AUGUST 2024

AI and the Evolution of Biological National Security Risks

Capabilities, Thresholds, and Interventions

Bill Drexel and Caleb Withers

About the Authors

Bill Drexel is a fellow for the Technology and National Security Program at the Center for a New American Security (CNAS). His work focuses on Sino-American competition, artificial intelligence, and technology as an element of American grand strategy. Previously, Drexel worked on humanitarian innovation at the UN (International Organization for Migration) and on Indo-Pacific affairs at the American Enterprise Institute. Always seeking on-the-ground exposure, Drexel has served as a rescue boat driver during Libya's migration crisis; conducted investigative research in the surveillance state of Xinjiang, China; and supported humanitarian data efforts across wartime Ukraine. He holds a BA from Yale University and master's degrees from Cambridge and Tsinghua universities.

Caleb Withers is a research assistant for the Technology and National Security Program at CNAS. Before CNAS, he worked as a policy analyst for a variety of New Zealand government departments. He holds an MA in security studies from Georgetown University with a concentration in technology and security, and a bachelor's of commerce from Victoria University of Wellington with majors in economics and information systems.

About the Technology & National Security Program

The CNAS Technology and National Security Program explores the policy challenges associated with emerging technologies. A key focus of the program is bringing together the technology and policy communities to better understand these challenges and together develop solutions.

About the Artificial Intelligence Safety & Stability Project

The CNAS AI Safety & Stability Project is a multiyear, multiprogram effort that addresses the established and emerging risks associated with artificial intelligence. The work is focused on anticipating and mitigating catastrophic AI failures, improving the U.S. Department of Defense's processes for AI testing and evaluation, understanding and shaping opportunities for compute governance, understanding Chinese decision-making on AI and stability, and understanding Russian decision-making on AI and stability.

Acknowledgments

The authors are grateful to Dr. Sonia Ben Ouagrham-Gormley and Dr. Nathan Paxton for their valuable feedback and suggestions on earlier drafts of this report, and to Michael Aird, Samuel Curtis, and Christian Ruhl, who all provided further useful inputs. This report would not have been possible without contributions from our CNAS colleagues, including Paul Scharre, Vivek Chilukuri, Melody Cook, Rin Rothback, Allison Francis, Jake Penders, Tim Fist, Josh Wallin, Michael Depp, and Noah Greene. The report was made possible with the generous support of Open Philanthropy.

As a research and policy institution committed to the highest standards of organizational, intellectual, and personal integrity, CNAS maintains strict intellectual independence and sole editorial direction and control over its ideas, projects, publications, events, and other research activities. CNAS does not take institutional positions on policy issues, and the content of CNAS publications reflects the views of their authors alone. In keeping with its mission and values, CNAS does not engage in lobbying activity and complies fully with all applicable federal, state, and local laws. CNAS will not engage in any representational activities or advocacy on behalf of any entities or interests and, to the extent that the Center accepts funding from non-U.S. sources, its activities will be limited to bona fide scholastic, academic, and research-related activities, consistent with applicable federal law. The Center publicly acknowledges on its website annually all donors who contribute.

TABLE OF CONTENTS

Executive Summary

Not long after COVID-19 gave the world a glimpse of the catastrophic potential of biological events, experts began warning that rapid advancements in artificial intelligence (AI) could augur a world of bioterrorism, unprecedented superviruses, and novel targeted bioweapons. These dire warnings have risen to the highest levels of industry and government, from the CEOs of the world's leading AI labs raising alarms about new technical capabilities for would-be bioterrorists, to Vice President Kamala Harris's concern that AI-enabled bioweapons "could endanger the very existence of humanity."[1] If true, such developments would expose the United States to unprecedented catastrophic threats well beyond COVID-19's scope of destruction. But assessing the degree to which these concerns are warranted—and what to do about them—requires weighing a range of complex factors, including:

- The history and current state of American biosecurity
- The diverse ways in which AI could alter existing biosecurity risks
- Which emerging technical AI capabilities would impact these risks
- Where interventions today are needed

This report considers these factors to provide policymakers with a broad understanding of the evolving intersection of AI and biotechnology, along with actionable recommendations to curb the worst risks to national security from biological threats.

The sources of catastrophic biological risks are varied. Historically, policymakers have underappreciated the risks posed by the routine activities of well-intentioned scientists, even as the number of high-risk biosecurity labs and the frequency of dangerous incidents—perhaps including COVID-19 itself—continue to grow. State actors have traditionally been a source of considerable biosecurity risk, not least the Soviet Union's shockingly large bioweapons program. But the unwieldiness and imprecision of bioweapons has meant that states remain unlikely to field large-scale biological attacks in the near term, even though the U.S. State Department expresses concerns about the potential bioweapons capabilities of North Korea, Iran, Russia, and China. On the other hand, nonstate actors—including lone wolves, terrorists, and apocalyptic groups—have an unnerving track record of attempting biological attacks, but with limited success due to the intrinsic complexity of building and wielding such delicate capabilities.

Today, fast-moving advancements in biotechnology—independent of AI developments—are changing many of these risks. A combination of new gene editing techniques, gene sequencing methods, and DNA synthesis tools is opening a new world of possibilities in synthetic biology for greater precision in genetic manipulation and, with it, a new world of risks from the development of powerful bioweapons and biological accidents alike. Cloud labs, which conduct experiments on others' behalf, could enable nonstate actors by allowing them to outsource some of the experimental expertise that has historically acted as a barrier to dangerous uses. Though most cloud labs screen orders for malicious activity, not all do, and the constellation of existing bioweapons norms, conventions, and safeguards leaves open a range of pathways for bad actors to make significant progress in acquiring viable bioweapons.

But experts' opinions on the overall state of U.S. biosecurity range widely, especially with regard to fears of nonstate actors fielding bioweapons. Those less concerned contend that even if viable paths to building bioweapons exist, the practicalities of constructing, storing, and disseminating them are far more complex than most realize, with numerous potential points of failure that concerned parties either fail to recognize or underemphasize. They also point to a lack of a major bioattacks in recent decades, despite chronic warnings. A more pessimistic camp points to experiments that have demonstrated the seeming ease of successfully constructing powerful viruses using commercially available inputs, and seemingly diminishing barriers to the knowledge and technical capabilities needed to create bioweapons. Less controversial is the insufficiency of U.S. biodefenses to adequately address large-scale biological threats, whether naturally occurring, accidental, or deliberate. Despite COVID-19's demonstration of the U.S. government's inability to contain the effects of a major outbreak, the nation has made limited progress in mitigating the likelihood and potential harm of another, more dangerous biological catastrophe.

New AI capabilities may reshape the risk landscape for biothreats in several ways. AI is enabling new capabilities that might, in theory, allow advanced actors to optimize bioweapons for more precise effects, such as targeting specific genetic groups or geographies. Though such capabilities remain speculative, if realized they would dramatically alter states' incentives to use bioweapons for strategic ends. Instead of risking their own militaries' or populations' health with the unwieldy weapons, states could sabotage other nations' food security or incapacitate enemies with public health crises from which they would be unlikely to rebound. Relatedly, the same techniques could create superviruses optimized for transmissibility and lethality, which may considerably expand the destructive potential of bioweapons. Tempering these fears, however, are several technical challenges that scientists would need to overcome—if they can be solved at all.

The most pressing concern for biological risks related to AI stems from tools that may soon be able to accelerate the procurement of biological agents by nonstate actors. Recent studies have suggested that foundation models may soon be able to help accelerate bad actors' ability to acquire weaponizable biological agents, even if the degree to which these AI tools can currently help them remains marginal.[2] Of particular concern are AI systems' budding abilities to help troubleshoot where experiments have gone wrong, speeding the design-build-test-learn feedback loop that is essential to developing working biological agents. If made more effective, emerging AI tools could provide a boon to would-be bioweapons creators by more dynamically providing some of the knowledge needed to produce and use bioweapons, though such actors would still face other significant hurdles to bioweapons development that are often underappreciated.

AI could also impact biological risks in other ways. Technical faults in AI tools could fail to constrain foundation models from relaying hazardous biological information to potential bad actors, or inadvertently encourage researchers to pursue promising medicinal agents with unexpected negative side effects. Using AI to create more advanced automated labs could expose these labs to many of the risks of automation that have historically plagued other complex automated systems, and make it easier for nonspecialists to concoct biological agents (depending upon the safety mechanisms that automated labs institute). Finally, heavy investment in companies and nations seeking to capitalize on AI's potential for biotechnology could be creating competition dynamics that prioritize speed over safety. These risks are particularly acute in relation to China, where a variety of other factors shaping the country's biotech ecosystem also further escalate risks of costly accidents.

Attempting to predict exactly how and when catastrophic risks at the intersection of biotechnology and AI will develop in the years ahead is a fool's errand, given the inherent uncertainty about the scientific progress of both disciplines. Instead, this report identifies four areas of capabilities for experts and policymakers to monitor that will have the greatest impact on catastrophic risks related to AI:

1. Foundation models' ability to effectively provide experimental instructions for advanced biological applications
2. Cloud labs' and lab automation's progress in diminishing the demands of experimental expertise in biotechnology
3. Dual-use progress in research on host genetic susceptibility to infectious diseases
4. Dual-use progress in precision engineering of viral pathogens

Careful attention to these capabilities will help experts and policymakers stay ahead of evolving risks in the years to come.

