RESEARCH MEMO
AUGUST 2024

AI GOVERNANCE INITIATIVE
OXFORD MARTIN SCHOOL
UNIVERSITY OF OXFORD

The Future of International Scientific Assessments of AI's Risks

Authors: Hadrien Pouget*, Claire Dennis*, Jon Bateman, Robert F. Trager, Renan Araujo, Haydn Belfield, Belinda Cleeland, Malou Estier, Gideon Futerman, Oliver Guest, Carlos Ignacio Gutierrez, Vishnu Kannan, Casey Mahoney, Matthijs Maas, Charles Martinet, Jakob Mökander, Kwan Yee Ng, Seán Ó hÉigeartaigh, Aidan Peppin, Konrad Seifert, Scott Singer, Maxime Stauffer, Caleb Withers, and Marta Ziosi

*Equal contribution. Name order randomised. Please cite as: 'Pouget, H., Dennis, C., et al. (2024)'

In partnership with

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

Contents

Executive Summary

Managing the risks of artificial intelligence (AI) will require international coordination among many actors with different interests, values, and perceptions. Experience with other global challenges, like climate change, suggests that developing a shared, science-based picture of reality is an important first step toward collective action. In this spirit, last year the UK government led twenty-eight countries and the European Union (EU) in launching the *International Scientific Report on the Safety of Advanced AI*.

The UK-led report has accomplished a great deal in a short time, but it was designed with a narrow scope, limited set of stakeholders, and short initial mandate that's now nearing its end. Meanwhile, the United Nations (UN) is now moving toward establishing its own report process, though key parameters remain undecided. And a hodgepodge of other entities—including the Organisation for Economic Co-operation and Development (OECD), the emerging network of national AI Safety Institutes (AISIs), and groupings of scientists around the world—are weighing their own potential contributions toward global understanding of AI.

How can all these actors work together toward the common goal of international scientific agreement on AI's risks? There has been surprisingly little public discussion of this question, even as governments and international bodies engage in quiet diplomacy. Moreover, the difficulty of the challenge is not always fully acknowledged. Compared to climate change, for example, AI's impacts are more difficult to measure and predict, and more deeply entangled in geopolitical tensions and national strategic interests.

To discuss the way forward, Oxford Martin School's AI Governance Institute and the Carnegie Endowment for International Peace brought together a group of experts at the intersection of AI and international relations in July. Drawing from that discussion, six major ideas emerged:

- **No single institution or process can lead the world toward scientific agreement on AI's risks.** There are too many conflicting requirements to address within a single framework or institution. Global political buy-in depends on including a broad range of stakeholders, yet greater inclusivity reduces speed and clarity of common purpose. Appealing to all global audiences would require covering many topics, and could come at the cost of coherence. Scientific rigor demands an emphasis on peer-reviewed research, yet this rules out the most current proprietary information held by industry leaders in AI development. Because no one effort can satisfy all these competing needs, multiple efforts should work in complementary fashion.

- **The UN should consider leaning into its comparative advantages by launching a process to produce periodic scientific reports with deep involvement from member states.** Similarly to the Intergovernmental Panel on Climate Change (IPCC), this approach can help scientific conclusions achieve political legitimacy, and can nurture policymakers' relationships and will-to-act. The reports could be produced over a cycle lasting several years and cover a broad range of AI-related issues, bringing together and addressing the priorities of a variety of global stakeholders. In contrast, a purely technical, scientist-led process under UN auspices could potentially dilute the content on AI risks while also failing to reap the legitimating benefits of the UN's universalist structure.

- **A separate international body should continue producing annual assessments that narrowly focus on the risks of "advanced"[1] AI systems, primarily led by independent scientists.** The rapid technological change, potential scale of impacts, and intense scientific challenges of this topic call for a dedicated process which can operate more quickly and with more technical depth than the UN process. It would operate similarly to the UK-led report, but with greater global inclusion, drawing data from a wider range of sources and within a permanent institutional home. The UN could take this on, but attempting to lead both this report and the above report under a single organization risks compromising this report's speed, focus, and independence.

- **There are at least three plausible, if imperfect candidates to host the report dedicated to risks from advanced AI.** The network of AISIs is a logical successor to the UK-led effort, but it faces institutional uncertainties. The OECD has a strong track record of similar work, though it remains somewhat exclusive. The International Science Council brings less geopolitical baggage but has weaker funding structures.

Regardless of who leads, all of these organizations—and others—should be actively incorporated into a growing, global public conversation on the science of advanced AI risks.

- **The two reports should be carefully coordinated to enhance their complementarity without compromising their distinct advantages.** Some coordination would enable the UN to draw on the independent report's technical depth while helping it gain political legitimacy and influence. However, excessive entanglement could slow or compromise the independent report and erode the inclusivity of the UN process. Promising mechanisms include memoranda of understanding, mutual membership or observer status, jointly running events, presenting on intersecting areas of work, and sharing overlapping advisors, experts, or staff.

- **It may be necessary to continue the current UK-led process until other processes become established.** Any new process will take time to achieve stakeholder buy-in, negotiate key parameters, hire staff, build working processes, and produce outputs. The momentum and success of the UK-led process should not be squandered after the first edition is presented at France's AI Action Summit in February.

Introduction

International coordination will be a necessary part of addressing AI's global impacts and effective coordination demands a shared, scientifically rigorous understanding of AI risks. While it does not guarantee international cooperation, shared scientific understanding has been a necessary precondition for progress in other globally significant policy domains like climate change,[2] biodiversity and ecosystems,[3] and radiation risks.[4]

However, several challenges make this task deeply complex in the context of AI. Areas of scientific agreement and disagreement are in constant flux as the technological frontier evolves quickly and AI systems are deployed in new ways. Information about systems' current capabilities and impacts is not systematically collected or published. Moreover, much of this crucial data is proprietary and closely guarded by private AI companies due to competitive commercial interests. Attempting to make predictions about the evolution of the technology is even more fraught due to the unpredictable nature of AI advancements. International efforts will be required to systematically and frequently assess the impacts of AI. To ensure this assessment reflects the distinct contexts of countries around the world, it requires accuracy and legitimacy needed to serve as a springboard for international action.

