POLICY FORUM

BIOSECURITY

AI and biosecurity: The need for governance

Governments should evaluate advanced models and if needed impose safety measures

By **Doni Bloomfield[1,2], Jaspreet Pannu[1,3,4], Alex W. Zhu[1], Madelena Y. Ng[5], Ashley Lewis[6], Eran Bendavid[3,4], Steven M. Asch[4], Tina Hernandez-Boussard[5,6], Anita Cicero[1], Tom Inglesby[1]**

Great benefits to humanity will likely ensue from advances in artificial intelligence (AI) models trained on or capable of meaningfully manipulating substantial quantities of biological data, from speeding up drug and vaccine design to improving crop yields (*1 3*). But as with any powerful new technology, such biological models will also pose considerable risks. Because of their general-purpose nature, the same biological model able to design a benign viral vector to deliver gene therapy could be used to design a more pathogenic virus capable of evading vaccine-induced immunity (*4*). Voluntary commitments among developers to evaluate biological models' potential dangerous capabilities are meaningful and important but cannot stand alone. We propose that national governments, including the United States, pass legislation and set mandatory rules that will prevent advanced biological models from substantially contributing to large-scale dangers, such as the creation of novel or enhanced pathogens capable of causing major epidemics or even pandemics.

Current models only provide blurry image[s] of novel bacterial genomes (*5*), are data limited (*6*), and require in vitro validation (*7*). However, the rapid progress of AI and the ever-larger resources being invested in computation and data generation for biological models suggests that capabilities are likely to accelerate (*8*). Researchers creating leading biological models recognize this dual-use danger. Baker and Church caution that protein-design technology is vulnerable to misuse and the production of dangerous biological agents (*9*). The creators of the genomic-prediction model Evo note that the ability to discern fitness associated with certain sequences can assist in the discovery of novel biomarkers or therapeutic targets, but can also catalyze the development of harmful synthetic microorganisms (*5*). Although they recognize that there remain barriers to the production of dangerous novel agents, they call for a proactive discussion involving the scientific community, security experts, and policy-makers...to prevent misuse [M.Y.N., A.L., and T.H.-B. were also co-authors of (*5*)]. Scientists have an ethical obligation to consider the broader effects of this line of work, including the potential for future misuse.

A global consortium of biological model developers has recently taken a crucial first step, adopting the Responsible AI x Biodesign statement of community values and commitments (*10*). Signatories committed to developing methods to evaluate models' dangerous capabilities, undertaking such evaluations before model release, and purchasing synthetic nucleic acids only from providers that screen orders for biosecurity purposes (A.C. and T.I. signed this statement as supporters).

NECESSARY BUT INSUFFICIENT

We do not rely on scientists voluntary ethical agreements alone to protect human subjects of scientific research or to ensure that live Ebola virus is handled safely in laboratories. The scientific community generally recognizes that some forms of research involve substantial risks that must be considered in a formal oversight process before a study moves forward. Given academic pressure to publish quickly, the lack of model developer expertise in biosecurity risks, and the absence of standard approaches to risk evaluation, voluntary risk assessment of new biological models will not be sufficient.

There are many types of biological models, and the number of models that qualify as such will continue to expand as general-purpose AI tools and combinations of models can increasingly be used for biology-relevant purposes. Within this broad range of models, only a narrow set of advanced biological models have characteristics that currently warrant oversight. Regulations should initially focus on models that (i) are trained with very large computational resources (for example, greater than 10^{26} integer or floating-point operations, a threshold used in the White House's recent Executive Order on AI) on very large quantities of biological sequence and/or structure data, or (ii) were trained with at least a lower quantity of computational resources on especially sensitive biological data that are not widely accessible (for example, new data that link viral genotypes to phenotypes with the capacity for pandemic spread). By focusing on these classes of models, officials are more likely to evaluate the models that pose the greatest risks without unduly hampering academic freedom. Officials would also thereby avoid attempting to control a vast swath of AI research, which would likely do more harm than good. The scope of concerning models may change over time. For example, it may become necessary to evaluate models capable of autonomously running laboratory experiments, even if such models were not trained on biological data directly [for example, (*2*)]. Regulators should therefore be given sufficient, although not unlimited, flexibility to modify the definition of biological models subject to oversight as technology and the nature of high-consequence risks change.

The US and UK governments have taken initial steps in this direction by creating new organizations tasked with designing safety evaluations for leading frontier models, including evaluations to better understand biological-weapons threats. Prerelease evaluations of advanced biological models, targeted to identify capabilities that present pandemic-level risks, should be required and standardized where possible.

