AN INTERNATIONAL CONSORTIUM FOR AI RISK EVALUATIONS

Ross Gruetzemacher[1,2,3], Alan Chan[4], Štěpán Los[2,5], Kevin Frazier[6,7], Siméon Campos[8], Matija Franklin[9], James Fox[10], José Hernández-Orallo[3,11,12], Christy Manning[1,2], Philip Tomei[13], and Kyle Kilian[2,14]

[1]Wichita State University [2]Transformative Futures Institute [3]Centre for the Study of Existential Risk, University of Cambridge [4]Mila [5]University of St Andrews [6]St. Thomas University [7]Legal Priorities Project [8]SaferAI [9]University College London [10]University of Oxford [11]Universitat Politècnica de València [12]Leverhulme Centre for the Future of Intelligence, University of Cambridge [13]Pax Machina [14]Center for the Future Mind, Florida Atlantic University

ABSTRACT

Given rapid progress in AI and potential risks from next-generation frontier AI systems, the urgency to create and implement AI governance and regulatory schemes is apparent. A regulatory gap has permitted labs to conduct research, development, and deployment with minimal oversight or guidance. In response, frontier AI evaluations have been proposed as a way of assessing risks from the development and deployment of frontier AI systems. Yet, the budding AI risk evaluation ecosystem faces significant present and future coordination challenges, such as a limited diversity of evaluators, suboptimal allocation of effort, and races to the bottom. As a solution, this paper proposes an international consortium for AI risk evaluations, comprising both AI developers and third-party AI risk evaluators. Such a consortium could play a critical role in international efforts to mitigate societal-scale risks from advanced AI. In this paper, we discuss the current evaluation ecosystem and its problems, introduce the proposed consortium, review existing organizations performing similar functions in other domains, and, finally, we recommend concrete steps toward establishing the proposed consortium.

1 INTRODUCTION

Existing AI and next-generation frontier AI systems pose many serious societal-scale risks (Critch and Russell 2023; Hendrycks et al. 2023; Gutierrez et al., 2023). In response, scholars and industry leaders have discussed visions for AI governance and regulatory regimes (Anderljung et al, 2023; Ho et al. 2023; Trager et al. 2023; Suleyman 2023) and have identified functions needed to evaluate and respond to extreme risks from frontier AI systems [1] (Shevlane et al. 2023; Anthropic 2023).

Continuous and thorough risk evaluations of frontier AI deserves significant attention from all stakeholders intent on effectively regulating it. This is because current approaches to building general-purpose AI have tended to produce increasingly large models with surprising and unforeseen capabilities (Ganguli et al., 2022; Wei et al. 2022). Risk evaluations of these systems can contribute to mitigating societal-scale risk from advanced AI (Shevlane et al. 2023), facilitating effective AI governance (Anderljung et al. 2023), and developing a regulatory framework for approving model training/deployment (Avin 2023). Regulation for risk evaluations is crucially time-sensitive as capabilities research continues at a rapid pace, quickly nearing critical thresholds[2] (Anthropic 2023). This paper proposes establishing an organization tasked with 1) coordinating risk evaluations of frontier AI and 2) producing standards intended to mitigate any unanticipated risks from it.

[1]The importance of testing/evaluation of AI is becoming more accepted in various organizations: the US military is creating robust protocols that even recognize the crucial role of evaluating frontier AI (National Academies of Sciences 2023); CSIS emphasizes evaluation/testing in developing trustworthy AI (Chin 2023).

[2]In a US Senate hearing, Anthropic's CEO warned that AI will precipitate grave terrorist threats from biological and chemical weapons from 2024 to 2026 (US Senate 2023); see Boiko et al. (2023) for context.

A consortium of AI risk evaluators is needed due to limitations of the status quo between third-party evaluators, AI labs, regulators, and other stakeholders. These limitations are most simply explained as a coordination problem, for which a consortium is a solution. Thus, the proposed institution will complement rather than replace other proposed regulatory efforts. We illustrate this coordination problem in Figure 1, and we elaborate on the need for a consortium in the following section.

