Geneva, Brussels, DC & beyond

INHR

advocacy ▪ development ▪ technology

Center for a New American Security

RECOMMENDATIONS TO GOVERNMENTS ON MITIGATING AIXBIO RISKS WHILE PURSUING THE PROMISE OF AI TECHNOLOGY FOR IMPROVEMENTS IN HEALTH AND WELLBEING

The participants of the INHR/CNAS trilateral dialogue and invited biosecurity experts from the Johns Hopkins Center for Health Security:

> Recognize that advances in artificial intelligence (AI) and biotechnology over the next decade have the potential to bring about transformative improvements for human health, animal and plant health, food security, the climate, and economic well-being;
>
> Recognize that the convergence of AI and biotechnology (AIxBio), when combined with AI-driven design and automation, could introduce profound risks that demand global attention in order to prevent the misuse or unintended consequences of such technologies, such as the development of dangerous synthetic pathogens and the enhancement of biological weapons;
>
> Recognize that it is especially imperative to prevent AI from leading to pandemic-level outcomes and risks to national security, economic security, and public health security, such as by AI significantly reducing the capability requirements for non-experts to design, synthesize, acquire, and use biological weapons;
>
> Acknowledge that countries and civil society would benefit from a better shared understanding of these potential risks as well as an improved awareness of best practices and effective measures to prevent or mitigate especially high-consequence outcomes that could have a global impact; and
>
> Appreciate the unique role that governments have in ensuring that their national policies foster continued innovation for AI and biotech technology while also implementing robust safeguards to prevent or mitigate particularly high-consequence risks, ensuring these transformative technologies are developed and deployed responsibly for the benefit of all.

Therefore, we have developed the below recommendations for consideration by governments. These recommendations are offered to inform the development of national level governance measures related to the convergence of AI and biotechnology.

The scope of these recommendations applies to both frontier AI models[1] as well as highly capable AI models trained substantially on biological data. The recommendations are also specifically focused on preventing or mitigating particularly high-consequence AIxBio outcomes that could have a global impact, rather than addressing all types and levels of AIxBio risk. The below recommendations represent collective insights of this group and are not endorsed by any country, organization, or individual in their official capacity.

Governments should explore and undertake actions in the following categories:

[1] For purposes of this document, "frontier AI models" are highly capable AI systems that are often trained with immense computational resources. Such models are sometimes referred to as general-purpose AI (GPAI) models.

Awareness Raising, Training and Human Capacity Building

1. Develop better technical capacity by investing in education and training for a professional workforce -- in government, law enforcement, the private sector, and academia -- prepared to address the intersection of biosecurity threats and artificial intelligence more comprehensively. This will require greater awareness, education, training, and the involvement of experts from a wide range of disciplines to assess and mitigate potential misuse risks of highly capable AI models.
2. Ensure safe and secure innovation of AI and beneficial uses of AI by collaborating with AI developers, biosecurity[2] experts, and other subject matter experts to continuously improve state-of-the-art practices for developing, conducting risk assessments, and safety testing AI models in order to prevent high-consequence AIxBio risks.
3. Encourage and, when appropriate, incentivize private sector actors and investors to provide training on biosecurity risks, red and blue-teaming, the development of effective guardrails, and on other security and safety measures to AI startups as a condition of funding.

Safety Evaluations, Testing, and Industry Best Practices

4. Involve AI and biotechnology companies in analysis and deliberations regarding future national governance measures, including codes of conduct[3] and any necessary regulations applicable to frontier AI models and highly capable AI models trained on biological data. Such measures should include standards for safety evaluations and responsible scaling programs focused on capabilities-based thresholds.
5. Understanding the challenge of identifying and mitigating all types of AIxBio risks, focus first on advancing risk assessments and safety testing standards to identify and mitigate AI model bio-capabilities of concern that could lead to particularly high-consequence harms with global impact.
6. Require these high-priority safety evaluations to become regularized, and actively consider appropriate consequences and necessary remedial actions when such capabilities of concern are identified. Develop and share guidelines for AI developers, deployers, and other actors to recognize when and how to mitigate dangerous capabilities identified.

---

[2] The term “biosecurity” is defined in various ways across governments, institutions, and international organizations. For purposes of this document, biosecurity refers to actions taken to prevent and respond to the theft, misuse, or the inadvertent or intentional release of dangerous biological agents. Biosecurity is an important component of national and international security.

[3] The Tianjin Biosecurity Guidelines for Codes of Conduct for Scientists, endorsed by the InterAcademy Partnership, is a useful example and precedent for providing guidelines aimed at preventing misuse of bioscience research without hindering beneficial outcomes.

7. Oversee or develop the capability inside governments to conduct red-teaming exercises of frontier AI models and highly capable AI models trained on biological data to identify capabilities of concern that could lead to global harms and blue teaming to address and remedy these vulnerabilities. Draw red- and blue-teamers from a cross-sectoral pool of human talent. Put precautions in place for red-teaming practices to avoid laboratory validation of potential risk when such validation could lead to the creation of genuinely dangerous biological constructs. Work with the private sector to develop safe proxy experiments if necessary to conduct evaluations. Use AI tools where effective to enhance and double-check human red- to identify vulnerabilities and blue-teaming to identify patches.
8. Develop and share best practices for safety evaluations and red-teaming that involve assessing risks across an interconnected ecosystem of AI models, robotics, and tools, rather than evaluating only isolated individuals models.
9. Consider and create appropriate incentives, including financial or other incentives, for industry and especially academic laboratories to develop safety and security mechanisms to reduce high-consequence AIxBio risks.
10. Analyze the potential benefits and potential risk vulnerabilities of open-source frontier models as well as highly capable AI models trained on biological data. Consider whether and to what extent certain specified types of open-source models should be regulated – including by limiting access to models or model weight information– to make it more difficult for malicious actors to circumvent safety guardrails and otherwise “jailbreak” safety measures. For example, consider whether and how to limit open access to model weights of certain AI models trained on high volumes of sensitive biological data.
11. Consider the utility of mandates and incentives where competitive pressures between private companies prevent voluntary, industry-wide implementation of best safety practices, provided they carefully balance public safety and risks of suppressing innovation against the positive social benefits of AI applications.

Protein Design Security

12. In order to prevent the creation in the laboratory of a dangerous biomolecule designed by AI, establish national protein design security policies applicable to manufactures of synthesized genetic sequences, users of such products, and manufacturers of desktop equipment for synthesizing nucleic acids. Such policies could include Know Your Customer regulations and order screening requirements, to ensure that nucleic acid synthesis technologies are appropriately used to advance beneficial outcomes in research and prevent misuse by malicious actors.

International Cooperation on Enhancing Safety for AI x chem-bio threats

13. Reaffirm and strengthen the international norm against the creation of biological weapons and bolster the Biological and Toxin Weapons Convention by considering the creation of a process or mechanism to provide expert and scientific support to States Parties on the risks of biological weapons hazards associated with AI and other emerging technologies.
14. Engage in international dialogues to investigate international agreements, institutions, or other international measures to prevent or otherwise address particularly high-consequence AIxBio risks.
15. Create or designate a national, authoritative focal point institution or agency that can be the official technical point of contact on AI safety and security issues between governments.
16. Develop and institutionalize international information sharing about AIxBio threats to support counter-terrorism cooperation and emerging threats to international peace and security.

NOTE: This revision to the original AI-bio recommendations from INHR's May 2024 workshop in Thailand was made possible by a grant from Founder's Pledge.