In July, the Oxford Martin School's AI Governance Initiative and Carnegie Endowment for International Peace co-hosted an expert workshop—held under the Chatham House rule—to explore options for the future of international scientific assessments of AI's risks including efforts led by the UK and the UN. Based on insights from that workshop, this paper explores the full potential ambition of international efforts, recognizes tensions between different goals, and makes three recommendations for balancing the need for timely, scientifically grounded updates with global inclusiveness and international legitimacy. It also takes particular note of the urgency and difficulty of assessing emerging risks from the most advanced AI systems, which have developed rapidly in recent years.

The International Scientific Report on the Safety of Advanced AI

Building a shared, science-based international understanding of AI risks is not an entirely new goal. This was one motivation for the launch in 2018 of Global Partnership on AI (GPAI), a body comprised of twenty-eight countries and the EU. GPAI has carried out research on a range of AI issues, aiming to bridge theory and practice.[5] However, the release of ChatGPT in late 2022 created new urgency and brought particular attention to the impacts of advanced AI systems, providing momentum for a dedicated effort.

At the AI Safety Summit held in Bletchley Park last November, the UK government, with support from twenty-eight countries and the European Union, commissioned a "state of the science" report on risks from advanced AI systems. Led by globally recognized AI expert Professor Yoshua Bengio, this initiative aimed to "facilitate a shared science-based understanding of the risks associated with frontier AI." An interim version of the report, the *International Scientific Report on the Safety of Advanced AI*, was presented at the AI Seoul Summit in May 2024, and the final version is expected before the French government's AI Action Summit in February 2025. The report primarily synthesizes academic research on AI safety, focusing on three key areas: 1) the expanding capabilities of advanced AI systems, which are a primary driver of risk; 2) the current state of technical abilities to evaluate models and mitigate risks; and 3) specific risk categories, including malicious use and systemic impacts.

The interim report represents, so far, the most significant step toward a global understanding of risks from advanced AI. It provided a sober yet prudent overview of key risks, acknowledging major areas of uncertainty and debate—no small feat, given the limited time and resources available. Importantly, the report process drew upon a diverse group of writers, representing various institutions, geographic regions, and areas of expertise within AI and related fields. An international expert advisory panel composed of representatives from thirty nations, as well as from the EU and the UN, had the opportunity to provide feedback (and dissenting views) during the writing process. Another group of senior advisors from academia, civil society, industry, and government bodies also contributed to the process.

However, the process had several limitations. First, the scope was kept intentionally narrow, focusing solely on "advanced" and "general-purpose" AI. These types of systems come with distinct risks that merit special attention, but they nevertheless comprise only a small slice of AI—omitting tools like recommendation algorithms and facial recognition systems that are already having large impacts on societies. Second, the report focused only on risks, even as many countries and companies care just as much or more about the potential benefits of AI.

The Irish representative to the expert panel published a brief dissent that "noted concern that the general tone of the report is excessively negative." Third, although the report incorporated a range of international perspectives, the process was not wholly inclusive. Only a few dozen countries were involved, largely high-income, and the UK government unilaterally established the procedures by which the report is currently being produced, granting full editorial control to the chair. Fourth, the report relied solely on "high-quality" published sources, with no mechanism to incorporate classified or proprietary data, or to receive and consider views from the public.

Limitations of this sort represent inevitable compromises, particularly for such a new and urgent endeavor. In fact, the report's one-year initial mandate is an implicit acknowledgment that future iterations might need to look different. Ultimately, the world will need a more enduring and institutionalized set of processes to promote common scientific understandings of AI risks.

An Emerging UN Process

Recent drafts of the Global Digital Compact (GDC) have proposed the establishment of an International Scientific Panel (ISP) on AI within the UN that would be tasked with producing annual scientific assessments of AI risks. While the compact is still under negotiation, early signs suggest the proposal for the International Scientific Panel will be included when the compact is finalized and adopted by member states for Summit of the Future in September 2024. This is in line with a similar recommendation from the interim report of the UN's High-Level Advisory Body on AI.

A UN process would be an incredibly significant development in this space, given the UN's unique position in world politics and its wide membership. The key question for this paper is therefore not *whether* the UN should do something, but *what* it should do, and how other actors can complement and enhance its work. For example, the UK-led report, or any direct successor, would need to be redefined with respect to the UN's effort.

A Task Too Ambitious for Any One Organization

An ideal approach would entail an annual (or more frequent) scientifically rigorous assessment that covers the full spectrum of pertinent issues. It would draw upon current data and diverse global perspectives, bringing together academics, policymakers, civil society, and members of industry—ultimately achieving buy-in from global policymakers. It would aim to be policy-relevant, in the sense that its content speaks directly to and is relied upon by policymakers on the major issues they face. Yet it should remain policy-neutral, not prescribing actions or solutions, to retain scientific credibility and avoid becoming entangled in global politics.[6]

Individually, these are challenging goals. Collectively, they are daunting. Inherent tensions among them would make it nearly impossible for any single organization to accomplish all these ambitions in a single process.

First, the basic decision of scope presents an immediate dilemma. Keeping the scope narrow could represent an implicit prioritization of advanced AI risks on the global stage, frustrating countries who are more concerned with the economic and political consequences of failing to take advantage and ownership of the technology. Such countries may comprise much of the world—including low-income countries, but also major powers like China and France who aspire to grow their AI ecosystems. Without the support of these countries, the report's findings may not translate into robust global action. However, widening the scope of any report to help build broader buy-in creates problems of its own. It would make it harder to move quickly and would offer more surface area for disagreement to arise and derail the process—for example, if countries like the United Kingdom and the United States become concerned that discussions of risk are being diluted.