For now, the following measures should be taken to curb emerging risks at the intersection of AI and biosecurity:

- Further strengthen screening mechanisms for cloud labs and other genetic synthesis providers
- Engage in regular, rigorous assessments of the biological capabilities of foundation models for the full bioweapons lifecycle
- Invest in technical safety mechanisms that can curb the threats of foundation models, especially enhanced guardrails for cloud-based access to AI tools, "unlearning" capabilities, and novel approaches to "information hazards" in model training
- Update government investment to further prioritize agility and flexibility in biodefense systems
- Long term, consider a licensing regime for a narrow set of biological design tools with potentially catastrophic capabilities, if such capabilities begin to materialize

Introduction

In 2020, COVID-19 brought the world to its knees, with nearly 29 million estimated deaths, acute social and political disruptions, and vast economic fallout.[3] However, the event's impact could have been far worse if the virus had been more lethal, more transmissible, or both. For decades, experts have warned that humanity is entering an era of potential catastrophic pandemics that would make COVID-19 appear mild in comparison. History is well acquainted with such instances, not least the 1918 Spanish Flu, the Black Death, and the Plague of Justinian—each of which would have dwarfed COVID-19's deaths if scaled to today's populations.[4]

Equally concerning, many experts have sounded alarms of possible deliberate bioattacks in the years ahead. There is some precedent: in the weeks following 9/11, letters containing deadly anthrax spores were mailed to U.S. lawmakers and media outlets, and the attack could have been considerably worse had the perpetrator devised a more effective dispersion mechanism for the anthrax. The episode could portend a future in which more widely available biological capabilities mean malicious individuals and small groups devastate governments and societies through strategic biological attacks. Jeff Alstott, former director for technology and national security at the National Security Council, warned in September 2023 that the classified record contained "fairly recent close-ish calls" of nonstate actors attempting to use biological weapons with "strategic scale."[5]

Accurately weighing just how credible such dire warnings are can feel next to impossible, and requires clear judgment in the face of opaque counterfactuals, alarmism, denialism, and horrific possibilities. But regardless of their likelihood, the destructive potential of biological catastrophes is undeniably enormous: history is littered with examples of societies straining and even collapsing under the weight of diseases—from ancient Athens's ruinous contagion during the Peloponnesian War, to the bubonic plague that crippled the Eastern Roman Empire in the 6th century, to the cataclysmic salmonella outbreak in the Aztec empire in the 16th century.[6] It is essential that U.S. leaders soberly address the risks of biological catastrophe—which many claim will change dramatically in the age of artificial intelligence.

Government and industry leaders have expressed grave concerns about the potential for AI to dramatically heighten the risks of catastrophic events in general, and biological catastrophes in particular.[7] In a July 2023 congressional hearing, Dario Amodei, CEO of leading AI lab Anthropic, stated that within two to three years, there was a "substantial risk" that AI tools would "greatly widen the range of actors with the technical capability to conduct a large-scale biological attack."[8] Former United Kingdom (UK) Prime Minister Rishi Sunak similarly expressed urgent concern that there may only be a "small window" of time before AI enables a step change in bioterrorist capabilities.[9] U.S. Vice President Kamala Harris warned of the threat of "AI-formulated bio-weapons that could endanger the lives of millions . . . [and] could endanger the very existence of humanity."[10] These are serious claims. If true, they represent a significant increase in bioterrorism risks. But are they true?

This report aims to clearly assess AI's impact on the risks of biocatastrophe. It first considers the history and existing risk landscape in American biosecurity independent of AI disruptions. Drawing on a sister report, *Catalyzing Crisis: A Primer on Artificial Intelligence, Catastrophes, and National Security*, this study then considers how AI is impacting biorisks across four dimensions of AI safety: new capabilities, technical challenges, integration into complex systems, and conditions of AI development.[11] Building on this analysis, the report identifies areas of future capability development that may substantially alter the risks of large-scale biological catastrophes worthy of monitoring as the technology continues to evolve. Finally, the report recommends actionable steps for policymakers to address current and near-term risks of biocatastrophes.

While the theoretical potential for AI to expand the likelihood and impact of biological catastrophes is very large, to date AI's impacts on biological risks have been marginal. There is no way to know for certain if or when more severe risks will ultimately materialize, but careful monitoring of several capabilities at the nexus of AI and biotechnology can provide useful indications, including the effectiveness of experimental instructions from foundation models, changing demands of tacit knowledge as lab automation increases, and dual-use AI-powered research into host genetic susceptibility to infectious diseases and precision pathogen engineering. Lest they be caught off guard, policymakers should act now to shore up America's biodefenses for the age of AI by strengthening screening mechanisms for gene synthesis providers, regularly assessing the bioweapons capabilities of foundation models, investing in a range of technical AI safety mechanisms, and preparing to institute licensing requirements for sophisticated biological design tools if they begin to approach potentially catastrophic capabilities.

The Current State of Catastrophic Biological Risks

To assess AI's emerging national security impacts on biological risks is difficult not only because of AI's unpredictable progress and varied applications in the field, but also because simply establishing a clear baseline of biorisk today is a challenge. Perceptions of existing biorisks vary widely, as do the sources of potential threats. The following sections provide an overview of the sources of catastrophic biorisk, evolving capabilities in biotech independent of AI tools, existing safeguards and gaps, and differing perceptions of risks. Taken together, significant, unaddressed biological risks to national security exist today independently of AI disruptions on the horizon, though some renditions of bioterrorist threats in particular are exaggerated due to an overly simplistic appreciation of the demands of bioweapons development.

Sources of Risk

The COVID-19 pandemic shocked public consciousness into an active, daily awareness of biological risks, an issue that previously was largely the purview of experts. Whether COVID-19 was the result of a lab leak or a naturally occurring event, the virus's vast disruptions and its likely death toll of nearly 29 million individuals constituted a catastrophe of global proportions.[12] For many in the biological sector who have been warning of the risks of pandemics, COVID-19 vindicated longstanding concerns over the vulnerability of both the United States and existing global response mechanisms to large-scale pandemics. Those concerns remain legitimate; in a highly connected world, conditions are ripe for devastating, fast-moving viruses to rapidly spread—a risk also highlighted, albeit less severely, by earlier pandemics such as severe acute respiratory syndrome (SARS) and swine flu. Future pandemics have the potential to be far worse. The 1918 Spanish Flu, for example, killed approximately 1 to 2 percent of the world's population—equivalent to 70 to 150 million people today. Moreover, the 1918 pandemic had peak mortality for prime–working age adults, resulting in severe economic damage as millions succumbed to the illness.[13] Before the discovery of modern antibiotics, bacterial pandemics such as plague would sometimes kill half or more of affected populations. The Black Plague, for example, killed around half of Europeans over a few years in the mid-1300s.[14]

The competing origin stories of COVID-19 both provide examples of catastrophic risk scenarios worthy of concern. Naturally occurring viruses can create catastrophes of devastating proportions independently of deliberate biological experimentation. Factors including increased travel, greater urbanization, climate change, changing interactions between humans and animals, and healthcare deficiencies in low- and middle-income countries all contribute to greater chances of extreme pandemics now and in the future.[15]

But biological experimentation, too, can be a source of catastrophic risk. Controversial forms of scientific research, such as gain-of-function research (also referred to as enhanced pandemic potential pathogen research) that sometimes entails altering existing viral strains and creating new ones, have the potential to enable biological catastrophes by accident.[16] The facts that the Wuhan Institute of Virology engaged in gain-of-function research, acted as a center of coronavirus research, and elicited safety warnings within the U.S. Department of State before COVID all mean that a potential lab leak with catastrophic consequences

Members of the St. Louis (Missouri) Red Cross Motor Corps on duty during the global influenza pandemic, October 1918. (Library of Congress)

Security personnel stand guard outside the Wuhan Institute of Virology as members of the World Health Organization team investigating the origins of COVID-19 visit the institute in China's central Hubei province, February 3, 2021. (Hector Retamal/AFP via Getty Images)

was possible there, and could also be possible in a number of biological labs around the world.[17] Indeed, a single biological lab in Beijing was the source of four known SARS leaks in early 2004.[18] Another lab in Lanzhou, China, leaked aerosolized Brucella to surrounding areas in 2019, leading to more than 10,000 individuals contracting the disease in what may be the largest lab leak to date.[19] As of 2023, a total of 69 biosafety level 4 (BSL4) laboratories worldwide—biolaboratories that require the highest safety standards to deal with extremely hazardous biological materials—are in operation, under construction, or planned.[20] In recent years, the number of such high-risk labs has dramatically increased, with three-quarters located in urban areas.[21] And between 1975 and 2016, there were more than 60 known accidents from lab researchers (BSL4 or otherwise) that resulted in individuals being exposed to highly infectious pathogenic agents.[22] The true number, including unreported or unknown incidents, is likely much higher.[23] Though many of these exposures did not come from BSL4 labs, and most were contained, the trend is not promising.

Onlookers are often baffled by the growth of high-risk biological research, but there are strong incentives to engage in it. Pioneering remedies for particularly dangerous diseases or other biological agents is often associated with scientific prestige, can have financial benefits related to monetized cures, and can represent a lifesaving contribution to society. Whatever the motive, such research typically requires working with, and sometimes manipulating, biohazards. As such, some element of risk is ultimately unavoidable in experiments to advance medical understanding of high-risk pathogens.

