ADVANCING GOVERNANCE MODELS FOR FRONTIER FOR AIXBIO:

Key Takeaways and Action Items from the Johns Hopkins Center for Health Security Meeting with Industry, Government, and NGOs
29 November 2023

Artificial intelligence (AI) systems, including foundation models such as large language models (LLMs), are rapidly becoming more powerful, and governments and industry are racing to better understand the potential benefits and risks of this technology. As models become more capable, developers and governments are developing strategies for reducing AI-related risks, including those that may present national security and catastrophic risks. Important initial progress has been made in the US, including the voluntary commitments made by industry[1] as well as the White House's Executive Order on AI (the Executive Order).[2] On November 29, 2023, the Johns Hopkins Center for Health Security (CHS) hosted a not-for-attribution meeting to consider and advance next steps for governance and oversight of the convergence of artificial intelligence and biotechnology (AIxBio). Specifically, the goal of the meeting was for participants to consider two questions:

> (1) What new legal and regulatory requirements are needed to reduce emergent high-consequence AIxBio risks from foundation models, beyond what is currently required by the White House AI Executive Order?
>
> (2) What is the best approach for ensuring developer and deployer accountability for protecting the public from high-consequence AIxBio threats?

In the meeting, CHS convened 51 participants, including AI company representatives from Amazon, Anthropic, Google DeepMind, Meta, Microsoft, and OpenAI. Also participating were US government officials from the White House National Security Council, White House Office of Science and Technology Policy, Department of Energy, Department of Commerce, Department of Defense, Department of State, and Department of Homeland Security. Individuals from the UK Cabinet Office charged with implementing recommendations from the UK Frontier AI Task Force joined the meeting remotely. Employees from other organizations, including the Centre for Long-Term Resilience, Gryphon Scientific, MIT, Rand Corporation, and Yale Law School, also attended. Please see the appendix for the complete list of participants and agenda.

The meeting was structured to create a candid and action-oriented exchange of views on the feasibility and effectiveness of potential US legal requirements that would specifically focus on preventing or

---

[1] https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

[2] EO #14110, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

mitigating future global catastrophic biological risks related to the development or deployment of frontier LLMs and other highly capable foundation AI models. The meeting discussions did not address potential threats from misuse of AI biological design tools in great detail, other than noting that such risks also deserve near-term attention. Discussions focused primarily on potential national efforts in the US but recognized that global approaches will ultimately be needed to adequately address AIxBio risks.

The takeaways below represent the views of CHS and do not represent consensus views or views of specific participants.

The takeaways primarily focus on addressing the risk that foundation models may substantially assist actors seeking to create pandemic-capable pathogens. Though this risk remains nascent, we are concerned that rapidly improving foundation models may, in the time ahead, provide a uniquely accessible combination of tacit knowledge, data, scientific expertise, and troubleshooting skills to actors seeking to create deadly and transmissible biological constructs. Increasing evidence that frontier foundation models can help scientists design and carry out biological experiments, and interact with both specialized models and robotic tools, bolster these concerns.[3] Although important empirical questions in this area remain unresolved, CHS believes it is important to establish required safety measures now that would prevent catastrophic risks emerging from highly capable AI models.[4]

KEY TAKEAWAYS

1. Agencies tasked in the Executive Order with duties related to AIxBio risks should prioritize mitigating pandemic-level biothreats.

- It is useful to clearly define consequential AIxBio risks that are most crucial to avoid, even while safety conversations, the sharing of best practices, and ongoing research on a broad range of threats continue. CHS judges the most consequential emergent AIxBio risks as those in which AI could be used—either deliberately or accidentally—to enable, accelerate, or simplify the creation of dangerous, pandemic-capable viruses, or other highly transmissible pathogens. Such outcomes could include the reintroduction of extinct or highly controlled pandemic viruses (such as smallpox) or the creation of entirely new biological variants or constructs that could start pandemics among people, animals, or plants.
- While other high consequence AI-enabled biological risks also deserve attention (such as nontransmissible bioweapons capable of mass casualties), it is CHS's view that in considering the

[3] See, eg, Sarah R. Carter et al., The Convergence of Artificial Intelligence and the Life Sciences, NTI (2023), https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/; Richard Moulange et al., Towards Responsible Governance of Biological Design Tools (working paper, 2023), https://arxiv.org/abs/2311.15936; The Impact of Large Language Models on Scientific Discovery: a Preliminary Study Using GPT-4, Microsoft Research (working paper, 2023), https://arxiv.org/pdf/2311.07361.pdf; Andres M. Bran et al., Augmenting Large Language Models With Chemistry Tools (working paper, 2023), https://arxiv.org/pdf/2304.05376.pdf; Sophie Rose & Cassidy Nelson, Understanding AI-Facilitated Biological Weapon Development, Centre for Long-Term Resilience (2023), https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio.

