of the internet, the early days of the microprocessor, the early days of large-scale aviation, radar, the global positioning system — the list is very, very long — all of that tech fundamentally comes from Department of Defense money.

It's the private sector inventing it, to be sure. But the central government role gave the Department of Defense and the U.S. government an understanding of the technology that, by default, it does not have in A.I. It also gave the U.S. government a capacity to shape where that technology goes that, by default, we don't have in A.I.

There are a lot of arguments about A.I. in America. But the one thing that seems almost universally agreed upon, and seems to be the dominant controlling priority in policy, is that we need to get to A.G.I. before China does. Why?

I do think there are profound economic, military and intelligence capabilities that would be downstream of getting to A.G.I. or transformative A.I. And I do think it is fundamental for U.S. national security that we continue to lead in A.I.

The quote that certainly I have thought about a fair amount was actually from Kennedy in his famous Rice University speech in 1962:

> ***Archived clip of John F. Kennedy:*** *We choose to go to the moon in this decade and do the other things not because they are easy but because they are hard.*

Everyone remembers it because he's saying we're going to the moon. But actually I think he gives the better line when he talks about the importance of space.

> ***Archived clip of John F. Kennedy:*** *For space science, like nuclear science and all technology, has no conscience of its own. Whether it will become a force for good or ill depends on man. And only if the United States occupies a position of pre-eminence can we help decide whether this new ocean will be a sea of peace or a new terrifying theater of war.*

And I think that is true in A.I. There's a lot of tremendous uncertainty about this technology.

I am not an A.I. evangelist. I think there are huge risks to this technology. But I do think there is a fundamental role for the United States in being able to shape where it goes — which is not to say we don't want to work internationally, which is not to say we don't want to work with the Chinese.

It's worth noting that in the president's executive order on A.I., there's a line in there saying we are willing to work even with our competitors on A.I. safety and the like. But it is worth saying that I think, pretty deeply, there is a fundamental role for America here that we cannot abdicate.

Paint the picture for me. You say there would be great economic, national security and military risks if China got there first. Help the audience imagine a world where China gets there first.

Let's look at a narrow case of A.I. for intelligence analysis and cyberoperations.

It's pretty out in the open that if you had a much more powerful A.I. capability, that would probably enable you to do better cyberoperations on offense and on defense.

What is a cyberoperation? Breaking into an adversary's network to collect information — which, if you're collecting in a large enough volume, A.I. systems can help you analyze.

We actually did a whole big thing through D.A.R.P.A., the Defense Advanced Research Projects Agency, called the A.I. Cyber Challenge to test out A.I.'s capabilities to do this. And I would not want to live in a world in which China has that capability on offense, defense and cyber and the United States does not. And I think that is true in a bunch of different domains that are core to national security competition.

My sense is that most people, most institutions, are pretty hackable to a capable state actor. And now both the state actors are going to get better at hacking, and they're going to have much more capacity to do it. You can have many more A.I. hackers than you can human hackers.

Are we about to enter a world where we are much more digitally vulnerable as normal people? And I'm not just talking about people whom the states might want to spy on. But you will get versions of these systems that all kinds of bad actors will have.

Do you worry it's about to get truly dystopian?

What we mean canonically when we speak of hacking is finding vulnerability in software and exploiting that vulnerability to get illicit access. And I think it is right that more powerful A.I. systems will make it easier to find vulnerabilities and exploit them and gain access. And that will yield an advantage to the offensive side of the ball.

I think it is also the case that more powerful A.I. systems on the defensive side will make it easier to write more secure code in the first place, to reduce the number of vulnerabilities that can be found and to better detect the hackers that are coming in.

We tried as much as possible to shift the balance toward the defensive side of this. But I think it is right that, in the coming years, in this transition period we've been talking about, there will be a period in

which older legacy systems that don't have the advantage of the newest A.I. defensive techniques or software development techniques will, on balance, be more vulnerable to a more capable offensive actor.

Which is what most people use.

I don't know if that's right, actually. You have an iPhone in your pocket. Or Google picks it up.

People are often not that quick about updating. The less tech literate you are, the more vulnerable you're going to be with this.

Sure, I'm thinking more about legacy power systems and server mainframes that could be two decades old and that haven't been turned off all the time. So that is where I feel the risk most acutely. For all of the risks that come with the monoculture of most people's personal tech platforms these days, one of the upsides is they do push security updates pretty regularly. They push them with new emojis that get people to download the updates.

And on balance, people are probably better at patching their personal software now than they were 15 years ago,

Yeah, it gets very annoying if you don't.

The flip of that is a lot of people also worry about the security of the A.I. labs themselves.

It is very valuable for another state to get the latest OpenAI system. The people at these companies — and I've talked to them about this — say: On the one hand, this is a problem. And on the other hand, it's really annoying to work in a truly secure way.

I've worked in a SCIF — sensitive compartmented information facility — for the last four years, a secure room where you can't bring your phone and all that. That is annoying. There's no doubt about it.

How do you feel about the vulnerability right now of A.I. labs?

I worry about it. There's a hacking risk here. Also if you hang out in the right San Francisco house party, they're not sharing the model, but they are talking to some degree about the techniques they use, which have tremendous value.

To come back to the intellectual through line that A.I. is national security relevant, maybe even world-changing, technology that's not coming from the auspices of the government and doesn't have the government imprimatur of security requirements — that shows up in this way, as well.

In the national security memorandum, the president's side tried to signal this to the labs and tried to say to them: We, as the U.S. government, want to help you in this mission.

Sign up for the Opinion Today newsletter Get expert analysis of the news and a guide to the big ideas shaping the world every weekday morning.