It is difficult to forecast precisely when these dangerous capabilities may surface. Without robust data on pathogen characteristics, AI models may struggle to carry out the most concerning tasks, such as increasing viral transmission. Models trained solely on data from the distribution of pathogens present in nature may be limited in their ability to generate novel pathogens with properties

[1]Center for Health Security, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [2]School of Law, Fordham University, New York, NY, USA. [3]Department of Health Policy, Stanford School of Medicine, Stanford University, Stanford, CA, USA. [4]Department of Medicine, Stanford School of Medicine, Stanford University, Stanford, CA, USA. [5]Department of Medicine (Biomedical Informatics), Stanford School of Medicine, Stanford University, Stanford, CA, USA. [6]Department of Biomedical Data Science, Stanford School of Medicine, Stanford University, Stanford, CA, USA. Email: tinglesby@jhu.edu

far outside existing evolutionary bounds. Currently, the field lacks automated, scalable ways to iteratively synthesize, manipulate, test, and generate data on novel pathogens. These limitations are not likely to restrain progress indefinitely, however. High-throughput data generation and autonomous robotics may in the future enable task-specific feedback loops that permit exploration of pathogenicity and transmission characteristics that extend beyond what is present in natural pathogen diversity. Moreover, there is evidence that AI models can extrapolate to novel biological design space (*11*, *12*). Biological models may not be able to substantially contribute to creating novel or enhanced pandemic-capable pathogens today. But scientists can already generate high-quality data on pathogenic characteristics and will likely soon be able to generate such data at higher volumes and at reduced cost. Researchers could then use such data to train models that are potentially capable of extrapolating to new biological constructs. The essential ingredients to create highly concerning advanced biological models therefore may already exist or soon will. Establishment of effective governance systems now is warranted.

In the United States, this would entail providing a federal agency with the authority to subject a narrow class of advanced biological models (as defined above), before release, to an independent biosafety and biosecurity evaluation and to impose safety measures if needed. Whichever agency is tasked with this oversight should work with the new US AI Safety Institute to design validated tools capable of assessing whether a biological model possesses high-consequence dangerous capabilities. The oversight process should be both narrowly tailored and adequately funded to ensure speedy review and limited interference with the scientific process. It is important that these rules extend to private companies, which are also developing biological models.

There is substantial precedent for this approach. For example, scientists and policy-makers have jointly developed principles and regulations to ensure that government-funded research that involves human subjects is conducted responsibly and have created rules to protect animals in research settings. A tiny fraction of life-sciences research that could lead to both meaningful benefit and harm referred to as dual-use research of concern is also subject to federal oversight. Research regulations are not limited to government grantees: the US Food and Drug Administration requires that human trials used to support pharmaceutical approvals comply with ethical standards.

BALANCING ACCOUNTABILITY AND OPENNESS

In creating new rules, policy-makers should prioritize mitigating the most consequential biosecurity risks—including the substantial facilitation of novel or enhanced pathogens capable of causing major epidemics or pandemics while preserving broad autonomy for researchers to develop models, publish, and collaborate (*13*). Policy-makers should craft narrowly tailored rules to retain the substantial benefits of scientific openness, including those that accrue to public health.

To that end, the United States should only subject an AI model to prerelease evaluation if it meets the advanced biological model criteria discussed above. Evaluations should attempt to elicit such models' abilities to substantially contribute to high-consequence risks. Officials designing evaluations can draw on longstanding recommendations developed to reduce risks from dual-use research of concern and pathogens enhanced in ways that make them more likely to cause an epidemic or pandemic (*13*). Experts have spent years evaluating proposed research work related to increasing pathogenicity, experience that can now be applied to assess AI models. Thus, for example, evaluations should test whether a biological model is able to plausibly increase a pathogen's transmissibility or virulence or its ability to evade the immune system and medical countermeasures.

Because physically testing a model's ability to design or synthesize new pathogens capable of causing a pandemic may be risky in itself, evaluations could also create tests that proxy for certain dangerous capabilities. For example, evaluators could test whether future AI-enabled robotic tools are capable of autonomous, de novo synthesis of benign pathogens, as a proxy assessment for the synthesis of more dangerous ones. We expect that current biological models would not pose substantial risk by these metrics, but that may change. Thresholds for when a meaningful improvement in model capability has occurred should be set in advance, and it should be clear when such thresholds have been crossed. In the above described example, thresholds might correspond to the percentage of steps in the synthesis process that have been fully automated by AI.