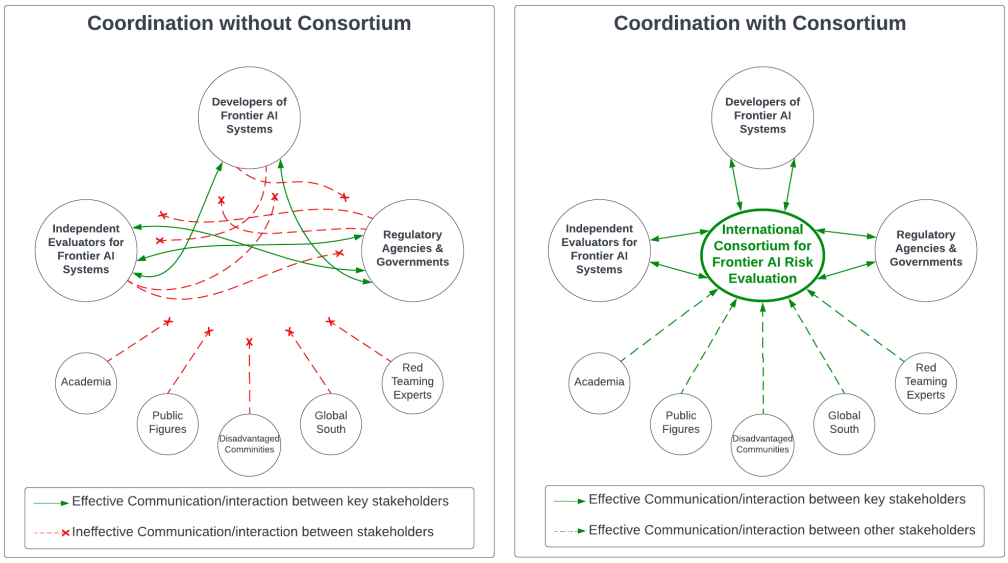

Figure 1: **(Left)** This illustrates the complexity of communication between large numbers of stakeholders that would result by default, without the creation of an intermediary. Many critical elements of a frontier AI regulatory framework may struggle to function effectively without optimal coordination between the three core groups of stakeholders involved. **(Right)** The complex and unmanageable coordination amongst the stakeholders of a frontier AI regulatory framework can be simplified and made tractable by the introduction of an international consortium for AI risk evaluations.

The remainder of the paper proceeds by first outlining the challenges posed by the status quo regarding the mitigation of societal-scale risks by system evaluators. We then identify functions that could address these challenges and review extant organizations with similar functions to draw upon for structuring and organizing the consortium. Finally, we outline a plan of action for this initiative.

2 The Need for a Consortium

2.1 The Current Ecosystem

Previous work on AI risk evaluation has focused on various risks, ranging from existing harms (Liang et al. 2022; Weidinger et al. 2022) to more **extreme AI risks** associated with dangerous capabilities of frontier AI (Shevlane et al. 2023). Additionally, there are structural risks involving the interactions of AI systems with powerful civil, political, or economic forces (Zwetsloot and Dafoe 2019). Here, we focus on **societal-scale AI risks**, which we define as AI-powered risks to large-scale systems (e.g., global supply chains, financial systems, geopolitical stability), nations, or other large groups if the negative outcomes (e.g., hot war, human rights violations, economic risks) are sufficiently widespread[3]. Thus, such risks include far-reaching existing harms, extreme AI risks, and structural risks; however, risk evaluators working with the frontier AI labs developing the highest-risk systems appear focused on extreme risks. National governments have come forward with risk

[3]While many AI risks constitute societal-scale risks, many do not. Consider risks from autonomous vehicles—these would not constitute societal-scale risks because the negative outcomes will not be sufficiently widespread. For example, malfunctioning autonomous vehicles will, at worst, pose risks equal to those of human drivers. Another example is AI-driven industrial control systems, where normal accidents can be expected, but the scope of the risks will be limited to those working directly with such systems. If these systems are widely used, other users will discontinue their use before the risk grows to constitute a societal-scale risk.

management frameworks for "mapping, measuring, and managing" risks from AI[4] (Tabassi, 2023; OECD, 2022), but they are intended only as roadmaps, and do not account specifically for frontier AI risks. We depict the relationship between extreme risks and societal-scale risks in Figure 2[5].