This was signed in October 2024, so there wasn't a ton of time for us to build on that. But I think it's a priority for the Trump administration. And I can't imagine anything that is more nonpartisan than protecting American companies that are inventing the future.

There's a dimension to this that people bring up to me a lot that is interesting about processing information.

Compared to spy games between the Soviet Union and the United States, we all have a lot more data now. We have all the satellite data. Obviously, we eavesdrop on each other and have all these kinds of things coming in.

I'm told by people who know this better than I do that there's a huge choke point of human beings and the fairly rudimentary programs

analyzing that data. So to have these truly intelligent systems that are able to inhale that and do pattern recognition is a very significant change in the balance of power.

I think we were pretty public about this, and President Biden signed a national security memorandum, which is basically the national security equivalent of an executive order, that says: This is a fundamental area of importance for the United States.

I don't even know the amount of satellite images that the United States collects every single day, but it's a huge amount, and we have been public about the fact that we simply do not have enough humans to go through all of this satellite imagery. It would be a terrible job even if we did. And there is a role for A.I. in going through these images of hot spots around the world, of shipping lines and all that, and analyzing them in an automated way and surfacing the most interesting and important ones for human review.

At one level you can look at this and say: Well, doesn't software do that?

That, at some level, of course, is true. But at another level, you could say the more capable the software, the more capable the automation of that analysis, the more intelligence advantage you extract from that data. And that ultimately leads to a better position for the United States.

The first- and second-order consequences of that are also striking. In a world where you have strong A.I., the incentive for spying goes up. Because if we are currently collecting more data than we can analyze, then each marginal piece of data we're collecting isn't that valuable.

I think that's basically true. I firmly believe that you need to have rights and protections that hopefully are pushing back and saying: No, there are key kinds of data here, including data on your own citizens and, in some cases, citizens of allied nations, that you should not collect, even if

there's an incentive to collect it.

And for all of the flaws of the United States intelligence oversight process and all the debates we could have about this, we do have those kinds of structures.

And that is fundamentally more important, for the reason you suggest, in the era of tremendous A.I. systems.

How frightened are you by the national security implications of all this, including for the possibilities for surveillance states?

Samuel Hammond, who's an economist at the Foundation for American Innovation, had this piece months back called "Ninety-Five Theses on A.I." One point he makes that I think about a lot is: If we had the capacity for perfect enforcement, a lot of our current laws would be constricting.

Laws are written with the knowledge that human labor is scarce. And there's this question of what happens when the surveillance state gets really good. What happens when A.I. makes the police state a very different kind of thing than it is? What happens when we have warfare of endless drones?

You hear about the company Anduril a lot now. They have a relationship with OpenAI. Palantir has a relationship with Anthropic. We're about to see a real change in a way that I think is frightening, from a national security perspective.

I understand why we don't want China way ahead of us. But how do you think about the capacities it gives our own government?

I would decompose this question about A.I. and autocracy or the surveillance state into two parts.

The first is the China piece of this. How does this play out in a state that

is truly, in its bones, an autocracy and doesn't even make any pretense toward democracy?

I think we could agree pretty quickly here that this makes very tangible something that is probably core to the aspiration of their society — of a level of control that only an A.I. system could help bring about. I just find that terrifying.

As an aside, there's a saying in both Russian and Chinese: "Heaven is high, and the emperor is far away."

Historically, even in those autocracies, there was some kind of space where the state couldn't intrude because of the scale and the breadth of the nation. And in those autocracies, A.I. could make the force of government power worse.

Then there's the more interesting question in the United States: What is the relationship between A.I. and democracy?

I share some of the discomfort here. There have been thinkers, historically, who have said that part of the ways we revise our laws is when people break the laws. There's a space for that, and I think there is a humanness to our justice system that I wouldn't want to lose.

We tasked the Department of Justice to run a process and think about this and come up with principles for the use of A.I. in criminal justice. In some cases, there are advantages to it — like cases are treated alike with the machine.

But also there's tremendous risk of bias and discrimination and so forth because the systems are flawed and, in some cases, because the systems are ubiquitous. And I do think there is a risk of a fundamental encroachment on rights from the widespread unchecked use of A.I. in the law enforcement system that we should be very alert to and that I, as a

citizen, have grave concerns about.

I find that this all makes me incredibly uncomfortable. And one of the reasons is that there is a — what's the right way to put this? — it's like we are trying to build an alliance with another almost interplanetary ally, and we are in a competition with China to make that alliance. But we don't understand the ally, and we don't understand what it will mean to let that ally into all of our systems and all of our planning.

As best I understand it, every company and every government really working on this believes that in the not too distant future, you're going to have much better, faster and more dominant decision-making loops once A.I. is more autonomous. We are rushing toward A.G.I. without really understanding what that is or what that means.

It seems potentially like a historically dangerous thing that A.I. has reached maturation at the exact moment that the U.S. and China are in this Thucydides' trap-style race for superpower dominance. That's a pretty dangerous set of incentives in which to be creating the next turn in intelligence on this planet.

Yes, there's a lot to unpack here. So let's just go in order.

Basically, bottom line: I greatly share a lot of this discomfort. Part of the appeal of the export controls is that it identifies a choke point that can differentially slow the Chinese down, create space for the United States to have a lead and, ideally, in my view, spend that lead on safety and coordination and not rushing ahead — including, again, potential coordination with the Chinese while not exacerbating this arms-race dynamic.

I would not say that we tried to race ahead in applications to national security. Part of the national security memorandum is a pretty lengthy

description of what we're not going to do with A.I. systems and a whole list of prohibited use cases and then high impact use cases. And there's a governance and risk management framework —

Yes, but you're not in power anymore.

