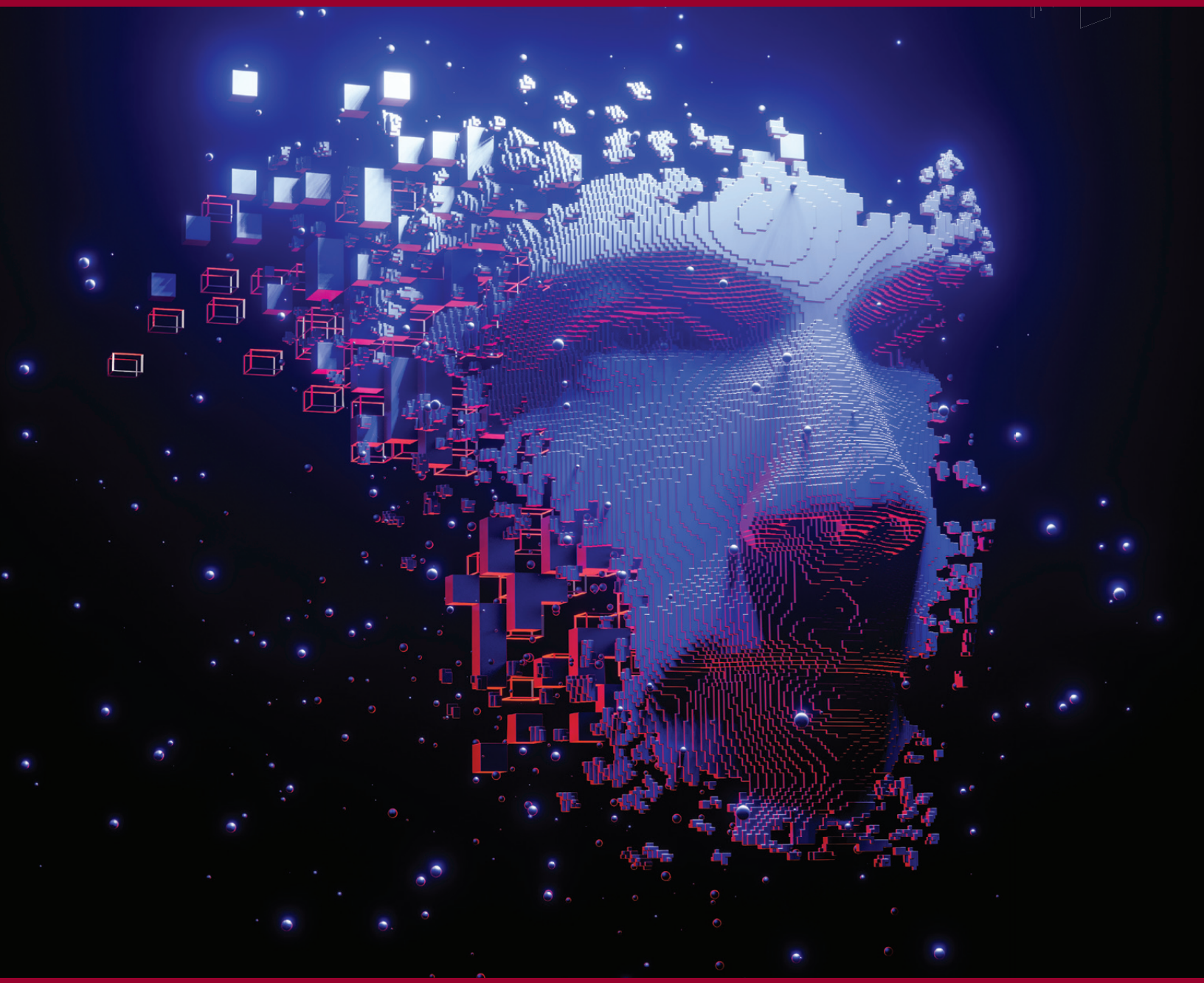
DECEMBER 2024

Defense Against the AI Dark Arts

Threat Assessment and Coalition Defense

Philip Zelikow, Mariano-Florentino Cuéllar,
Eric Schmidt, and Jason Matheny



A PUBLICATION OF THE HOOVER INSTITUTION

HOOVER INSTITUTION

Defense Against the AI Dark Arts

Threat Assessment and Coalition Defense

Philip Zelikow, Mariano-Florentino Cuéllar, Eric Schmidt, and Jason Matheny

An intelligence and defense agenda for artificial intelligence (AI) will likely take shape in 2025. The present agenda for "AI safety" is now commonly equated with checking the safety of products from private firms that use AI. The new agenda we propose goes far beyond that. The United States and its partners must get ready for what the worst people in the world might do with the most advanced AI models.

That agenda will be separate from ongoing arguments about how the US military should adapt to more use of AI or guide the use of autonomous combat systems. Those military plans are not the focus of this paper.

We focus on the overarching geopolitical risk: What might happen if America and its friends became vulnerable to world-changing surprise as powerful enemies wield *their* AI tools?

Many Americans assume that the United States is far ahead in the development of AI. Such beliefs are too complacent. For instance, the world's best open-weights model might now be Chinese. That is the takeaway from a recent Tencent paper that introduces Hunyuan-Large. In a broad range of benchmarks, the model outperforms Facebook's Llama 3.1 405B parameter model.[1] Even newer Chinese models, like Qwen 2.5 and DeepSeek R1, are also impressive.

In other words, competence is widespread; it just may be the available computing power that matters. And countries such as China can offset some (probably just temporary) weaknesses in that computing power by the way in which they organize their networks, by their access to energy resources, and by their preexisting advantages in accessing data without some of the legal constraints (like copyright law) found elsewhere.

In other words, *the United States must now start working very hard with allies to secure democratic advantage in the domain of frontier AI*. We suggest a way to manage the

A Hoover Institution Essay

December 2024

convergence of three great vectors: private sector–led innovation, emerging threats, and international efforts.

How to balance these three conditions? An essential starting point is to build a defensive agenda, join private innovation with public needs, and rally an international effort. The time to start shaping that agenda constructively has arrived.

The agenda must be designed for coalition intelligence and defense, not just America's "national security." American leaders may put America first. Other leaders will also understandably put their own countries' interests first. Yet, as Benjamin Franklin once put it, we and at least some of our friends "must, indeed, all hang together or, most assuredly, we shall all hang separately."

Consider the worldwide talent pool; the worldwide networks that produce or power the technology, including chips, tools, cloud infrastructure, and software; and the worldwide flows of information, finance, and commerce that sustain private firms at the frontier. That analysis should drive a vision for an allied ecosystem of AI threat assessment and defenses.

That core allied ecosystem must be more than just a coalition of the willing. It has to be a coalition of the ready, willing, and able. Several governments will face strategic choices about who should participate and what participants must do, which is all the more reason to clarify what the core intelligence and defense agenda should be.

We are uncertain about how the technology will develop. AI is becoming a general-purpose technology. Its progress will therefore be driven not only by developments in the AI field but also in how AI is linked to a variety of other technical breakthroughs and many possible applications—including how AI may recursively advance the AI frontier itself.