[4] Though AIxBio capabilities are trending in the direction of increasing risk, the extent to which foundation models will contribute to lowering barriers to creating bioweapons is still unclear. As the government seeks to reduce global catastrophic biological risks from AI, it should continue to assess the nature of the threat to ensure that its regulations are tailored to robustly reduce risk without unnecessarily slowing the beneficial uses of AI. Continued research on the nature of AI-created biosecurity risks will also allow the government to better assess the importance of additional policy tools, such as regulating biological service providers.

imposition of standards and legal requirements, governments and developers should start by focusing on substantial pandemic biothreats that are potentially unstoppable and could result in global loss of life.

2. The US government should require safety measures along with reporting and oversight requirements to eliminate or mitigate substantial pandemic risks.

- The US government should not permit new or updated versions of highly capable AI models to be released until they have been properly evaluated and red-teamed for substantial pandemic risks, and any identified risks have been adequately mitigated. Independent third-party evaluations should be conducted and should include evaluators with bio expertise. Because models may gain capabilities when users provide new data, engage in fine-tuning, or incorporate additional applications, evaluations should consider the potential for increasing substantial pandemic risks after release. Safety measures should include mandatory no-fault incident reporting and auditing by independent teams. The government should conduct spot-check evaluations and audits.
- The US government, with technical input from developers and biosecurity experts, needs to more clearly define the characteristics and parameters of highly capable models that will subject them to safety mandates.
- The US government should evaluate additional measures, such as dataset controls and model-access controls, for feasibility and effectiveness at reducing substantial pandemic risks from foundation models.
- Congress should provide agencies with the authorities they need to establish a regulatory framework to mandate safety requirements.

3. The Executive Branch should establish mechanisms to facilitate real-time exchange of important AIxBio information among foundation model developers, deployers, and relevant civil society experts in biosecurity.

- AI developers and industry are currently best positioned to understand the power, complexities, and technical capabilities of their models, while government and nongovernmental biosecurity experts are best positioned to understand the nature and likelihood of substantial pandemic threats. Over time, AI developers need to build more expertise to improve their biorisk assessments, just as the government needs to build and sustain AI expertise through workforce development efforts.
- To address the most concerning AIxBio risks, companies must receive clear biosecurity priorities from government and should partner with biosecurity experts within and outside of government to obtain more detailed technical information regarding emerging biorisks and trends. Both the government and developers should quickly seek to create effective evaluation and red-teaming requirements.
- The US government and/or civil society should create a sustained public/private forum to share sensitive risk-related information as well as safety-relevant information on model capabilities, such as the results of red-teaming exercises.

4. **Congress should develop a liability framework that establishes a statutory standard of care and mandatory governance, imposing liability for developers and deployers that fall below that standard, with an appropriate safe harbor to reward responsible practices without creating blanket immunity.**

- A statutory, unified liability framework would provide companies with clarity about their responsibilities, create a compliance baseline, and potentially provide a best practice standard for other countries.
- Such a federal law could include a preemption provision to override potential conflicting state laws as well as common law tort claims, thus providing more certainty for companies and the public.
- Various liability standards should be considered, including strict liability for catastrophic outcomes, joint and several liability, and other approaches to ensure that proper incentives are in place for industry to act responsibly.

CHS Next Steps

The Center for Health Security plans to continue its outreach to government, industry, biosecurity experts, and other key stakeholders to advance agreement around future safety requirements aimed at reducing the potential for foundation models to increase the most concerning pandemic risks. With an ongoing focus on such high-end biorisks, we plan to launch workstreams to:

- Investigate and, if possible, facilitate the creation in the US of a public-private forum for appropriately sharing technical information and sensitive information related to biosecurity risks and red-teaming results;
- Propose a regulatory framework that defines mandatory practices, reporting, and oversight of foundation models; and
- Propose a fair and appropriate statutory liability framework to incentivize developers and deployers of in-scope foundation models to meet a clear standard of care.

To advance these workstreams in 2024, CHS will reach out to meeting participants and other key stakeholders to seek additional input and gauge interest in particular activities moving forward.