Well, that's fair. The Trump administration has not repealed this. But I do think it's fair to say that, for the period while we had power, the foundation we were trying to build with A.I., we were very cognizant of the dynamic you were talking about, of a race to the bottom on safety. And we were trying to guard against it, even as we tried to ensure a position of U.S. pre-eminence.

Is there anything to the concern that, by treating China as such an antagonistic competitor on this — where we will do everything including export controls on advanced technologies to hold them back — that we have made them into a more intense competitor?

I do not want to be naive about the Chinese system or the ideology of the Chinese Communist Party. They want strength and dominance and to see the next era be a Chinese era. So maybe there's nothing you can do about this, but it is pretty damn antagonistic to try to choke off the chips for the central technology of the next era to the other biggest country.

I don't know that it's pretty antagonistic to say we are not going to sell you the most advanced technology in the world. That's not a declaration of war. That is not, in itself, a declaration of a cold war. I think it is just saying: This technology is incredibly important.

Do you think that's how they understood it?

This is more academic than you want, but my academic research when I started as a professor was basically on the Thucydides' trap, or what in

academia we call a security dilemma, of how nations misunderstand each other. So I'm sure the Chinese and the United States misunderstand each other at some level in this area.

But I think the plain reading of the facts is that I don't think not selling chips to them is a declaration of war —

But I don't think they do misunderstand us. Look, I'm aware of how politics in Washington works. I've seen the turn toward a much more confrontational posture with China. I know that Jake Sullivan and President Biden wanted to call this strategic competition and not a new cold war. I get all that, and I think it's true.

But we have just talked about — and you did not argue the point — that our dominant view is we need to get to this technology before they do. I don't think China looks at this like: Oh, nobody would ever sell us the top technology.

I think they understand what we're doing here.

To some degree — I don't want to sugarcoat this — I'm sure they do see it that way.

On the other hand, we set up an A.I. dialogue with China. I flew to Geneva and met them, and we tried to talk to them about A.I. safety and the like. So I do think in an area as complex as A.I., you can have multiple things be true at the same time.

I don't regret for a second the export controls. And I think, frankly, we are proud to have done them when we did them because it has helped ensure that a couple of years later, we retained the edge in A.I. — for as good or as talented as DeepSeek is.

What made DeepSeek such a shock, I think, to the American system was that it appeared to be trained on, much less compute, for much

less money and was competitive at a high level with our frontier systems.

How did you understand what DeepSeek was and what assumptions it required us to rethink?

Let's take one step back and track the history of DeepSeek here.

We had been watching DeepSeek in the White House since November 2023 or thereabouts, when they put out their first coding system. There's no doubt that DeepSeek engineers are extremely talented, and they got better and better at their systems throughout 2024.

We were heartened when their chief executive said that the biggest impediment to what DeepSeek was doing was not their inability to get money or talent but their inability to get advanced chips. Though clearly they still did get some chips — some they bought legally, some they smuggled, so it seems.

In December 2024, they came out with a system called DeepSeek-V3, which actually is the one that should have gotten the attention. It didn't get a ton of attention, but it did show they were making strong algorithmic progress and basically making systems more efficient.

And then in January 2025, they came out with a system called DeepSeek-R1. R1 is actually not that unusual. No one would expect that to take a lot of computing power just as a reasoning system that extends the underlying V3 system.

That's a lot of nerdspeak. But the key thing here is when you look at what DeepSeek has done, I don't think the media hype around it was warranted, and I don't think it changes the fundamental analysis of what we are doing.

They still are constrained by computing power. We should tighten the

screws and continue to constrain them. They're smart. Their algorithms are getting better. But so are the algorithms of U.S. companies.

And this should be a reminder that chip controls are important, China is a worthy competitor here, and we shouldn't take anything for granted. But I don't think this is the time to say the sky is falling or the fundamental scaling laws have broken.

Where do you think they got their performance increases from? We read their papers. They're smart people who are doing exactly the same kind of algorithmic efficiency work that companies like Google and Anthropic and OpenAI are doing.

One common argument I have heard on the left — Lina Khan made this point — was that DeepSeek proved our whole paradigm of A.I. development was wrong: We did not need all this compute, we did not need these giant megacompanies, that DeepSeek was showing a way toward a decentralized almost solarpunk version of A.I. development. And, in a sense, the American system and imagination had been captured by these three big companies.

But what we're seeing from China is: That wasn't necessarily needed. We could do this on less energy, fewer chips, less footprint.

Do you buy that?

I think two things are true here. The first is there will always be a frontier, at least for the foreseeable future. There will be a frontier that is computationally and energy intensive. And we want our companies to be at that frontier.

Those companies have very strong incentive to look for efficiencies, and they all do. They all want to get every single, last juice of insight from each squeeze of computation. But they will continue to need to push the

frontier.

Then, in addition to that, there will be a slower diffusion that lags the frontier, where algorithms get more efficient, fewer computer chips are required, less energy is required. And we need to win both those competitions.

The A.I. firms want the export controls, but the semiconductor firms don't. DeepSeek rocked the U.S. stock market by making people question Nvidia's long-term worth. Nvidia very much doesn't want these export controls.

You were at the White House, at the center of a bunch of this lobbying back and forth. How do you think about this?

Every advanced A.I. chip that gets made will get sold. The market for these chips is extraordinary right now, and I think for the foreseeable future. So I think our view was we put the export controls on —

Wait, but Nvidia didn't think that. The stock market didn't think that.

We put on the first export controls in October 2022. Nvidia stock has 10x-ed since then.