Many AI evaluation approaches are now automated and standardized, features that can enable fast and fair assessment. US government officials should work with experts in machine learning, infectious disease, ethics, and biosecurity, along with the governments of other nations, to devise a frequently updated battery of tests to probe probable risks from existing and novel pathogen families. On the basis of the results of those tests, officials can determine—with a presumption toward openness—whether model restrictions such as limitations on model access or required use of application-programming interfaces should be required.

Oversight policies will need to address special risks associated with publications that release a model's weights, which define how the model operates. Such releases allow third parties to modify a model's capabilities. For example, creators of the

Geographic distribution of biological model developers

To analyze the geographic distribution of biological model developers, we compiled the country-level location of the authors of the 59 relevant publications in the most comprehensive database of such models to date (*8*). The authors of (*8*) collected data on 60 publications concerning 90 models (some publications described multiple models). According to their report, the authors collected their data by surveying the biological sequence model literature, leaderboards of common benchmarks used to evaluate models and frequently downloaded open-source protein language models (*8*). We supplemented their data by hand-collecting country-level location data for each publication author in their database (see supplementary materials). We excluded a report of one model published only on a newsletter website; that model was not otherwise described in the database. Overall, 83% of publications included authors located in just four countries: the United States, United Kingdom, China, and Germany. A majority (58%) of publications included a US author, more than publications involving authors from the United Kingdom (19%), China (14%), or Germany (14%). The US role is even more pronounced when examining the models trained by using the most substantial computing resources. Among the top 20 models by compute training measured as the number of floating-point operations per second carried out to create the final model all but two involved a US author. In one of those exceptions, the biological model was a modified version of LLama-7B, an open model released by Meta Platforms, a US company.

Evo model sought to guard against misuse by excluding viral genomes that infect eukaryotic hosts" from the model s training data. They then published the model weights, as is currently the norm in academia (*5*). Within weeks of Evo s release, other scientists had refined the model s published weights with data on viruses that infect humans (*14*). The data in that case involved a benign virus family, but the case highlights that fine-tuning an open model is often much cheaper than training a new large model, and so oversight policies will need to account for the possibility of postrelease fine tuning. As the Responsible AI x Biodesign signatories recognized, "[p]ractices for limiting access and distribution should be followed for AI systems that present identified meaningful and unresolved risks" (*10*). In addition, the US government should create best practices for responsible sharing of large-scale new data on pathogenic characteristics of concern, such as viral genomic data matched with transmissibility characteristics. In this area, officials can draw on past policies for responsible data distribution, such as those applied by the National Institutes of Health to certain human genomic data.

EVALUATION REQUIREMENTS INDEPENDENT OF FUNDING SOURCES

Because considerable biological model research happens outside of US government funded laboratories, prerelease evaluation requirements for concerning advanced biological models should be required irrespective of funding source. The private sector has created some of the world s leading biological models—for example, Google DeepMind s AlphaFold series of protein-folding models (Google DeepMind recently released an AI safety framework that includes biosecurity evaluation commitments). Because the computational resources used to train biological models are increasing exponentially, it is likely that a growing proportion of leading biological models will be created in industry, as in other domains of AI research (*8*). Government AI biosecurity policies should therefore apply to advanced biological models that pose potential high-consequence threats, whether or not they were developed with federal funding assistance.

COMMON INTERNATIONAL APPROACHES

For now, a considerable portion of biological model development takes place in the United States (see the box). Although this makes the United States a promising place to initiate regulatory discussions, researchers are also developing leading biological models outside the United States. Governments around the world should create rules to ensure that advanced biological models, wherever developed, are evaluated before release. The UK AI Safety Summit was a promising early indication that governments may be able to find common ground on AI safety. During the summit, many countries signed the Bletchley Declaration acknowledging the potential risks of AI, including in domains such as biotechnology and cybersecurity. Given the nearly universal antipathy toward biological weapons expressed in the Biological Weapons Convention, and the InterAcademy Partnership s recent adoption of biosecurity guidelines for scientists, there may be a path toward international cooperation on biological model risks as well. Cooperation on this front should extend beyond countries responsible for the majority of biological model development to date, given that risks will be borne globally.

Policy-makers should strive to harmonize oversight standards to avoid regulatory arbitrage and conflicts over digital trade, problems that have plagued the encryption and data privacy arenas. At the same time, officials in countries with the most advanced AI technology should prioritize effective evaluations, even if they come at some cost to international uniformity.

