CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

JULY 2024

Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance

Jon Bateman, Dan Baer, Stephanie A. Bell, Glenn O. Brown, Mariano-Florentino (Tino) Cuéllar, Deep Ganguli, Peter Henderson, Brodi Kotila, Larry Lessig, Nicklas Berild Lundblad, Janet Napolitano, Deborah Raji, Elizabeth Seger, Matt Sheehan, Aviya Skowron, Irene Solaiman, Helen Toner, and Polina Zvyagina

Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance

Jon Bateman, Dan Baer, Stephanie A. Bell, Glenn O. Brown,
Mariano-Florentino (Tino) Cuéllar, Deep Ganguli, Peter Henderson, Brodi Kotila,
Larry Lessig, Nicklas Berild Lundblad, Janet Napolitano, Deborah Raji, Elizabeth Seger,
Matt Sheehan, Aviya Skowron, Irene Solaiman, Helen Toner, and Polina Zvyagina

Carnegie does not take institutional positions on public policy issues; the views represented herein are those of the author(s) and do not necessarily reflect the views of Carnegie, its staff, or its trustees.

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

Contents

Introduction

As policymakers around the world grapple with the rise of artificial intelligence (AI), much of their attention has focused on highly capable foundation models—those with advanced capabilities across a wide range of tasks, to include the generation of words, images, sounds, and video. Companies, governments, and civil society organizations are urgently debating how to govern such models, as well as the models' components, supply chains, and the deployed AI systems they ultimately power.[1]

For much of the last eighteen months, debate about so-called open models[2] has been especially vigorous. While this term has been used in various ways, models are often described as open when their key components are publicly released for download.[3] Among these components, the release of model weights has received outsized attention. Model weights are the statistical parameters that drive a model's core behavior, so their public release can be an important factor in the ongoing advancement and broad dispersion of AI capabilities.

Open foundation models and weight release have been celebrated as a promising pathway to hasten innovation, reduce market concentration, increase transparency, and combat inequality. At the same time, there have been warnings that open models can empower bad actors, make it harder to detect or thwart misuse, and increase the risk that humans eventually lose control of AI.[4] These parallel benefits and risks have previously led to heated debates about what kinds of foundation models should be openly released and who should decide. Until fairly recently, debate would sometimes devolve into ideological conflict between two deeply entrenched camps.

Thankfully, there have been signs in recent months of an emergent shift toward depolarization and an increase in fresh thinking by all "sides" of the debate on open models. This reset is welcome and should be more widely recognized and analyzed, especially by policymakers. The world needs more productive and actionable discussions of how to govern highly capable foundation models—both open and closed.

Within industry, a growing number of major AI labs have embraced mixed release strategies for foundation models—releasing some as open and others as closed, depending on the properties of each. In the expert community, several papers and workshops have sought to complicate the picture of an open/closed binary; highlight decisions beyond just weight release; and broaden the focus from individual models to larger ecosystems (including social institutions).[5] Government has also helped to stimulate new discourse: the U.S. National Institute of Standards and Technology's call for public comments on model weight governance has drawn a wide range of commentary, much of it nuanced.[6]

But as encouraging as these developments are, more work remains to clarify, consolidate, and build upon positive trends. One important task is to identify and document the areas of emerging consensus with some precision—so that these ideas can be further refined, acted upon, and used as a springboard for tackling harder issues. Another task is to frame the key open questions that need further research and debate—so that policymakers can be aware of current gaps, and researchers and advocates can focus their attention on most urgent or promising areas of the next governance frontier.

To that end, the Carnegie Endowment for International Peace hosted a convening in late April at the Rockefeller Foundation's Bellagio Center in Italy. It brought together a diverse set of experts—from leading AI labs, universities, and civil society organizations—who represent a wide range of perspectives on open models and foundation model governance. Several days of intensive, structured discussions resulted in this document, co-signed by attendees.

Our two major conclusions:

- It is no longer accurate or productive to cast decisions about model and weight release as an ideological debate between rigid "pro-open" and "anti-open" camps. Rather, different camps have begun to converge on the shared recognition of open model release as a positive and enduring feature of the AI ecosystem, even as it also brings potential risks and limits. Part 1 of this paper aims to capture this emerging consensus in seven points.

- At the same time, many key governance debates remain unresolved, and new challenges are rapidly emerging. Part 2 of this paper suggests an agenda for further research and discussion in the form of seventeen open questions.

Areas of Emerging Consensus

We suggest the following seven areas as points of an emerging consensus among diverse perspectives on the governance of highly capable and open foundation models. To be sure, this paper can only speak for its authors. There may be elements of these principles that still merit refinement or revision. Even so, we hope to highlight the existence of what seems to be substantial new common ground in a historically fractious debate. Such common ground, if durable, can provide a solid foundation for addressing more difficult open questions (described in Part 2).

To clarify, these points are not commitments for action by the authors or their organizations. Nor do they claim to comprehensively cover everything important and relevant. Rather, these are intended as general reference points to help frame governance discussions.

1. It is no longer accurate or productive to cast decisions about model weight release as an ideological debate between rigid "pro-open" and "anti-open" camps. Open and closed foundation models both have legitimate, positive, and important roles to play. They will inevitably co-exist in a hybrid ecosystem, as diverse forms of AI models and systems interact with each other, with non-AI technologies, and with human institutions.

However, a select subset of foundation models—in particular, some future models—could pose risks that warrant more restrictive modes of release.

For the vast majority of foundation models released to date, broad public access—including but not limited to the public release of weights—has not produced known harms that outstrip their apparent benefits. The spread of advanced AI capabilities can yield countless upsides, including the enablement of scientific and medical advances and the further democratization of knowledge and technological power.

On the other hand, many foundation models are general-purpose or dual-use, meaning they have both beneficial and harmful applications (like other technologies). As a result, models with certain advanced or particularly harmful capabilities pose risks that could outpace available safeguards and benefits. This category may include a small number of current models as well as an unknown number that might be developed in the future. In such cases, "precautionary friction" is prudent. This can include staged release, where a model is initially held closely but gradually released more and more openly, and structured access, where external parties have a certain degree of access that is designed to facilitate specific goals.[7]

The major policy debate is not about whether foundation models in general should or shouldn't be openly released, but rather, how to draw practical lines in specific cases. This more practical debate should be premised on the notion that a wide range of "open," “closed,” and hybrid strategies are acceptable for today’s technology. Choice of strategy can be guided by model developers’ individual priorities, within broad parameters that account for the substantive and procedural interests of society at large. Over time, the advisability of and approaches to release may evolve a great deal based on factors such as business models, accumulating real-world data on model impacts, societal adaptation, evolving cost structures, and technology trends—all of which are nascent.

2. “Openness” is a multifaceted spectrum encompassing various options, values, and goals. Weight release can be one important element of openness, but it will not always be necessary or sufficient to achieve all the different benefits of openness in its various interpretations.