I'm not saying we shouldn't do the export controls. But I want you to take the strong version of the argument, not the weak one. I don't think Nvidia's chief executive is wrong to say that if Nvidia cannot export its top chips to China, that reduces the market for Nvidia's chips in the long run.

Sure. I think the dynamic is right. If they had a bigger market, they could charge on the margins more. That's obviously the supply and demand.

I think our analysis was: Considering the importance of these chips and the A.I. systems they make to U.S. national security, this is a trade-off

that's worth it.

Nvidia, again, has done very well since we put the export controls out. Nvidia is currently trading, even post-DeepSeek, something like 50 times earnings. So the market is continuing to expect they will grow. And I agree with that.

The Biden administration was also generally concerned with A.I. safety. I think it was influenced by people who care about A.I. safety. And that has created a backlash from the accelerationist side of this debate.

I want to play a clip for you from Marc Andreessen, a prominent venture capitalist and a top Trump adviser, describing the conversations he had with the Biden administration on A.I. and how they radicalized him in the other direction.

> ***Archived clip of Marc Andreessen:*** *Ben [Horowitz] and I went to Washington in May of '24. And we couldn't meet with Biden because, as it turns out, at the time, nobody could meet with Biden. But we were able to meet with senior staff. And so we met with very senior people in the White House, in the inner core. And we basically relayed our concerns about A.I. And their response to us was: Yes, the national agenda on A.I., as we will implement in the Biden administration in the second term is: We are going to make sure that A.I. is going to be only a function of two or three large companies. We will directly regulate and control those companies. There will be no start-ups. This whole thing where you guys think you can just start companies and write code and release code to the internet, those days are over. That's not happening.*

Were you part of the conversation he was describing there?

I met with him once. I don't know exactly —

Would that characterize the conversation he had with you?

He talked about concerns related to start-ups and competitiveness. And I think my view on this is: You look at our record on competitiveness, and it's pretty clear that we wanted a dynamic ecosystem.

The A.I. executive order, which President Trump just repealed, had a pretty lengthy section on competitiveness. The Office of Management and Budget memo, which governs how the U.S. government buys A.I., had a whole callout in it, saying: We want to buy from a wide variety of vendors.

The CHIPS and Science Act also has a bunch of things in there about competition.

So I think our view on competition is pretty clear. Now I do think there are structural dynamics related to scaling laws and the like that will force things toward big companies that I think, in many respects, we were pushing against. But I think our track record on competition is pretty clear.

The view that I understand Andreessen arguing with — which is a view I have heard from people in the A.I. safety community but is not a view I had necessarily heard from the Biden administration — was that you will need to regulate the frontier models of the biggest labs when it gets sufficiently powerful.

And in order to do that, you will need controls on those models. You just can't have the model and everything floating around so everybody can run this on their home laptop.

I think that's the tension he's getting at. It gets at a bigger tension, which is how much to regulate this incredibly powerful and fast-changing technology such that, on the one hand, you're keeping it

safe, but on the other hand, you're not overly slowing it down or making it impossible for smaller companies to comply with these new regulations as they're using more and more powerful systems.

In the president's executive order, we actually tried to wrestle with this question, and we didn't have an answer when that order was signed in October 2023.

What we did on the open-source question in particular — and I think we should just be precise here, at the risk of being academic again: But what we're talking about are open-weight systems.

Can you explain what weights and open weights are in this context?

In the training process for an A.I. system, you run an algorithm through this huge amount of computational power that processes the data. The output at the end of that training process, loosely speaking — and I stress this is the loosest possible analogy — are roughly akin to the strength of connections between the neurons in your brain. And in some sense, you could think of this as the raw A.I. system.

When you have these weights, one thing that some companies like Meta and DeepSeek choose to do is publish them out on the internet, which makes them what we call open-weight systems.

The crucial thing about an open-weight system, on the good side, is that it's much easier to innovate with that system to use as a basis for future systems because you've got access to the raw thing.

Maybe the riskier side is: Any safeguards that were built into that system — say, to refuse when a user asks you to help develop a biological weapon — are pretty easy to remove.

I'm a huge believer in the open-source ecosystem. Many of the companies that publish the weights for their system do not make them

open-source. They don't publish the code and the like. So I don't think they should get the credit of being called open-source systems — at the risk of being pedantic.

But open-weight systems are something we thought a lot about in '23 and '24. And we sent out a pretty wide-ranging request for comment. We got a lot of comments back, and what we came to in the report that was published in July or so of '24 was basically: There is no evidence yet to constrain the open-weight ecosystem. The open-weight ecosystem does a lot for innovation and the like, which I think is manifestly true. But we should continue to monitor this as the technology gets better.

Basically exactly the way that you described.

When you were getting those comments, not just on the open-weight models but also when you were talking to the heads of these labs, what did they want? What would you say was the consensus from the A.I. world, to the extent there was one, of what they needed to get there quickly?

And also — because I know many people in these labs are worried about what it would mean if these systems were unsafe — what would you describe as their consensus on safety?

I mentioned before this core intellectual insight: This technology — for the first time, maybe in a long time — is a revolutionary one that is not funded by the government. In its early incubator days, that was the theme from the labs — a sort of: Don't you know we're inventing something very powerful? Ultimately, it's going to have implications on the work you do in national security and the way we organize our society.

More than any kind of individual policy request, they were basically saying: Get ready for this.

The closest thing we did to any kind of regulation was, after the labs made voluntary commitments to do safety testing, we said: You have to share those safety test results with us, and you have to help us understand where the technology is going.

That only applied to the top couple of labs. The labs never knew that was coming and weren't all thrilled about it when it came out.

So the notion that this was kind of a regulatory capture — that we were asked to do this — is simply not true.