Naturally occurring pandemics and lab leaks are not the only sources of risks for biological catastrophes. State and nonstate actors could create bioweapons with wide-ranging—and potentially catastrophic—effects. Since the first documented use of biological weapons in the 14th century BCE, when the Hittites sent diseased rams to infect their enemies, armies have employed a wide range of now primitive tactics to use biological agents for strategic effects. These have included infected or poisoned arrows and catapulting diseased corpses into besieged cities.[24] A major turning point in the history of states' use of biological weapons arrived at the end of the 19th century, as Louis Pasteur and Robert Koch provided the foundations for microbiology, thereby opening new possibilities to understand and develop biological weapons.[25] These new capabilities were explored in the first half of the 20th century by a range of countries, including France, the United Kingdom, Italy, Canada, Belgium, Poland, Germany, Japan, and the United States, and ultimately reached their apex in the Soviet Union's staggering bioweapons ecosystem.[26] The Soviet Union created the largest bioweapons program in history, with roughly 15,000 scientists, technicians, and support staff directly working to produce hundreds of tons of biological weapons agents—including shocking efforts to make some of the world's most deadly diseases more lethal and resistant to treatment.[27]

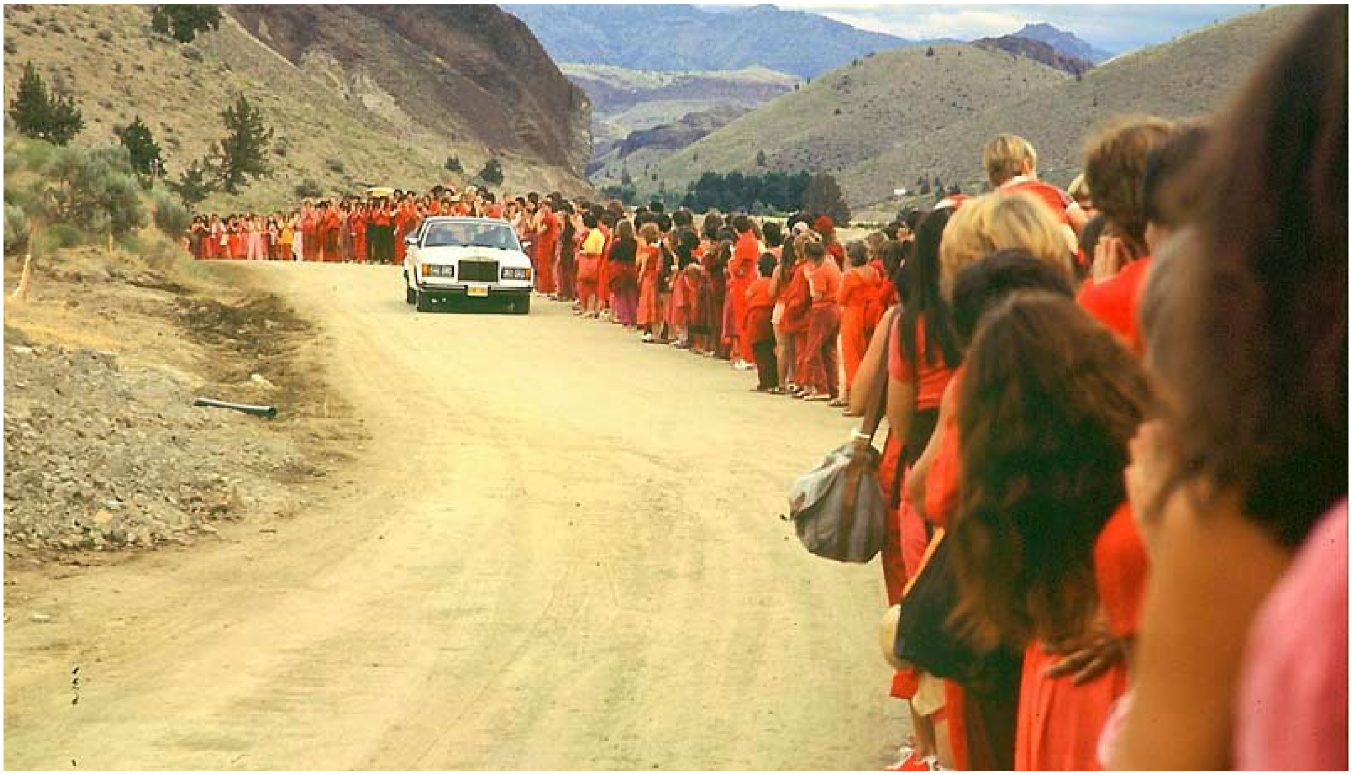
Despite the Soviet Union's extraordinarily productive program, however, biological weapons remain relatively unattractive to most states due to their uncontrollable nature, except for limited specialized operations such as assassinations or special operations' sabotage efforts.[28] Since 1915, a total of 23 states have had known or suspected bioweapons programs, nearly all of which have been shuttered (if they existed at all).[29] Japan's World War II–era bioweapons program involved considerable fatalities, including thousands who were killed for experimental purposes.[30] Japan's primary offensive deployments were executed in China, where Japanese forces reportedly poisoned more than 1,000 water wells with cholera or typhus and distributed plague-infested fleas across several Chinese cities, among other activities.[31] The 2002 International Symposium on the Crimes of Bacteriological Warfare, convened in China, estimated that the number of casualties from Japan's bioweapons program in China amounted to 580,000 individuals at a minimum.[32]

Though a range of governments have accused adversaries of deliberate state uses of bioweapons since Japan's World War II operations, there is little evidence to substantiate most of these claims.[33] A notable exception is recently uncovered documentation suggesting that the Israeli military used typhoid and dysentery during the 1948 Arab-Israeli War, which it intended as a nonlethal means to deter Arab militiamen from returning to villages and towns they had been driven from and to impede the progress of invading Arab troops.[34] Today, the U.S. State Department's Bureau of Arms Control, Verification, and Compliance assesses that North Korea and Russia currently maintain offensive biological capabilities, while China and Iran engage in concerning biological research that may suggest they too maintain secret offensive capabilities of unknown size, or could quickly stand up such programs.[35]

Notably, a state need not launch a biological attack for a catastrophe to occur. As with private scientific labs, state-run bioweapons labs can also have consequential accidental leaks, such as in the accidental 1979 outbreak of anthrax from the Soviet Union's Sverdlovsk lab. The secret bioweapons facility emitted a plume of spores that were carried by wind over adjacent communities, killing 68 by official records—though the true number was likely greater.[36] Another Soviet leak of highly lethal and transmissible smallpox from a bioweapons research center on Vozrozhdeniye Island in 1971 could have been even more devastating, had it not been swiftly contained.[37] Given that the products of bioweapons labs are dangerous by design, such leaks are comparatively more hazardous than those from most other scientific facilities.

Terrorist groups or lone wolf individuals pursuing mass-casualty bioweapons also pose risks, though recent efforts of this kind have thankfully had limited impact. In recent memory, the United States was subjected to two bioweapon attacks, though neither reached catastrophic proportions. In 1984, a religious commune in Oregon systematically contaminated local salad bars with salmonella in an attempt to incapacitate non-commune voters in a local election. In 2001, shortly after the 9/11 attacks, a suspected lone wolf perpetrator mailed letters filled with anthrax to news outlets in Florida and New York, as well as a congressional office building in Washington, DC, resulting in the deaths of 5 individuals and illness in 17 others. Both incidents could have been much worse. The orchestrators of the Oregon attack bought and considered using much more severe pathogens, but decided salmonella would be sufficient for their purposes.[38] The powder used in the anthrax attack was only of a low grade and failed to disperse effectively, limiting its impact.[39]

A hazmat worker sprays colleagues after an anthrax search at Dirksen Senate Office Building on November 18, 2001, in Washington, DC. Authorities closed two Senate buildings to test for anthrax spores after investigators discovered a contaminated letter addressed to Sen. Patrick Leahy (D-VT). (Alex Wong via Getty Images)



Devotees watch their leader drive by the Rajneeshpuram commune in 1982. Two years later, members of the commune contaminated local salad bars with salmonella to incapacitate non-commune voters in a local election. (Samvado Gunnar Kossatz)

Some terrorist groups have held ambitions for still more far-reaching bioattacks that have thankfully not panned out, but could have been catastrophic. Al Qaeda and the Islamic State of Iraq and Syria (ISIS) have both sought mass-casualty biological weapons, as has the apocalyptic Japanese cult Aum Shinrikyo, infamous for its successful use of sarin gas, a chemical agent, to kill 13 individuals and injure more than 6,000 on the Tokyo Metro.[40] These failures to effectively develop and deploy bioweapons reflect the delicate nature of biological agents as opposed to chemical weapons or more conventional weapons.

Taken together, the sources of potentially catastrophic biological risks span natural origins, accidents from legitimate scientific experiments, and intentional production of biological agents from states, terrorist organizations, lone wolves, and apocalyptic groups. Each potential progenitor of biological catastrophe is subject to a distinct set of sometimes unpredictable forces and incentives that could alter the likelihood of developing dangerous pathogens or other bioweapons. Whereas state actors must respond to complex strategic incentives and deterrence dynamics, lone wolves and apocalyptic groups are largely impervious to such considerations—with terrorist groups operating somewhere in between.[41] Even those attempting to develop bioweapons for strategic purposes can face challenges in managing them, raising the risk of potentially catastrophic incidents beyond their control, just as scientific motivations to explore dangerous diseases for medical progress can run the risk of costly accidents. But much as the motives and incentives around man-made biothreats vary, evolving technical capabilities help shape evolving risks from each source.