The history of technology is a history of surprises in how it is used. Often, when technological change accelerates, those most surprised were the pioneers at the frontier. Accordingly:

- Our approach takes no position about the imminence of, or dangers posed by, artificial general intelligence (AGI). By AGI we mean systems that reach or begin to exceed the upper limits of human capabilities across all or most cognitive domains and may rapidly accelerate the development of even more advanced AI that substantially exceeds human capabilities.

- Our approach nonetheless takes seriously the prospect of AGI, even though we don't know right now precisely what the risks will be. We also recognize that some AI breakthroughs may bear little or no connection to AGI or large language models. They may instead involve AI adapted to performing highly specialized tasks.

What we must do is design a policy approach that can cope with uncertainty and defend our societies. To that end, we take up five topics, as follows:

• • •

I. DEFENSIVE MISSIONS OF A COALITION SECURITY ENTERPRISE

A. ASSESS POTENTIAL THREATS, INCLUDING OF STRATEGIC SURPRISE

The US government will be formally responsible for authoritatively evaluating two kinds of threats.

- First, there are the threats that could be posed by the worst people and governments in the world using the most advanced possible models. These include bio and cyber risks or other kinds of strategic surprise that might be developed by hostile states with models or developmental approaches quite different from anything being developed by private companies in the free world.

- Second are threats that could be posed by loss of control of a misaligned, highly capable model, which may be a model or system with AGI capabilities.

Governments will need to establish or designate one or more institutions with lead responsibility for conducting the most sensitive threat evaluation work and thus guiding much of the work on countermeasures and preparedness. This institution or group becomes an intermediary that must be able to work inside and outside the frontier companies, understanding and leveraging their work while protecting the companies' intellectual property (IP).

Fortunately, the British and American governments have pioneered institutions to do this and to start building a workforce and skills. The UK AI Safety Institute is associated with Britain's Department for Science, Innovation and Technology. The US AI Safety Institute, part of America's Department of Commerce, has followed suit.

In a wise move, the United States and United Kingdom have signed a memorandum of understanding (MoU) so their institutes can work cooperatively and transnationally. They have already begun conducting pre-deployment evaluations of frontier models being developed by three of the leading AI companies—a constructive learning experience for both sides.

An October 2024 presidential national security memorandum on AI created a complex interagency process for evaluating the more serious national security threats. It set up the US AI Safety Institute as a kind of hub that can also do classified work in coordination with the various national security agencies.

The British government had already created an analogous process to work with its security services. The British government may be further along in actually conducting, and learning

how to conduct, such sensitive security evaluations. The 2024 presidential national security memorandum requires various reports to be delivered in 2025. The incoming Trump administration will have to decide whether to maintain or accelerate these plans.

It is not clear whether the United States is planning to make the broad investments that will be needed to detect and evaluate frontier AI threats. There are current plans to emphasize the buildup of secure compute capabilities at the Department of Energy's Oak Ridge National Laboratory in Tennessee and its Lawrence Livermore National Laboratory in California.

B. CREATE NECESSARY COUNTERMEASURES

Those who study the defense threats—either from misuse of private products or more deliberate adversary designs—are usually those most expert in identifying the requirements to counter them. In biology, for example, vaccine design depends on scientists working with—and sometimes developing—the dangerous pathogens that vaccines must counter. This is why much vaccine work on pathogens with pandemic potential is conducted in costly Biosafety Level (BSL)-3 or BSL-4 labs with elaborate safeguards.

The same may be true with AI safety, where the countermeasures will probably also involve advanced AI. Thus, both the threat evaluation work and the countermeasures work may also require the assistance of frontier companies, again doing dangerous work at the outer limits, subject to suitable safeguards.

The countermeasures work, like some threat evaluations, may require the use of special data available only to governments, including classified national security data, law enforcement information, or telemetry collected by government systems. Such data might then need to be intermingled with the very costly foundation models, or those models may need to be fine-tuned with such data.

C. UNDERSTAND AND RESOURCE NECESSARY PREPAREDNESS

Preparedness can cover a wide range of measures to help society cope with novel, emerging dangers. Some of these may be domain specific, reflecting the general-purpose nature of AI systems and the difficulty of preparing against every conceivable threat posed by advanced AI.

In cybersecurity, AI may exacerbate dangers now posed only by the most skilled or well-resourced hackers by increasing the productivity of less-skilled attackers. Also, despite the COVID-19 pandemic, the United States remains poorly prepared against biological dangers; these risks are likely to escalate as AI models increasingly become linked to further breathtaking advances in biotechnology.

D. COPE WITH TECHNOLOGICAL UNCERTAINTY

Global business investment in AI is enormous and growing—and AI model capabilities have quickly come to equal or exceed human capabilities in areas like image classification, software development, and linguistic reasoning. These changes have come faster and with greater intensity than many knowledgeable observers predicted even four years ago. They underscore the enormous uncertainty about the pace and extent of change at the technological frontier.

It is hard to overstate the significance of this uncertainty. The computer itself was originally thought to have very limited uses. This modest view was held even by the original innovators and companies building the machines. "We could, if we liked, amuse ourselves indefinitely at the failure of earlier generations to see the obvious, as we see it today," Nathan Rosenberg wrote in a classic essay nearly thirty years ago.

"But," Rosenberg added, "that would be a mistaken conceit. . . . Much of the difficulty," he argued, "is connected to the fact that new technologies typically come into the world in a very primitive condition. Their eventual uses turn upon an extended improvement process that vastly expands their practical applications."[2]

The binary notion that there is some great invention followed by mere adaptation is misconceived. It is more useful to think of an evolutionary process, where some initial innovation is followed by a succession of improvements, some of which may then become transformational.

Also, innovations are often paired across fields and disciplines in unexpected ways, as with the way lasers and glass fibers were paired to revolutionize communications or the way biotech has exploded from the combination of the discoveries of DNA and the genome into the use of digital technologies in gene-splicing or powerful processing of genomic data.

For security policy, the problem of coping with uncertainty requires a fundamental judgment about the possible significance of AI as a general-purpose technology likely to experience substantial evolution and progress in the coming years.

Coping with uncertainty, governments have to acquire an *independent* capacity to evaluate the most dangerous risks, including risk from enemies. They should start with much more experimentation of how even current models can be used or misused. To acquire the full ability to evaluate the most dangerous risks, governments will have to articulate the scope of work for the project and carefully specify the resources and capabilities necessary to perform that work.

II. SCOPE OF WORK FOR THE COALITION SECURITY ENTERPRISE

A. THE STATUS QUO: PRIVATE PRODUCTS AND GRAVE DANGERS OF MISUSE

Evaluating the product safety risk and evaluating the defense challenge are two different challenges. The key point to notice is that they overlap. They overlap in the evaluation work, and they overlap in the role private companies may play in actually doing the work, whether on safety or on defense. Product safety work is already under way and at the moment, that is what is driving the rise of safety research and norms.

The British and American AI Safety Institutes and AI safety institutes (AISIs) being established in other countries are brand-new public institutions, little more than a year old. They were brilliantly improvised rapid responses to startling progress in private-sector AI innovation. The 2023 Bletchley Park AI Safety Summit helped launch the AISI precedent. At a successor summit in 2024 in Seoul, ten countries and the EU agreed to create an international network of national AISIs. China did not sign that commitment but appears to be planning to establish an AISI as well.