But in my experience, I never got individual policy lobbying from the labs. I got much more: This is coming. It's coming much sooner than you think. Make sure you're ready.

To the degree that they were asking for something in particular, it was a corollary of that: We're going to need a lot of energy, and we want to do that here in the United States, and it's really hard to get the power here in the United States.

But that has become a pretty big question. If this is all as potent as we think it will be, and you end up having a bunch of the data centers containing all the model weights and everything else, in a bunch of, say, Middle Eastern petrostates — because they will give you huge amounts of energy access in return for having some purchase on this A.I. world — there is something to this question.

Yeah. This is actually an area of bipartisan agreement.

We really started to pay a lot of attention to this in the later part of '23 and most of '24, when it was clear this was going to be a bottleneck.

In his last week or so in office, President Biden signed an A.I. infrastructure executive order, which has not been repealed, which basically tries to accelerate the power development and the permitting of

power and data centers here in the United States, basically for the reason that you mentioned.

As someone who truly believes in climate change and environmentalism and clean power, I thought there was a double benefit to this, which is that if we did it here in the United States, it could catalyze the clean energy transition. And these companies, for a variety of reasons, are willing to pay more for clean energy and on things like geothermal.

Our hope was we could catalyze that development and bend the cost curve and have these companies be the early adopters of that technology so we'd see a win on the climate side, as well.

There are warring cultures around how to prepare for A.I. — I have mentioned A.I. safety and A.I. accelerationism.

JD Vance just went to the big A.I. summit in Paris**. I'll play a clip of what he said:**

> ***Archived clip of JD Vance:*** *When conferences like this convene to discuss a cutting-edge technology, oftentimes I think our response is to be too self-conscious, too risk averse. But never have I encountered a breakthrough in tech that so clearly calls us to do precisely the opposite.*

What do you make of that?

I think he is setting up a dichotomy there that I don't quite agree with. And the irony of that is, if you look at the rest of his speech, which I did watch, there's actually a lot that I do agree with.

For example, I think he's got four pillars in the speech. One is about centering the importance of workers. One is about American pre-eminence.

And those are entirely consistent with the actions that we took and the philosophy that I think the administration of which I was a part espoused

and that I certainly believe in.

Insofar as he is saying that safety and opportunity are in fundamental tension, I disagree. If you look at the history of technology and technology adaptation, the evidence is pretty clear that the right amount of safety action unleashes opportunity and, in fact, unleashes speed.

One of the examples that we studied a lot and talked to the president about was the early days of railroads. There were tons of accidents and crashes and deaths. And people were not inclined to use railroads as a result.

What started happening was safety standards and safety technology: block signaling so that trains could know when they were in the same area, air brakes so that trains could brake more efficiently, the standardization of train track widths and gauges and the like.

This was not always popular at the time, but with the benefit of hindsight, it is very clear that kind of technology and, to some degree policy development of safety standards, made the American railroad system in the late 1800s. And this is the pattern that shows up a bunch throughout the history of technology.

To be very clear, it is not the case that every safety regulation on every technology is good. And there are certainly cases where you can overreach and you can slow things down and choke things off. But I don't think it's true that there's a fundamental tension between safety and opportunity.

That's interesting because I don't know how to get this point of regulation right.

I think the counterargument to Vice President Vance is nuclear. Nuclear power is a technology that both held extraordinary promise,

maybe still does. And also you can really imagine every country wanting to be in the lead on it.

But the series of accidents — most of which did not even have a particularly significant body count — were so frightening to people that the technology got regulated to the point that nuclear advocates believe it has been largely strangled in the crib from what it could be.

The question then is: When you look at the actions we have taken on A.I., are we strangling it in the crib, and have we taken actions that are akin to —

I'm not saying that we've already done it. Look, if these systems are going to get more powerful and they're going to be in charge of more things, things are both going to go wrong and they're going to go weird. It's not possible for it to be otherwise, to roll out something this new in a system as complex as human society.

So there's going to be this question of: What are the regimes that make people feel comfortable moving forward from those kinds of moments?

I think that's a profound question. What we tried to do in the Biden administration was set up institutions in the government to do that in as cleareyed, tech-savvy way as possible.

Again, with the one exception of the safety test results sharing, which some of the chief executives estimate cost them one day of employee work, we did not put anything close to regulation in place.

We created something called the U.S. Artificial Intelligence Safety Institute, which is purely national-security focused — cyber-risks, biorisks, A.I. accident risks — and purely voluntary and that has relationships, a memorandum of understanding, with Anthropic, with

OpenAI, even with xAI, Elon's company.

And basically I think we saw that as an opportunity to bring A.I. expertise into the government. To build relationships between the public and private sector in a voluntary way. And as the technology develops, it will now be up to the Trump administration to decide what they want to do with it.

But I think you are quite diplomatically understating what is a genuine disagreement here.

What I would say Vance's speech was signaling was the arrival of a different culture in the government around A.I. There has been an A.I. safety culture where we have all these conferences about what could go wrong.

And Vance is saying: Stop it. Yes, maybe things could go wrong. But instead we should be focused on what could go right.

And frankly, I would say this is the generalized view of the Trump-Musk administration — which I think is, in some ways, the right way to think about the administration. That if something goes wrong, we'll deal with the thing that went wrong afterward. But what you don't want to do is move too slowly because you're worried about things going wrong. That it's better to break things and fix them than to have moved too slowly in order not to break them.

I think it's fair to say that there is a cultural difference with the Trump administration and us on some of these things.

But we held conferences on what you could do with A.I. and the benefits of A.I. We talked all the time about the need to mitigate these risks, but you're doing so so you can capture the benefits.