Evolving Capabilities

The risk profiles of biological catastrophes—regardless of their source—are heavily shaped by the capabilities of available biological tools and techniques. These capabilities can be thought of in two categories: "classic" or conventional biological capabilities, those based on naturally occurring agents; and synthetic biological capabilities, those dependent on artificial manipulation of genetic code.

A clear, if small-scale, example of a classic biological attack is the aforementioned 1984 salmonella attack in Oregon, organized by a religious commune that aimed at incapacitating local voters in an election.[42] Several naturally occurring biological agents, most notably anthrax and botulinum, could be leveraged to catastrophic effect with the right dispersion mechanisms.

Groups such as al Qaeda and Aum Shinrikyo, as well as the Soviet bioweapons program, have focused on these more accessible agents that could have plausibly led to mass-casualty events, albeit with greatly varying degrees of success.[43] Though conventional biological capabilities have a more established history, leveraging naturally occurring biological agents still requires considerable expertise in terms of cultivating, sustaining, and dispersing biological agents effectively (though the difficulty of each of these areas, too, varies depending on which agent is used).

Synthetic biological capabilities rely on the techniques of synthetic biology more broadly, a field of research in which the genetic material of organisms is read, edited, and rewritten. Though synthetic biology has its origins in the 1970s, recently there has been considerable acceleration in the speed, cost-effectiveness, and sophistication of synthetic biology. This increases the ease with which various actors can produce bespoke biohazards, as the Soviet Union's bioweapons program once aspired to do. The range of potential synthetic biological risks is potentially endless but could include altering or designing pathogens to be more lethal, more transmissible, less treatable, or less detectable.

The proliferation of synthetic biology tools is accelerating, primarily driven by a desire to revolutionize biology for the good of humanity. These tools have considerable potential to create a more robust bioengineering ecosystem, which could facilitate rapid, collaborative breakthroughs in medicine, agriculture, biomanufacturing, and other applications. A key component of this push is "democratizing" access to advanced biological capabilities, which proponents hope will enable the field to achieve some of the fast-paced, collaborative success that has marked the software industry.[44] But the flip side of democratizing access to constructive synthetic biology applications is, perhaps, democratizing access to destructive synthetic biology capabilities. Some of the new services, technologies, and tools associated with synthetic biology create a patchwork of capabilities that can be strung together in a range of combinations to produce biological agents or toxins with catastrophic potential.[45]

The biological tools and services powering advancements in synthetic biotechnology are complex and variated, but some key advancements include:[46]

- **CRISPR gene editing techniques**, which allow greater precision in DNA editing than previously possible and are rapidly falling in cost. CRISPR enables biologists to manipulate pathogens in a variety of new ways at relatively low cost.
- **New DNA synthesis tools** that allow scientists to order, combine, or "print" genomic sequences of organisms, including pathogens, with lower costs and increasing ease.
- **Improvements in genome sequencing**, which permit biologists to sequence DNA with increasing speed, accuracy, and cost-effectiveness—an important element of testing and verifying potential biological agents.

These three areas of technical advancement provide the foundation for a more dynamic biotech ecosystem but also augur new risks.

One final element of the changing risks, for both conventional and synthetic capabilities, is the rise of cloud labs—laboratories that conduct biological processes and experiments on others' behalf. Emerald Cloud Lab, for instance, describes itself as a "remotely operated research facility that handles all aspects of daily lab work—method design, materials logistics, sample preparation, instrument operation, data acquisition and analysis, troubleshooting, waste disposal, and everything in between—without the user ever setting foot in the lab."[47] Since the first robotic biolab was launched 2012,

The Carnegie Mellon University Cloud Lab is a remotely operated, automated lab that gives researchers access to more than 200 pieces of scientific equipment. (Carnegie Mellon University) rapidly growing companies such as Emerald Cloud Lab, Gingko Bioworks, and Synthego have offered services that reduce the need for biology professionals to manage the minutiae of conducting physical experiments themselves in favor of simply designing experiments for outsourced execution. Cloud labs represent an increasingly important element of the "digital-to-physical barrier," through which digital designs or plans for biological production are made physical realities. By centralizing, standardizing, and automating biolab resources and procedures, cloud labs aim to make considerable gains in life sciences' efficiency and experimental reproducibility.

Much as cloud labs aim to make biological experimentation faster and cheaper, they also lower barriers to entry, including for potentially malicious parties. Rather than having to source biological lab equipment and cultivate the experimental skills that would have usually been required to develop biological agents, motivated actors could in principle outsource some or all of their laboratory needs to cloud labs, assuming they could circumvent cloud labs' safety mechanisms. Nonetheless, there remain a number of hurdles to successfully automating sophisticated experiments (see "Tacit Knowledge," page 21), and cloud labs maintain obvious incentives to not facilitate malicious activities and develop their policies and safeguards accordingly.[48]

Even so, cloud labs' ambitions to consistently automate an increasingly broad range of biological capabilities could represent a transformation in catastrophic biological risks. As Dr. Sonia Ben Ouagrham-Gormley, an expert in the history of biological weapons, has demonstrated, experimental expertise and organizational culture have typically acted as the most significant inhibitors to success for bioweapons production. From the Soviet bioweapons megaproject to the smaller, clandestine programs of Iraq, South Africa, and Aum Shinrikyo, attempts to produce bioweapons have been most stymied by a combination of organizational challenges and gaps in "tacit knowledge"—subtle expertise in the minutiae of experimentation that can be difficult or impossible to articulate.[49] While the social and political hurdles that often accompany bioweapons programs may persist, cloud labs may alter the demands of tacit experimental knowledge by outsourcing some of the tacit knowledge needed to build biological agents.[50]

The tacit knowledge in question can be as subtle as the air pressure in a lab chamber, the speed of swirling together a mixture, or minuscule variations in the pH level of water used in an experiment. Often, it can be difficult for researchers to identify such variances in experimental conditions between labs, making the challenge seemingly that much harder to automate in robotic labs. Cloud labs have the advantage of being able to iterate on experiments more rapidly and over longer stretches of time than traditional lab technicians, allowing them to accumulate a repertoire of highly precise and replicable methods, even if their dexterity pales in comparison to conventional technicians. The compounding effects of these automation techniques may add up, as more and more steps of experiments could be reliably strung together. That said, the degree to which this will be the case is ultimately unknown, and the reliable operation of such complex machinery and experimental operation may introduce its own forms of tacit knowledge that must be mastered, erecting new barriers to synthetic biology. Additionally, there are several important forms of tacit knowledge—not least those that relate to the intricacies of cooperation among technicians—with which automation will be very unlikely to help (for a more thorough exploration of the changing dynamics of tacit knowledge and automation, see "Tacit Knowledge," page 21).[51]

Ultimately, it may be too early to assess the impact of cloud labs on biorisks. Just as such cloud labs could make bioweapon production easier for lone wolves, terrorist organizations, or apocalyptic groups, they could also centralize lab expertise under more controlled—and monitorable—bottlenecks of biological production. But given that not all cloud labs maintain robust safety monitoring systems for their orders, they could also raise the risks of malicious actors gaining access to potentially catastrophic biological capabilities.

Safeguards and Gaps

The primary safeguards against biological catastrophes of relevance to the United States include international organizations that aim to curb the deliberate production of high-risk biological agents; a variety of bodies, practices, and mechanisms designed to diminish the biological risks of experimental research; and the U.S. government's dedicated organs and law enforcement resources for biodefense. Taking each in turn, the primary international organizations of relevance include the following:

- **The Biological Weapons Convention (BWC)**, a treaty that came into force in 1975, currently has 185 states parties that have committed to not produce or stockpile bioweapons, and to conduct only biodefense-related research in relation to bioweapons. With limited enforcement mechanisms and funding, the BWC's primary significance in biocatastrophe mitigation is in having helped establish an international norm with relatively few instances of noncompliance among states.

- **The Australia Working Group** is an informal grouping of members of the BWC that meet annually to establish guidance on materials and tools worthy of export controls due to their ability to empower malicious actors. While this mechanism helps to constrain critical materials and tools that could exacerbate biorisks globally, there remain demonstrable gaps in these constraints that could be exploited.[52]

- **The International Gene Synthesis Consortium (IGSC)** is a voluntary group of gene synthesis companies, including cloud labs, that commit to screening both orders and customers for hazardous requests. While the consortium works to reduce risks, malicious actors can in theory still circumvent screening protocols to benefit from gene synthesis companies.[53] Additionally, although nearly all of the world's largest companies making high-quality, gene-length DNA are members of the IGSC, there is at least one notable outlier in China, and many other labs and companies maintain more limited gene synthesis capabilities.[54]

In addition to these international bodies, various national practices including ethics review boards, biosafety committees, and funding of due diligence mechanisms contribute to global biosecurity efforts. In the United States, such entities are informed by guidance from the National Institutes of Health. In the wake of COVID-19 and concerns over the American government's funding of potentially risky research, the risk tolerance that such mechanisms should exhibit has been a recent source of contention, especially in regards to the oversight of research involving enhanced potential pandemic pathogens.[55] Moreover, some labs, especially private ones, may not be subject to the same oversight as government-funded or academic laboratories engaged in high-risk research. On occasion, some scientists may also simply circumvent ordinary oversight mechanisms, as with the 2018 He Jiankui incident in China, in which Dr. He conducted illegal heritable human genome editing in three human fetuses.[56]

Biodefense and preparedness measures provide a final set of safeguards against deliberate, accidental, and naturally occurring biological risks. Several governmental bodies work to establish wide-ranging biodefenses, notably:

- **The Biomedical Advanced Research and Development Authority (BARDA)**, which was established in 2006 in the Department of Health and Human Services to "develop medical countermeasures that address the public health and medical consequences of chemical, biological, radiological, and nuclear (CBRN) accidents, incidents and attacks, pandemic influenza, and emerging infectious diseases" through a range of initiatives.[57]

- **The National Biosurveillance Integration Center**, housed in the Department of Homeland Security's Countering Weapons of Mass Destruction Office. It manages and analyzes important information about biological events among agencies to help enable better informed responses.