Openness is an idea with a long and celebrated pedigree in the history of technology. However, its precise meaning varies. It can convey a range of values, such as transparency, access, freedom, inclusion, and reciprocity.[8] These can be pursued for the sake of various practical goals—such as promoting innovation, increasing competition, reducing inequality, bolstering security and safety, spreading technical knowledge, and enhancing the agency of everyday people in their individual lives and in societal decisionmaking. Such values and goals often correlate but sometimes conflict with each other or themselves. For example, releasing model weights can have the potential to improve safety in some ways (like empowering independent researchers to identify and help fix design flaws) or erode it in other ways (like enabling bad actors to strip out safeguards via fine-tuning).[9]

Model weights are just one of several key model components or artifacts, including architecture, code, and training data, that may be externally released. For each artifact, a range of options—such as staged release, structured access, and varying amounts of documentation—is available to shape who receives what information, when, and how.[10] The process for deciding among these options can also be more or less open, in the sense of being inclusive, transparent, and accountable. The simple downloadability of model artifacts, moreover, does not guarantee that all actors can truly benefit from them. Rather, this may depend on the actors’ practical access to enabling resources such as cloud infrastructure, technical training, and language expertise.

Both open- and closed-weight models can contribute, or not contribute, to different kinds of openness depending on the circumstances. For example, an open-weight model may have highly permissive licensing terms but limited non-English capability and poor documentation. Conversely, a closed-weight model could be developed by a transparently governed NGO that provides extensive free support to users in low-income countries. It is important to clarify the goals of openness and to link weight-release decisions with a larger strategy for achieving them—which could then be assessed, debated, and adjusted over time.

3. Model weight release can have special impacts. In principle, it tends to exaggerate both the positive and negative potential of a model (though not always symmetrically).

But such theoretical tendencies don't necessarily translate to specific real-world cases. Some open models have benefits or risks more commonly associated with closed models, and vice versa. Scrutiny of weight release decisions should not overshadow the many other design and implementation factors that can be equally or even more significant in shaping a model's effects.

Model-weight release gives anyone with sufficient skills and resources a particularly wide-ranging and permanent ability to apply, alter, adapt, and learn from a foundation model. *All else being equal*, the open release of a foundation model's weights will tend to further empower its users and third-party developers. This has countless social benefits, such as accelerating scientific and commercial innovation, democratizing access to information and power, and improving public knowledge of AI systems.

However, model-weight release can potentially also have harmful consequences, including the empowerment of bad actors—such as criminals and adversarial states—and the possibility of AI systems acting autonomously in ways unintended by their creators or users. The harmful effects of open-weight models can be irrevocable and extremely difficult if not impossible to monitor or restrict with current technology. Because of its potentially heightened benefits and risks, open-weight release continues to merit special attention from all stakeholders (even as the broader spectrum of openness must also be better mapped and explored).

Yet it's important to recognize that real-world consequences of open- (or closed-) weight release can differ markedly from these theoretical tendencies. For example, highly customizable closed-weight models that lack adequate usage monitoring or enforcement of terms would not achieve the safety and governance potential of closed-weight release. Likewise, open-weight models with high inference costs would not fulfill their potential for broadening access. These architectural, design, and implementation factors deserve much more scrutiny in governance conversations—not just for open-weight models, but for closed-weight models as well.[11]

4. Model release decisions should depend on the assessment of marginal risks and marginal benefits. More capable and potentially impactful models should have broader and more rigorous pre-release evaluations, as well as post-release monitoring and enforcement. This requires dedicated resources, creativity, and a consideration of the model within its larger human and technological environment.

The notion of marginal risk helps to focus attention on how the release of a new model or system—including open release—compares to a baseline. For example, the preexisting risk associated with other widely available information and tools (like search engine results) can serve as one baseline. But there are other possible approaches to defining the relevant baseline. How to choose a meaningful baseline, and how to estimate the baseline and marginal risks, are important areas of research and discussion. A similar assessment of marginal benefits can also help to inform model release decisions.

The scope and rigor of pre-release evaluations and post-release monitoring should be proportional to the model's apparent capability and potential impact. Because model evaluation remains in its infancy, its sophistication should keep pace with—or ideally, grow faster than—rising model capabilities and impacts.

Rather than seeing a model in isolation, evaluations should account for the broader environment—including other AI models, other technologies, and, above all, the human element. Analysis at the ecosystem level is extremely difficult due to its complexity, scale, and resistance to measurement. Yet it is necessary. An important task is to find ways of addressing this challenge, including through large and long-term investments in the build-up of research infrastructures, and by crafting governance structures that are explicitly designed to operate with imperfect information.

5. At this time, we lack compelling evidence that any major AI company has released an open foundation model which, in retrospect, it clearly shouldn't have.

However, this judgment is tentative at best because it's still unclear how to assess the overall impact of open and closed foundation models. Small models, some specifically designed for harm, also need closer scrutiny. The evaluation ecosystem must quickly mature in capability, capacity, standardization, and accessibility to address these gaps.

The impacts of any model release are always uncertain. This can be magnified with open-weight release because of its irreversibility and monitoring challenges. So far, though, we lack evidence that any prominent or popular foundation model, including those developed by the leading AI companies, should not have been released. But this should not provide any false assurances about the efficacy or robustness of pre-evaluation and post-release monitoring, both of which remain embryonic at best. Better assessments are also needed for other aspects of a model's impact, such as energy usage and worker treatment, that arise from the model's initial development rather than its subsequent release. And small models, especially those intended to be harmful, are a known but neglected problem.

Future foundation models that push the so-called frontier of capabilities and/or bring fresh kinds of risks will need more advanced evaluations and other risk assessment methods than have been developed to date. The frontier is a fuzzy and contentious line, to be sure. Fundamentally, developers who intend to create or market new generations of more powerful and potentially impactful models—particularly those with new modalities—should make parallel and proportional strides in the evaluations of such models. Advances in evaluation practice must include both technical and nontechnical dimensions, such as the sophistication and realism of threat models. "Frontier" models should, in many cases, undergo some form of "precautionary friction" to shore up understanding of marginal risks and benefits prior to any open-weight release.

Much more investment is needed in evaluation tools, infrastructure, institutions, standards, best practices, and policymaking. This includes incident monitoring, harms discovery, and documentation practices.[12] It is important not only to develop these resources but also to improve access to them—for example, by actively assisting and funding civil society organizations, academic researchers, small and third-party developers, and stakeholders in low- and middle-income countries.

6. Because release decisions are based on imperfect information, they are shaped by the risk tolerance and appetite for uncertainty of decisionmakers. These are value judgments. Regardless of one's values, responsible judgments must be well-reasoned, consistently applied, informed by a range of considerations (beyond just profit or self-interest), and transparently explained.

In practice, most release decisions for highly capable models have so far been made by a relatively small number of for-profit companies, predominantly but not exclusively based in Western countries. They have generally based these decisions on an internal, largely private weighing of factors such as commercial incentives and emerging norms. Direct regulation in areas such as privacy and consumer protection has had some impact on release decisions. The shadow of the law—such as ongoing and potential tort and copyright litigation, draft bills, and the prospect of future regulation—has also mattered.