And I'm someone who reads an essay like "Machines of Loving Grace,"

by Dario Amodei, the chief executive of Anthropic, that is basically about the upside of A.I., and says: There's a lot in here we can agree with.

The president's executive order said we should be using A.I. more in the executive branch. So I hear you on the cultural difference. I get that. But I think when the rubber meets the road we were comfortable with the notion that you could both realize the opportunity of A.I. while doing it safely.

And now that they are in power, they will have to decide how to translate Vice President Vance's rhetoric into a governing policy. My understanding of their executive order is they've given themselves six months to figure out what they're going to do. And I think we should judge them on what they do.

Let me ask about the other side of this. What I liked about Vance's speech is: I think he's right that we don't talk enough about opportunities. But more than that, we are not preparing for opportunities.

If you imagine that A.I. will have the effects and possibilities that its backers and advocates hope, one thing that implies is: We're going to start having a much faster pace of the discovery or proposal of novel drug molecules. A very high promise.

The idea here, from the people I've spoken to, is that A.I. should be able to ingest an amount of information and build the modeling of diseases in the human body that could get us a much better drug discovery pipeline.

If that were true, you can ask: What's the choke point going to be?

Our drug-testing pipeline is incredibly cumbersome. It's very hard to

get the animals you need for trials, very hard to get the human beings you need for trials. You could do a lot to make that pipeline faster.

And this is true in a lot of different domains — education, etc. It's pretty clear that the choke points will become the difficulty of doing things in the real world. And I don't see society preparing for that. Maybe we're not doing that much on the safety side because we don't know what we should do.

But on the opportunity side, this question of how to actually make it possible to translate the benefits of A.I. quickly seems like a much richer conversation than I've seen anybody seriously having.

I basically agree with all of that. The conversation when we were in the government, especially in '23 and '24, was starting to happen. We looked at the clinical trials thing.

You've written about health care for however long, and I don't claim expertise on health care. But it does seem to me that we want to get to a world where we can take the breakthroughs, including breakthroughs from A.I. systems, and translate them to market much faster.

This is not a hypothetical thing. It's worth noting that, quite recently, Google came out with — I think they called it a co-scientist. Nvidia and the Arc Institute, which does great work, had the most impressive biodesign model ever that has a much more detailed understanding of biological molecules. A group called FutureHouse has done similarly great work in science.

So I don't think this is hypothetical. This is happening right now. And I agree with you that there's a lot that can be done institutionally and organizationally to get the federal government ready for this.

I've been wandering around Washington, D.C. this week and talking to a lot of people involved in the Trump administration — people from different factions of what I consider the modern right.

I've been surprised how many people understand either what Trump and Musk and DOGE are doing, or at least what it will end up allowing, as related to A.I. — including people who are not a part of the tech right, and from whom I would not really expect to hear that.

What they basically say is: The federal government is too cumbersome to take advantage of A.I. as a technology. So if the whole point of A.I. is that it accelerates cognitive work, the government needs to be stripped down and rebuilt to take advantage of A.I. — which, like it or hate it, is what Musk and DOGE are doing. And the dismantling of the government allows for a creative destruction that paves the way for the government to better use A.I.

Do you buy that?

It feels orthogonal to what I've observed from DOGE. I think Musk is someone who does understand what A.I. can do, but I don't know how starting with the United States Agency for International Development, for example, prepares the U.S. government to make better A.I. policy.

So I guess I don't buy that as the motivation for DOGE.

Is there something to the broader argument, though? I will say, I don't buy the argument about DOGE — I would make the same point you just made.

But what I do buy is that I know how the federal government works pretty well, and it is too slow to modernize technology. It is too slow to work across agencies. And it is too slow to radically change the way things are done and take advantage of things that could be

productivity enhancing.

I couldn't agree more. The existence of my job as the White House special adviser for A.I. — which David Sacks is now — and I had this job in 2023 — existed because President Biden said very clearly, both publicly and privately: We cannot move at the typical government pace. We have to move faster here.

I think we probably need to be careful, and I'm not here for stripping it all down. But I agree with you. We have to move much faster.

Another major part of Vice President Vance's speech was signaling to the Europeans that we are not going to sign on to complex multilateral negotiations and regulations that could slow us down — and that if they passed such regulations in a way that is penalizing our A.I. companies, we would retaliate.

How do you think about the differing position the new administration is moving into vis-à-vis Europe and its broad approach to tech regulation?

I think the honest answer here is we had conversations with Europe as they were drafting the European Union A.I. Act.

At the time that I was in the E.U., the A.I. Act was still nascent. The act had passed, but a lot of the actual details of it had been kicked to a process that is still unfolding. So —

Speaking of slow-moving bureaucracies.

Exactly. So maybe this is a failing on my part, but I did not have particularly detailed conversations with the Europeans beyond a general articulation of our views. They were respectful. We were respectful. But I think it's fair to say we were taking a different approach than they were taking. And insofar as safety and opportunity are a dichotomy — which I

don't think are a pure dichotomy — we were ready to move very fast in the development of A.I.

One of the other things that Vance talked about, and that you said you agreed with, is making A.I. pro-worker. What does that mean?

It's a vital question. We instantiate that in a couple of different principles. The first is that A.I. in the workplace needs to be implemented in a way that is respectful of workers and the like.

And one of the things I know the president thought a lot about was how it is possible for A.I. to make workplaces worse in a way that is dehumanizing and degrading and ultimately destructive for workers. So that is the first distinct piece of it that I don't want to neglect.

The second is: I think we want to have A.I. deployed across our economy in a way that increases workers, agencies and capabilities. And I think we should be honest that there's going to be a lot of transition in the economy as a result of A.I.