APPENDIX: MEETING ATTENDEES AND AGENDA

ADVANCING GOVERNANCE MODELS FOR FRONTIER AIXBIO
29 November 2023

Jeff Alstott
The RAND Corporation

Jennifer Alton
Pathway Policy Group

Samantha Anderson
US Department of State

Rachel Appleton
Anthropic

Chris Austin
Johns Hopkins University

Rachel Azafrani
Microsoft

Bill Beaver
US Department of Defense

Isabel Bennett
National Security Secretariat, UK Cabinet Office

Doni Bloomfield
Johns Hopkins Center for Health Security

John Branscome
Meta

Ben Buchanan
White House National Security Council

Josh Bunce
National Security Secretariat, UK Cabinet Office

Anita Cicero
Johns Hopkins Center for Health Security

Teddy Collins
White House National Security Council

Tim Cranton
Watson-Kelly

Helen Cui
US Department of Energy National Nuclear Security Administration

Natalie de Graaf
White House National Security Council

Sarah Domnitz
US Department of Energy

Gerald Epstein
Johns Hopkins Center for Health Security

Robert Fisher
White House National Security Council

Paul Friedrichs
White House Office of Pandemic Preparedness and Response Policy

Anjali Gopal
Massachusetts Institute of Technology

Chandresh Harjivan
White House Office of Pandemic Preparedness and Response Policy

Melissa Hopkins
Johns Hopkins Center for Health Security

Heidi Carmen Howard
Google DeepMind

Tom Inglesby
Johns Hopkins Center for Health Security

Michael Kaiser
US Department of Homeland Security

Saif Khan
US Department of Commerce

James King
National Security Secretariat, UK Cabinet Office

Dillon Leet
Johns Hopkins Center for Health Security

Margaret Lentz
US Department of Energy, Advanced Scientific Computing Research

Gary Lopez
Microsoft

David Luckey
The RAND Corporation

Greg McKelvey
The RAND Corporation

Ayaz Minhas
Meta

Vanessa Murdock
Amazon

Sella Nevo
The RAND Corporation

Louise Owen
National Security Secretariat, UK Cabinet Office

Jassi Pannu
Stanford University

Tejal Patwardhan
OpenAI

Fiona Pollack
US Department of Defense

Lala Qadir
White House Office of Science and Technology Policy

Ketan Ramakrishnan
Yale Law School

Sophie Rose
The Centre for Long-Term Resilience

Margaret Rush
Gryphon Scientific

Gopal Sarma
White House Office of Science and Technology Policy

Daniel Singer
White House National Security Council

Shankar Sundaram
White House National Security Council

Christian Troncoso
Amazon Web Services

Kirsten Weand
US Department of State

Alex Zhu
Johns Hopkins Center for Health Security

10:00am – 3:00pm
Freedom Room (Lobby Level)
Hotel Washington, 515 15th Street, NW, Washington DC

10:00-10:30AM WELCOME, INTRODUCTIONS, AND GOALS OF MEETING
Tom Inglesby, Johns Hopkins Center for Health Security

The purpose of the meeting is for participants to consider and offer answers to the following:

Q: What new legal and regulatory requirements are needed to reduce emergent high-end AIxBio risks from LLMs and similar models, beyond what is required by the EO?

Q: What is the best approach for ensuring developer and deployer accountability?

10:30-11:00AM PRIORITY SETTING: WHAT AIXBIO RISKS ARE THE MOST CRITICAL TO PREVENT?

11:00-11:30AM LESSONS LEARNED FROM GOVERNANCE OF EMERGING TECHNOLOGY

11:30-12:45PM GROUP DISCUSSION 1: HOW DO WE ENSURE THAT AI IS DEVELOPED AND DEPLOYED WITHOUT SIGNIFICANT BIORISK?

- What are the key characteristics that should subject an AI model or system to mandatory evaluations, red-teaming, and audits?
- What are the most critical elements of evaluations, red-teaming, and audits to successfully reduce high-end biorisks?
- What are the most critical steps to be taken to effectively implement the federal government's evals, red-teaming, and audits under the EO?

12:45-1:00PM BREAK TO PICK UP LUNCH

1:00-2:00PM GROUP DISCUSSION 2: BEYOND EVALS, RED-TEAMING, AND AUDITS

- What, if any, requirements should be set for safe development and deployment beyond evals, red-teaming, and audits (e.g., limiting datasets, export controls, compute governance, etc.)?
- How important is a common international approach to governance? What would be the critical next steps toward reaching that goal?
- What risk tolerance is appropriate for high-end AIxBio risks? Should the standard be to keep risks "as low as is reasonably achievable"? Other alternatives?

2:00-2:50PM GROUP DISCUSSION 3: ACCOUNTABILITY AND CONSEQUENCES

- What kinds of mitigation measures should be required if evaluations, red-teaming exercises, and audits find significant biorisks (e.g., model refinement; use of API; mandatory development pause, etc.)?
- Beyond mitigation measures, what are the most effective levers to ensure compliance with new requirements (e.g., fines, remedial action plans, etc.)?
- What, if any, additional forms of liability are needed in the event that compliant models cause harm?

2:50-3:00PM CLOSING COMMENTS AND NEXT STEPS