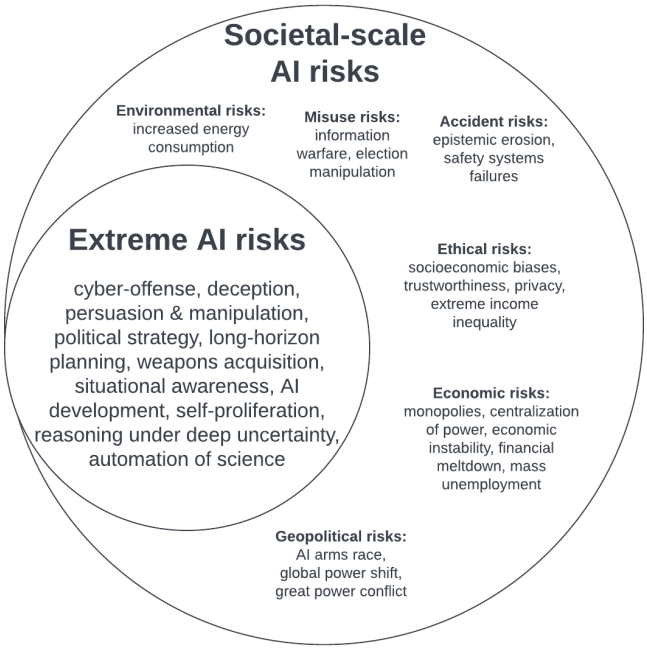

Figure 2: This figure depicts societal-scale AI risks as a superset of more extreme AI risks (Shevlane et al. 2023). Not all societal-scale AI risks are capabilities-related; some are structural risks while others include existing harms from AI systems. A full discussion is beyond the scope of this paper, but it is important to note that frontier AI evaluation should not be limited to extreme AI risks.

Evaluation of frontier AI systems[6] for societal-scale risks is critical to addressing risks from advanced AI (Anderljung et al. 2023). Indeed, there appears to be a consensus regarding the need for risk evaluations as essential to any comprehensive frontier AI regulatory framework (Engler 2023; Habuka 2023; Suleyman 2023; Wallin et al. 2023). Such frontier AI evaluations can be conducted by AI labs developing frontier AI systems, by regulatory authorities, or by independent third parties. Our discussion focuses primarily on third-party organizations devoted to evaluating frontier AI systems for societal-scale risks given the importance placed on such evaluations by frontier AI regulatory frameworks (Avin 2023). While it would be ideal to have a rich ecosystem of AI risk evaluators able to provide rigorous risk assessment across a broad spectrum of societal-scale risks, at present, the ecosystem for frontier AI evaluations for societal-scale risks is underdeveloped.

2.2 Problems with the Current Ecosystem

Limited Independence and Diversity of Evaluators: Many third-party organizations conducting AI risk evaluations have strong ties to the AI labs developing the systems being evaluated. These connections may lead evaluators to minimize some risks, like those involving economic or political concentrations of power (Chan et al. 2023), and may invite skepticism from regulatory agencies.[7]

As full independence is hard, diversity in socio-technical backgrounds and interests is also desirable. Nevertheless, it is difficult to get more diverse parties to engage in risk evaluations of frontier AI

[4]The NIST AI Risk Management Framework emphasizes the importance of model testing, evaluation, verification, and validation. France's LNS also emphasizes reliable standards and evaluations for AI systems.

[5]The first ten extreme risks in the figure were taken directly from Shevlane et al. (2023).

[6]We define frontier AI systems as highly capable AI systems, like foundation models (Bommasani et al. 2021), that push the boundaries dangerous state-of-the-art capabilities sufficient to pose societal-scale risks.

[7]Evaluation gaming involves evaluators optimizing for demonstrating safety by privileging specific evaluations, and, whether intentional or not, can diminish the validity of evaluation metrics (Fist et al., 2023).

systems. Although OpenAI has a public researcher access program, the admission process is opaque. Other frontier AI labs like Anthropic and DeepMind do not have such researcher access programs. This status quo privileges those evaluators who already have personal connections to the frontier labs. Moreover, a regulatory effort led exclusively by an unrepresentative set of technical experts is unlikely to receive the public's support nor earn and sustain its trust (Stern 2011).

Scalability of Evaluations: Given that increased scaling is likely to require increasingly more rigorous evaluations as frontier AI capabilities grow (Anthropic 2023), and that the latest computing hardware will make scaling cheaper (Hobbhahn and Besiroglu 2023), regulation will need to be able to keep up with the speed of the technology driving progress and of the businesses seeking to capitalize on it. Consider that GPT-4 (OpenAI 2023) was trained with 2020 technology (Patel and Wong 2023), and that the current state-of-the-art hardware only became available on the cloud—and presumably to many others—at the end of summer 2023 (Shah 2023). Moreover, within 18 months some organizations will have the computational resources to train systems two orders of magnitude larger[8] than GPT-4 (Harris and Suleyman 2023). Thus, it is likely that soon more actors will be able to train dangerous frontier systems that require more intensive risk evaluations (Anthropic 2023).