Then there are the technical challenges. Staying abreast of the latest developments in the AI world would prove difficult even for an assessment narrowly focused on risks from advanced AI systems, let alone AI in general. Painting an up-to-date picture of the risks would involve a significant effort to draw on a wider range of sources than traditional science-based official reports, which have typically relied on peer-reviewed literature and data from member states.[7] The traditional approach works less well for studying advanced AI because the scientific landscape is so fast-moving that rigorous, groundbreaking research is often published outside of the peer review process to optimize for impact. The UK-led report explains that "not all sources used for this report are peer-reviewed" for exactly this reason. The private sector's especially large role in AI's scientific community further complicates things, as much of the most up-to-date data on the frontier of AI capabilities and on the use and impacts of the systems is proprietary.[8]

Keeping the report current would therefore require pulling on diverse sources of data, notably from academia and the private sector, but also governments and civil society around the world. Doing so would entail navigating complex relationships with companies where conflicts of interest, legal concerns, and competition could color the data they choose to share. It will also require rigorously vetting the sources of data and information that are used in general, if academic peer-review cannot be fully relied upon, and it could include drawing on inputs in many different languages and formats, posing a practical challenge. Taken together these challenges would make tracking the rapidly evolving scientific understanding of the risks and writing yearly scientific reports an intensive process.

Addressing these technical challenges might require empowering a sizable group of highly competent scientists to act independently. However, that approach could run at cross-purposes with the ultimate goal of having global policymakers acknowledge the assessment's findings as a legitimate basis for coordinating international action. Scientific rigor, by itself, is insufficient to achieve political buy-in and perceived legitimacy—there must also be meaningful representation of the variety of existing perspectives. Involving policymakers in the drafting process is a powerful tool for building engagement and acceptance of the findings, but of course, this can slow the process and potentially dilute the findings as political interests come into play. This is a particular risk in areas of great uncertainty, such as assigning likelihoods to future scenarios. The IPCC process, for instance, has been accused of making overly optimistic predictions about climate impacts in the pursuit of political compromise.

It is difficult to envision a single organization or institution, including the UN, that could accomplish these competing tasks and resolve these inherent tensions. However, two or more entities working together, or in parallel, could potentially complement each other by leaning into their respective comparative advantages. The next sections explore ways of doing this.

Recommendations

Recommendation 1: The UN Process Should Focus on Meaningfully Engaging Member States

In setting up the process and mandate for the International Scientific Panel on AI, we recommend the UN lean into its comparative advantage of bringing together global policymakers. It should aim to boost global buy-in through the co-creation of its report between scientists and policymakers in a manner similar to the IPCC, rather than having a completely independent scientific process. In addition to the findings of the report, the consensus-building process is politically valuable in and of itself, particularly given that actors involved in the report will intersect with other international fora, domestic policy processes, and AI development. Of course, a member state–driven process will be slow and less scientifically independent. But these gaps can be addressed with a separate, complementary process purpose-built for the task (see Recommendation 2). Trying to accomplish every goal with a single UN process risks a muddled outcome that falls short on all scores.

To bring member states together, we believe the scope of issues tackled by the UN process should be broader than risks from advanced AI systems, and could include a wide range of AI risks and benefits. What precisely is covered will ultimately be a compromise between member states, although we would suggest focusing on the issues that most require international coordination[9] and being mindful of topics already covered by international processes to avoid competing or duplicating. The UN could split the report into different working groups to ensure that the process remains relevant and valuable to all member states, regardless of their level of AI development. These different groups should form external partnerships with organizations best suited for each issue.

Drawing Inspiration From Existing Organizations

While we recommend aiming for a model like the IPCC, it is worth recognizing that in practice, this may be too ambitious. Similar efforts have since failed to gain the same traction as the IPCC, the IPCC itself has struggled with growing politicization, and intergovernmental institutions in general have been increasingly gridlocked. In designing the details of the report-writing process, the UN can draw inspiration from its multitude of existing and associated scientific bodies. We briefly outline the IPCC as well as a couple of other models for how member states can be engaged.

IPCC Model

The IPCC is perhaps the most famous example of such a process. Its reports feature a summary for policymakers (SPM) which requires line-by-line approval by government representatives from 195 member states. Appointed scientists collaborate in working groups on thematic areas (for example, physical science bases, impacts and adaptation, and mitigation), relying on peer-reviewed academic literature to produce the deeper scientific report upon which the SPMs are based. The high level of political buy-in facilitates direct incorporation of findings into national and international policies, but the whole process typically takes five to seven years and requires significant investment and expertise from participating states.

IPBES Model

Another possible model is the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES), which is structured similarly to the IPCC but provides more flexibility and opportunity for diverse input.[10] Member states participate in the initial scoping process, where the objectives, scope, and outline of the assessment are determined. While the SPMs are also approved by member states, the process is typically less detailed than the IPCC's line-by-line scrutiny. Working groups are formed and dissolved according to demand, rather than being fixed, and draw from diverse sources including indigenous and local knowledge (ILK) and "grey literature" such as government reports, nonprofit publications, and other non-peer-reviewed sources when they are deemed relevant and credible. IPBES also more proactively and formally seeks multistakeholder input throughout its processes than the IPCC. However, this increased flexibility came at the cost of consistency and policy relevance, and a 2019 review suggested it needed to do more to engage strategic stakeholders and feed into policy.

UNSCEAR Model

Finally, the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) represents a third model, characterized by lighter touch member-state involvement which we believe would work less well for AI. UNSCEAR studies the levels and effects of exposure to ionizing radiation through a committee of scientists appointed by the UN General Assembly (UNGA) and supported by a secretariat. Governments provide data, commission reports, and review drafts, but do not approve them. Nonetheless, the legitimacy of UNSCEAR's reports is reinforced by UNGA resolutions acknowledging the reports, and UNSCEAR is unique among the organizations we mention in reporting directly to the UNGA. However, this setup is in part possible because UNSCEAR deals with a relatively stable field of radiation science, explicitly stays away from economic benefits of the technology or safety standards, and has only thirty-one countries on its scientific committee—in contrast with the rapidly evolving field of AI, which is deeply enmeshed with economic and security issues and requires global attention.