While each of these early dynamics has begun to shift in different ways, the fundamental issue is what governance should look like going forward. Who should be involved in model-release decisions and how? What information and criteria should they use? What authority and enforcement power should they wield? What is the proper divide between democratically adopted regulations and private entrepreneurial freedoms?[13] Such questions will be debated for some time and pit competing interests and values against each other. Still, a few core answers are already clear. At a minimum, release decisions should adhere to basic standards of procedural soundness, such as non-arbitrariness, consistency, transparency, and the conscious weighing of a breadth of factors (including societal interests and views) in model release decisions.

7. “Open models” should not be conflated with “open-source software.” The two ideas are distinct, though they also have significant connections that merit further exploration.

“Open-source software” has strict definitions that were developed by a community of practitioners over many years. In general, open-source software must have a license that permits nearly any distribution, modification, and use of the software, code, and derived works, for both for-profit and not-for-profit uses.[14] Although open foundation models often have fairly permissive licenses, many of them do not meet this definition.[15] Additionally, a key benefit of open-source software is that code transparency enables the full analysis of a program’s behavior. AI models cannot yet be analyzed as fully and readily as traditional code, even when the weights and other artifacts are publicly released and extensively documented. Another difference is that open model release, unlike traditional open-source software, can sometimes involve publication of vast amounts of data. In recognition of these differences and others, new definitions of open-source AI are currently being developed.[16]

Despite these distinctions, there are several important connections between AI models and open-source software. First, a broad array of open-source software tools (such as PyTorch, Trion, scikit-learn, and TensorFlow) have contributed in many key ways to the development and use of open and closed models. In turn, developers of these models have contributed much to the open-source software ecosystem. Second, advocates of “openness” in AI—including but not limited to open-weight release—often cite many of the same basic values and goals long associated with open-source software. Third, regulatory efforts such as the European Union’s AI Act have drawn on some aspects of the open-source software definition—while omitting or contravening others—in their regulation of AI models.[17] Overall, the relationship between open models and open-source software is a frequent source of confusion and should be further clarified by both communities.

Key Open Questions

The seven areas of emerging consensus offer a starting point for policymakers. However, they leave much unresolved. For this reason, we further propose seventeen open questions as priority areas for research and debate.[18] These questions cover several different areas: the benefits and risks of foundation and open models, ways of getting better data, tripwires and risk tolerances, domestic governance structures, and global developments.

Benefits and Risks of Highly Capable Foundation Models and Open Model Release

Question 1: How important is model weight release in shaping a foundation model's benefits and harms?

Though model weight release can have special importance for benefits and harms in some cases, it isn't clear how well this generalizes across a broad variety of models and circumstances. For example, while open-weight release greatly limits the theoretical ability of model developers to monitor and restrict harmful uses, these safety practices are themselves immature and unevenly (and opaquely) applied. It is therefore hard to tell how often the theoretical safety advantages of closed models are borne out in practice. Similar questions can be raised about the special benefits of open release, such as the prospect of helping low-income countries and marginalized communities partake in AI innovation. Data on this point remains anecdotal. More research is needed to understand the significance of weight release within the larger spectrum of openness.

Additionally, several current trends have the potential to lessen the impact of any one model's open-weight release. Open foundation models are growing in number and in capabilities. At the same time, small and narrow models are also proliferating and becoming more economically important. If these trends continue, the marginal benefits and risks of new open releases (and closed releases, for that matter) may decline. Instead, the key dynamic would be an overall ecosystem with gradual evolution in general capabilities, punctuated by periodic jumps in specific areas due to the development of niche models and systems. Of course, this is just one scenario, but it highlights the importance of monitoring ecosystem-level trends to better assess the benefits and risks of specific open models.

Question 2: Should stakeholders try to agree on a common taxonomy and prioritization of benefits and harms?

Highly capable foundation models and open releases are associated with a range of potential benefits and risks, yet there is wide variation in how stakeholders estimate, rank, and even conceptualize these varied impacts. For example, some stakeholders are primarily concerned with present-day harms—such as the AI creation of nonconsensual intimate imagery (NCII) of women—that have well-documented impacts on individuals or marginalized groups. Others mainly focus on larger-scale but more speculative future scenarios, like the potential loss of human control over AI. This second category is often called "catastrophic risks," yet many argue that present-day AI harms should be seen as catastrophic to those directly affected. The example highlights the role of divergent frames of reference in the conflicts over governance priorities.

A clear, shared taxonomy of benefits and harms could help to guide measurement, policymaking, and debate. It might, for example, draw upon existing scholarship that classifies known patterns of social impact from periods of disruptive technological diffusion. Ideally, a common taxonomy could then help to highlight any areas of shared priority across diverse stakeholder groups, rallying action in those areas. For example, preventing the creation of NCII and child sexual abuse material seems to be an increasingly important focus for many different actors.

On the other hand, a number of credible taxonomies have already been proposed and have not yet resulted in broad agreement on how to understand benefits and risks.[19] Stakeholders are divided not only by their empirical estimates of different impacts but also by their values and interests, which heavily shape priorities. Perhaps full agreement on how to understand AI benefits and risks is not a true precondition for individual and even collective action, including internationally. Even so, researchers should consider whether more progress can be made on mapping harms and benefits in coherent frameworks—at a minimum, with the aim of helping different stakeholder communities better understand each other.

Question 3: How can more attention be focused on crosscutting, complex AI impacts that don't fit easily into traditional categories?

Foundation model governance is often dominated by discussions of specific, intentional uses by good and bad actors. These include the purposeful use of AI in scientific innovation, as well as in malicious activities like hacking and disinformation. Yet foundation models and open release can have more diffuse, complex effects that transcend what individual users (or developers) intend. Such effects may ultimately be more significant to society, and they therefore deserve more focused attention.

One example is AI accidents. History is replete with tragedies—such as the Challenger disaster, the Boeing 737 MAX crashes, and the Three Mile Island nuclear meltdown—caused by the flawed application of otherwise sound technology. As foundation models become embedded in more and more parts of society and the economy, errors of engineering, planning, and communication will multiply. By the same token, there will also be surprising benefits that emerge when scientific innovators and society's leading institutions combine foundation models with other technologies. Many past innovations—such as commercial drones and the app-based platform economy—resulted from various technological building blocks being integrated in ways not anticipated by the inventors of those building blocks.

Another kind of complex, overlooked impact might be called "boiling frog" scenarios, after the apocryphal notion that a frog in a pot will fail to notice a deadly but gradual rise in water temperature. In these scenarios, the growing influence of foundation models has massive effects, but they accumulate so incrementally that decisionmakers struggle to notice and respond to them. Many environmental problems fit this pattern, including climate change, biodiversity loss, micro- and nano-plastic pollution, and space junk. In the AI context, the rise of powerful AI agents could cause a gradual loss of individual or societal agency. In the same way that complex and large-scale human institutions (like stock markets or electoral systems) induce self-reinforcing behavior and thereby resist change, the embedding of powerful AI systems throughout society could create path dependencies that become deeply rooted over time.