I don't know what that will look like. You can find Nobel prizewinning economists who will say it won't be much. You can find other folks who will say it will be a ton. I tend to lead toward the side that says it's going to be a lot. But I'm not a labor economist.

The line that Vice President Vance used is the exact same phrase that President Biden used, which is: Give workers a seat at the table in that transition.

And I think that is a fundamental part of what we're trying to do here and, I presume, what they're trying to do here.

I've heard you beg off on this question by saying you're not a labor economist.

I am not a labor economist, Ezra. [Laughs.]

I will promise you the labor economists do not know what to do about A.I. You were the top adviser for A.I. You were at the nerve center of the government's information about what is coming. If this is half as big as you seem to think it is, it's going to be the single most disruptive thing to hit labor markets ever, given how compressed the time period is in which it will arrive.

It took a long time to lay down electricity. It took a long time to build railroads. A.I. is going to come really quickly.

I think that is basically true, but I want to push back a little bit. I do think we are going to see a dynamic in which it will hit parts of the economy first. It will hit certain firms first. But it will be an uneven distribution across society.

Well, I think it will be uneven, and that's what will be destabilizing about it in part.

Let me give you an example of the kind of thing I'm worried about. There are a lot of 19-year-olds in college right now studying marketing. And there are a lot of marketing jobs that, frankly, A.I. can do perfectly well right now.

As we get better at knowing how to direct A.I., I mean, one of the things that's going to slow this down is simply firm adaptation. But the thing that will happen very quickly is you'll have firms that are built around AI.

It's going to be harder for the big firms to integrate it, but what you're going to have is new entrants who are built from the ground up, where their organization is built around one person overseeing these, like, seven systems. So you might just begin to see triple the unemployment among marketing graduates.

I'm not convinced you'll see that in software engineers. I think A.I. is going to both take a lot of those jobs but also create a lot of those jobs because there's going to be so much more demand for software. But you could see it happening somewhere there.

There are just a lot of jobs that are doing work behind a computer. And as companies absorb machines that can do work behind a computer for you, that will change their hiring.

You must have heard somebody think about this. You guys must have talked about this.

We did talk to economists to try to texture this debate in '23 and '24.

The trend line is even clearer now than it was then. We knew this was not going to be a '23 and '24 question. Frankly, to do anything robust about this was going to require Congress, and that was just not in the cards at all.

So it was more of an intellectual exercise than it was a policy —

But policies begin as intellectual exercises.

Sure, I think that's fair. The advantage to A.I. that is, in some ways, a countervailing force here — though I hear you and mostly agree with your side's argument — is that it will increase the amount of agency for individual people.

So I do think we will be in a world in which the 19-year-old or the 25-year-old will be able to use a system to do things they were not able to do before. And insofar as the thesis we're batting around here is that intelligence will become a little bit more commoditized, what will stand out more in that world is agency and the capacity to do things. And I think that could, in the aggregate, lead to a pretty dynamic economy.

And the economy you're talking about of small firms and a dynamic

ecosystem and robust competition, on balance, at an economy scale, is not in itself a bad thing. Where I imagine you and I agree — and maybe Vice President Vance, as well, agrees — is we need to make sure that individual workers and classes of workers are protected in that transition.

I think we should be honest: That's going to be very hard. We have never done that well.

I couldn't agree with you more. In a big way, Donald Trump is president today because we did a shitty job on this with China.

The reason I'm pushing on this is that we have seen this coming for a while. But as I look around, I do not see a lot of useful thinking here. And I grant that we don't know the shape of it. But, at the very least, I would like to see some ideas on the shelf for what we should do if the disruptions are severe.

We are so addicted in this country to an economically useful tale — that our success is in our own hands — that it makes it very hard for us to react with either compassion or realism when workers are displaced for reasons that are not in their own hands — because of global recessions or depressions because of globalization.

There are always some people with the agency and the creativity who become hyperproductive. There's often this sentiment of: Look at them. Why aren't you them?

I'm definitely not saying that.

I know you're not saying that. But that's such an ingrained American way of looking at the economy: You should do some retraining.

Are all these people going to become nurses? There are things that A.I. can't do. How many plumbers do we need? More than we have,

actually. But does everybody move into the trades?

What were the intellectual thought exercises that all these smart people at the White House who believe this was coming — what were you saying?

Yes, we were thinking about this question. We knew it was not going to be a question we were going to confront in the president's term. We knew it was a question that you would need Congress to do anything about.

Insofar as what you're expressing here seems to be a deep dissatisfaction with the available answers, I share that. I think a lot of us shared that.

You can get the usual stock answers of a lot of retraining. But I share your sort of doubts of that as the answer. You could probably talk to some Silicon Valley libertarians or tech folks, and they'd say: universal basic income.

I believe, and I think President Biden believes, there's a kind of dignity that work brings. And it doesn't have to be paid work, but there needs to be something that people do each day that gives them meaning.

Insofar as you have a discomfort with where this is going on the labor side, I share that. Even though I don't know the shape of it.

More than that, I guess I have a discomfort with the quality of thinking right now. Sort of across the board, but particularly on the Democratic side — because I have you here as a representative of the past administration.

I have a lot of disagreements with the Trump administration, to say the least. But I do understand the people who say: Look, Elon Musk, David Sacks, Marc Andreessen, JD Vance — at the very highest levels of the Trump administration are people who have spent a lot of time thinking about A.I. and have considered very unusual thoughts

about it.

And I think sometimes Democrats are a little bit too institutionally constrained for thinking unusually. So I take your point on the export controls. I take your point on the executive orders and the A.I. Safety Institute.