Another issue is that the demand for talent in evaluation organizations will be in competition with the demand for talent in the labs developing frontier systems. As a result, third-party evaluators and regulators will need to provide competitive compensation with AI labs, but there will likely be limits to what the public sector is able to support[9] (Clark and Hadfield 2021).

Suboptimal Allocation of Effort: The status quo may make it difficult to allocate evaluation effort so as to maximally reduce societal-scale AI risks. As the number of model evaluators grows, there is a risk of unnecessary duplication of work. Given that evaluating societal-scale risks from frontier AI systems is still an early science[10], verification of results will be important.

At the same time, research foci should be optimally spread amongst the various societal-scale AI risks. For example, it would be suboptimal if half of the frontier AI evaluation organizations were working on deception evaluations but none were working on other salient risks like concentration of power or long-horizon planning. Therefore, an effort needs to be made to coordinate evaluations to ensure that at least one organization is advancing the science of evaluating each salient risk[11] whilst guaranteeing the capacity required to evaluate all salient risks for all frontier systems requiring evaluation (e.g., of a certain size). Moreover, organizations should not be incentivized as to only focus on more tractable risks (e.g., because they are rewarded for making progress faster) as this would lead to the neglect of more speculative risks[12] and increase the possibility of 'unknown unknowns'.

Additionally, efforts by independent AI labs to create risk evaluation networks can be counterproductive to collective safety efforts by focusing interested volunteers on a single organization.[13] Efforts to increase participation in red teaming or evaluation of AI systems needs to be coordinated by an independent party so that efforts are not concentrated on a single lab.

Races to the Bottom: Races to the bottom are also a significant risk. Suppose there is a licensing scheme wherein AI systems must be evaluated by an accredited organization before deployment. In such a market, risk evaluators with lower standards get more customers as they are less demanding yet provide the same perceived value. Hence, the markets get captured by inferior evaluators.

Races to the bottom can also occur across countries, where countries that regulate less and are less risk-averse capture a greater share of the value of frontier AI or attract more companies. Since there is currently no authoritative voice on AI risk evaluations and no standards for their evaluation, countries can defer to the least demanding standards when developing their regulations.

Barriers to Knowledge Sharing: There is currently no public communication forum for frontier AI risk evaluations, resulting in two significant consequences. First, policymakers have no authoritative,

[8] i.e., foundation models (Bommasani et al. 2021) larger in the number of model parameters.

[9] A remedy could be tax exemptions (e.g., European Space Agency employees are exempt from income tax).

[10] Given the nascent state of research on frontier AI evaluations, much can still be learnt from reviewing successes and failures of fields like psychometrics, comparative and developmental psychology, the extant literature on AI testing and evaluations, or software testing and other areas of engineering and computer security.

[11] This requires incentivizing scientists to work on risks they are both motivated and suited to work on.

[12] Independent evaluators *should* be encouraged to allocate some resources to exploring speculative risks.

[13] OpenAI's Red Teaming Network is a potential example (https://openai.com/blog/red-teaming-network).

impartial voice to look to about evaluations when crafting legislation. Second, model evaluators lack an easy way to share information about research results, such as methodological problems or the presence of dangerous capabilities in a model, beyond personal channels, academic publication, or haphazard search through masses of preprints.

2.3 Objective and Scope of a Consortium

To address these challenges, we propose the creation of an **international consortium for AI risk evaluation**. This proposed consortium would coordinate frontier AI risk evaluations amongst the three core groups of stakeholders, and other stakeholders, providing responsive expert guidance to government regulators in response to evolving AI risks with increasingly large and capable systems.

Objectives of a Consortium: To address the challenges we discussed above, we identify three objectives for the proposed international consortium of third-party evaluators:

1. Act as an intermediary between third-party risk evaluators, frontier AI developers, governments and regulators, and other stakeholders (e.g., academics, civil liberties groups);
2. Provide a means to set and implement standards quickly while minimizing bureaucratic challenges;
3. Serve as an advisory body for regulators 1) in developing their own risk assessment capacity and 2) in verifying frontier AI evaluators' risk assessment abilities (via certification)

If implemented effectively to achieve these objectives, the proposed consortium would complement some of the previously proposed models (Anderljung et al. 2023; Avin 2023; Suleyman 2023) for governing frontier AI while offering an alternative approach to others (Ho et al. 2023; Traeger et al. 2023). In the case of the latter, the proposed consortium would provide a simpler and more easily implemented solution for the issues of standard-setting and coordinating evaluations by not requiring that all advanced AI governance functions be housed in a single organization.