A third category is the multidimensional interactions between foundation models and the human structures—the economy, the labor market, democracy, and information platforms—that shape the AI industry and are shaped by it in turn. Such feedback loops can be very powerful but are poorly understood. For example, the growing use of AI-enabled communication and persuasion technologies by a broad range of political and commercial actors may gradually reshape the character of public discourse—including discourse on AI policy itself. Indeed, social processes such as democratic discourse are poorly understood and managed to begin with, irrespective of AI disruption.[20] Fundamentally, analysts should consider a wide range of long-term, complex causal processes when assessing and designing governance mechanisms for foundation models. Such systemic effects have been the dominant impacts of other recent major technologies, such as social media and smartphones.

Question 4: How can governance systems account for rapid structural changes in the AI sector?

The AI landscape will look radically different in the future than it does today due to shifts in technology, commercial incentives, and societal responses, among many other factors. While the outlines of some changes can already be anticipated, their consequences are very unclear. Several structural trends appear likely over the next five years.

First, the number and diversity of AI models and systems will likely explode as compute costs fall, smaller models gain traction, and ways of combining or conjoining multiple models are further developed. In the near future, today's era may be remembered as the "pre-Cambrian" moment just prior to a rapid acceleration of AI's growth, evolution, and differentiation. If so, the future AI ecosystem will be harder to understand, predict, and govern than what exists today.

Second, the embedding of AI into all kinds of companies and infrastructures will increase the number of relevant decisionmakers and societal impacts. Whereas foundation models built by frontier labs are often seen as the key innovations today, future business value may be largely driven by countless companies developing context-specific applications. The growing number of AI applications would likely cause unexpected AI events—whether damaging accidents or innovative breakthroughs—to increase in frequency, diversity, and scope. Governance and evaluation would thus need to involve a much larger universe of actors, actions, and potential outcomes.

Third, the AI market structure will almost certainly change a great deal. Already, the AI industry's growing demand for chips, energy, data, and talent is driving changes in multiple business sectors. There are new kinds of corporate partnerships, financial innovation, industrial policy, and more. It still isn't clear what kind of market structure will, or should, take shape in the future. For example, it's impossible to predict which parts of the AI value chain will prove most profitable, dominant, or scarce. While foundation models are currently a key node, the future of AI could conceivably depend more on upstream nodes (such as semiconductor design and fabrication, pre-training data, and even land for data center siting) or downstream nodes (like fine-tuning data, AI application development and distribution, and the integration of AI with other business lines). It is also unclear how shifting commercial incentives, norms, and technical developments will affect model release practices such as structured access.

Finally, public debates about AI risks and benefits will look very different in the future as more elements of society are drawn into these debates. AI has captured the attention of publics throughout the world during the last eighteen months, and many national leaders have sought to make their mark. Yet a number of key policy conversations, including how to govern foundation model release, have remained dominated by experts. This will undoubtedly change soon as a much broader range of constituencies, interest groups, and voices—including those directly impacted by AI—come to the table. The democratization of discourse could surface valuable new information about how different groups are affected by foundation models, but it could also bring familiar challenges like polarization, populism, and gridlock.

In sum, the design of governance systems today should somehow account for the likelihood of a radically altered AI landscape just a few years from now. Some changes will be logical and predicable outgrowths of current trends, while others will be unexpected and emergent. Governing through such changes is a difficult puzzle that researchers and stakeholders should actively try to solve. To paraphrase the great hockey player Wayne Gretzky, we should skate to where the puck is likely to be, rather than where it is right now.

Getting Better Data on Model Risks and Benefits

Question 5: What kinds of foundation model evaluations are needed and why?

Model evaluation initially developed as a primarily scientific project aimed at better understanding and characterizing the technical evolution of AI architectures. In recent years, the rapid commercialization of AI has led to a surge of interest in evaluating models for more practical needs such as product design, marketing, and governance. But the practices of model evaluations have not yet fully come into alignment with this evolving, broader set of purposes. For example, today's evaluations often focus more on technical than on sociotechnical aspects, even as the sociotechnical sphere grows in importance due to deepening human interactions with AI systems.[21] More fundamentally, there is not yet a clear sense of the full array of evaluations that are needed and what role each will play.

This is partly because the decisionmaking structures that evaluations can support are themselves underdeveloped. For example, the U.S. AI executive order requires certain red-teaming results be shared with the government, but it does not spell out whether or how this data will inform any particular decision, in part because U.S. agencies are still developing their policy and research agendas.[22] More fundamental discussion is needed on the various purposes of evaluations—including policymaking as well as pure science. Answers to this question can help to guide investments in the scientific and commercial infrastructure for model (and system) evaluation. It can also help the consumers of evaluations, such as policymakers, understand their strengths and limitations.

Question 6: What will be accomplished by improving evaluations over time?

The current state of model evaluations remains very limited. Evaluations are typically nonstandard, nonrepeatable, noncomprehensive, not reflective of real-world conditions, and require significant human judgment. These problems are widely recognized among experts, though not always among political leaders and the broader public. However, there is significant disagreement about what progress to expect, over what timelines, and with what investments.

Incident reporting is one example. Some argue that better systems of reporting harmful AI incidents would enable the rapid accumulation of data to identify key real-world trends. Others are more pessimistic, noting that past efforts have not yielded high-quality data, nor did they lead to actual fixes in identifiable problem areas. A proper AI incident reporting mechanism would require major investments in technical and other infrastructure, but the bottom-line outcomes are difficult to predict. Evaluation reproducibility is another area where stakeholders disagree about the products for further improvement. It's clear that frequent, opaque version changes in models are a challenge for evaluation reproducibility—but it's not clear how to solve this problem and whether other causes also exist.

As these examples illustrate, evaluations have many limitations, but it isn't always apparent whether they are fixable. At one extreme, some limitations are inherent and could persist indefinitely. For example, pre-release evaluations of general-purpose models cannot, by definition, give comprehensive predictions of future use cases. Moreover, large foundation models are fundamentally complex systems whose impacts will always depend a great deal on chaotic interactions with human systems—which are themselves poorly understood.

On the other end of the spectrum, some evaluation problems seem readily solvable with the right resources and incentives. Inadequate domain expertise within the AI industry would be an example. And in the middle, there are generational limitations of evaluations that are tied to the field's current immaturity and can hopefully be addressed over time. For example, standardized metrics have not yet emerged for many benefits and harms, but they might well be defined in the future. More research is needed to discern the addressability of different kinds of limitations.

A related but even more fundamental question is whether better evaluations can bring real clarity to difficult large-scale policy decisions. Stakeholders disagree about this. Some hope that stronger evaluations can help to show, for example, how close AI capabilities are to crossing into a specified danger zone. Others believe that the nature and location of the danger zone itself will remain disputed, forcing decisionmakers to resort to simpler heuristics rather than clear technical data. One such heuristic is the "precautionary principle" developed by environmental scientists and advocates, which calls for acting to prevent the possibility of great harms "even if some cause and effect relationships are not fully established scientifically."[23]

It's also possible that the increasing prominence of evaluations—and wider awareness of their limits—will be strategically leveraged by those with a vested interest in slowing down decisionmaking. This has been a major barrier in the regulation of greenhouse gas emissions, tobacco, toxic chemicals, and much else. While investments in better evaluation science are crucial, decisionmaking should be designed to function in the face of persistent knowledge gaps. Stakeholders should actively debate how much information is really needed to make various governance decisions: the perfect should not be the enemy of the good.