A meeting of the Biological Weapons Convention (BWC) was held in Geneva in December 2014. The BWC is the primary multilateral agreement in effect to constrain the development, production, and stockpiling of biological weapons internationally. (Eric Bridie/U.S. Mission Geneva)

- **The Defense Threat Reduction Agency** in the Department of Defense, which works to deter, prevent, reduce, and counter weapons of mass destruction (WMDs) and emerging threats, including biothreats.

These purpose-built entities aim to directly address biological catastrophic scenarios, and to complement the broader efforts of the Centers for Disease Control and Prevention within the Department of Health and Human Services, which would take a leading role in addressing any biological catastrophe in the United States. National intelligence and law enforcement agencies also work to identify and prosecute actors who might try to build malicious bioweapons. But despite the wide range of organizations and institutions that work to mitigate biological threats to national security—and a reasonably clear picture of the sources and changing biotechnology capabilities that shape those threats—perceptions of the overall risks continue to vary considerably.

Perceptions of Risk

Assessments of the state of biological risks range considerably, most of all with regard to catastrophic bioterrorist threats. These worst-case scenarios represent the most extreme cases that often animate the biorisk conversation in public discourse, and as such are a fitting place to start to establish the broader contours of the debate about contemporary biorisks. Much of experts' divergences in opinion boil down to how they weigh different factors' influence in the likelihood of threats emerging, including the availability of bioweapons information, technological capabilities, and experimental experience.[58] For those who believe catastrophic bioterrorism poses a severe risk, the rapidly advancing information and technology available to potential bad actors make worst-case scenarios increasingly plausible. Such scenarios, like the strategic release of a highly lethal and contagious virus, could dwarf the impact of COVID-19. Others contend that the primary barrier to such worst-case scenarios is and always has been organizations' challenges in wielding specialized experimental expertise, which continue to greatly constrain the potential of would-be bioterrorists today. This challenge, they contend, continues to significantly limit the capabilities of potential bioterrorists. This section considers each factor in turn, before exploring the comparatively less controversial issue of American vulnerability to catastrophic biological events, whether man-made or naturally occurring.

Those who harbor strong bioterrorism concerns have warned for years that increasingly accessible biotech capabilities and widely available information on potentially catastrophic biological agents constitute a recipe for disaster. In this view, the United States is on borrowed time: the absence of major incidents in recent decades owes more to luck than to effective risk management. By comparison with Aum Shinrikyo or the perpetrator of the 2001 anthrax attacks, actors today have access to far more powerful resources and readily available information on how to make biological agents and on how biological weapons programs fail.

> **Much of experts' divergences in opinion boil down to how they weigh different factors' influence in the likelihood of threats emerging, including the availability of bioweapons information, technological capabilities, and experimental experience.**

In this view, new biological tools are easing the barriers to malicious actors building highly dangerous pathogens. Perhaps the clearest indication of this trend comes from a controversial experiment conducted in 2016, in which a private lab successfully constructed a horsepox virus from scratch by stitching together DNA segments that the lab legally purchased from a commercial company. At a cost of $100,000—a larger sum than would be necessary today—the lab was able to recreate from scratch a nearly extinct virus using entirely commercially available inputs. Troublingly, the horsepox virus is a cousin of the virus that causes smallpox—a disease that has been eradicated but that, if unleashed, would have catastrophic consequences due to its combination of high transmissibility and lethality, in a world of widespread lack of immunity.[59]

Tellingly, in the face of backlash against the execution and publication of this research, the principal investigator's primary defense was simply that there were no legal or informational barriers to conducting the experiment. Therefore, he maintained, the experiment itself did not substantially alter the risks of a malicious actor constructing smallpox by the same methods. Some have disagreed, arguing that clarifying the process and providing proof of concept are significant steps in the wrong direction.[60] Regardless of how much the experiment did or did not impact overall risks, it at least demonstrated that the door to engineering powerful, smallpox-like viruses is open, at least as far as technical capabilities are concerned. To make matters worse, Jeff Alstott, former director for technology and national security at the National Security Council, warned in September 2023 that the classified record contained "fairly recent close-ish calls" of nonstate attempts to produce and scale bioweapons for strategic use, suggesting that the likelihood of groups attempting to field bioweapons may be more severe than some imagine.[61] Taken together, proponents of this view argue that mounting risks of bioterrorism require immediate attention, lest—similar to COVID-19—experts' warnings go unheeded, with catastrophic results.

By contrast, those less concerned about recent technological advancements that might enable bad actors point to the historical failures by states, terrorist organizations, apocalyptic groups, and lone wolves to create powerful biological agents. This track record, they argue, suggests that the risks are exaggerated, or at least misguided. Even if the theoretical knowledge and materials needed to make dangerous bioweapons are freely available, the barriers to producing working, effective biological agents at scale are very high, as any pharmaceutical company knows well from the difficulty of developing and reliably producing biopharmaceuticals. Cultivating the organizational effectiveness and experimental expertise needed to develop, sustain, and deliver biological agents is extremely challenging, especially under the conditions of secrecy that bioterrorist actors require—not to mention the extreme sociological conditions under which extremist groups operate. And even if new technologies make biological tinkering more cost-effective and less onerous, learning to navigate new, delicate tools and systems that are constantly evolving tends to introduce new barriers to execution even as traditional ones diminish.[62] By this logic, the aforementioned horsepox example is a case in point: that the lab in question had years of accumulated organizational and experimental experience enabling it to recreate the virus was more central to its success than was the availability of the tools and commercial resources that resulted in the virus. A terrorist group or lone wolf actor would struggle to do the same, as they have historically.

This camp also stresses that the only U.S. incident of note in the past quarter century, the 2001 anthrax attacks, had a highly limited impact, despite the high-level concern on these issues for several decades.[63] If biological risks were as grave as supposed, the thought goes, the dearth of significant large-scale bioterrorism events in the past century, both at home and abroad, suggests these risks may be exaggerated—a symptom of overemphasizing technological capacity as the principal driver of risks to the exclusion of sociotechnical factors such as experimental expertise and organizational dynamics.[64] From this perspective, more alarmist concerns may also be inflected by biosecurity experts' incentives to stress worst-case scenarios, or, like experts in any field, an inflated sense of their work's importance.

There is undoubtedly truth to both of these views: however quickly biotechnology tools are improving, the likelihood of would-be bioterrorists successfully fielding bioweapons will continue to be constrained by the complexities of getting secretive organizations to conduct extremely delicate experimental processes effectively. And while history shows that such coordination and expertise are very difficult to achieve, it is also true that the availability of tools and information matters, and that dramatic changes to both in this sector will affect the ease with which nonstate actors can develop and use bioweapons.

This debate will likely continue to evolve as new biological tools emerge and proliferate.

This debate will likely continue to evolve as new biological tools emerge and proliferate. While useful to many actors, they may have outsized implications for the nonstate bioweapons threat. State and other advanced programs generally have greater existing capacity to work with hazardous biological agents, making the impact of new tools and information less dramatic. Consequently, discussions surrounding research accidents and state-level risks often revolve around familiar concerns like lab safety, regulations, and bioweapons policies. In contrast, more cantankerous debates about bioterrorism threats often focus on the accessibility of new tools and sensitive information, and their relative impact on risks.

But whether from malicious state or nonstate actors, scientific accidents, or naturally occurring diseases, there is growing agreement among many experts that the United States' current biodefenses are insufficient to effectively manage large-scale biological crises.[65] To a degree, COVID-19 bore this concern out: though Operation Warp Speed was able to greatly accelerate the speed of vaccine development and delivery, efforts to contain the virus were largely ineffective, in terms of both protecting the United States from contagion beyond its borders and within its society. If COVID-19 had been more lethal, or had a similar virus been strategically deployed by an adversary, the United States likely would have suffered more severe losses in lives, economic vitality, and strategic maneuverability, with little ability to shape outcomes.