Some companies have made a major effort to reassure anxious publics and governments. At the international level, a G7 Hiroshima AI Process agreed in 2023 to articulate an evolving "international code of conduct" for companies. The Bletchley Park process and the Organisation for Economic Co-operation and Development (OECD) have created similar opportunities for companies to make voluntary commitments that they will monitor the risks in frontier AI development. At the Seoul summit, sixteen companies agreed to publish their approaches to assessing risk and promised not to publish models that exceeded thresholds the companies regarded as "intolerable."

The product evaluation, sometimes called "dangerous capability evaluations" (DC evals for short), should have at least four objectives:

1. Assess Risks in Relation to Utility

The products cannot be evaluated only in their intended uses. They will be misused.

There is a distinction between models that use application programming interfaces (APIs) and models that share the model weights with users. In the latter case, it is already obvious that bad actors will strip away guardrails, cannibalize the published model, fine-tune it with malicious data to serve their purposes, and ignore the seller's warnings or licensure requirements. It may also be possible to fine-tune models accessed through APIs, though in theory such misuse could be monitored by the company hosting the platform.

Yet many essential products are risky. Tools now commonplace in modern life—e.g., steam engines, chemicals, electric power, railroads, automobiles, and aircraft—all presented or still present serious dangers, accompanied by much loss of life.

So evaluation has to spot risks, but it has to put them in the context of social benefits. The evaluators themselves are not the right people to strike the balance. Democratic governance should do that. The evaluators should inform that governance.

2. Assess Manageability of Risks and Evolving Best Practices

Creators of the main large language model products in the United States and Europe have gone through costly voluntary efforts to improve safety. They do this through their model design, in the data used for training, and constraints in how the models are used.

Over time, the practice of risk (and utility) assessment along with advances in technical knowledge can help us enhance the practice of safety and security evaluation. That we can expect better and more precise knowledge of AI capabilities and threats over time is not a reason to delay such assessment or slow down investment in this area. It is why governments must get going with such assessments with private-sector partners, assisted by research centers or nonprofits like Apollo Research, METR, and others. Knowledge will improve with practice.

A major topic for further research and action will be the potential for technical risk mitigation. This can include certain design features of the expensive hardware to limit misuse. The hardware is difficult to replace, and so it may be harder to circumvent such measures. Work on these ideas is still in a very early stage.

Governments may want to develop a playbook of options they can employ as they approach a world of meaningful recursive self-improvement (RSI) of AI systems that may then accelerate the pace of work toward AGI. This could happen in 2025. AI is already boosting progress in the development of both hardware and software.

To have such a playbook of options, governments will need enough ongoing knowledge of, and access to, the most advanced models so that they can decide when and whether to take extraordinary measures to secure them.

3. Establish Guidances, Protocols, Standards, and Safe Harbors

Whether they use standards of negligence or strict liability for unreasonably dangerous products, courts usually end up with calculations that balance risk and utility. They get to these calculations as they weigh whether producers took reasonable or available precautions in manufacture, design, or warning. It will be difficult to make these assessments in the early cases involving novel AI products, whose inner workings can be mysterious even to their makers. As courts help police the allocation of risk among AI users, technology companies, insurers, third parties, and the government (among others), early decisions in different jurisdictions may contribute to a period of uncertainty before courts begin to converge.

The AISIs can help evaluate the risk-utility trade-offs. Judicial decisions about liability often pivot on the content of expert analyses or industry standards that define reasonable behavior in a particular field, even when these are not binding. The AISI contributions can also be incorporated into the work of law-oriented standard-setting organizations such as the American Law Institute, which is working on a project defining AI liability standards.

In other words, if the AISI analyses show catastrophic risks that well outweigh any likely utility, those findings would create enormous liability risks for any makers of such products. Those risks could be on a scale that might bankrupt even the biggest companies.

The AISIs can quickly provide publicly released analyses and voluntary best-practices guidance and protocols that can evolve with the technology. These guides and protocols would then contribute to the design of *explicit safe harbors* that reduce uncertainty and standardize what counts as reasonable behavior in the development of frontier AI.

Where appropriate, the guides and protocols could be agreed upon among more than one government through forums such as the AI safety summits. In time these protocols can form the basis for formal technical standards.

Formal standard-setting is a complex and technical process. The International Organization for Standardization that sets ISO standards is a nongovernmental forum that has been bringing together communities of experts since 1946 to suggest various categories of voluntary standards that have been widely used. One such community in the United States, for example, is the Institute of Electrical and Electronics Engineers.

So our suggestion is: Put voluntary best-practices guidance and protocols first. Use those to help establish international norms and safe harbors. Then work to set these into technical standards over time as AI safety research matures.

4. Further Define "Independent" Evaluation

Companies should not grade their own homework. But government agencies may not have the expertise or infrastructure to do all the work themselves. Analogous problems have arisen in the pre-deployment testing of several kinds of products from microscopically prepared pharmaceuticals to large aircraft and spacecraft.

Whether or not governments decide to enact more mandatory requirements, they have to build the capacity to make their own independent evaluations of risk. Almost all AISIs are not yet well funded, and their efforts are in their infancy. They are only now beginning to learn how to evaluate those private products, partly building on the best practices pioneered by some conscientious early frontier companies.

Most of the money, workforce, and expertise to evaluate AI product safety still resides in a handful of private companies. With relatively modest legal authorities so far, the AISIs

are trying to keep pace. In at least three cases, companies have allowed some independent pre-deployment evaluation of frontier large language models. Both private and public actors are still feeling their way.

Such companies have created a private Frontier Model Forum to share information with one another about best practices, requiring a commitment to safety as a condition of membership.

Also, while it is widely assumed that only a few frontier companies can produce potentially dangerous applications of AI, that assumption may not turn out to be right. Open-weights models are proliferating in America and China and may soon follow in Europe. Other kinds of AI models may eventually be trained narrowly for specific use cases that also present dangers. It is too soon to judge how broadly the net of safety evaluation may need to be cast and what communities of expertise may grow up to serve the needs of public authorities and private insurers.

The AISIs may not want, or have the capacity, to duplicate the costly evaluations the companies conduct. It may be a waste to ask them to do for themselves what they cannot do best. But the AISIs must at least be able to rigorously evaluate the quality of the private work. They must come to their own judgments about that work. They must communicate those judgments in terms that are meaningful to anxious citizens and their representatives.

Thus, governments will need to make extensive investments in research and testing to build an ecosystem that supports and informs third-party evaluation and a workforce to do it.

B. THE NEW: POTENTIAL HOSTILE TOOLS AND WEAPONS; PREVENT STRATEGIC SURPRISE

A new scope of threat evaluation identifying potential hostile tools and weapons was plainly envisioned in the October 2024 presidential national security memorandum. The incoming Trump administration is likely to reaffirm such a mission, at the very least.

Yet this kind of threat evaluation requires a major effort that goes well beyond what governments are doing now. Conducting these evaluations—and building up dangerous capabilities in order to do these evaluations—will require work with the most advanced models that we have made or that others could make. In technical terms this is sometimes called "capability augmentation" or "capability elicitation." In other words, the evaluators will have to make those models explore things that enemies could train or manipulate them to do.