Question 7: What is the state of post-release monitoring and enforcement, and how can these be improved?

Pre-release evaluations have received a great deal of attention, but there has been comparably less discussion of post-release practices for monitoring uses and enforcing relevant policies. This is a significant omission, particularly in the context of foundation model governance and open-release decisions. The availability of post-release practices represents a key potential advantage of closed models, yet there is little public information about how common and effective these practices are.

Improved monitoring will require the active involvement of actors beyond the foundation model developers. These include social media platforms (where AI-generated content often reaches mass distribution) and model customers (who may not want to share data about their usage).

It will also require new metrics and reporting standards. For example, good policymaking on AI and the labor market will require rigorous, clear, and measurable definitions of AI-related job displacement and job creation. Even so, pre-release and post-release research should not be limited to quantifiable metrics. These can be gamed, and they fail to capture impacts on intangible goods such as privacy and democracy. Conversations, perhaps in the form of ongoing focus groups, with those directly affected by AI models and systems can provide important qualitative information.

Question 8: What is the proper role of government in foundation model evaluation?

Governments are pursuing a range of different roles in foundation model evaluation. Some, like the UK, have described a core focus on national security concerns, where governments have unique data and capacity. The national security mission tends to be more politically achievable and is somewhat more insulated from politicization than other areas of concern. National security evaluations may need to remain largely classified, creating the potential for information silos. But this is a familiar challenge, which can be mitigated via mechanisms like declassification and the selective clearing of outside experts.

However, there are a number of pitfalls and question marks with government-led evaluation. The proper scope of these missions hasn't been clearly defined—creating risks of overinclusion, underinclusion, or mission creep. A national security focus will inevitably exclude many important stakeholders, and its opacity makes it harder to trust that proper methods are being followed. More generally, governments may end up captured by private sector interests rather than independently assessing them. And very few governments actually have the capacity to perform useful evaluations. Even those with a latent capacity may fail to resource this mission adequately or sustainably. Additionally, leaders may politicize the evaluation process or simply not understand it. If so, governments and publics could draw inappropriate conclusions from the results.

More discussion will be needed about the appropriate role of governments in studying and evaluating foundation models. Policymakers should view their initial efforts, including recently launched national AI Safety Institutes, as experiments. These will require ongoing readjustments and, at some point, a wholesale review. Lessons can be learned from governmental involvement in other domains, such as the regulation (or nonregulation) of critical infrastructure cybersecurity.

Tripwires for Concern and Risk Tolerance Levels and Processes

Question 9: How effective are early efforts to define and enforce risk thresholds?

In recent months, several leading AI labs have issued public documents that define their internal thresholds for concern and outline plans for monitoring and responding to identified risks. This trend helped set the stage for a public commitment by sixteen companies, at the AI Seoul Summit in May 2024, to release risk thresholds and mitigation frameworks by next year's summit in France.[24]

The creation of formal risk thresholds, although immature, has already had positive effects. It has helped to raise the internal status of evaluation work within the labs, aiding in the recruitment of safety talent and in internal collaboration. The public articulation of risk thresholds has also contributed to broader discussions about regulation, standards of care, and professional norms. Legislatures, litigants, and advocacy groups have begun to cite these documents in their own attempts to hold companies accountable and to develop their own governance proposals. For example, the earliest risk thresholds published by a handful of companies provided a concrete benchmark that assisted the UK and South Korean governments in designing and negotiating formal commitments from a larger group of companies.

However, these kinds of policies are in their earliest phase of development and still lack significant detail, leaving many questions unanswered. Published risk thresholds have generally been articulated at high levels of abstraction, so it isn't clear how each company will interpret them in practice. Internal enforcement mechanisms aren't fully specified, raising questions about the impact of commercial conflicts of interest on future enforcement decisions. There is no common understanding of what systems and investments are needed to monitor the identified risks—to include clarifying the distinct but complementary roles of pre-release evaluations and post-release observations. Actual monitoring practices seem to vary significantly. Some categories of risk—such as persuasion and autonomy—remain quite undertheorized.

In sum, much is unknown about how well this first generation of risk thresholds is performing. Answering these questions will be important for helping design follow-on policies—whether voluntary or mandatory—that are more comprehensive, measurable, and enforceable.

Question 10: What is the right format for risk thresholds?

Risk thresholds can be quantitative, qualitative, or both. At one extreme, the U.S. executive order uses a strict quantitative threshold, based on the amount of compute used in training, to define "dual use foundation models." At the other end of the spectrum, the Partnership on AI's Guidance for Safe Foundation Model Deployment uses qualitative language to define "paradigm-shifting or frontier models."[25] There are also mixed approaches. Anthropic's Responsible Scaling Policy and OpenAI's Preparedness Framework propose an initial number of quantitative metrics as illustrative indicators within larger, primarily qualitative frameworks.[26]

More discussion and experimentation are needed on the full range of approaches. Quantitative measures offer specificity and accountability but can be reductive and misleading given the current state of the science of evaluations, while qualitative measures are more adaptable and inclusive but leave significant room for interpretation and discretion. In general, more complex tiered and multivariable approaches should be explored to incorporate the best of both approaches. There is also a great deal that the AI industry and its regulators can learn from other fields, such as (re)insurance, financial stability oversight, and critical infrastructure protection, which have faced similar challenges in defining and predicting serious harms.

Question 11: How can the menu of risk mitigations be expanded?

More stakeholders are shifting focus from a limited set of binary governance options—such as whether or not to release a model, and whether weights should be open or closed—to a more varied menu. Recently, for example, companies and experts have converged on the idea of "precautionary friction," which refers to staged or structured release strategies that help to build confidence in initial risk assessments before a fully open release is considered.

Much more should be done to explore the full menu of risk mitigations and governance options. Technical researchers are exploring how to make model safeguards harder or costlier to remove.[27] Others are exploring the notion of "kill switches," which might leverage on-chip governance or model weight encryption.[28] Models themselves could perhaps be made intentionally brittle, self-destructing in the event of certain scenarios—much like the U.S. government reportedly sought to contain the Stuxnet virus.[29] Technical solutions like these must also be weighed against their feasibility for all types of release on the spectrum. In addition, a vast risk mitigation space exists outside the models themselves. For example, public and private investments in societies' biodefenses and cybersecurity can mitigate the threat of AI-empowered bioweapons or cyberweapons.[30] There is also a need for more classic contingency planning, such as table-top exercises, to stress-test decisionmaking processes and communication links.

In general, it is necessary to develop a larger and more flexible variety of risk mitigations that involves a broader array of actors. This fuller menu will allow for better tailoring, increased effectiveness, and lower cost of risk mitigation.

Domestic Decisionmaking and Governance Structures

Question 12: What is the right mixture of roles and responsibilities across private and public sector actors?