Though there have been some efforts to address America's vulnerability in the wake of COVID-19, notably the release of the U.S. National Biodefense Strategy and Implementation Plan in October 2022 and the establishment of the National Security Commission on Emerging Biotechnology, some experts fear that actions remain incommensurate with evolving risks.[66] A range of organizations, including the Bipartisan Commission on Biodefense, the Nuclear Threat Initiative, and the Helena Institute, have advocated for the development of stronger measures to shore up American biological preparedness to cope with growing vulnerabilities from various potential originating sources.[67]

AI Safety and Biosecurity

To consider how AI will impact preexisting biological risks, this report draws upon its sister study, *Catalyzing Crisis: A Primer on Artificial Intelligence, Catastrophes, and National Security*, which distills the literature on AI and catastrophic risks into four dimensions of AI safety of relevance to high-impact domains:[68]

Dimension	Question
New capabilities	What dangers arise from new AI-enabled capabilities across different domains?
Technical safety challenges	In what ways can technical failures in AI-enabled systems escalate risks?
Integrating AI into complex systems	How can the integration of AI into high-risk systems disrupt or derail their operations?
Conditions of AI development	How do the conditions under which AI tools are developed influence their safety?

When applied to biosecurity, the most significant concerns around new capabilities center on the potential of AI-powered biological design tools (BDTs) to help develop more sophisticated biological weapons, and foundation models' improving abilities to potentially help bad actors create bioweapons more easily. In terms of technical safety challenges, the related challenge of effectively constraining foundation models' abilities to assist bad actors has dominated discussions so far, but there are other concerns worthy of note, including the development of AI tools with ill-understood risks for therapeutics development. The integration of AI tools into broader biological systems could have a distinct set of impacts on risks, both from safety challenges that tend to emerge with automation, and related to the reduction of tacit knowledge barriers for less-experienced actors. Finally, corporate and geopolitical competitive pressures are exerting significant influence on the development of AI and biotechnology, and this could shape safety outcomes, particularly in China, where other conditions of biological and AI development exacerbate risks.

New Capabilities

Emerging AI capabilities hold tremendous promise for the biological sciences in two ways. First, AI tools are uniquely suited to turbocharge synthetic biology by providing novel, powerful means to interpret and manipulate genetic information toward specific ends.[69] Though some of these capabilities are still matters of conjecture, there is good reason to think that AI holds tremendous potential to enable unprecedented advancements in biology and medicine—with significant implications for catastrophic risks. Second, AI foundation models may have the potential to amplify the biological capabilities of individuals with limited biological knowledge or expertise. Though current AI tools' effect in this area has been marginal so far, if such tools' biological capabilities continue to improve, they may increase risks from nonstate actors, albeit with some important caveats. This report considers each of these AI capabilities in turn.

AI AND BIOLOGICAL DESIGN TOOLS

Though it has long been understood that human DNA encodes various genetic diseases and contributing factors to diseases, it remains beyond existing capabilities to fully understand the diversity of constellations of DNA segments at the root of different conditions. At 3.2 billion base pairs in length, and with vast variations from person to person, the human genome contains too much information for scientists using conventional methods to understand the precise dynamics of genes' downstream effects. But such data-intensive, multivariable problems are precisely where AI excels: building complex models and detecting patterns and correlations across vast troves of data. AI holds tremendous potential to unlock unprecedented capabilities in the world of biology—in exploring not only the human genome, where much scientific research now focuses, but also the genetic material of pathogens and other organisms.

Together with advances in CRISPR gene editing methods and gene sequencing technologies, AI's likely ability to discern genetic patterns with greater precision could act as a watershed development in synthetic biology, allowing unprecedented precision in manipulating genetic information toward deliberate goals. AI has already enabled significant advances in solving complex biological problems, such as in protein folding, where AI has reduced the time it takes researchers to understand many proteins' shape from weeks or months to seconds, predicting structures with near-experimental accuracy.[70]

Researchers have also used machine learning to help identify modifications to viral capsids (the protein shells of viruses) to better evade the immune system. This is a potential boon for gene therapy using non-pathogenic viral vectors, but it also raises concerns about the possibility of transferring relevant knowledge or methods to the engineering of pathogenic viruses.[71] Similarly, researchers have used machine learning to develop models that predict the zoonotic and human-infectivity potential of viruses and pathogenicity of bacterial DNA, as well as mutations that help viruses such as SARS-CoV-2 overcome immunity—of potential use for targeting anticipatory research and surveillance but also with obvious potential for misuse.[72]

The AI tools used to accomplish these feats are narrow systems referred to as biological design tools and will be at the heart of the coming biological revolution.[73] Although at the moment BDTs' abilities are still nascent, future BDTs may hold the potential for highly sophisticated design or editing functionality—including for pathogens. Editing the genetic material of pathogens to achieve particular effects is not new, but the potential of BDTs to accomplish this feat with greater precision could augur a step change in biological threats.

Most concerningly, in principle it may be possible to design a more dangerous pathogen than has yet existed or that nature could produce on its own. Many biologists have suggested that there may be a naturally occurring evolutionary tradeoff between transmissibility and severity of diseases in naturally occurring pathogens. Some experts contest this hypothesis, and a host of factors and caveats complicate the idea, but reductively it suggests that because viruses rely on living hosts to spread, natural selection tends to diminish the severity of the most lethal pathogens over time.[74] Because a pathogen that is too severe will quickly die out, other less severe variants survive, multiply, and dominate. But if a BDT were used to modify or build a pathogen to optimize for lethality, transmissibility, and a long incubation period, in theory the resulting pathogen could transcend the natural pressures away from severity and result in a biological agent of unprecedented destructive power.

Such an AI-enabled "supervirus," or anything approaching it, would constitute a risk of catastrophe of the highest order. Given that the BDT or, more likely, BDTs able to produce such a pathogen would have to be extremely advanced, it is likely that should such a capability arise, it would first be available only to highly advanced biolabs and state actors. However, only the most deranged actors intent on causing maximum uncontrolled destruction would be motivated to deliberately create and release such a pathogen. Groups with such apocalyptic motivations exist, for example Aum Shinrikyo, and sufficiently advanced BDTs could in principle enable small groups or even individuals to design such a supervirus. Thus, the potential for AI to greatly escalate worst-case scenarios in biological catastrophes is—theoretically—vast.

However, advancing from current BDTs, such as those used for protein design, to future BDTs that can edit pathogens to produce novel, specific effects—if such BDTs are possible at all—involves complex steps. There would undoubtedly be significant hurdles to overcome. Ensuring that the bioweapon remains potent over time and through diverse geographies would also be a significant challenge, given how delicate pathogens often are. Similarly, it may be impossible to predict how such a pathogen would interact with human populations over time; it may regress to a less-lethal predominant variant. Even now, despite the considerable attention focused on COVID-19, virologists remain unable to reliably anticipate the impacts of new strains of the virus.[75]

Another possibility for AI-powered BDTs to alter biological risks would be helping to design pathogens with more targeted effects in specific geographic areas or genetic populations. In principle, there is reason to believe future BDTs with sufficient biodata could alter the horizons of possibility. Given that many viruses can only thrive under specific environmental conditions, such as temperature, humidity, and air pressure, it stands to reason that if AI can identify with greater precision exactly what elements of genetic information predispose viruses to environmental strengths and weaknesses, it may be possible to optimize biological agents to work in particular locales. More disturbingly, given that different genetic populations have differing susceptibility to some viruses, it may also be possible to optimize viruses to target specific populations or avoid others. Zhang Shibo, former president of China's National Defense University and former general in the People's Liberation Army, noted as early as 2017 that advanced biology techniques could enable new offensive capabilities, including "specific ethnic genetic attacks."[76] There are also some less-obvious high-impact applications of such capabilities, including the creation of pathogens genetically targeted to induce crop failures in a country's critical food supply chains, offering the potential to strategically disrupt adversaries' food security.

Like ultra-lethal superviruses, there remain many technical hurdles to overcome before viral engineering for geographic or genetic targeting is feasible. Arguably, such engineering is more tentative than superviruses, which have precedents in Soviet bioweapons and gain-of-function research, both of which made progress in enhancing the transmissibility or severity of viruses to humans without BDTs. There may also be unforeseen limits or tradeoffs to just how precisely biological agents can be targeted to either geographic conditions or genetic groups. Given that viruses mutate over time in unpredictable ways, ensuring that a BDT-engineered virus remains both potent and targetable would be a considerable—and perhaps intractable—problem.

Despite these caveats, if geographically or genetically targeted biological agents are ever achieved, the result will profoundly alter the incentives and deterrents to using bioweapons. For state actors, the *imprecision* of conventional biological agents has been the primary disincentive to employing them. A world in which such weapons could be targeted raises the specter of greater incentives to incapacitate enemy forces with a weapon that may be able to be administered with subtlety—at least initially—and to devastating effect.

Though the most consequential impacts of BDTs remain theoretical, the threat they pose raises the potential destructive capacity of biological agents dramatically from what was already a grave baseline. Existing biological agents such as smallpox, anthrax, and botulinum hold the potential to catalyze catastrophes with millions of victims. Biological agents optimized toward even greater destruction could achieve exponentially more devastation. Similarly, precision bioweapons would dramatically escalate the incentives for states and other groups to develop and deploy bioweapons. If such capabilities do emerge, advanced state actors or advanced laboratories will most likely be the first to realize the potential of BDTs to create specialized biological agents.

Given the degree to which BDTs could enable the most advanced biological actors to raise the ceiling of harm possible from biological agents, the introduction of advanced AI holds the potential to greatly expand the scope of biological catastrophic risks. However, these capabilities are still theoretical, and the timing and conditions under which such possibilities are realized—if ever—are exceedingly hard to predict, as is the extent to which BDTs will ultimately be able to achieve these prospective hazardous capabilities.