With the UK Frontier AI Taskforce now signaling an intention to take an active role in the evaluations ecosystem, it is possible that the proposed consortium could come to meet all three of the designated objectives, or that it takes on a role complementary to other new organizations like this in achieving these objectives. Thus, and as mentioned in the previous subsection, there are many ways in which a consortium or international institution for model evaluators could be implemented[14].

3 Plan for Action

This brief paper[15] describes motivations for an international consortium for AI risk evaluations and objectives for the consortium. To quickly establish the consortium, we suggest the follow-up steps:

- Collect feedback from AI researchers, think tanks, and other potential stakeholders
- Continue research to better understand existing organizations with similar functions
- Conduct a workshop with experts and stakeholders on how to structure the consortium
- Incorporate feedback and workshop results to prepare a detailed plan of action for establishing and quickly scaling the consortium
- Solicit funding for the detailed plan of action
- Form the consortium

As we proceed with this plan of action, it will be prudent to remain cognizant of potential failure modes. Some examples include the proposed consortium being launched with the best intentions but ultimately gravitating away from the original intent and either 1) becoming a lobbying group or advocacy group, or, 2) becoming a standards-setting organization that lacks the levers necessary for enforcement. The former could occur if too much power is ceded to the independent AI risk evaluators as members, and the latter could occur if, like the Institute of Electrical and Electronic Engineers (IEEE), coordination with regulators is insufficient or ineffective.

[14] Appendix A discusses existing organizations that offer guidance for implementing the consortium.

[15] An extended version can be found at https://arxiv.org/abs/2310.14455.

4 REFERENCES

Avin. S. 2023. Frontier AI Regulation Blueprint. Blog. Centre for the Study of Existential Risk, University of Cambridge. https://www.cser.ac.uk/news/frontier-ai-regulation-blueprint/

Anthropic, 2023. Anthropic's Responsible Scaling Policy. https://www.anthropic.com/index/anthropics-responsible-scaling-policy

Anderljung, M., Barnhart, J., Leung, J., Korinek, A., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D. and Chang, B., 2023. Frontier AI Regulation: Managing Emerging Risks to Public Safety. arXiv preprint arXiv:2307.03718.

Boiko, D.A., MacKnight, R. and Gomes, G., 2023. Emergent autonomous scientific research capabilities of large language models. arXiv preprint arXiv:2304.05332.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E. and Brynjolfsson, E., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M. and Lin, M., 2023, June. Harms from Increasingly Agentic Algorithmic Systems. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 651-666).

Chin, C. 2023. Navigating the Risks of Artificial Intelligence on the Digital News Landscape. Blog. CSIS. https://www.csis.org/analysis/navigating-risks-artificial-intelligence-digital-news-landscape

Clark, J., and Gillian K.H. 2019. Regulatory markets for AI safety. arXiv preprint arXiv:2001.00078.

Critch, A. and Russell, S., 2023. TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI. arXiv preprint arXiv:2306.06924.

Engler, A., 2023. A comprehensive and distributed approach to AI regulation. Brookings Institute. https://www.brookings.edu/articles/a-comprehensive-and-distributed-approach-to-ai-regulation/

European Commission, 2023. Sectorial AI Testing and Experimentation Facilities under the Digital Europe Programme. https://digital-strategy.ec.europa.eu/en/activities/testing-and-experimentation-facilities

Fist, T., Depp, M., Withers, C., 2023. Response to OSTP "National Priorities for Artificial Intelligence Request for Information". Center for a New American Security. July 20, 2023. https://www.cnas.org/publications/commentary/ostp-national-priorities-for-artificial-intelligence

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A., Conerly, T., Dassarma, N., Drain, D., Elhage, N. and El Showk, S. 2022, June. Predictability and surprise in large generative models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1747-1764).

Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C. C., and Franklin, M. 2023. A proposal for a definition of general purpose artificial intelligence systems. Digital Society, 2(3), 36.

Maas, M.M. and Villalobos, J.J., 2023. International AI Institutions: A Literature Review of Models, Examples, and Proposals. AI Foundations Report, 1.