Recommendation 2: An Annual, Independent Scientific Report on Advanced AI Risks Should Be Maintained

A UN process along the lines described in Recommendation 1 would be a hugely ambitious project. Even so, it would not be fully sufficient by itself. A deliberate and inclusive UN process that addresses a breadth of AI risks and benefits would, in particular, struggle to give adequate focus to risks from advanced AI—an especially challenging subject. A UN report would potentially take too long given the pace of scientific advancement, be too vulnerable to politicization, and be too broad to tackle the incredibly demanding scientific tasks. We therefore propose another process, intended to follow and build on the UK-led report, which would complement and support the UN work.

Like the current UK-led process, the new process on risks from advanced AI would produce reports at least yearly, with editorial control given to an international group of independent academics. Policymaker input would be limited—for example, endorsing the scope of questions to be considered. A high level of independence and technical competence would enable writers to make connections with a wide range of stakeholders, including industry, and vet different sources of information against a high scientific standard for inclusion in the report. It would also allow the report to tackle areas of greater scientific uncertainty, such as forecasting or painting scenarios of future developments, which could be more easily politicized if governments were directly involved in the drafting.

Unlike the current UK-led process, the new process would have a permanent institutional home, a long-term commitment of resources, and an enduring mandate. This probably means an international organization would need to serve as a secretariat, hosting and organizing the independent academics who would ultimately run the report—in short, playing the role the UK is playing for the current report. This report would continue to synthesize existing research rather than conduct new studies.

Importantly, this process and the resulting report would be designed to complement and feed into the UN's reports, ideally through a formal mechanism (see Recommendation 3). But it would retain an independent existence in another organization, and the published product would be accessible to the public and capable of standing on its own merits.

Choosing a Host

The ideal host would meet four criteria:

1. **Competence**—in the form of resources, connections, expertise, and experience.
2. **Independence**—drawing insights from multiple governments and private companies without becoming captured by any group or perspective.
3. **Robustness**—operating without distortion by any political and commercial disputes involving the host organization or its members or funders.
4. **Global inclusiveness**—drawing participation and support from a broad and diverse group of countries, and as an absolute minimum, the countries most involved in developing and deploying advanced AI systems.

Below we evaluate a few prominent candidates based on these criteria: the network of AISIs, the OECD, or an independent scientific organization like the International Science Council (ISC). To be sure, all of these organizations—and many more—can and should make contributions to global understanding of advanced AI risks. But there is merit in identifying one organization that could exercise primary leadership by collecting insights in a single place, while the others contribute in complementary ways.

Overall, there is no single perfect candidate, because even a process focused entirely on the risks of advanced AI cannot escape all the basic tensions described in this paper. Table 1 provides a summary of our evaluations, and the Appendix gives more detail. Making such evaluations is a deeply nuanced task, and these evaluations should be taken only as initial indications to guide further, more detailed analysis.

We do not recommend the UN itself as the host for this independent scientific report on advanced AI risks, although it could conceivably do so with an approach similar to the UNSCEAR example outlined above. It appears unlikely that the UN could achieve the same level of technical depth on risks from advanced AI systems as a dedicated external partner could. The need to cover a wider range of AI-related issues could spread its resources thinner, and it lacks as deeply knowledgeable a secretariat as the UK AISI—which was a significant boon to the current report, and would be difficult to replicate within the UN system. Having the UN run the independent report as well as a more political track could also compromise its independence (as explored in the discussion of Recommendation 3). Finally, having the UN focus on engaging member states (as in Recommendation 1) while an independent body produces the scientific report allows each process to lean into its comparative advantages. This approach captures more value than having the UN attempt to do both. That said, if only one process were possible, a UN-led scientific report would likely be the most viable option.

Table 1. Potential Hosts for an Advanced AI Safety Report

	Competent	Independent	Robust	Globally Inclusive
AISI Network	Uncertain	Mixed	Strong	Weak
OECD	Strong	Mixed	Mixed	Mixed
International Science Council	Uncertain	Strong	Mixed	Strong

AISI Network

Because the UK AISI successfully hosted the initial report, it makes sense to consider the emerging network of AISIs as a host for subsequent reports. Currently, ten countries and the EU have committed to establishing AISIs,[11] and recently announced their intention to form an international AISI network. It is currently unclear how formal or structured this network will be. Relying on AISIs would maintain the role of experts with intimate and cutting-edge knowledge. Transitioning to a network approach would help to share the resource burden and provide greater direct involvement for a larger number of countries than before.

However, there are significant uncertainties. AISIs themselves are young organizations that vary in size and function, making their potential for collaboration uncertain.[12] It's possible that some AISIs could face resourcing or domestic political constraints that either prevent them from contributing to this work, or cause their participation to cannibalize resources from other core tasks.

Also, the countries that currently have AISIs are a small group and geopolitically friendly. It remains to be seen whether China, for example, forms an AISI or equivalent organization—and whether and how a Chinese AISI would be welcomed into the existing AISI network. Given China's role in developing cutting edge AI systems, its inclusion would be a crucial component of the report's success.

More generally, most countries will have no AISI for the foreseeable future; in fact, the AISI network is currently smaller than the group of countries that joined at Bletchley to launch the UK-led scientific assessment process. An AISI-driven approach would therefore need to offer significant opportunities for broad international input to be seen as globally legitimate. This could include efforts to fund and establish AISIs in more countries, or regional AISIs in less-resourced contexts.

OECD

The OECD is arguably a more reliable choice. It has extensive substantive experience in AI.[13] Perhaps even more important, it has a demonstrated record of producing yearly, technical, expert-led reports tracking global issues, such as its "Economic Outlook" reports. In some cases, these reports have involved formal partnership with the UN.[14]

The main weakness of the OECD is limited global inclusiveness. As a group of thirty-eight generally wealthy and Western-aligned nations, it is more exclusive than the UN (though broader than the AISI network). In particular, it is unlikely that China could be organically included as an equal partner, although the OECD has mechanisms for outreach beyond its members: its "Key Partner" status has been allotted to Brazil, China, India, Indonesia, and South Africa; and GPAI, which it recently absorbed, boasts broader inclusion than the OECD, with India as its current chair. However, attempts to further internationalize OECD work have not always been successful because of its limited membership and reputation as a club of rich nations.[15] Also, as an intergovernmental body generally operating on consensus from all thirty-eight member-states, the OECD remains vulnerable to political fractures—less so than the UN, but perhaps more so than the AISI network.