Critically, uncertainty about the timing and extent of BDT's contributions to catastrophic risks cuts both ways: some risk scenarios may prove to be unfeasible entirely; but some may arrive much more suddenly than expected. In one study, for instance, an American pharmaceutical company found that it was unexpectedly able to use an AI tool usually used to help develop medicinal drugs to instead design new potential chemical weapons (see "The MegaSyn Experiment," page 16). Similarly, some BDTs used for legitimate medical research could harbor unexpected hazardous capabilities that could be surfaced with minimal tinkering, though the greater complexity of biological agents compared to their chemical counterparts may make this less likely.

FOUNDATION MODELS AND DEMOCRATIZING RISK

If narrow-use BDTs hold the potential to dramatically escalate the impacts of biocatastrophes, general-purpose foundation models may also raise the likelihood of biocatastrophes by helping to diffuse relevant expertise to a broader population.

Foundation models are AI tools trained on a large corpus of data to accomplish a broad range of tasks.[79] Another common term is "frontier models," which are highly capable general-purpose models that could pose considerable safety risks.[80] Foundation models include large language models (LLMs) such as OpenAI's GPT-3.5, Meta's Llama 2, and Anthropic's Claude 2, all of which can dynamically engage with users in natural language to communicate information, generate content, and even build websites and programs. Many leading AI labs are also building multimodal models, which can engage users not only with written text, but also with photos, videos, and audio. The most widely used multimodal model today is OpenAI's GPT-4o, with which users on the web platform ChatGPT can engage via text, images, and audio.

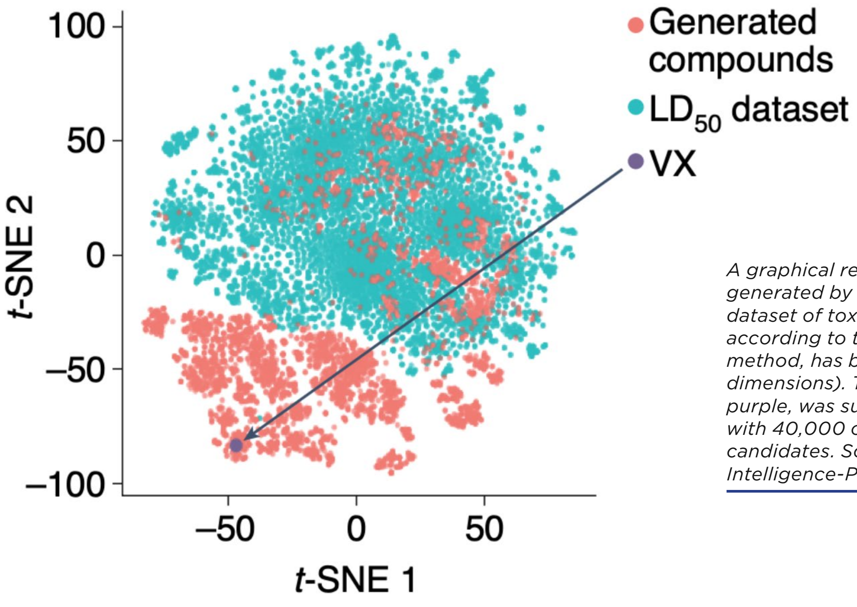
THE MEGASYN EXPERIMENT: A GLIMPSE INTO FUTURE BDT RISKS

Though the current risks of BDTs are limited, an experiment by a U.S. company offers an example of how AI can impact chemical weapons in a similar way to how BDTs may eventually impact biological weapons. In 2022, researchers from Collaborations Pharmaceuticals, a North Carolina–based pharmaceutical firm, published findings from an experiment in which they used a drug design AI model called MegaSyn to design, in less than six hours, 40,000 molecules with potential for use as chemical weapons. Though not all the system's proposed molecules would work as chemical weapons, MegaSyn was able to successfully generate several agents known to be highly effective in that capacity, including VX—one of the most toxic chemical agents known. There is also reason to believe that the model pioneered entirely new classes of neurotoxins with previously unknown weapons potential.[77]

As disturbing as the results of the study are, the unexpected ease with which the company was able to generate these results is perhaps even more unnerving. When a biosecurity conference invited the company to probe how drug discovery technology could be misused, Collaborations Pharmaceuticals' researchers wondered if they could simply use their MegaSyn drug discovery AI model. Usually they used the tool to avoid predicted toxicity in candidate drugs, but in this case they could instead optimize for it. By simply reversing the system's goal, and with virtually no AI engineering involved, the scientists were immediately able to generate 40,000 potential chemical weapons. They noted that many hundreds more companies use similar AI tools for drug design, and that their "commercial tools, as well as open-source software tools and many datasets that populate public databases, are available with no oversight."[78] Notably, the AI model in question was not particularly complex; it ran on a 2015 MacBook.

The MegaSyn experiment highlights practically how information technology is easing the accessibility of dangerous information, and the theoretical potential of BDTs to enable similar breakthroughs in biological weapons. AI tools such as MegaSyn can help avoid the toxicity of potential drugs and are an accelerant for more effective healthcare. But these tools are dual use. If they can avoid toxicity, they can also pursue it. Given how focused the creators of these tools are on medicinal uses, they may not even recognize latent, weaponizable capabilities. The researchers behind MegaSyn, for example, were surprised—and alarmed—by the ease with which their tool could generate toxic chemicals and thought to attempt it only after being prompted to do so. If well-intentioned researchers in the future can use BDTs to alter virulence and transmissibility in genetic code, they may also inadvertently create novel capabilities to design more destructive pathogens. As in the MegaSyn case, repurposing AI systems designed with good intentions into weapons may be an unnervingly easy proposition.

Biological weapons are considerably more complex than chemical weapons, making the capabilities currently available in chemical applications only a loose analogy for prospective capabilities in biology. Additionally, designing bioweapons is only one step in a much broader and more complex process of finding ways to produce, store, and disseminate viable bioweapons. Even so, the MegaSyn experiment provides a precedent for exactly the sort of risks that many fear from future, more capable BDTs.



A graphical representation of a selection of molecules generated by MegaSyn (in salmon) and an existing dataset of toxic molecules (in turquoise), clustered according to their structural similarity (t-SNE, a statistical method, has been used to simplify the clustering to two dimensions). The chemical weapon VX, represented in purple, was successfully generated by MegaSyn along with 40,000 other potentially viable chemical weapon candidates. Source: Urbina et al., "Dual Use of Artificial Intelligence-Powered Drug Discovery."

Among general-purpose AI systems' many capabilities is the ability to interactively distill scientific information, including biological information, into actionable steps to achieve particular experimental results. While the capabilities of today's general-purpose AI systems are relatively limited in this regard, some experts fear that taken to their extreme, future, more capable foundation model systems could guide bad actors to build powerful biological agents. If so, making such tools widely available could dramatically expand the pool of individuals and groups able to cause a biological catastrophe.

Some experts view the risks of foundation models helping bad actors develop a bioweapon as a pressing, even urgent, issue. Unlike sophisticated BDTs, general-purpose large language models already offer proof of concept in helping to accelerate dangerous biological activities, albeit with marginal, if any, benefits for success when compared with conventional internet assistance. Five recent experiments hint at the extent to which existing LLMs could accelerate bad actors' acquisition of dangerous biological agents.

First, in April 2023, researchers at Carnegie Mellon University created a system of interconnected LLMs that—with access to the internet, code execution capabilities, hardware documentation, and remote control of an automated cloud laboratory—was able to achieve a surprising level of experimental proficiency. The system was "capable of autonomously designing, planning, and executing complex scientific experiments" without human intervention, the most complex of which included successfully performing a cross-coupling reaction, a chemical process of several steps that would ordinarily require significant chemistry expertise.[81] The system agreed to autonomously synthesize a common date rape drug and phosphene, a chemical weapon used in World War I. Only after a web search did the system refuse to synthesize other concerning compounds, including methamphetamine; sarin; and VX, an extremely toxic nerve agent (the system could autonomously search the internet to gain information). Though the experiment required considerable technical expertise to create the system, the long-term ambition of many leading AI labs is to develop general-purpose foundation models with even greater capabilities in scientific experimentation. Likewise, though the results of this experiment were limited to chemical agents—generally much simpler to produce than biological agents—some experts fear that rapid improvements in general-purpose AI systems in combination with rapidly improving biological cloud labs could mean that similarly powerful systems could soon be produced that are able to experiment with biological agents.

Second, in a June 2023 paper, MIT researchers explored how LLMs might assist nonexperts in causing a pandemic by having nonscientist students use the models. In one hour, the LLMs proposed four potential pandemic-causing pathogens, explained how they could be created from synthetic DNA, suggested several DNA synthesis companies that were not likely to screen DNA orders, and proposed detailed protocols to assemble the pathogens—including troubleshooting measures. The researchers argued that the "results suggest that LLMs will make pandemic-class agents widely accessible as soon as they are credibly identified, even to people with little or no laboratory training."[82] Notably, the students were able to circumvent the safety measures in place on some of the platforms they were engaging with to access the sensitive information. Three important caveats mitigate these seemingly alarming results. First, having information about building pathogens is only the initial step in successfully building said pathogens. Additionally, one of the pathogens proposed was almost certainly too complex to be feasibly produced by amateurs, and others may not pose much of a pandemic threat due to preexisting immunity, even if they were achieved.[83] And finally, while the information culled from the LLMs would have certainly accelerated bad actors' progress, the same information is readily available on the internet, though it might take more time to locate and distill into actionable plans.