The right balance of governance responsibility, as between industry and government, is not yet clear for foundation models. Industry is more knowledgeable about the technology and more agile in responding to new developments, yet companies' narrow self-interest and competitive pressures can lead to short-sighted decisions not in the best interest of society as a whole. Government is charged with a broader public mission and has deep practical knowledge from previous governance activities, yet it is often slow, ham-fisted, and subject to capture by politics or special interests.

Rather than accept a false choice between two inadequate options, policymakers should design intelligent combinations that allow the strengths of one mode to counterbalance the weaknesses of the other. For example, companies could join with civil society actors to establish independent foundations that make certain governance decisions, such as the standard for model release. Companies could also be encouraged or required to purchase large amounts of insurance, with the insurer then serving as a soft regulator of potentially risky behavior. Litigation funds could be established to combat model abuse that violates license restrictions. New channels for protected whistleblowing could encourage insiders to make responsible disclosures of activities that concern them.

These are just a few options that could be explored. In general, the boundaries of the policy space need to be expanded with creative and pragmatic solutions to known governance challenges.

Question 13: What aspects of the governance patchwork are most in need of further research, experimentation, scale-up, or restraint?

In this early phase of foundation model governance, policymakers at a range of institutions are considering and implementing many different measures. An important task, then, is to assess the overall governance picture and identify persistent gaps, question marks, or pain points. At a high level, this means identifying the key objectives (such as reducing catastrophic risk), actors (including legislatures, government agencies, AI labs, other business sectors, investors, and more), and mechanisms (for example, regulation, standards, reporting, licensing, and norm-setting.). The goal should be a healthy governance tapestry, with minimal gaps or seams, to address the full range of risks while also enabling diverse approaches to innovation.

Compared to this future ideal, today's governance reality looks quite patchy. In the United States, for example, fractured governmental authority has forced federal policymakers to invoke ill-fitting powers—such as the Defense Production Act—rather than start from first principles. Congressional leaders like Senator Chuck Schumer have larger ambitions and some promising ideas,[31] but Congress as a whole has increasingly struggled in recent decades to pass major legislation. Stakeholders should consider a long-term effort to draft model U.S. federal laws, to prepare for the possibility that political will unexpectedly emerges. This could happen very abruptly—for example, due to a court decision or an AI-related catastrophe.

There are a number of other governance players with potentially interesting roles. In the United States, state governments have independent legal powers and often show greater political agility than their federal counterparts. Traditional regulatory agencies have preexisting authorities that could apply to new technologies as much as old ones. Model hosting services perform an array of governance actions, such as content moderation and platform design, that affects model availability and community norms. More thought is needed on how all of these actors can complement each other. Additionally, many institutions will need new resources to build the adequate capacity to understand, adapt, and respond to emerging governance challenges.

Question 14: How can governance expertise and public participation both be expanded?

Foundation model governance requires many different kinds of expertise. This includes technical expertise on AI systems themselves, sociotechnical expertise on how different human communities interact with AI, commercial expertise on the evolving business environment, and policy expertise on the hard and soft law mechanisms available—among many other key areas. It is very difficult to assemble all of this knowledge together in one institution. Governments, in particular, often struggle to acquire and retain top technical talent. Stakeholders need to consider how to enhance their own organizations' expertise while at the same time contributing to a national and international build-up of governance talent. This may involve education, fellowships, and other mechanisms.

At the same time, there is danger in relegating foundation model governance to AI experts and AI policymakers alone. Ordinary people have important expertise on how AI concretely affects their lives. Furthermore, the rise of AI comes at a time when people in many different countries already feel that powerful decisionmakers are no longer responsive to their needs and that the direction of society is increasingly beyond their ability to influence. AI has the potential to worsen this problem—but it can also help alleviate it if properly governed. Sustainable and effective governance should build a substantial role for public participation. This doesn't mean holding plebiscites over technical matters or creating public vetoes of private decisionmaking. But neither does it mean empty processes with no real impact. Careful thought should be given to this problem.

Global Developments and International Governance

Question 15: What aspects of governance should be internationalized, and what aspects should remain primarily domestic?

There are many different visions for the role of international governance in foundation models. At the high end, some expect the eventual creation of a multidimensional "regime complex," which could include robust supranational regulation and technology transfer akin to the Treaty on the Non-Proliferation of Nuclear Weapons.[32] At the low end, the world might continue much as it is today—with broad-based but fairly modest efforts to share scientific assessments of AI, alongside ad hoc collaborative projects by smaller groupings of like-minded states. There are many possibilities in the middle. For example, it is conceivable that major powers could agree on a thin but powerful set of international regulations that narrowly target loss-of-control scenarios—perhaps facilitated by on-chip governance.

The viability and desirability of these visions remain highly indeterminate. While catastrophic risks most clearly call for international governance, the fractious state of geopolitics makes this pathway a narrow and long one. Good policy ideas must be married with careful and patient diplomacy to build political will over time. Stakeholders should look for ways to set reasonably high aspirations that are nevertheless achievable.

For many other policy challenges, domestic governance may make more sense for the foreseeable future. National and subnational governments have traditionally been the prime movers when policymaking is highly sensitive to cultural values and local conditions, or it requires the balancing of competing constituencies within a political community. AI policy goals such as limiting bias, promoting fairness, and countering misinformation may well belong in this category. While technical and diplomatic exchanges among like-minded countries can still be valuable, not all AI governance challenges are ripe for internationalization.

The growing number of international initiatives on AI governance—at the United Nations, the AI Summits, the G7 group of advanced economies, and elsewhere—have brought helpful focus to key AI policy challenges and built useful networks of thinkers and decisionmakers. But they have also fractured the attention of stakeholders and generated a large amount of diplomatic and compliance work. It remains to be seen which, if any, of these bodies should be a focal point for sound policymaking on foundation models in general and open models in particular.

Question 16: How can familiar patterns of international competition and sclerosis be overcome?

Ambitious ideas for the global governance of AI, and even modest efforts at international collaboration, inevitably run against familiar frustrations. Tension between the United States and China is only the most obvious example. There is simply no satisfying global structure to organize the world on common interests. Rather, there are a dizzying array of international initiatives and forums vying for relevance, risking fracture and distraction.

An important question is whether and how the particular dynamics of AI create opportunities to break these well-worn patterns. For example, the novelty and obvious significance of AI has created new discussion space in relationships where this can be hard to find—most notably, with recent direct talks between the United States and China. More broadly, the time and attention that national leaders, diplomats, and international forums have devoted to AI within the last eighteen months is truly remarkable—no doubt the envy of stakeholders in other important issue areas. To make progress on the international dimensions of AI and foundation model governance, players will need to sustain this energy somehow and resist the normal gravity of geopolitics. Otherwise, the logic of national self-interest and zero-sum gamesmanship will lead to internationally fragmented governance that fails to unlock the full benefits or mitigate the real risks of AI.

Question 17: How should Global North actors engage responsibly with the Global South?

Although Global South economies are sometimes characterized merely as consumers and bystanders in the global AI and foundation model ecosystem, a range of actors in these countries are participating actively in multiple ways.[33] Local entrepreneurs are building on top of leading open models, and increasingly, they are training new foundation models native to their own languages. Behind the scenes, workers in the Global South provide data cleaning, content moderation, and human feedback services to international companies.