Governments now do not have the technical teams or the infrastructure to do this work. In the United States there are indications that such capabilities are beginning to be built up, especially at the Department of Energy's Oak Ridge National Laboratory and, perhaps, at the Lawrence Livermore National Laboratory.

Governments will have to build partnerships with the private sector to leverage their capabilities while also augmenting their own workforces so that they are not helplessly dependent on private firms. Serious national security risks may arise not that far in the future, so a massive public effort to build at least a base capability for high-level national security evaluation and countermeasures development must start immediately.

That is not all. To be plain: Specific, highly advanced frontier AI capabilities will become part of the defense industrial base of the free world, indispensable in evaluating dangers and developing countermeasures. Governments will have to rely on the private sector to help build up the most advanced capabilities in the world, so that governments can then access, supplement, train, or fine-tune those capabilities to defend against dangers that may go beyond anything the companies conceive for their own private purposes.

That conclusion implies another: There is a compelling public interest in assuring that the private firms on which governments will rely can innovate and scale up in a flourishing, competitive private sector.

1. Beyond Design of Known Company Models

To understand or replicate the most dangerous pathways that adversaries might pursue is a mission that goes beyond just evaluating the safety of the models in the form the companies have trained and designed or how the guardrails on those models might be broken. This scope involves independent work on problems that have little to do with commercially available models and nothing to do with the business plans or use cases envisioned by the companies.

It is not work the company labs are currently incentivized or best positioned to undertake. It is not likely to be profitable. And the companies lack the threat intelligence or other classified information governments may have to perform this task adequately.

2. Beyond the Training or Training Data Available to Companies

Other governments—or even possible nonstate actors—could train AI models that they have designed for their purposes and with their data. This data could take many forms: military, law enforcement, proprietary, criminal, behavioral, or telemetry from many kinds of sensors.

A major part of the multinational security enterprise seems likely to require the most advanced technical work with AI models specifically designed, trained, and tested for their capacity to do terrible harm. This is a fraught, humbling, and morally challenging necessity. It is borne from a practical judgment that development of such models may be the best way to prevent or counter such harm.

Development of such models may also be the best way to learn how to identify indicators that others are doing such work. Those indicators can guide intelligence collection and cue warnings. Forewarned is forearmed.

Both of these scopes of work, targeting the "status quo" and the "new," can become quite sensitive and dangerous. This essay will later address the nature of the government institutions to do this work. For now, we stress that the government will need to regulate the safety and security of the evaluation work, as it has for work on the most extreme biohazards. RAND Corporation researchers have already begun outlining a possible framework for systems at Security Level 4 (SL4) and Security Level 5 (SL5). They detail why SL5 precautions will probably be necessary and that SL5 "does not seem feasible without significant government assistance."[3]

C. THE NATIONAL SECURITY BASE CAMP AND THE PUSH TOWARD THE AGI SUMMIT

Precisely because AI is a general-purpose technology and its development is so uncertain, it is too soon to judge how, and how directly, governments should own, guide, or sponsor the mission of attaining AGI. Nor does this paper assess what role governments should play in guiding, distributing, or regulating all the many possible positive applications of AI or AGI for the benefit of humanity around the world.

What we present here is more modest, though it is still hugely ambitious in relation to where we are now. There is a compelling case to get massive action under way in 2025 to build the defensive "base camp" that can at least evaluate the worst dangers and cue vital countermeasures. To do just that, governments will have to rely on and supplement some necessary base of extremely advanced and costly capabilities in the private sector.

A separate question involves when, whether, and how governments must push themselves or companies to reach the AGI summit. This push may or may not be necessary just for coalition defense, or the defensive enterprise may follow the push made by others, rather than try to lead it. Governments will want to reconsider which institutions will lead such a push and how to manage those benefits and risks.

But in any case, we must get started building the base camp. That base camp will have to work through the next stage of relations between governments and industry at the AI frontier and more seriously consider how to sustain this part of the twenty-first-century defense industrial base.

If governments help sustain the companies essential to their defensive work, they may ask companies to accept certain obligations to protect the public interest and preserve a healthy, competitive environment for innovative applications of AI. These understandings should start being fashioned whether or not the governments also decide to push on toward the AGI summit.

III. INTERNATIONAL ORGANIZATION FOR AI SECURITY

A. BORN AS A COALITION EFFORT: A SMART START

The US-UK MoU and joint operations have created an important precedent. The intermediary should be designed as a coalition enterprise that can operate in several countries. This would be politically wise and functionally necessary, given the nature of the companies, their workforces, and their supply chains.

In practice, we already see the emergence of a coalition effort. Expertise, research results, energy resources, and required capital can be pooled.

This defense partnership and its threat evaluators should not serve as the industry's regulators. Those are national responsibilities. Also, the intermediary must not only protect a company's intellectual property from competing companies; it must have an arm's-length relationship from the various government regulators, which may be responsive to several different national authorities.

The intermediary must be able to provide relevant information to regulatory authorities. But those authorities have their own interests, many stakeholders, and their own processes for making decisions in a democratic government.

B. PROBLEMS WITH OLDER ARMS CONTROL ANALOGIES

Unfortunately, past arms control precedents are not a very good template for this problem, though their history is instructive. Why is this case so different?

- The capabilities are still controlled by companies, not governments. And governments have not yet seriously evaluated how adversaries might threaten them with advanced AI tools.

- Suppliers' groups may be feasible, but the circumstances are exceptional. The hardware supply chains outside of China are dominated by the United States, Japan, Taiwan, South Korea, the Netherlands, and Israel. The software development and other "infrastructure as a service" operations, like data centers, have a different pattern. Governments' relationships to the key companies and the technology (like the issues of open weights) are still unclear.

- It is premature and unwise to lean into atomic energy analogies that formalize divides between "haves" and "have nots." The supposed "have" countries have not yet worked out their own core principles and institutions.

- As dangers surface, the track record of effectiveness for universal, inclusive controls and governance is bad. The nuclear nonproliferation agreements offer both encouraging and discouraging precedents. The relevant precedents and processes of

the Biological Weapons Convention, the World Health Organization, the Chemical Weapons Convention, and the Organisation for the Prohibition of Chemical Weapons are discouraging. Efforts to improve biotech controls after the COVID pandemic have also been discouraging. Attempts to improve oversight of highly risky research have been disappointing on a national basis, even in the United States.

- Past arms control successes were based on some understandings about the strategic requirements of the opposing sides along with common understandings about the practicality of international control or verification mechanisms. In the US-Soviet nuclear case, it took eighteen years to get even to the first agreement (the Limited Test Ban Treaty), and that was only after the high point of confrontation had passed.

C. A SMALL CIRCLE: COALITION FOR AI DEFENSE

Questions about whom to include in a core defense partnership using world-changing technology echo some historical arguments about allied sharing of nuclear weapons know-how. Debates about how to share knowledge of world-changing technology have mixed in worries about how countries could protect themselves as well as intense and well-justified concerns about how widely to share atomic energy secrets.