International Science Council

Finally, an entity like the International Science Council (ISC)—an international non-governmental organization dedicated to coordinating many national and international scientific bodies—could represent a path away from intergovernmental processes and their geopolitical baggage. The ISC already benefits from deep connections with the UN, for which it serves as a formal convener of scientific expertise, as well as being a key UN partner on narrower issues, like its Scientific Committee on Antarctic Research. The potential for its contribution here, though, would depend in part on its funding. As ISC funding often comes on a per-project basis, the reliability and independence of the ISC's efforts would be to some extent contingent on funding structures. It's also not clear that the ISC could consult with industry from a position of strength relative to organizations with deeper governmental ties.

Ultimately, these three organizations will make their own choices about whether to take on any role in leading the world toward shared understanding of risks from advanced AI. However, partnership with the UN (whether formal or informal) and recognition from key stakeholders could be a decisive factor in the global political impact of any scientific report. As explained below, the goal would be to marry the political imprimatur of the UN with the independence and technical strengths of a complementary institution. The advanced AI risk report could also be enhanced by presenting some form of continuity with the UK's current report, given the strong overlap in form and function.

Regardless of the host, implementing this recommendation will require careful planning, and the momentum of the current report should not be dropped after France's AI Action Summit in February. It may be necessary, in the interim, to maintain the current process until another is able to take over.

Relying on Data From Industry

As emphasized above, much of the data on the most advanced AI systems is proprietary. This data could include results of model evaluations, anonymized data on model capabilities, deployment statistics, observed impacts, and more qualitative elements (like those included in the G7's Hiroshima reporting framework). While all the actors listed above could establish some form of working relationship with industry in the context of preparing this report, there is also a need for a more general and established mechanism to relay information from industry to the scientific community. This could ensure data is shared more broadly, and offer more consistency in how industry shares data.

A potential solution would be to establish trusted intermediaries to play this role. One workshop participant noted there is currently meaningful interest from industry in proactively engaging in such processes. Precedents exist in other issue areas for trusted intermediaries being established both in the public sector and as nonprofit industry coalitions.[16] Equivalent organizations have been proposed and explored for risks from advanced AI systems, with proposals for how industry actors might be incentivized to provide accurate information, such as through reporting requirements and safe harbor regulations. It will be crucial to establish guidelines for how information will be used to maintain trust if these intermediaries are to share the data publicly, and intermediaries could filter, aggregate, and anonymize sensitive data before publication to preserve confidentiality.

AISIs appear well-placed to play a role (although those with regulatory authorities, like the EU's AI Office, may have different relationships with industry than others). This paper does not explore this in more detail but underscores the importance of such a mechanism as a source of data.

Recommendation 3: These Two Reports Should Be Carefully Coordinated Without Creating Overdependence

For the two reports to complement one another effectively, the relationship between them must be defined carefully. The goal should be to preserve the comparative advantages of each report while enabling them to be mutually supportive. This requires a careful balance between coordinating the reports and keeping them distinct.

On the one hand, tight coordination would offer benefits to both efforts. The UN's work would be able to draw from the depth of the independent scientific report, while the latter would gain political legitimacy and policymaker influence from a strong relationship with the UN. However, excessive entanglement between the two could undermine their distinct strengths, allowing challenges in one to affect the other. For example, tying publication of the independent report to the results of a UN process would create risks of slowdowns, and pressure to harmonize findings or recommendations could affect the independence of the report's findings.

Conversely, the UN process should also avoid overreliance on the independent report as a single source of input for its own analysis of risks from advanced AI, as this may cause some member states to feel excluded from the conversation and reject the end product. Rather, the UN should still seek to bring to bear the full extent of international perspectives it has access to. Of course, a well-executed independent scientific report on the risks of advanced AI would ideally be so compelling and fair that any reasonable UN process would give it great weight.

Mechanisms for Coordination

There are many models of mechanisms by which international organizations can coordinate complementary efforts of this kind. These mechanisms play both functional and signaling roles, bringing organizers and participants together while messaging to others the significance of the partnership. The UN and the organization leading the independent report could sign a memorandum of understanding (MoU), for example, through which the UN formally and publicly requests the yearly reports as inputs into its own work.[17] This can come with detailed agreement on how the external findings will be integrated into the UN's efforts and help establish clear scopes and boundaries for the partnership.

Through the MoU, the partners can also express a joint commitment to interface and contribute to each other's work. They can offer each other membership or observer status for direct cooperation and establish regular meetings. More organic mechanisms can also be used to increase the facetime of members of the different groups, including jointly running events and presenting on intersecting areas of work—the AI Summit series could provide useful hooks for doing so. In some cases, the scientists contributing to both reports are also likely to intersect, further helping align priorities and communicate key insights.

Finally, it is also important to note that an external report can influence the UN process in the absence of any formal coordination if it is of high quality, caliber, and relevance. This was the case with the Stern Review, a comprehensive analysis of the economic impacts of climate change commissioned by the UK government and published in 2006.[18] The review was widely acknowledged, and the UN's incorporation of it in broader processes proved mutually beneficial—enhancing the credibility and impact of both the Stern Review and the UN's own reports.

Conclusion

The rapid advancement of AI presents unprecedented policy challenges for the global community. As such, it is crucial to establish an enduring, scientifically grounded, and globally legitimate mechanism to assess and address these emerging challenges. This paper has outlined one path forward that balances the need for scientific rigor, political buy-in, and timely action.

Our recommendations for a two-track approach—combining a UN-led process with an independent scientific report—aim to leverage the strengths of various stakeholders while mitigating potential pitfalls. The UN's unique position and convening power can provide the necessary global legitimacy and political engagement, while an independent scientific track ensures the continued production of timely, in-depth analyses of advanced AI risks.