Marius H. and Tamay B. 2022, "Trends in GPU price-performance". Blog. https://epochai.org/blog/trends-in-gpu-price-performance

Habuka, H., 2023. The Path to Trustworthy AI: G7 Outcomes and Implications for Global AI Governance. Center for Strategic and International Studies. June 6, 2023. https://www.csis.org/analysis/path-trustworthy-ai-g7-outcomes-and-implications-global-ai-governance

Harris, S., Suleyman, M. 2023. Can We Contain Artificial Intelligence. Podcast 322. August, 2023. https://www.samharris.org/podcasts/making-sense-episodes/332-can-we-contain-artificial-intelligence

Hendrycks, D., Mantas M., and Thomas W. 2023. An Overview of Catastrophic AI Risks. arXiv preprint arXiv:2306.12001 (2023).

Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., Chowdhury, R., Dafoe, A., Hadfield, G., Levi, M. and Snidal, D., 2023. International Institutions for Advanced AI. arXiv preprint arXiv:2307.04699.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A. and Newman, B., 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

National Academies of Sciences, Engineering, and Medicine. 2023. Test and Evaluation Challenges in Artificial Intelligence-Enabled Systems for the Department of the Air Force. Washington, DC: The National Academies Press. https://doi.org/10.17226/27092

OECD (2022). OECD Framework for the Classification of AI Systems. OECD Digital Economy Papers. No. 323, OECD Publishing, Paris. https://doi.org/10.1787/cb6d9eca-en

OpenAI, R., 2023. GPT-4 technical report. arXiv, pp.2303-08774.

Patel, D., Wong, G. 2023. GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. Blog. Semi Analysis, https://www.semianalysis.com/p/gpt-4-architecture-infrastructure

Shah, A. 2023. Google TPU v5e AI Chip Debuts after Controversial Origins. HPC Wire. https://www.hpcwire.com/2023/08/30/google-tpu-v5e-ai-chip-debuts-after-controversial-origins/

Shavit, Y., 2023. What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring. arXiv preprint arXiv:2303.11341.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N. and Ho, L., 2023. Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.

Stern, P., 2011. Design principles for global commons: Natural resources and emerging technologies. International Journal of the Commons, 5(2).

Suleyman, M. 2023. The Coming Wave. Penguin Random House.

Tabassi, E. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Trager, R., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L., Kreps, S., Lall, R., Larter, O., Ó hÉigeartaigh, S., and Staffell, S., 2023. International Governance of Civilian AI: A Jurisdictional Certification Approach. arXiv preprint arXiv:2308.15514.

U.S. Senate Subcommittee on Privacy, Technology and the Law, 2023. "Oversight of A.I.: Principles for Regulation." Recording of hearing, July 25th, 2023. https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-principles-for-regulation

Wallin, J., Drexel, B., Depp, M., Withers, C., 2023. CNAS Responds: Oversight of A.I.: Rules for Artificial Intelligence. Center for a New American Security. May 17, 2023. https://www.cnas.org/publications/commentary/ostp-national-priorities-for-artificial-intelligence

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D. and Chi, E.H., 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A. and Biles, C., 2022, June. Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214-229).

Zwetsloot, R. and Dafoe, A., 2019. Thinking about risks from AI: Accidents, misuse and structure. Lawfare. February, 11, p.2019.

A APPENDIX A: EXISTING ORGANIZATIONS

How the proposed consortium or a similar organization is structured will have a significant bearing on the development of the risk evaluations ecosystem, but a full discussion of the structuring possibilities of such an organization is beyond the scope of this work. In this section we review existing organizations that offer insights for how the proposed consortium, or some variant thereof, might function, focusing on the implications of the extant organizations on each of the proposed consortium's three objectives. We describe some existing organizations, their activities, and their stakeholders in Table II, and we use the remainder of the section to discuss the relevance of these organizations to the consortium. Yet, there are many limitations to this discussion relevant to organizing and implementing a consortium, for example, how such an organization will be funded or what levers are available to enforce the organization's guidelines and recommendations. A more thorough discussion of the topics touched on in this section is a very important direction for future work.

The first objective concerns coordination between a variety of stakeholders, which, importantly, include both the auditors and auditees. Of the organizations we have reviewed, the ICAO is the most exemplary because it employs in-house auditors who audit the member countries' aviation capacities, so both the auditors and auditees are members. The PCI SSC is very similar because regulatory bodies and standard-setting organizations, as well as the companies who voluntarily adopt these standards, are all members; however, the PCI SSC is concerned with standards and not auditing. GEM is yet another example because it has both policymakers and natural-hazards researchers as members.