Between 1942 and 1963, the nuclear sharing debate reached peak intensity at the highest levels among the United States and its allies. As with AI now, the initial phase saw a tight collaboration between the Americans and the British. In the nuclear case, though, the US-British nuclear relationship then became torn before it was eventually patched up. Beyond the British, the most important sharing issues involved France and West Germany. For reasons particular to their eras, the Eisenhower administration moved toward more nuclear sharing, while the Kennedy administration finally retreated from it. Those decisions, along with choices by West Germany and Japan, set in motion the chain of events that led to the Nuclear Non-proliferation Treaty that was opened for signature in 1968.

Considering present circumstances—such as the character of the involved companies, their workforces, and the distribution of key scientists—we think it will be best to stick with the coalition already being built in the US and UK AI Safety Institutes and expand it.

Extending the enterprise to the Five Eyes—the intelligence alliance among the United States, United Kingdom, Canada, Australia, and New Zealand—should not be an automatic reflex. That institution has its own history, originating mainly in the Second World War and arising from the particulars of how to cooperate in the collection of signals intelligence. The scope of the AI coalition defense effort should be considered on its own merits, not reflexively.

What the Five Eyes precedent can offer is a heritage of how to grow trust across borders and set institutional precedents to implement security and counterintelligence cooperation.

The most difficult and important choices in 2025 may be about whether and how to make a major political approach about core participation to other US allies. These would involve challenging decisions and negotiations, in different ways.

To be clear: These are not choices about where to make chips or construct data centers, though those choices are also important. These are choices about participation in a coalition defensive enterprise that may involve highly dangerous work to understand and evaluate frontier AI dangers and to develop countermeasures.

D. A LARGER CIRCLE: RESPONSIBILITIES OF PRODUCERS AND SUPPLIERS

The coalition AI defense enterprise, doing very sensitive national security work, would not be broadly inclusive. It may yield knowledge and insights of very broad value, however, for advancing wider international cooperation.

Precisely because the high-end AI safety work will not be inclusive, parallel institutions should foster global dialogue about what everyone is learning. Those institutions should help disseminate emerging guidance, protocols, and eventually standards for safety and security.

1. A Suppliers' Group

A wider circle of cooperation flows from the existence of choke points in the infrastructure that will produce and sustain high-level AI. That supply chain may make it feasible to create an AI Compute Suppliers' Group.

An interesting analogy from nuclear history is President Eisenhower's Atoms for Peace initiative of 1953. In the original conception, the United States (and the Soviet Union in its own parallel initiative) offered to help with peaceful uses of nuclear technology. In exchange, countries agreed to forgo military uses. Both the assistance and the monitoring were provided by a new international agency, the International Atomic Energy Agency, which was established in 1957. This effort has experienced both notable successes and notable failures.

In one AI-inspired reconception of this idea, the leading AI-producing countries would share the benefits from safe AI systems with other member states. Those states would then agree to accept safeguards, regulating their domestic AI activities to reduce risks to public safety and global security. The agreement might be enforced by export controls of key hardware or efforts to constrain cloud computing and reliance on insecure data centers.[4] In a variation, companies might agree to share the benefits or provide services in exchange for the government support they will need if they are pushing toward AGI.

Both of these ideas are worth considering. We do not take a position on either of them yet, since the relevant governments have not yet gotten to "base camp." They have not yet

worked through the fundamental questions of how they will relate to the companies that still exercise effective control over the technology.

2. A Wider AISI Community

An issue today is whether to include China in efforts to build cooperation among AISIs being created around the world. Some argue that China should not be included because of the sensitivity of some of the safety research.

This issue can perhaps be avoided with clear articulation of a coalition defense approach with its small circle. AISIs that cooperate on defense and intelligence issues could then still participate in wider meetings to discuss common issues of AI safety.

If China does create an AISI, China should be welcome to join a broader suppliers' group. There are AI safety issues of common concern. For example, allies that cooperate on allied defense, including nuclear weapons issues, also participate in the wider Nuclear Suppliers Group that was an outgrowth of the Atoms for Peace initiatives mentioned above.

E. LARGER STILL: SERVING THE WIDER INTERNATIONAL COMMUNITY

In sum, a national security–oriented AI project led by the United States and centered on securing democratic advantage should still encourage technical dialogue, help make the benefits of AI broadly available, and strengthen safety cooperation with countries that are not involved in the most sensitive defensive work.

A broadly inclusive, cooperative approach is already demonstrating some value. Leading experts are organizing to share updated scientific knowledge and foster high-level discussions. In the early nuclear age, at the height of the Cold War, forums of scientists—like the Pugwash Conferences that began in 1957—played a valuable role and later began to feed into high-level political processes.

In this context the most constructive development is the panel that produces the *International Scientific Report on the Safety of Advanced AI*.

Created by the British government and chaired by Yoshua Bengio, the panel currently includes thirty countries and is another product of the Bletchley Park safety summit process. This inclusive process welcomes China. The *International Scientific Report* was inspired by the precedent of the Intergovernmental Panel on Climate Change (IPCC) for sharing scientific understanding and then convening leaders to discuss it.

There is pressure, including from the Chinese, to move this process into the UN system, where the IPCC also lives. The IPCC has had some success within the UN framework. But there were specific reasons for that success, and the process was often slow and cumbersome.

The international panel on AI safety may not fare so well in the UN framework unless an agreement is found to make its scientific synthesis independent of political pressure and the process sufficiently agile when compared with the IPCC process. Recent developments indicate that a compromise could result in an arrangement that allows the panel to operate more swiftly than it might were it entirely within the UN system, by partnering with the OECD and then having its work feed into the UN process. Another good option could be to house the Secretariat for the *International Scientific Report* within the emerging network of AISIs.

Under either of these options, a solution would need to be found to maintain the ongoing involvement of China in the process, given that China is not a member of the OECD and is not yet part of the network of AISIs. Ultimately, policymakers focused on AI and national security in the United States and other leading democracies will benefit from a functional, multilateral scientific evaluation process about the progress of AI.

IV. DESIGN OF A HISTORIC PUBLIC-PRIVATE PARTNERSHIP

We see no prospect in the foreseeable future that public institutions in the free world will be able to match the scale of capital, workforce, and institutional knowledge in private companies. The vital public purposes will have to be addressed by a novel public-private partnership on a historic scale.

Governments should try to build their internal capacities through targeted investments in institutions like the American national laboratories. They can also invest in university research in several related fields. But the scope of work we have described above (see Scope of Work) will have to be undertaken principally by private companies working with governments.

A. BASIC CHOICES: CIVILIAN? COALITION?

We explained earlier that the British and American governments initially gave the institutional lead to their AISIs and that these institutes were set up as essentially civilian enterprises in civilian departments—the Department of Commerce in the American case and, within Commerce, connected to the National Institute of Standards and Technology.

We have stressed the emergence of a very demanding defensive security mission to evaluate the worst possible threats and guide work on countermeasures. So far, as this mission has become visible to leading officials in Washington and London, they prefer to keep the civilian AISIs in the lead. They are working on ways to connect these institutes to the departments and agencies that have essential intelligence and defense information and responsibilities and to ensure the institutes get the support they need.

We are reluctant to second-guess these choices. Outsiders do need to be aware of the strains and the challenges to effective interagency management. Outsiders also need to

anticipate the strains as national security pressures encounter the coalition character of the defensive enterprise.

