JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH
Center for
Health Security

Summary of Paper:
Prioritizing High-Consequence Biological Capabilities in Evaluations of AI Models[1]

The Problem

Artificial intelligence (AI) technologies could revolutionize how healthcare is organized and delivered, how medicines and vaccines are developed, how diseases are diagnosed, and the speed with which new outbreaks are detected. The vast majority of biological and life sciences research using AI can be done in ways that pose minimal to no risk to society.

However, some new AI models expected to emerge in the near future could increase the risk of high-consequence outcomes resulting from accidents or misuse of biotechnology and the life sciences. As model capabilities increase, it is anticipated that there will be a commensurate increase in the ability to engineer and manipulate existing pandemic pathogens and possibly create new ones. Researchers will also be able to combine rapidly improving AI models with wet-lab advances to facilitate, accelerate, and augment this work.

AI researchers and policymakers have not yet broadly agreed upon what AI model features or uses most increase significant biosecurity risks to the public—or what forms of risks are most worth mitigating. Some large language model (LLM) developers have used red teams to evaluate the biosecurity risks of their models in the absence of concrete government guidance, but they have varied in content and methods. No unified framework for the content of evaluations exists, and there is no shared understanding regarding the degree of concern warranted for a particular capability level.

As a result, the limited published biosecurity studies of AI models done to date (which have only assessed LLMs) test for different risks and use differing assumptions regarding which threats should be guarded against. This in turn reduces the potential impact of mitigation efforts.

Because it's impossible to evaluate AI models for their ability to contribute to *any* possible biology-related accident or misdeed, some level of prioritization is needed. Merely asking whether a model increases the risk of “bioweapons planning,” for example, is an insufficient evaluative question—it is ambiguous, under inclusive, and difficult to extend beyond LLMs. The ultimate purpose of biosecurity assessments should be to determine whether a model meaningfully increases the likelihood of high-consequence risks to the public, regardless of human intent.

The Solution

Thankfully, a ready parallel exists for the identification and prioritization of AI-related biology harms. This can be found in existing policies and practices governing scientific research intended for benefit but with the potential for significant harm, known as “dual-use research of concern” (DURC) and research intended to create pathogens with enhanced pandemic potential (PEPP). Scientists and policymakers have studied DURC and PEPP extensively for more than a decade and have developed detailed guidance and practices to address the potential risks to public health and safety. These previously identified dual-use capabilities[2] and practices should inform and help identify potential harms and testable components for AI model evaluations.

[1] For more detail, see the full paper https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4873106.
[2] https://www.phe.gov/s3/dualuse/Pages/USGOversightPolicy.aspx.

Indeed, the recently released *United States Government Policy for Oversight of Dual Use Research of Concern and Pathogens with Enhanced Pandemic Potential*[3] "recognizes [this parallel in the context of] the rapidly evolving nature of computational biology and the increasing use of computational models and approaches, including the use of artificial intelligence, that potentially contributes to the production of dual-use biological knowledge, information, technologies and products." The policy encourages the assessment of the dual-use potential of *in silico* research conducted through computer simulation and development of risk mitigation plans.

Applying previously identified dual-use capabilities in the life sciences to AI models, testable components include the ability of an AI model to:

- Enhance the harmful consequences of the agent or toxin
- Disrupt immunity or the effectiveness of an immunization against the agent or toxin without clinical or agricultural justification
- Confer to the agent or toxin resistance to clinical or agriculturally useful prophylactic or therapeutic interventions against that agent or toxin or facilitate their ability to evade detection methodologies
- Increase the stability, transmissibility, or the ability to disseminate the agent or toxin
- Alter the host range or infectiousness abilities of the agent or toxin
- Enhance the susceptibility of a host population to the agent or toxin
- Generate or reconstitute an eradicated or extinct agent or toxin

Examples of emerging AI-enabled capabilities of greatest concern associated with these capabilities include AI models that help predict viral resistance to neutralizing antibodies, models that can design protein shells around a virus such that the virus is not affected by either vaccines or the natural immune system, and genomic foundation models capable of designing viral traits (such as an ability to infect lung cells) and engineering routes to producing these traits.[4]

Governments and industry should develop model evaluations that assess the extent to which these capabilities of concern produce high-consequence biological outcomes. Two potential harms that are extraordinarily important to prevent are AI models or AI tools that could currently or in the near to mid-term future, either on their own or when paired with other emerging or existing capabilities:

(1) Greatly accelerate or simplify the reintroduction of dangerous extinct viruses or dangerous viruses that only exist now within research labs that could have the capacity to start pandemics, panzootics, or panphytotics; or
(2) Substantially enable, accelerate, or simplify the creation of novel variants of pathogens or entirely novel biological constructs that could start pandemics, panzootics, or panphytotics.

These are not the only potential AI-enabled biological harms that should be governed, but governance efforts should prioritize and address them at a minimum. If these specific large-scale harms are initiated by an AI model, there may be limited opportunity to stop them from having a global impact. We strongly recommend that governments and model developers establish targeted, standardized evaluations such that they assess the above capabilities and these 2 potential harms.

[3] https://www.whitehouse.gov/wp-content/uploads/2024/05/USG-Policy-for-Oversight-of-DURC-and-PEPP.pdf.

[4] For a non-exhaustive list of additional AI-enabled capabilities of concern, see Table 3. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4873106.