The second objective concerns the flexible development and implementation of standards. Here, FIRST is the most relevant example because it must prioritize the quick adoption of standards in response to new developments in the fast-paced domain of cybersecurity. The W3C is another strong example of an effective standards-setting organization, although it typically operates with less urgency than FIRST. Finally, IAIS is a strong example as it is a voluntary membership institution of insurance supervisors and regulators that functions as the international standards-setting body for the insurance industry. Additionally, the ICAO and PCI SSC are useful examples of organizations that effectively function to manage standards-setting in addition to fostering multi-stakeholder coordination, achieving the first and second objective of the proposed organization.

The final objective concerns the consortium acting as an advisory body regarding best practices for frontier AI risk evaluation as well as the authority to certify frontier AI risk evaluators. We could not find examples of an organization functioning in each of these capacities, or in either of these capacities in combination with the first two objectives. However, with respect to advising government agencies, INTERPOL is an example of an international institution that advises policymakers and national law enforcement bodies in their preparations for future security challenges. Other international institutions' core functions concern accreditation of auditors. This alone may still be a valuable function of a consortium or similar institution in the frontier AI evaluations ecosystem if this effort to create an organization working toward all three proposed objectives fails. Either way, it may be possible to learn from these examples:

- International Audit Practice Consortium (IAPC)
- The International Register of Certified Auditors (IRCA)
- International Organization of Supreme Audit Institutions (INTOSAI)

Another useful example might be the IAEA—similar to the ICAO as a model of coordination of policy and regulation (Maas and Villalobos 2023)—which is often referenced in the context of frontier model regulation (Ho et al. 2023; Shavit 2023). Additionally, the European Union's establishment of four Testing and Experimentation Facilities (TEFs) for AI (European Commission 2023) could serve as a good model for how the academics could be engaged in AI risk evaluations.

These examples—and potentially other examples (Maas and Villalobos 2023)—will be useful to draw on in developing an understanding of the best practices required to meet all three of the objectives designated for the proposed consortium. Future efforts should do this, and build on the cursory analysis described here in order to inform the architects of the proposed consortium when they begin

to address the looming challenge of crafting an effective and scalable organizational structure and governance.

Table A1: Existing Organizations with Similar Functions

Organization	Brief description	Main activities	Main stakeholders
International Civil Aviation Organization (ICAO)	Specialized agency of the United Nations observing the administration and governance of civil aviation	International coordination, establishing guidelines/standards, conducting compliance audits and evaluations	Signatories of UN Chicago Convention, orgs including civil professional associations
Payment Card Industry Security Standards Council (PCI SSC)	Global forum developing data security standards for safe payments	Conducting compliance audits with PCI standards, standard-setting, professional training, developing new security solutions	Industry companies, national and regional standard-setting organizations
Global Earthquake Model (GEM)	Global partnership to reduce risk from earthquakes and natural hazards	Providing data, open-source models, risk assessment software and expertise	Governments, organizations (private, public, professional, nonprofit), and individuals
Forum of Incident Response and Security teams (FIRST)	A global forum connecting incident response teams	Standard-setting for cybersecurity, knowledge-sharing between professionals	Private and public incident response teams
World Wide Web Consortium (W3C)	Global consortium ensuring the interoperability, safety and transparency of web technologies	Developing and publishing web standards, information exchange between stakeholders	Technology companies, liaison partnerships with civil and professional associations and country governments
International Association of Insurance Supervisors (IAIS)	International organization of professionals that covers 97 percent of global insurance premiums	Insurance standard setting; works to promote, train, and peer review observance of standards	Insurance firms, and partners like the World Bank, the IMF, and the International Organisation of Securities Commissions
The International Police Organization (INTERPOL)	Intergovernmental organization, connecting police teams and law enforcement agencies	Globally connecting and coordinating law enforcement teams, capacity building, running databases and housing expert teams and response teams	Member countries' law enforcement teams
The International Atomic Energy Agency (IAEA)	Intergovernmental organization ensuring safe nuclear energy development and monitoring nuclear weapons proliferation	Inspecting nuclear facilities; providing information and developing standards; acting as a hub for knowledge-sharing; ensuring peaceful use of nuclear	Member states of the Nuclear Nonproliferation Treaty