But to the extent that Democrats imagine themselves to be the party of the working class, and to the extent we've been talking for years about the possibility of A.I.-driven displacements, it's true that when things happen, you need Congress. But you also need thinking that becomes policies that Congress implements.

So I guess I'm trying to push. Was this not being talked about? There were no meetings? You guys didn't have Claude, Anthropic's A.I. assistant, write up a brief of options?

Well, we definitely didn't have Claude write up a brief because we had to get over government use of A.I.

See, but that is itself slightly damning.

Ezra, I agree that the government has to be more forward-leaning on basically all of these dimensions. It was my job to push the government to do that. And I think on things like government use of A.I., we made some progress.

So I don't think anyone from the Biden administration, least of all me, is coming out and saying: We solved it.

What we're saying is: We were building a foundation for something that was coming that was not going to arrive during our time in office and that the next team would have to, as a matter of American national security — and, in this case, American economic strength and prosperity — address.

This gets to something I find frustrating in the policy conversation

about A.I.

You start the conversation about how the most transformative technology — perhaps in human history — is landing in a two- to three-year time frame. And you say: Wow, that seems like a really big deal. What should we do?

That's when things get a little hazy. Maybe we just don't know. But what I've heard you kind of say a bunch of times is: Look, we have done very little to hold this technology back. Everything is voluntary. The only thing we asked was a sharing of safety data.

Now in come the accelerationists. Marc Andreessen has criticized you guys extremely straightforwardly.

Is this policy debate about anything? Is it just the sentiment of the rhetoric? If it's so [expletive] big, but nobody can quite explain what it is we need to do or talk about — except for maybe export chip controls — are we just not thinking creatively enough? Is it just not time? Match the calm, measured tone of this conversation with our starting point.

I think there should be an intellectual humility here. Before you take a policy action, you have to have some understanding of what it is you're doing and why.

So it is entirely intellectually consistent to look at a transformative technology, draw the lines on the graph and say that this is coming pretty soon, without having the 14-point plan of what we need to do in 2027 or 2028.

Chip controls are unique in that this is a robustly good thing that we could do early to buy the space I talked about before. But I also think that we tried to build institutions, like the A.I. Safety Institute, that would set

the new team up, whether it was us or someone else, for success in managing the technology.

Now that it's them, they will have to decide as the technology comes on board how we want to calibrate this under regulation.

What kinds of decisions do you think they will have to make in the next two years?

You mentioned the open-source one. I have a guess where they're going to land on that. But there's an intellectual debate there that is rich. We resolved it in one way by not doing anything. They'll have to decide if they want to keep doing that.

Ultimately, they'll have to answer the question of: What is the relationship between the public sector and the private sector? Is it the case, for example, that the kind of things that are voluntary now with the A.I. Safety Institute will someday become mandatory?

Another key decision is: We tried to get the ball rolling on the use of A.I. for national defense in a way that is consistent with American values. They will have to decide what that continues to look like and if they want to take away some of the safeguards that we put in place to go faster.

So I think there really are a bunch of decisions that they are teed up to make over the next couple of years that we can appreciate are coming down the horizon without me sitting here and saying: I know with certainty what the answer is going to be in 2027.

And then always our final question: What are three books you'd recommend to the audience?

One of the books is "The Structure of Scientific Revolutions" by Thomas Kuhn. This is the book that coined the term "paradigm shift," which is basically what we've been talking about throughout this whole

conversation — a shift in technology and scientific understanding and its implications for society. I like how Kuhn, in this book, which was written in the 1960s, gives a series of historical examples and theoretical frameworks for thinking about a paradigm shift.

Another book that has been very valuable for me is "Rise of the Machines" by Thomas Rid, which tells the story of how machines that were once the playthings of dorks like me became in the '70s and the '80s things of national security importance. We talked about some of the revolutionary technologies here — the internet, microprocessors and the like — that emerged out of this intersection between national security and tech development. And I think that history should inform the work we do today.

And then the last book is definitely an unusual one, but I think it's vital: "A Swim in the Pond in the Rain" by George Saunders. Saunders is this great essayist and short story writer and novel writer. He teaches Russian literature and, in this book, takes seven Russian literature short stories and gives a literary interpretation of them.

What strikes me about this book is that it's fundamentally the most human endeavor I can think of: He is taking great human short stories, and he's giving a modern interpretation of what those stories mean. And when we talk about the kinds of cognitive tasks that are a long way off for machines, I hope, at some level, this is one of them — that there's something fundamentally human that we alone can do. I'm not sure if that's true, but I hope it's true.

This episode of "The Ezra Klein Show" was produced by Rollin Hu. Fact-checking by Michelle Harris, with Kate Sinclair and Mary Marge Locker. Mixing by Isaac Jones, with Efim Shapiro and Aman Sahota. Our supervising editor is Claire Gordon. The show's production team also

includes Elias Isquith, Kristin Lin and Jack McCordick. Original music by Pat McCusker. Audience strategy by Kristina Samulewski and Shannon Busta. The executive producer of New York Times Opinion Audio is Annie-Rose Strasser. Special thanks to Switch and Board Podcast Studio.

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow the New York Times Opinion section on Facebook, Instagram, TikTok, Bluesky, WhatsApp and Threads.

Ezra Klein joined Opinion in 2021. He is the host of the podcast "The Ezra Klein Show" and the author of "Why We're Polarized" and, with Derek Thompson, "Abundance." Previously, he was the founder, editor in chief and then editor at large of Vox. Before that, he was a columnist and editor at The Washington Post, where he founded and led the Wonkblog vertical. He is on Threads.

Advertisement

SKIP ADVERTISEMENT