Still, it remains unclear whether current trajectories will provide significant long-term economic benefits to Global South countries and workers. The main economic value of the growing global AI economy is being captured elsewhere.[34] The boom in data enrichment work has largely occurred in settings with poor protections for workers, highlighting a need for better monitoring, oversight, and regulation of the AI data supply chain. In fact, AI disruption may even threaten preexisting economic development pathways for Global South economies—for example, if middle-income jobs at international call centers are replaced by AI voice assistants and chatbots, or entry-level web development, data analysis, and more are replaced by AI applications. While optimists hope that an AI revolution will bring a rising tide of abundance that lifts all boats, precedent from past technologies has been more mixed.

In this context, Global North governments, companies, and civil society organizations need to find ways of partnering effectively and responsibly with those in the Global South. A range of possibilities have been suggested, drawing on classic notions of development assistance, technology transfer, and import substitution: helping build localized data centers, creating national versions of foundation models, or investing in AI-relevant education and human capital development. However, these ideas are untested and largely undertheorized. There must be more active and robust dialogue between development economists and AI experts in the Global South and the Global North, followed by a period of intense experimentation.

Ultimately, development pathways will vary depending on specific conditions across and within each country. Politics and power are also part of the equation. The Global South includes a wide variety of countries, some of which are not democracies and/or have poor human rights records. In such places, civil society groups (local or international) may be more appropriate partners than governments at times. Special thought should also be given to the world's 746 million people living without electricity and to the 2.6 billion people who lack Internet access. These groups will not be able to directly access or shape AI systems, but they may well indirectly shape the systems (through data collected about them) and be shaped by them (through AI-driven systems operated by others, which are already affecting pricing and other negotiations in informal economies, for instance). Much more work is needed to identify sensitive, pragmatic ways for national and international agencies, civil society organizations, and companies to engage these populations and account for their interests in AI governance.

About the Authors

Jon Bateman (corresponding author) is a senior fellow at the Carnegie Endowment for International Peace, where he focuses on global technology challenges at the intersection of national security, economics, politics, and society.

Dan Baer is senior vice president for policy research and director of the Europe Program at the Carnegie Endowment for International Peace.

Stephanie A. Bell is the chief programs and insights officer at the Partnership on AI, where she leads programs and research for a multistakeholder nonprofit dedicated to ensuring artificial intelligence benefits society.

Glenn O. Brown is a consultant and fractional executive to startups, public companies, and NGOs. He is a senior adviser at MIT's Center for Constructive Communication and a board member of Creative Commons and The Texas Tribune.

Mariano-Florentino (Tino) Cuéllar is the president of the Carnegie Endowment for International Peace.

Deep Ganguli is a research scientist at Anthropic focusing on the interpretability, fairness, transparency, and societal impacts of AI.

Peter Henderson is an assistant professor at Princeton University with appointments in the Department of Computer Science and the School of Public and International Affairs. Previously, he received his PhD in computer science and JD from Stanford University.

Brodi Kotila is international AI governance lead at the RAND Corporation's Technology and Security Policy Center.

Larry Lessig is the Roy L. Furman Professor of Law and Leadership at Harvard Law School.

Nicklas Berild Lundblad is strategic adviser and director of global public policy at Google DeepMind. He has worked in public policy and technology for more than twenty-five years with Google, Stripe, and other expert and advisory roles. He is a member of advisory groups in Spain, Sweden, and the OECD.

Janet Napolitano is a professor of public policy at the University of California, Berkeley. She is a former president of the University of California, secretary of homeland security, and governor of Arizona.

Deborah Raji is a senior Trustworthy AI fellow at the Mozilla Foundation and a computer science PhD student at the University of California, Berkeley, who is interested in questions on AI auditing and evaluation. She works closely with civil society, investigative journalists, policymakers, and corporations on various projects to investigate, assess, and better understand AI deployments. She has been named to the Forbes 30 Under 30, MIT Tech Review 35 Under 35 Innovators, and TIME 100 Most Influential in AI.

Elizabeth Seger, PhD, is the director of digital policy at Demos, the UK's leading cross-party think tank. Elizabeth is internationally recognized for her work on open-source model sharing policy, epistemic security, and AI democratization.

Matt Sheehan is a fellow at the Carnegie Endowment for International Peace, where his research focuses on global technology issues, with a specialization in China's artificial intelligence ecosystem.

Aviya Skowron is head of policy at EleutherAI. They translate EleutherAI's expertise in machine learning into policy recommendations informed by the latest LLM research and promote best practices among open source and "open-ish" AI developers.

Irene Solaiman is the head of global policy at Hugging Face, where she leads public policy and conducts social impact research. Irene serves on the Partnership on AI's Policy Steering Committee, the Center for Democracy and Technology's AI Governance Lab Advisory Committee, and the Aspen Institute's AI Elections Initiative Advisory Council.

Helen Toner is the director of strategy and foundational research grants at Georgetown University's Center for Security and Emerging Technology.

Polina Zvyagina is currently the global AI policy and governance director at Meta. She's also a privacy and product lawyer who has supported product and AI-ML teams at Airbnb, Uber, and Apple.

Acknowledgments

Carnegie is grateful for input and feedback on this project from Matt Clifford, Jason Kwon, and Hadrien Pouget.

Notes

[1] "Governance" means the full range of decisionmaking structures used by private and public actors, to include internal controls, external voluntary commitments, market incentives, professional norms, and binding government regulations.

[2] It is important to recognize that "open" and "closed" are ambiguous, nonstandardized terms in the context of AI models. This is discussed more extensively below, in the section on areas of consensus, particularly points two and seven.

[3] Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei et al, "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives," Centre for the Governance of AI, 2023, https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf.

[4] Sayash Kapoor and Rishi Bommasani et al, "On the Societal Impact of Open Foundation Models," February 27, 2024, https://arxiv.org/pdf/2403.07918v1.

[5] Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," February 5, 2023, http://arxiv.org/pdf/2302.04844; Seger, Dreksler, Moulange, Dardaman, Schuett, Wei et al, "Open-Sourcing Highly Capable Foundation Models; and Adrien Basdevant et al, "Towards a Framework for Openness in Foundation Models," May 21, 2024, https://arxiv.org/pdf/2405.15802.

[6] "NTIA AI Open Model Weights RFC: Document Comments," National Telecommunications and Information Administration, February 26, 2024, https://www.regulations.gov/document/NTIA-2023-0009-0001/comment.

[7] Solaiman, "The Gradient of Generative AI Release"; and Toby Shevlane, "Structured Access: An Emerging Paradigm for Safe AI Deployment," April 11, 2022, https://arxiv.org/abs/2201.05159.

[8] Nik Marda, "Technical Readout - Columbia Convening on Openness and AI," Mozilla and Columbia University, March 27, 2024, https://foundation.mozilla.org/en/research/library/technical-readout-columbia-convening-on-openness-and-ai/; Udbhav Tiwari, "Policy Readout - Columbia Convening on Openness and AI," Mozilla and Columbia University, March 27, 2024, https://foundation.mozilla.org/en/research/library/policy-readout-columbia-convening-on-openness-and-ai/; and Adrien Basdevant et al, "Towards a Framework."