We explained earlier that it has been vital, as a practical matter, to conceive of the defensive enterprise as a coalition effort. Understandable habits of national security work and some of their related families of laws and regulations, like the ITAR family (International Traffic in Arms Regulations), are bound to obstruct multinational coalition work.

The current US-British AISI partnership reflects such an understanding. Thus, the institutional choice to keep the AISIs in the lead is a fundamental design choice that may keep the civilian orientation at the forefront. Some of the work with some of the institutes may include a very sensitive defensive mission as well, with complex relations to the agencies that have operational responsibilities for evaluating and responding to threats. But the "civilian-led" design choice may reflect the breadth of concerns the AISIs must consider and help sustain political support in a coalition.

B. WHAT THE PARTNERING COMPANIES NEED

Right now, it appears that the highest-level work at the AI frontier can only be done by a handful of companies. The high-end capabilities to create AI may involve a natural oligopoly because of the computing power, infrastructure, energy, and safety efforts they require.

There are several past precedents for dealing with such problems in vital industries, as in the history of telecommunications, electric power, railroads, and aerospace. The national security requirements of the emerging aerospace industry did much to create the foundation of today's Silicon Valley and the growth of Southern California. The American pattern is to design partnerships that serve both private and public interests. Sometimes these have been done well. There is not a single template for how to do this.

But, at least in the American context, the partnerships should not be overly reliant on instruments like nationalization or blunt coercion. Policymakers and leaders of the companies working at the AI frontier should instead try to create a situation in which their private partners can flourish if they can perform.

Even companies that wish to help governments conduct this defensive threat assessment may find this a strain, adding to others they are already likely to face, such as a need for greater capital investment to pay for data centers and electricity and the risks of copyright infringement liability.

- Top talent is scarce. Companies can outbid one another for it, for their preferred purposes.

- Top talent is multinational. Enough workers will have to believe their mission makes many countries safer and does not merely advance the interest of one of them.

- Top-level compute and data center use are scarce and costly resources. There may be trade-offs between company missions and security work in terms of allocating time and capacity to do training runs for either. Also, governments will have to reckon with the costs of training runs for the security work.

The participating companies may choose to create specialized subsidiaries to do some of the national security work, though this would be a costly solution for them, and it might fragment and complicate the management of their organizations.

C. SOME OF THE ISSUES ON THE TABLE

1. Some Relevant Public Authorities

In the United States, the baseline statutory authority is in the Defense Production Act (DPA). In that act, 50 U.S.C. § 4533(a)(1)(D)(ii), provides: "To create, maintain, protect, expand, or restore domestic industrial base capabilities essential for the national defense, the President may make provision—(D) for the increased use of emerging technologies in security program applications and the rapid transition of emerging technologies—(ii) from commercial research and development to national defense applications."

This is sufficient authority to establish a public-private partnership to perform the scope of work described above. It is an authority under Title III of the Act that gives the president coercive powers (if he or she chooses to use it) to make firms participate.

2. Company Duties and Possible Limits on Vertical or Horizontal Integration

Governments would have to consider what kind of services the participating companies are obliged to offer (or withhold) to the many companies and countries wishing to apply advanced AI. More than a hundred years ago, the law already recognized that people were engaged in "common callings" or formed "public service companies." When the United States started building railroads, governments started writing laws discussing the duties of "common carriers." Those analogies are only suggestive of reciprocal duties. Governments and companies might consider the right formula for this new era of transformative technology.

If certain companies play a special role in the emerging industry and receive special government support to help with the defense of society, those companies should not be able to leverage this support for unfair competitive advantage. In the past, at least in the United States, the government therefore managed the antitrust objectives with limits on vertical or horizontal integration.

3. Money—Direct and Indirect

Obviously, governments can contract for services. That is a negotiation about "direct" money. Part of that negotiation, however, will be about longer-term contracting, less

vulnerable to the ups and downs of annual appropriations, in order to justify work that requires a lot of capital investment.

The frontier companies currently generate enormous requirements for private capital investment. To secure democratic advantage, either the public or private investment flow has to be sufficient compared to what adversaries can do. If most of the money is coming from private capital, then government revenue probably won't be enough to sustain that flow.

Therefore, the governments have an indirect interest in the profitability and investment flows of the private companies on whom they choose to rely. To the extent governments become providers of capital, they can also gain a further voice in how the frontier companies navigate the risks and opportunities of their new technologies.

4. Access to Required Hardware

To perform its part in the multinational security missions proposed in this paper, the US government has considerable legal authority to commandeer what it needs from private companies. The president could, as needed, use his coercive prioritization power under Title I of the DPA to be sure the partnership has access to available hardware and resources, as laid out in 50 U.S.C. § 4511(a).

In the United States, this power allows the government to make claims, even exclusive claims, on the resources (such as energy), hardware (such as advanced chips and connectors), and software products needed to do its defense work. Such claims could include commandeering production from chip companies like Nvidia.

The United States may also be able to exercise jurisdiction within its borders over foreign-produced items that incorporate US-origin design or technology.

The DPA also authorizes the president to "prescribe such regulations and issue such orders as the President may determine to be appropriate to carry out this Act" (50 U.S.C. § 4554[a]). The October 2024 presidential national security memorandum on AI is one example—and thus carries the force of law unless rescinded by the next president. These regulations would naturally include the safety and security rules that would accompany the government's AI safety research. The president also has broad authority to obtain information about what any company is doing in this space (see 50 U.S.C. § 4555).

Such authorities might be employed if the US government, for its defense purposes, needed to pool the compute resources being deployed by more than one company or had to redirect production of chips or network hardware more quickly, or in a more concentrated way, than the market was likely to do on its own.[5]

5. Land Use Permitting

Construction of data centers and their related electricity supplies obviously involves serious permitting requirements at the federal, state, tribal, county, or local levels. These requirements, above all the analyses of environmental impact required under the National Environmental Policy Act, have long hindered the construction of major infrastructure projects throughout the United States, adding years of delays and much higher costs. For at least ten years, laws and executive orders have tried to manage these concerns. Now the pressing need to build AI infrastructure has come knocking at the door.

In September 2024 the Biden administration launched a new White House Task Force on AI Datacenter Infrastructure. It was supposed to work with a unique agency, the Federal Permitting Improvement Steering Council (which was created to manage the bewildering array of permitting issues among its nine participating agencies), and the US Army Corps of Engineers. The new task force promised to help identify possible government financing, provide agency points of contact, and prioritize AI datacenter development. Throughout, the goals were to build in the United States, with American workers, and using "clean" energy.

That was a start. Congress is already considering bills to address permitting issues, with support from members of both parties, and a legislative remedy will probably be necessary. OpenAI supports the creation of "AI Economic Zones" to give states incentives to speed up permitting and approvals for AI infrastructure.[6] That proposal appears to understate the difficulties.

We are adding two other factors to the mix. First, we are stressing a particular defensive mission that may be quite urgent. Rather than argue, a bit quixotically, for a general ease of permitting for any private-sector AI building, it may be more feasible to prioritize a small number of projects that can be completed quickly and are related to the allied defense. Most governments have some exceptional capacities to clear a path for urgent defense work.

