Geneva, New York, DC & beyond

INHR

advocacy ▪ development ▪ technology

# RECOMMENDATIONS TO GOVERNMENTS ON MITIGATING AIxBIO RISKS

December 2025

Developed by INHR experts and endorsed by Ailurus Ltd (UK), AI Safety Asia (Hong Kong), Alethia (Poland), Centre du Commerce Internationale pour le Développement, CECIDE (Guinea, Switzerland), Concordia AI (China, Singapore), Council on Strategic Risks (USA), Horizon Insights Center (China), INHR Geneva (USA, Switzerland), International Biosecurity and Biosafety Initiative for Science, IBBIS (Switzerland), Johns Hopkins University Center for Health Security (USA), Photosynthesis and Environment Laboratory, Center of Excellence in Molecular Plant Sciences, Chinese Academy of Sciences (China), Pour Demain (Belgium, Switzerland), Sentinel Bio (USA), Tianjin University Center for Biosafety Research and Strategy (China), and the United Services Institution of India (India).

# Recommendations to Governments on Mitigating AIxBio Risks

These recommendations are based on the input and insights provided by participants of the INHR/CNAS trilateral dialogue. This US-China-International dialogue, focused primarily on the safety of AI military systems, includes experts such as retired generals, diplomats, subject matter experts, and private sector professionals from China, Denmark, France, India, Ireland, Korea, Norway, Poland, Russia, Sweden, Switzerland, the United Kingdom and the United States. It is open to participation from additional member states. In May 2024, the dialogue convened an AIxBio working group in Thailand and the United States, as well as online, specifically to consider the convergence of AI with biotechnology and to offer proposals to governments for addressing consequential AIxBio risks without unduly limiting AIxBio technology development. Throughout 2025, this working group met with AIxBio experts at the AI4Good Summit and the Biological Weapons Convention in Geneva, at the World AI Conference in Shanghai, as well as online. Working group members participated in their individual capacities. The recommendations below are informed by expert discussions and input from members of the INHR/CNAS trilateral dialogue and the working group, but they do not necessarily reflect the views of all participants.

> Recognizing 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;
>
> Recognizing that the convergence of AI and biotechnology (AIxBio), especially AI-driven biological design and research 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;
>
> Recognizing that it is especially imperative to prevent AI from enabling high-consequence widespread harms to plants, animals, or humans and creating risks to national security, economic security, and public health security, such as by AI significantly lowering barriers to design, synthesize, acquire, and use biological weapons;
>
> Acknowledging that the biotechnology research and development ecosystem of governments, research laboratories, and industry 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; and
>
> Appreciating the unique role that governments have in ensuring that their national policies foster continued AI-enabled biotechnology innovation 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.

The undersigned organizations recommend governments consider the following measures intended 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 highly capable general purpose AI models as well as highly capable biological AI models trained substantially on biological datasets and intended for biological tasks. 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.

Governments should explore and undertake actions in the following categories:

## Awareness Raising, Training and Human Capacity Building

1. Develop better technical capacity by investing in education and training for a professional workforce -- in government, the private sector, and academia -- prepared to address AIxBio risks 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, biosafety and biosecurity experts, and other subject matter experts to continuously improve state-of-the-art practices for developing, conducting risk assessments, and 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[1] and any necessary regulations applicable to highly capable general purpose AI models and biological AI models possessing capabilities of concern. Such measures should include standards for pre-deployment safety evaluations and responsible scaling programs focused on capabilities-based thresholds. For biological AI models

[1] 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.

specifically, measures could include pre-development biosecurity risk assessment processes to identify, evaluate, and mitigate high-consequence risks.

5. Understand 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 the risks of high consequence outcomes identified.
7. Oversee or develop the capability and capacity inside governments to conduct red-teaming exercises of highly capable general purpose AI models and biological AI models to identify capabilities of concern that could lead to high consequence outcomes and to conduct 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-teaming to identify vulnerabilities and blue-teaming to identify patches.
8. Develop and share best practices for standardized safety evaluations and red-teaming that involve assessing risks across an interconnected ecosystem of AI models, robotics, and tools, rather than evaluating only isolated individual 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 different release approaches (including fully open-source releases) for general purpose AI models and biological AI models possessing capabilities of concern that could lead to high-consequence outcomes. Consider under what circumstances regulating open-source models possessing capabilities of concern might be appropriate – including by limiting access to model source code, training data, weights and documentation– to make it more difficult for malicious actors to circumvent security measures and otherwise “jailbreak” safety guardrails. 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 such interventions carefully balance public safety and risks of suppressing innovation against the positive social benefits of AI applications.