As efforts move forward, it is essential that they remain coordinated and mutually reinforcing. The formal mechanisms we've proposed in Recommendation 3 for aligning the UN process with external scientific reports will be crucial in ensuring they complement each other, rather than compete.

Ultimately, the goal of these recommendations is to structure a clear process for establishing shared understanding of AI risks and thereby facilitating coordinated international action. By building on the foundation laid by the *International Scientific Report on the Safety of Advanced AI* and expanding it into a more robust and inclusive global framework, we think this process can endure and serve as a fundamental pillar in the global AI architecture in the long term.

Appendix

This appendix provides further details for the assessments made in Recommendation 2. Making such evaluations is a deeply nuanced task, and these evaluations should be taken only as initial indications to guide further, more detailed analysis. The scores are also intended to be relative, rather than an absolute measure of an organizations' competency.

As a reminder, the criteria are:

Competence—in the form of resources, connections, expertise, and experience.

Independence—drawing insights from multiple governments and private companies without becoming captured by any group or perspective.

Robustness—operating without distortion by any political and commercial disputes involving the host organization or its members or funders.

Global inclusiveness—drawing participation and support from a broad and diverse group of countries, and as an absolute minimum, the countries most involved in developing and deploying advanced AI systems.

The Network of AISIs

Over the past year, a number of governments have established AI Safety Institutes (AISIs). While they come in many shapes and sizes, most are at least in part motivated by a desire to advance and promote scientific understanding of advanced AI systems and encourage the uptake of safe practices.[19] A group of ten countries and the EU have announced their intention to form a collaborative network of AISIs, which will be further explored at a U.S.-led summit in late 2024. If the network were to launch a report, the exact structure would depend on whether the network has a secretariat. If it does, the secretariat could collect funding from AISIs in the network and could play a similar role to the UK in the current report.

Criterion 1: Competent	**Uncertain.** AISIs are still young institutions and vary significantly in their resourcing, mandates, and organizational structures. The structure of the emerging network of AISIs is even less defined, making it difficult to judge whether they would be able to coordinate and support this report. However, the UK government's commitment to the first report could be an indication of its ability and willingness to support future iterations. AISIs will also likely collect vital knowledge and expertise on advanced AI systems–the British and American AISIs have had success attracting top technical talent. They will also have close connections to industry and the frontier of AI development, putting them in a position to contribute a uniquely grounded perspective to the report—although this may depend on the AISI.
Criterion 2: Independent	**Mixed.** Governments established AISIs deliberately to provide independent oversight of AI companies. However, as government-funded institutions, AISIs ultimately remain subject to domestic interests and political influences.
Criterion 3: Robust	**Strong.** The network of AISIs has relatively limited membership and remains largely focused on technical and scientific endeavors, making dramatic disagreements more unlikely. Nonetheless, some caution is required, as they remain government bodies.
Criterion 4: Globally Inclusive	**Weak**. AISIs lack widespread global political legitimacy as they represent a small, primarily wealthy group of countries. To promote broader representation over time, arrangements could be made to include representatives from countries without AISIs, and/or regional AISIs could be established to allow governments to pool resources.

The OECD (or Other Intergovernmental Organization)

The OECD has a history of engaging on AI, having produced the influential OECD AI principles and established an AI policy observatory that tracks how the principles are put into practice. The OECD already has a stake in international governance of advanced AI through its work piloting a reporting framework for frontier AI companies' adherence to the G7 Hiroshima AI Process voluntary Code of Conduct.

Criterion 1: Competent	**Strong.** The OECD has a long track record of producing yearly, technical, expert-led reports tracking global issues, such as its yearly "Economic Outlook" reports. As outlined above, it already has experience with the safety of advanced AI. The OECD.AI Network of Experts includes, among policymakers, academics, and representatives from civil society, many experts from industry—although the extent to which this network can be substantially engaged is unclear. It is also piloting a reporting framework, working with leading AI labs.
Criterion 2: Independent	**Mixed.** The OECD demonstrates a capacity to balance industry and other perspectives through its expert working groups. However, as an intergovernmental organization comprising thirty-eight member states, the OECD's ultimate decisionmaking process requires consensus among these governments. This limits its ability to act independently of collective government interests.
Criterion 3: Robust	**Mixed.** While the OECD has wider membership than the AISIs, it is still significantly narrower than the UN's. Those who are members tend to be more aligned, lowering the risk of paralysis. Nonetheless, the report would remain at the mercy of an intergovernmental process, and fault lines in approaches to advanced AI have already started to form between OECD member states.
Criterion 4: Globally Inclusive	**Mixed.** Because of its limited membership, the OECD's attempts to internationalize its work have not always been successful (for example, in recent negotiations on a global tax treaty). However, its AI work streams recently absorbed the Global Partnership on AI, which boasts broader membership, and has indicated it may admit more countries to this effort. Still, neither the OECD nor GPAI include China as a member, and it is unclear whether China would be invited to join this work on an equal footing.

The International Science Council (or Similar Organization)

Another approach would be to have an independent scientific organization produce the report. This section focuses on the International Science Council, but another or new organization could also be considered. The ISC collects scientific organizations and promotes science as a public good, helping coordinate its members and running its own work. Its members include many private, nonprofit research organizations with histories of independently informing governments, such as the National Academies in the United States. It serves as a formal convener of scientific expertise for the UN as well as being a key UN partner on narrower issues, like its Scientific Committee on Antarctic Research.