Second, we are stressing the coalition character of the work. The United States does not have to, and may not be able to, solve this problem alone. Companies are already looking overseas. OpenAI discusses "a North American Compact for AI" and nods toward cooperation with countries like the United Arab Emirates.[7] Since the data centers and networks may involve very sensitive work, only a few governments may be able to undertake the precautions necessary to safeguard the data and the technology.

6. Access to Required Electric Power

The growing gigawatt-level electricity requirements for frontier AI are well-known. The most recent International Energy Agency (IEA) data show that these requirements still make up

only a very small portion of global electricity demand, a good deal less than, for instance, the requirements of primary aluminum production.[8] But the data networks and data center demands are highly concentrated in specific regions, which can create local bottlenecks.[9]

In the United States, for example, data centers have created enormous demand growth (up by more than 10,000 GWh) over the last five years in Virginia and in Texas, with rapid rises also occurring in Arizona and North Dakota. Globally, the data demands are especially significant in countries including Ireland and Singapore.

Governments are already restricting access to the electricity grid in several areas, including in the United States. There are many new proposals to regulate data center energy use. Meanwhile, unusable electricity surpluses (and negative prices) are occurring in regions like Southern California and South Australia that have expanded solar and wind generation.

Government policy is a critical variable in meeting possible energy needs of frontier AI. National, state, and local governments have to decide whether or how to modernize their grids. They have to develop ideas, still embryonic, for how they can create new gigawatt-level net baseloads of electric power in specific geographic regions. The US Department of Energy is creating an AI datacenter engagement team to convene the stakeholders. It is also helping companies look at repurposing closed coal sites for new energy infrastructure.

OpenAI is proposing a National Transmission Highway Act to address not only electricity demand but also the higher demands for wireless connectivity and spectrum allocation.[10] The idea could try to link demands for "highways" that enable transmission of electricity, data, and natural gas to combine and smooth permitting and financing for all of them. Even if the Trump administration decides to waive all the regulations it can, the usual problems extend to state, local, and tribal authorities, as well as to courts. If a legislative fix is sought, it might be more likely to pass if it is targeted relatively narrowly—again the highest priority can fairly be placed on the rights of way to facilitate urgent defense work.

7. Workforce Issues

The leading companies rely on multinational talent. They will therefore need to engage with governments to address their immigration law issues and the scope of government security restrictions.

8. Access to Required Training Data

Governments collect and own data of their own, of many kinds, much of which is available only for government-approved uses. Governments also set rules on data privacy and determine exceptions to their copyright protections.

9. Security and Safety Standards

We have already mentioned the possibility of using government standards to create a "safe harbor" to help companies manage their liability risks. Tort law is almost entirely a long-standing domain of state activity and may be better addressed through guidelines that can create a de facto safe harbor without implying a restriction on a traditional state law domain. Copyright, on the other hand, is a federal issue, and companies at the AI frontier may be keen to explore how "fair use" issues can be addressed more systematically instead of being resolved one case at a time.

Governments also have expertise and capabilities in how to protect data, insulate facilities from outside penetration, and vet people. On their good days, governments also know how to conduct complex counterintelligence investigations, sometimes on a multinational scale.

Companies have their own strengths and weaknesses in protecting their intellectual property. Both sides, public and private, must manage trade-offs in data sharing and competitiveness.

10. Defense of Data Centers and Networks

Data centers and networks are easy for adversarial states or even nonstate actors to identify and target. If they become key components of power, the incentives to target them will grow. Companies that invest in such centers may have a greater need to forge partnerships with governments in order to understand the threats against them and will look to governments to actively defend them.

V. THE OPEN-WEIGHTS AND OPEN-SOURCE ENVIRONMENT

The recent National Telecommunications and Information Administration (NTIA) report from the US Commerce Department weighed the risks and benefits of what it calls "dual-use foundation models with widely available model weights." These are models trained on broad data that contain at least tens of billions of parameters, usable on many kinds of problems. In the report's main conclusion, "The potential future outcomes are so uncertain that effective, definitive, long-term AI strategy-setting is difficult."[11]

Therefore, the agency argued, governments should not yet come down hard to favor either restriction or openness. They should instead "continuously evaluate the dual-use foundation model ecosystem and build & maintain the capacity to effectively respond."[12]

A. THE NECESSITY OF INDEPENDENT RISK EVALUATION BEFORE RELEASE

While sifting the evidence in the evolving debate about restriction and openness, we still hold to our premise that governments must be able to conduct *independent* evaluation of

model risks. In the case of release of model weights, it is even more vital that the independent evaluation occur *before* the model is released. Without independent evaluation, there should be a presumption not to release "frontier" model weights for the following reasons:

- Competing militaries are using open models released by the United States. For example, the Chinese People's Liberation Army is currently using Llama 3.

- Open models are difficult to monitor for misuse. Some of the largest companies that operate platforms have built workforces and capabilities to track and evaluate threats. They publish frequent reports about how the threats evolve and who seems to be behind them, as with Google's Threat Analysis Group.[13] One can't learn as much about the threats without directly watching them.

- Open-model guardrails can easily be removed after release.[14] Companies could invest in technical innovations that can reduce the ease with which open-weights models can be retrained to bypass their guardrails. Digital watermarking techniques to track misuse have not progressed very far.

- Open models can be fine-tuned on malicious data. For example, an open model can be fine-tuned on virology publications so that it is a more effective digital tutor for creating biological weapons.

- Open models are irreversible. Once released, an open model is out there forever. If a risk is discovered later, there's no way to reel it back in. Because open-model releases are irreversible, prerelease testing is even more important than it is for closed models. Prerelease testing could be more realistic than it is today. For example, tests should include having guardrails removed and having models fine-tuned on data that adversaries might use.

- Open models can introduce capability overhang that is exploited later. As OpenAI's o1 shows, users can significantly increase a system's capabilities with access to the weights and post-training improvements well beyond what might have been evident in evaluations at the time of deployment.

As they learn more, better-informed governments may choose to constrain who can create more advanced AI, seek oversight over the training of new models, or regulate the dissemination of model weights. It is hard to argue that governments should never be allowed to make those choices or that governments should not become informed enough to make those choices intelligently.

B. DIRECT AND INDIRECT SECURITY RISKS IN THE OPEN-WEIGHTS ENVIRONMENT

Direct threats are the threats of misuse, which the NTIA has classified as threats to public safety. These are threats to facilitate production of chemical, biological, radiological, and nuclear weapons or develop offensive cyber operations. These threats are not theoretical. The dark web already features jailbroke versions of the earlier Llama 2 "to facilitate malicious code generation."[15]

The more indirect threats, which the NTIA report calls "geopolitical," have received less attention. The availability of open-weights models may help adversaries develop advanced systems in much less time and at much less cost than would otherwise be feasible. The US government is already trying to restrain national security risks from AI competitors through export controls. The open-weights environment may help competitors circumvent many effects of these restrictions. Giving tyrannies a competitive boost may not gravely concern the companies developing and selling open-weights products, since their business models rely on sales and licensure mainly in free societies.

On the other hand, in theory, adversaries could steal the weights anyway from closed-model companies if they cannot be adequately secured. More work is needed to assess this threat. But there is a good argument that efforts to secure the distribution of open-weights models must be accompanied by efforts to secure the model weight data at other frontier companies as well.