[9] Andy Zou et al, "Universal and Transferable Adversarial Attacks on Aligned Language Models," December 20, 2023, https://arxiv.org/abs/2307.15043; and Pranav Gade, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish, "BadLlama: Cheaply Removing Safety Fine-tuning from Llama 2-Chat 13B," May 28, 2024, https://arxiv.org/abs/2311.00117.

[10] Basdevant et al, "Towards a Framework."

[11] Peter Henderson et al, "Safety Risks from Customizing Foundation Models via Fine-Tuning," Stanford Institute for Human-Centered Artificial Intelligence, January 11, 2024, https://hai.stanford.edu/policy-brief-safety-risks-customizing-foundation-models-fine-tuning.

[12] "ABOUT ML Resources Library," Partnership on AI, https://partnershiponai.org/about-ml-resources-library/; Henderson et al, "Safety Risks"; and Basdevant et al, "Towards a Framework"; and Shayne Longpre et al, "The Responsible Foundation Model Development Cheatsheet: A Review of Tools & Resources," June 26, 2024, https://arxiv.org/pdf/2406.16746.

[13] Elizabeth Seger et al, "Democratising AI: Multiple Meanings, Goals, and Methods," AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (August 29, 2023), https://dl.acm.org/doi/fullHtml/10.1145/3600211.3604693.

[14] "The Open Source Definition," Open Source Initiative, February 16, 2024, https://opensource.org/osd.

[15] Open models often have licensing terms that prohibit certain uses. Additionally, AI models and systems are comprised of more than just source code—which is often the primary or sole focus of traditional "open source" software definitions.

[16] See "The Open Source AI Definition – draft v. 0.0.8," Open Source Initiative, accessed July 18, 2024, https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-8.

[17] For example, the EU AI Act refers to "AI models that are released under a free and open-source licence that allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available." The traditional definition of open-source software includes additional parameters not mentioned by the Act, such freedom to create derived works under the same license, and non-discrimination against people or fields of endeavor (e.g., users or use cases). See EU AI Act, Article 53, Paragraph 2, available at https://artificialintelligenceact.eu/article/53/; and "Open Source Definition," Open Source Initiative.

[18] These open questions can help shape individual research efforts as well as broader attempts to define global research agendas. One such attempt is being made now by United States Agency for International Development and the U.S. Department of State, in collaboration with the Department of Energy and the National Science Foundation. "Global AI Research Agenda," USAID, March 14, 2024, https://www.federalregister.gov/documents/2024/03/14/2024-05357/global-ai-research-agenda.

[19] Kapoor and Bommasani et al, "Societal Impact."

[20] Alicia Wanless, *The More Things Change: Understanding Conflict in the Information Environment Through Information Ecology*, King's College London, April 1, 2023, https://kclpure.kcl.ac.uk/portal/en/studentTheses/the-more-things-change.

[21] There are promising efforts underway, including Reva Schwartz et al, "The Draft NIST Assessing Risks and Impacts of AI (ARIA) Pilot Evaluation Plan," NIST, June 5, 2024, https://ai-challenges.nist.gov/aria/docs/evaluation_plan.pdf.

[22] White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Section 4.2(i)(C), October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[23] David Kriebel et al, "The Precautionary Principle in Environmental Science," *Environmental Health Perspectives* 109, no. 9 (2001), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1240435/pdf/ehp0109-000871.pdf.

[24] "Frontier AI Safety Commitments, AI Seoul Summit 2024," UK Department for Science, Innovation & Technology, May 21, 2024, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024.

[25] "PAI's Guidance for Safe Foundation Model Deployment A Framework for Collective Action," Partnership on AI, 2023, https://partnershiponai.org/modeldeployment/.

[26] "Anthropic's Responsible Scaling Policy: Version 1.0," Anthropic, September 19, 2023, https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf; and "Preparedness Framework (Beta)," OpenAI, December 18, 2023, https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

[27] Peter Henderson et al, "Self-Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundation Models," Sixth AAAI/ACM Conference on AI, Ethics, and Society (August 9, 2023), https://arxiv.org/abs/2211.14946.

[28] Onni Aarne, Tim Fist, and Caleb Withers, "Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing," Center for a New American Security, January 2024, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-Tech-Secure-Chips-Jan-24-finalb.pdf; and Girish Sastry et al, "Computing Power and the Governance of Artificial Intelligence," Centre for the Governance of AI, February 14, 2024, https://cdn.governance.ai/Computing_Power_and_the_Governance_of_AI.pdf.

[29] Sydney J. Freedberg, Jr., "Cyber Command Lawyer Praises Stuxnet, Disses Chinese Cyber Stance," Breaking Defense, March 12, 2012, https://breakingdefense.com/2012/03/cyber-command-lawyer-praises-stuxnet-disses-chinese-cyber-stanc/.

[30] For an overview of AI's potential impact on cybersecurity, see Jenny Jun, "How Will AI Change Cyberoperations?," War on the Rocks, April 30, 2024, https://warontherocks.com/2024/04/how-will-ai-change-cyber-operations/.

[31] "Driving U.S. Innovation in Artificial Intelligence: A Roadmap for Artificial Intelligence Policy in the United States Senate," The Bipartisan Senate AI Working Group, May 2024, https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf.

[32] Emma Klein and Stewart Patrick, "Envisioning a Global Regime Complex to Govern Artificial Intelligence," Carnegie Endowment for International Peace, March 21, 2024, https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en.

[33] Aubra Anthony, Lakshmee Sharma, and Elina Noor, "Advancing a More Global Agenda for Trustworthy Artificial Intelligence," Carnegie Endowment for International Peace, April 30, 2024, https://carnegieendowment.org/research/2024/04/advancing-a-more-global-agenda-for-trustworthy-artificial-intelligence?lang=en.

[34] Bhaskar Chakravorti, Ajay Bhalla, and Ravi Shankar Chaturvedi, "Charting the Emerging Geography of AI," Harvard Business Review, December 12, 2023, https://hbr.org/2023/12/charting-the-emerging-geography-of-ai; and Mauro Cazzaniga et al, "Gen-AI: Artificial Intelligence and the Future of Work," International Monetary Fund, January 14, 2024, https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2024/01/14/Gen-AI-Artificial-Intelligence-and-the-Future-of-Work-542379.

Carnegie Endowment for International Peace

In a complex, changing, and increasingly contested world, the Carnegie Endowment generates strategic ideas, supports diplomacy, and trains the next generation of international scholar-practitioners to help countries and institutions take on the most difficult global problems and advance peace. With a global network of more than 170 scholars across twenty countries, Carnegie is renowned for its independent analysis of major global problems and understanding of regional contexts.

Technology and International Affairs Program

The Technology and International Affairs Program develops insights to address the governance challenges and large-scale risks of new technologies. Our experts identify actionable best practices and incentives for industry and government leaders on artificial intelligence, cyber threats, cloud security, countering influence operations, reducing the risk of biotechnologies, and ensuring global digital inclusion.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org