Criterion 1: Competent	**Uncertain.** The ISC would likely need additional external funding to take on a project of this magnitude, and so its success would depend on whether countries, international organizations, or philanthropies could step in to fund it. The ISC collects organizations with deep AI expertise, but it has only recently been ramping up its work on AI, and it is not clear how it would best take on this work. It is also unclear whether the ISC could engage with industry from as strong a position as government bodies.
Criterion 2: Independent	**Strong.** As a nongovernmental organization, the ISC would be the most independent of the organizations we consider. However, since funding tends to be on a project-by-project basis, the ISC's work may still be sensitive to its funders. Diversifying funding between governments and philanthropies could assuage this.
Criterion 3: Robust	**Mixed.** Again, this would largely depend on its sources of funding, and diversification could help prevent disagreements between funders and other funders, or funders and report writers from putting the project at risk.
Criterion 4: Globally Inclusive	**Strong.** The ISC has a history of collaboration with the UN, as well as wide and geographically diverse membership.

Example: The Scientific Committee on Antarctic Research

The Scientific Committee on Antarctic Research (SCAR) represents a useful example of how the ISC could support large independent scientific endeavors. SCAR's membership is made up of countries with an interest in Antarctic research. Members fund the organization, and their delegates set its main directions. SCAR also has an option for nonvoting associate membership for countries or organizations with a significant interest in Antarctic research, but less-developed research programs. SCAR's work is then driven by scientists and researchers from member countries. Engagement is primarily through scientific contributions and research collaboration rather than direct governmental negotiation or policymaking.

SCAR's main goal is to facilitate international scientific research in Antarctica. It focuses on promoting and coordinating high-quality research and providing scientific advice on Antarctic issues. While it does provide scientific advice to the Antarctic Treaty System, its primary role is not to influence global policy directly but to support and inform scientific understanding and collaboration.

This structure could allow for quicker adaptation to fast-moving AI developments, but it risks reducing political commitment to outcomes and potentially creating a disconnect between scientific recommendations and policy implementation. For example, while SCAR's scientific findings on climate change in Antarctica have been robust, the translation of these findings into binding international policy has often been slow, as seen in the ongoing negotiations over Marine Protected Areas in the Southern Ocean since 2002.

Structure:

- Executive Committee: Composed of elected scientists who manage the organization.
- Standing Committees: Focus on specific scientific areas such as geosciences, life sciences, and physical sciences.
- National Committees: Coordinate activities at the national level. Countries can join as full or associate members. National representatives from the forty-six member states participate in the biennial meetings where they discuss and plan SCAR's scientific agenda and elect the Executive Committee.

About the Authors

Hadrien Pouget and Claire Dennis contributed equally to the paper and should be jointly cited in references. Name order was randomized. If possible, please cite as follows: Pouget, H., Dennis, C., et al. (2024) 'The Future of International Scientific Assessments of AI's Risks,' Carnegie Endowment for International Peace. Jon Bateman and Robert Trager were core contributors and instrumental in convening the July workshop.

Given the large number of authors, authorship does not imply agreement with every point made in this paper.

Hadrien Pouget is an associate fellow in the Technology and International Affairs Program at the Carnegie Endowment for International Peace.

Claire Dennis is a Research Scholar at the Centre for the Governance of AI, affiliate of the Oxford Martin School AI Governance Initiative, and Senior AI Adviser to the UN Center for Policy Research. Her background in diplomacy informs her work on global AI governance strategies across academic, policy, and international contexts.

Jon Bateman is co-director of the Technology and International Affairs Program at the Carnegie Endowment for International Peace, where he focuses on global technology challenges at the intersection of national security, economics, politics, and society.

Robert F. Trager is Co-Director of the Oxford Martin AI Governance Initiative, International Governance Lead at the Centre for the Governance of AI, and Senior Research Fellow at the Blavatnik School of Government at the University of Oxford. He is a recognized expert in the international governance of emerging technologies and regularly advises government and industry leaders on these topics.

Renan Araujo is a research manager at the Institute for AI Policy and Strategy. He is a Brazilian lawyer with a background in comparative policy, and previously worked with AI governance at Rethink Priorities and the Institute for Law and AI.

Belinda Cleeland is the Director of Policy at the Simon Institute for Longterm Governance. She leads the development of SI's policy recommendations, working to synthesize research into concrete applications for the multilateral system to ensure the safe development of transformative technologies.

Malou Estier is a Law and Policy Associate at the Simon Institute for Longterm Governance. Her work focuses on the multilateral system, analysing UN processes and researching new and existing institutional mechanisms for safe and beneficial AI governance.

Gideon Futerman is an existential risk researcher who has primarily worked on the interaction of solar geoengineering and climate change and global catastrophic risk, as well as exploring the role of pluralism in existential risk studies. He is a student in Earth Sciences at the University of Oxford.

Oliver Guest is a research analyst at the Institute for AI Policy and Strategy. His research focuses on the international governance of advanced AI.

Carlos Ignacio Gutierrez is an AI policy researcher with a Ph.D. and M.Phil. in Policy Analysis from Pardee RAND Graduate School. He has contributed to national and international AI policy formation, including work on the EU AI Act and the NIST AI Risk Management Framework, and has led partnerships with stakeholders like the UN and IEEE.

Vishnu Kannan is an incoming student at Stanford Law School. He was the advisor to the president at the Carnegie Endowment for International Peace, where he led executive office strategic initiatives and the president's research team.

Matthijs Maas is a Senior Research Fellow at the Institute for Law & AI, and an associate fellow with the Leverhulme Centre for the Future of Intelligence.

Casey Mahoney is an associate political scientist at RAND. He conducts policy research on international AI governance and the geopolitics of emerging technology. Mahoney holds a Ph.D. in political science.

Charles Martinet is a Research Affiliate at the Oxford Martin AI Governance Initiative and a Summer Fellow at the Centre for the Governance of AI. His work aims to deliver research-based and operational policy advice for the international and European governance of advanced AI.

Jakob Mökander is the Director of Science and Technology Policy at the Tony Blair Institute and a research fellow at Yale Digital Ethics Center. He holds a DPhil from Oxford Internet Institute and has been a visiting scholar at Princeton Center for Information Technology Policy.

Kwan Yee Ng is a Senior Program Manager at Concordia AI, a Beijing-based social enterprise focused on AI safety and governance. She was also one of the writers for the International Scientific Report on Advanced AI Safety.