C. ANOTHER INDIRECT EFFECT, ON GOVERNMENT ACQUISITION AT THE FRONTIER

Companies are already developing and testing products on a scale vastly ahead of what governments have built or can currently afford. If governments need to rely on those hugely expensive capabilities, then governments have an incentive to sustain very large private capital investments. Otherwise, public investments will have to fill the gaps, if they can be filled at all.

The private sector is currently battling to see if closed- or open-model approaches will attract the capital and revenue they need to achieve market dominance. The investment flow to the losers may dry up. Governments may therefore soon need to judge what kind of ecosystem will best satisfy the public needs they must serve by sustaining state-of-the-art threat evaluation and countermeasures.

D. PRACTICALITY OF RESTRICTIONS

The NTIA report acknowledges that if there is evidence of significant risk, "there are many different ways to implement a structured access program that restricts access to model

weights" with government guidelines. New legislation may be needed. At least some international alignment on a common approach may be vital. International alignment may be easier if application programming interface (API) access to model use is sufficiently available in their countries.

But even without new laws, and even if model weights are stolen, authorities can spotlight and reinforce existing civil liability for making or selling dangerous products, as we discussed above.

These measures can mitigate proliferation, not prevent it. Both governments and transnational criminal networks will probably be able to obtain some of the capabilities they think they need.

E. POLICY PLANNING FOR AN UNRESTRICTED ENVIRONMENT

If it proves impossible to limit the proliferation of AI-powered abilities to do terrible harm, governments will have an even greater need to evaluate all the possible threats at the frontier and acquire all the AI capacities needed to develop countermeasures. The necessity for a public-private partnership becomes even greater.

ACKNOWLEDGMENTS

This report reflects many group conversations. We shared drafts of this paper with officials in multiple governments and with representatives from leading companies. We sifted their suggestions and are grateful for them.

We are especially grateful to Anja Manuel and the leaders of the Aspen Strategy Group, a program of the Aspen Institute, where some of these ideas were aired at a conference they held in July 2024. Both the Carnegie Endowment for International Peace and the RAND Corporation have sponsored other meetings that gathered some gifted thinkers on the AI security issues.

Yoshua Bengio, who chairs the *International Scientific Report on the Safety of Advanced AI*, was kind enough to give us multiple readings of this draft. We also greatly benefited from our attendance at the first AI Safety Summit, held in Bletchley Park in November 2023, with the sponsorship of the United Kingdom.

NOTES

1. Xingwu Sun, Yanfeng Chen, Yiqing Huang, et al., "Hunyuan-Large: An Open-Source MoE Model with 52 Billion Activated Parameters by Tencent," arXiv 2411:02265, version 3, November 6, 2024. Jack Clark helped call our attention to this work.

2. Nathan Rosenberg, "Uncertainty and Technological Change," in *The Mosaic of Economic Growth*, ed. Ralph Landau, Timothy Taylor, and Gavin Wright (Stanford University Press, 1996).

3. Brodi Kotila, et al., "Designing US Government-Lab Joint Projects to Build AI," draft paper, November 2024, 6.

4. See, e.g., Cullen O'Keefe, "Chips for Peace: How the US and Its Allies Can Lead on Safe and Beneficial AI," *Lawfare*, July 10, 2024; this was also discussed in our conversations with Colin Kahl and Lennart Heim. The "variation" has also been suggested by Kahl.

5. See, for example, the discussion in Kotila, et al., "Designing US Government-Lab Joint Projects," 11–14.

6. OpenAI, "OpenAI's Infrastructure Blueprint for the US," November 13, 2024.

7. OpenAI, "Infrastructure Blueprint."

8. International Energy Agency, "Electricity Mid-year Update," July 2024, https://iea.blob.core.windows.net/assets/234d0d22-6f5b-4dc4-9f08-2485f0c5ec24/ElectricityMid-YearUpdate_July2024.pdf.

9. For a set of useful but more incremental suggestions, see the report of the Secretary of Energy Advisory Board (headed by Arun Majumdar), "Recommendations on Powering Artificial Intelligence and Data Center Infrastructure," July 30, 2024, https://www.energy.gov/sites/default/files/2024-08/Powering%20AI%20and%20Data%20Center%20Infrastructure%20Recommendations%20July%202024.pdf.

10. This idea appears to build on an existing process that allows the US Department of Energy to designate "National Interest Electric Transmission Corridors." These designations unlock sources of federal financing support and permitting powers of the Federal Energy Regulatory Commission.

11. National Telecommunications and Information Administration (NTIA), "Dual-Use Foundation Models with Widely Available Model Weights," US Department of Commerce, July 2024, 33, https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf.

12. NTIA, "Dual-Use Foundation Models," 36.

13. Google, "Threat Analysis Group (TAG)," https://blog.google/threat-analysis-group/.

14. See, e.g., Dmitrii Volkov, "Badllama 3: Removing Safety Finetuning from Llama 3 in Minutes," arXiv 2401:01376, version 1, July 1, 2024. Hugging Face has also posted pre-trained versions of the Llama model.

15. Zilong Lin, Jian Cui, Xiaojing Liao, and XiaoFeng Wang, "*Malla*: Demystifying Real-World Large Language Model Integrated Malicious Services," arXiv 2401:03315, version 3, August 19, 2024.

30 29 28 27 26 25 24 7 6 5 4 3 2 1

Author photo credit: Eric Schmidt (Ben Gibbs)

ABOUT THE AUTHORS

PHILIP ZELIKOW

Philip Zelikow is the Botha-Chan Senior Fellow at the Hoover Institution. He held a chaired professorship in history at the University of Virginia for twenty-five years and was an associate professor at Harvard University. An attorney and former career diplomat, Zelikow worked across government in five presidential administrations and directed three successful bipartisan national commissions.

MARIANO-FLORENTINO CUÉLLAR

Mariano-Florentino (Tino) Cuéllar is the tenth president of the Carnegie Endowment for International Peace and serves on the boards of the William and Flora Hewlett Foundation and Inflection AI.

ERIC SCHMIDT

Eric Schmidt is an accomplished technologist, entrepreneur, and philanthropist. He served as Google's chief executive officer and chairman from 2001 to 2011, as well as executive chairman and technical advisor. In 2021 Schmidt founded the Special Competitive Studies Project, a nonprofit initiative focused on strengthening America's long-term AI and technological competitiveness in national security, the economy, and society.

JASON MATHENY

Jason Matheny is president and CEO of the RAND Corporation. He previously led technology and national security policy at the National Security Council and the Office of Science and Technology Policy. He was founding director of the Center for Security and Emerging Technology at Georgetown University and director of the Intelligence Advanced Research Projects Activity.

Synopsis

The United States must now start working very hard with allies to secure democratic advantage in the domain of frontier AI. We suggest how to manage the convergence of three great vectors: private sector–led innovation, emerging threats, and international efforts. An essential starting point is to build a defensive agenda, build a historic public-private partnership, and design overlapping circles of international cooperation. The time to start shaping the national security agenda for AI has arrived.

Hoover Institution, Stanford University
434 Galvez Mall
Stanford, CA 94305-6003
650-723-1754

Hoover Institution in Washington
1399 New York Avenue NW, Suite 500
Washington, DC 20005
202-760-3200

HOOVER INSTITUTION