GENOME SYNTHESIS SCREENING

Last, although it is critical and necessary to require providers of synthetic nucleic acids to screen purchases to prevent the unjustified distribution of highly dangerous nucleic acids, that policy will not by itself be sufficient to protect against AI biological model risks. The argument has been made that oversight of the digital-to-physical divide will be sufficient to reduce biosecurity risks because the digital information produced by biological models is harmless by itself (*9*). Unfortunately, no country has legal requirements for such genome synthesis screening in place. We strongly support such a legal requirement, and the White House has recently issued a directive to prevent recipients of federal funding from purchasing synthetic nucleic acids from providers that do not screen orders. However, those requirements will not apply to the private sector, and other countries do not have similar requirements in place. Even if they are put in place, such requirements are not sufficient to eliminate the risk of malicious actors synthesizing dangerous new constructs without detection (*15*). Chemical methods for creating short, single-stranded nucleic acid sequences (oligonucleotides) have been known for more than four decades and are now routine around the world. Oligonucleotides in turn can be stitched together to form whole viral genomes and then (with relevant expertise and resources) booted up into infectious viruses by using mammalian cells (*15*). Global oversight of nucleic acid synthesis and viral rescue, although critical, should be complemented by prerelease review of advanced biological models that may be able to generate highly concerning pathogen information or automate substantial portions of the synthesis and rescue process. In addition, policy-makers should develop and seek international agreement regarding best practices for the responsible sharing of new and important data on pathogenic characteristics of concern.

CONCLUSION

Biological model developers have taken an important step in adopting voluntary commitments to reduce risk. As they have before with the advent of new and powerful tools, researchers and other parties should now begin to work with government officials to implement those principles and inform government oversight policies and regulations that balance biological models substantial benefits with their greatest risks. ■

REFERENCES AND NOTES

1. Microsoft Research AI4Science, Microsoft Azure Quantum, arXiv:2311.07361 [cs.CL] (2023).
2. D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, *Nature* **624**, 570 (2023).
3. J. T. Rapp, B. J. Bremer, P. A. Romero, *Nat. Chem. Eng.* **1**, 97 (2024).
4. J. B. Sandbrink, E. C. Alley, M. C. Watson, G. D. Koblentz, K. M. Esvelt, *Gene Ther.* **30**, 407 (2023).
5. E. Nguyen *et al.*, *bioRxiv* [Preprint] (2024); https://doi.org/10.1101/2024.02.27.582234.
6. S. Lyu, S. Sowlati-Hashjin, M. Garton, *Nat. Mach. Intell.* **6**, 147 (2024).
7. N. N. Thadani *et al.*, *Nature* **622**, 818 (2023).
8. N. Maug, A. O Gara, T. Besiroglu, *Epoch AI* **17**, https://epochai.org/blog/biological-sequence-models-in-the-context-of-the-ai-directives (2024).
9. D. Baker, G. Church, *Science* **383**, 349 (2024).
10. Responsible AI x Biodesign; https://responsiblebiodesign.ai.
11. R. Krishna *et al.*, *Science* **384**, eadl2528 (2024).
12. B. Ni, D. L. Kaplan, M. J. Buehler, *Chem* **9**, 1828 (2023).
13. J. Pannu *et al.*, *SSRN*, 25 June 2024; http://dx.doi.org/10.2139/ssrn.4873106.
14. K. Workman, *LatchBio*, 20 March 2024; https://blog.latch.bio/p/engineering-aavs-with-evo-and-alphafold.
15. National Academies of Sciences, Engineering, and Medicine; Division on Earth and Life Studies; Board on Chemical Sciences and Technology; Board on Life Sciences; Committee on Strategies for Identifying and Addressing Potential Biodefense Vulnerabilities Posed by Synthetic Biology, *Biodefense in the Age of Synthetic Biology* (National Academies Press, 2018).

ACKNOWLEDGMENTS

D.B. and J.P. contributed equally to this work. The authors acknowledge support from Open Philanthropy (D.B., A.C., T.I., and A.W.Z.); Effective Giving (A.C., T.I., and A.W.Z.); Horizon Institute for Public Service (J.P.); and National Center for Advancing Translational Sciences of the National Institutes of Health, award UL1TR003142 (T.H.-B.). The content is solely the responsibility of the authors and does not necessarily represent the views of any funder or funders.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adq1977

10.1126/science.adq1977