Seán Ó hÉigeartaigh is Director of the AI: Futures and Responsibility Programme at the University of Cambridge. His work focuses on foresight, risk and governance relating to advanced AI systems.

Aidan Peppin is Policy and Responsible AI Lead at Cohere For AI, Cohere's non-profit research lab. Previously, he was a Research Lead at the Ada Lovelace Institute and program chair for the UK Fringe events to the Bletchley AI Safety Summit UK

Konrad Seifert is the co-CEO of the Simon Institute for Longterm Governance

Scott Singer is a Visiting Scholar at the Carnegie Endowment for International Peace, co-founder of the Oxford China Policy Lab, and affiliate of the Oxford Martin School AI Governance Initiative.

Maxime Stauffer is co-founder and -CEO of the Simon Institute for Longterm Governance. He advises governments and international organizations on the governance of emerging technologies, with a focus on extreme risk prevention. His background is in international relations and computational mathematics.

Caleb Withers is a research assistant for the Technology and National Security Program at the Center for a New American Security (CNAS). Before CNAS, he worked as a policy analyst for a variety of New Zealand government departments.

Marta Ziosi is a Postdoctoral researcher at the Oxford Martin AI Governance Initiative, University of Oxford. She is also the Head of AI for People, a non-profit organisation whose mission is to learn, pose questions, and take initiative on how AI can be used for social good.

Acknowledgments

The authors would also like to thank Adam Day, Owen Larter and Lucia Velasco for their valuable input and feedback.

Notes

[1] While definitions of "advanced" AI vary, this paper takes it to mean both general-purpose AI systems that exhibit strong performance across a range of tasks, and narrow systems that out-perform humans on specific tasks. Matthijs M. Maas, *Architectures of Global AI Governance: From Technological Change to Human Choice* (Oxford, UK: Oxford University Press, forthcoming), chapter 2.

[2] With the Intergovernmental Panel on Climate Change.

[3] With the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.

[4] With the UN Scientific Committee on the Effects of Atomic Radiation.

[5] GPAI was recently absorbed into the Organisation for Economic Co-operation and Development (OECD).

[6] Throughout this paper, we assume the international efforts we outline should focus on long-term trends and scientific understanding. This could include measured impacts of AI systems, their evolving capabilities, the efficacy of mitigations, and predictions for how each of these could evolve. For comparison, the 6th International Panel on Climate Change synthesis report has three sections which cover similar ground: "Current Status and Trends" which covers measured impacts and the effectiveness of current mitigatory efforts, "Long-Term Climate and Development Futures," and "Near-Term Responses in a Changing Climate." Notably, the reports we propose would not seek to predict or provide early warning of specific AI incidents.

[7] The IPCC requires literature to be published by scientific journals, for example, and a section of the UNSCEAR 2019 report explains "peer-reviewed studies published in the scientific literature and publications of relevant international organizations" were in the scope of their review.

[8] According to the 2024 AI Index Report, in 2023, fifty-one "notable" machine learning models were developed in industry, to academia's fifteen. The gap has been widening since 2016. In addition to the leading AI labs, much of the scientific work on AI is now done by an emerging ecosystem of auditors, and the impacts are most visible to the many companies using and deploying the systems around the world.

[9] The AI-related issues countries face vary in the amount of international coordination needed to address them. No country alone can counter the malicious use of powerful AI systems or shrink the gap in global access to the benefits of AI. However, other issues might benefit from international coordination but not require it to the same extent. Handling algorithmic discrimination, for example, depends intimately on social and legal conceptions of the problem that will differ between countries and cultures.

[10] The Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services is not strictly a UN organization, but it is closely associated and receives secretariat support from the UN Environmental Programme.

[11] The countries that have committed to establishing AI Safety Institutes include: Australia, Canada, the European Union, France, Germany, Italy, Japan, the Republic of Korea, the Republic of Singapore, the United States of America and the United Kingdom.

[12] The U.S. and UK AISIs are focused chiefly on research and development, including conducting model evaluations on frontier models, while the EU AI Office is primarily a regulator, tasked with implementing the EU AI Act. The U.S. AISI also has a budget of $10 million, while the UK AISI's is £100 million.

[13] The OECD has led the development of AI principles, the AI policy observatory that tracks how the principles are put into practice, and the pilot reporting framework for frontier AI companies' adherence to the G7 Hiroshima AI Process voluntary Code of Conduct.

[14] The OECD and FAO (Food and Agriculture Organization of the UN), for example, have collaborated on a yearly report for the last twenty years.

[15] The OECD has historically led discussions on global tax, but its limited membership has caused tensions, and the UN has recently started taking a (larger) role in response.

[16] For government actors, see, for example, the activities of the U.S. Cybersecurity and Infrastructure Security Agency. For non-profit industry coalitions, see the Information Sharing and Analysis Centres for critical infrastructure, or the World Association of Nuclear Operators for nuclear power plants. In the EU, the Digital Services Act has adopted a model where government actors mediate requests for data from researchers to companies.

[17] Memorandums of understanding are sometimes called a "Letter of Intent" or "Joint Statement of Co-operation." Many examples can be found on the OECD's website such as MOUs for OECD-ILO, OECD-Asian Development Bank, and OECD-UNHCR.

[18] Stern, N. (2006) Stern Review: The Economics of Climate Change. Cambridge University Press, Cambridge.

[19] The UK AISI's goals include "test advanced AI systems," "foster collaboration to . . . mitigate risks," and "strengthen AI development practices." Similarly, the U.S. AISI's goals include "advancing the science, practice, and adoption of AI safety.""

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.

Oxford Martin AI Governance Initiative

The AI Governance Initiative is co-led by Robert Trager, a social scientist specialising in international relations and frontier AI regulation, and Michael Osborne, a specialist in machine learning. Housed in the Martin School of the University of Oxford, AIGI is one of the few centres in the world focused on the governance of AI from both technical and policy perspectives. The initiative aims to anticipate and mitigate lasting risks from AI through (1) impactful research that is rigorously grounded in the social and computational sciences, (2) decision-maker education campaigns, and (3) training the next generations of technology governance leaders.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org