**National Governance**

12. Develop a national policy framework to prevent or mitigate high consequence AIxBio security risks. Such a framework could help to inform intelligence agencies and support counterterrorism, and could help to guide regulators, response agencies and compliance with national policies among AI developers and life sciences practitioners utilizing AI-enabled tools.
13. Create or designate a national, authoritative institution or agency that can be the official technical point of contact on AI and AIxBio safety and security issues between national security, national health security and disease control, and national AI security institutions.
14. With input from AI developers, biosafety/biosecurity experts, intelligence agencies, and public health officials, evaluate levels of AIxBio security risks and design national policy interventions and best practices.
15. In designing national policy interventions and best practices, governments should differentiate between AIxBio risks which emerge from considerations such as biosafety and biosecurity, general purpose AI models and biological AI models, and other such differentiating parameters in a granular approach.
16. In order to prevent the creation in the laboratory of a dangerous biomolecule designed by AI, establish nucleic acid sequence screening policies applicable to manufactures of synthesized nucleic acid 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**

17. 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.
18. Engage in international dialogues to investigate international agreements, institutions, or other international measures to prevent or otherwise address particularly high-consequence AIxBio risks.

19. Create or designate a national, authoritative focal point institution or agency that can be the official technical point of contact on AI and AIxBio safety and security issues between governments.
20. Develop and institutionalize international information sharing about AIxBio threats to support counter-terrorism cooperation and emerging threats to international peace and security.

These recommendations are based on consultation with and endorsed by the following organizations:

- INHR Geneva (Switzerland, United States of America)
- Tianjin University Center for Biosafety Research and Strategy (China)
- Johns Hopkins University Center for Health Security (United States of America)
- United Services Institute of India (India)
- Alethia XAI (Poland)
- Ailurus Ltd (United Kingdom)
- Sentinel Bio (United States of America)
- Centre du Commerce Internationale pour le Développement (Guinea, Switzerland)
- Concordia AI (China, Singapore)
- The International Biosecurity and Biosafety Initiative for Science (Switzerland)
- Horizon Insights Center (China)
- The Council on Strategic Risks (United States of America)
- Photosynthesis and Environment Laboratory, Center of Excellence in Molecular Plant Sciences, Chinese Academy of Sciences (China)
- Pour Demain (Belgium, Switzerland)
- AI Safety Asia (Hong Kong)

## Glossary

Solely for purposes of these recommendations, the following definitions are provided to ensure clarity and consistency. It is understood that in the realm of AI and AIxBio, many terms used do not yet have a globally agreed definition.

Biological AI Model: An AI model trained on biological data (e.g., genomic sequences, protein structures, epidemiological datasets, metabolic pathways) for tasks in the life sciences.

Biosafety: Containment principles, technologies, measures and practices that are implemented to prevent unintentional exposure to biological agents or their inadvertent release.[2]

In the context of AI, biosafety includes AI safety principles, technologies, measures, and practices to mitigate known shortcomings of AI models such as hallucination or misaligned outputs that can lead to unintended potentially high consequence outcomes.

Biosecurity: Principles, technologies, measures and practices that are implemented for the protection, control and accountability of biological agents, data or equipment, biotechnologies, skills and information related to their handling. Biosecurity aims to prevent their unauthorized access, loss, theft, misuse, diversion or release.[3]

In the context of AI, biosecurity includes AI security principles, technologies, measures, and practices to prevent unauthorized access, loss, theft, misuse, diversion or release of sensitive biological data, model weights, or AI tools that could materially assist misuse.

General Purpose AI Model**:** An AI model that is trained on a large amount of broad data at scale, that displays significant generality, and that is capable of performing a wide range of tasks.[4]

NOTE: These recommendations were made possible through the generous support to INHR, the U.S. registered parent organization of INHR-Geneva, from Founder's Pledge which enabled the INHR/CNAS Trilateral Dialogue on AI and the military. CNAS is the Center for a New American Security which is co-host of the dialogue and whose experts participated in the drafting process, but which, as a matter of policy, does not make institutional endorsements.

[2] Definition used in WHO Global Guidance Framework for the Responsible Use of the Life Sciences.
[3] Definition used in WHO Global Guidance Framework for the Responsible Use of the Life Sciences.
[4] Definition adapted from the EU AI Act's definition of "General Purpose AI Models" in Article 3(63).

## Endorsed by...

Johns Hopkins Bloomberg School of Public Health
Center for Health Security

Tianjin University (Peiyang University) 1895

Ailurus

Sentinel Bio

Estd. 1870

植物系统生物学
PLANT SYSTEMS BIOLOGY

COUNCIL ON STRATEGIC RISKS

IBBIS
International Biosecurity and Biosafety Initiative for Science

Horizon Insights Center

安远 AI
CONCORDIA AI

Alethia
Vision AI. Made Real.

AISA
AI SAFETY ASIA

CECIDE

POUR DEMAIN

Geneva, New York, DC & beyond
INHR
advocacy ▪ development ▪ technology

日内瓦、纽约、华盛顿以及其他地区
宏理国际
健康 ▪ 技术 ▪ 能力建设

Find the most up to date version with all civil society endorsers at INHR.org/AI