Results of a third experiment were published in July 2023, when Anthropic, a leading AI company, released an overview of an internal testing of their LLM for biological risks. Though they did not publish a detailed methodology, their study involved spending 150 hours with biosecurity experts to test their model's ability to communicate biological information with potentially dangerous applications. They found that "current frontier models can sometimes produce sophisticated, accurate, useful, and detailed knowledge at an expert level," though infrequently in most areas.[84] They also suspected that "models gaining access to tools could advance their capabilities in biology."[85] Ultimately, this led the investigators to conclude that if left unmitigated, within two to three years, "LLMs could accelerate a bad actor's efforts to misuse biology relative to solely having internet access, and enable them to accomplish tasks they could not without an LLM."[86] Though this implies that current improvements over internet access are at most marginal, the authors noted that as general-purpose AI systems improve, their biological expertise also expands, suggesting that in the years ahead, AI systems will likely offer a substantive edge over traditional internet-based research.

Fourth, in January 2024, the RAND Corporation released the results of a study on the risks of large-scale biological attacks. To assess such risks, RAND tasked 14 small teams of researchers with devising operational biological attack plans. Each team was given a maximum of 80 hours of effort per team member over the course of seven weeks to craft viable biological attack plans for large-scale effects. To evaluate the relative impact of LLMs, four control groups were given access only to the internet and were forbidden from using LLMs to augment their efforts. The resulting plans were graded by experts on both operational and biological feasibility. The researchers found that there was "no statistically significant difference in the viability of plans generated with or without LLM assistance."[87] It should be noted that the tested LLMs included safeguards, meaning these results may not be applicable to users given access to the raw models without safeguards (see "Technical AI Safety Challenges," page 19).

Fifth, OpenAI released results from a study in May 2024 examining whether GPT-4 could significantly enhance access to information necessary for creating bioweapons compared to internet access alone. The study involved 100 participants, divided into expert and student cohorts, each randomly assigned to groups with either GPT-4 access or internet-only access. Researchers observed slight improvements in metrics such as accuracy and completeness, with experts experiencing approximately 0.9-point increases on a 10-point scale. When analyzing specific subtasks, no individual task showed statistically significant increases after controlling for multiple comparisons.[88] However, the authors noted that if they had assessed the aggregate uplift in accuracy across all tasks, the result would have been significant.[89] Contrasting this with the RAND study, which reported an unambiguous negative result, the OpenAI researchers emphasized methodological differences, such as using a model without safety guardrails, a larger sample size, and varied task designs—suggesting these factors might explain the discrepancy.[90]

These research efforts—the OpenAI study in particular—suggest that while today's systems at most only marginally impact biological risks, foundation models may soon be able to improve the information available to malicious actors seeking to acquire bioweapons, especially nonstate groups. Such a development may be significant, but its effects can be easily underestimated or exaggerated, albeit for different reasons. To properly assess the import of LLMs' potential for nonstate bioweapons threats, it is best to consider them in the context of the previously mentioned debate about the perceptions of bioterrorism risks (see "Perceptions of Risk," page 11).

To focus first on how the impacts of LLMs can be underestimated, it is important to appreciate that the information needed to develop a bioweapon is far more complex than a simple list of instructions. The value of foundation models in crafting bioweapons is not just in distilling complex biological information scattered across a wide range of sources into actionable steps, though that alone could represent a significant advantage over conventional methods that make use of the internet. Of equal importance is the budding ability of general-purpose AI systems to help triage how experimental processes have gone wrong, accelerating the design-build-test-learn feedback loop that is essential to developing working biological agents.[91] To use a loose analogy, getting a good recipe for a rare, delicate baked good may be difficult, but even more important is having an experienced chef to help inform how and why one's attempts at following the recipe did not go according to plan. Foundation models may hold the potential to eventually provide both recipes for bioweapons and expert advice on missteps in following those recipes. AI systems have already demonstrated an ability to successfully triage and correct where chemical experiments have gone wrong, suggesting a precedent for future biological experiments.[92] As foundation models move from language to more multimodal capabilities—able to interpret visual and other inputs—their ability to help triage where experiments have gone awry will likely grow. This enhanced assistance could be important across the entire bioweapons lifecycle, including the processes of effectively storing, sustaining, and deploying bioweapons.

At the same time, however, there are often-overlooked limits to what sophisticated foundation models could contribute to nonstate actors' biological production efforts. First, as previously mentioned, information availability is often overemphasized as an element of bioweapons production to the exclusion of other important factors that can pose barriers to success such as organizational characteristics.[93] Regardless of how dynamically foundation models are able to convey information to aspiring bioterrorists, the sociological conditions that frame their efforts may still make it exceedingly difficult to effectively use that information, especially across a multifaceted and lengthy bioweapon lifecycle. Additionally, for the foreseeable future, foundation models simply will not be able to wield certain types of information that are often essential for successful experimentation. Some tacit knowledge, for example, is not articulated or absorbed verbally, and ordinarily must be passed on through apprenticeship or developed through accumulated trial-and-error experience (for a more thorough exploration of these themes, see "Tacit Knowledge," page 21).[94] To stretch the baking metaphor, an experienced chef who can correct errors is much more helpful than a simple recipe, but if that chef is nonetheless unable to physically coach apprentices through the delicate, somatically intensive techniques required for success, success may never be fully achieved.

Taken together, to the extent that foundation models prove to be a revolution in information assistance generally, they may also have the potential to provide a revolution in information assistance for nonstate actors' bioweapons development. But to the extent that actually developing and deploying bioweapons depends on several other indispensable factors, even a revolution in bioweapons information assistance is unlikely to directly equate to a revolution in bioterrorists' capabilities.

> **But to the extent that actually developing and deploying bioweapons depends on several other indispensable factors, even a revolution in bioweapons information assistance is unlikely to directly equate to a revolution in bioterrorists' capabilities.**

Finally, in addition to concerns about foundation models in isolation, it is also important to consider their evolving relationship with the risks of BDTs. As foundation models become more sophisticated, some experts believe that the issues associated with general-purpose AI systems and BDTs may converge. This could occur either because general-purpose AI systems begin to acquire BDT-level expertise in synthetic biology in their own right, or, more likely, because they lower the barriers to using BDTs by acting as an interface that allows users to wield BDTs much more easily. In a worst-case scenario, general-purpose foundation models may eventually compound the risks of BDTs, so that BDTs raise the ceiling on the destructive potential of biological catastrophes, while foundation models expand the number of individuals who can wield advanced BDTs or BDT-like capabilities. Such a scenario is far from certain, given the uncertain trajectory of both technologies, including new safety mechanisms that may emerge before such a convergence of risks.

Technical AI Safety Challenges

Technical safety challenges intrinsic to AI tools could exacerbate biological risks in a variety of ways. The most obvious relates to foundation models' potential to accelerate bioweapons production, as explored in the previous section. To mitigate the chances of their systems being used for malicious purposes, AI developers have attempted to create guardrails within AI systems that would prohibit their use for nefarious ends. But even while industry-leading foundation models are trained to refuse dangerous or harmful requests, foolproof techniques to reliably constrain systems' outputs have yet to be developed. With sufficient effort, models can be induced to bypass safeguards through intentional crafting of "jailbreaks" or "adversarial prompts," requests that are able to fool the system into ignoring instructions or training to refuse to answer certain questions.[95]

The difficulty in developing robust safeguards for advanced foundation models is related to the methods by which they are trained—using gargantuan data sets culled from the internet and elsewhere. Because these datasets are so large, the full extent of the knowledge that the models can acquire is difficult to identify, let alone control. The leading method to constrain outputs of information that creators would rather not be relayed is reinforcement learning from human feedback (RLHF). This technique iteratively fine-tunes a model based on human ratings of how its outputs meet the relevant policies and objectives of the creator—including not revealing harmful information. While this method works to a degree, it is a surface-level fix to the deeper issue. It reduces the model's tendency to produce harmful results, but by and large does not alter the fundamental capabilities of the system, thus allowing clever tinkerers to find other ways to coax the desired illicit information from the model. While many advanced general-purpose AI system providers allow their tools to be accessed only behind an online user interface—in effect barring users from tampering with the system itself—some, such as Meta and Hugging Face, give users direct access to the raw models themselves. As researchers have demonstrated, users can then functionally undo the RLHF safeguards that would otherwise constrain the outputs of these models at low cost.[96]

Researchers are working on more sophisticated ways to guide foundation models' outputs using techniques to instill more robust guardrails.[97] But as foundation models become more complex, including by integrating multimodal data, it is possible that the difficulty of ensuring reliably controllable outputs may rise commensurately.[98]

The difficulty of constraining AI systems' outputs has understandably dominated discussion of AI and biorisk, but a range of other general technical AI safety issues could have impacts on biotechnology risks. These include explainability, over- and underfitting to training data, challenges in generalizing beyond the training distribution, and failures to ground predictions with causal mechanisms and real-world physics.[99] Such issues can certainly pose threats to the usefulness of systems, and can result in clinical dead ends or patient harm. But they are unlikely to contribute to full-scale catastrophic scenarios.

One possible exception could be systems that are effective at identifying promising biological candidates for medical or therapeutic uses but unreliable in predicting the full impacts of such candidates. For example, in 2001 researchers inserted a gene for interleukin-4 (a protein that supports immune response) into the mousepox virus, in hopes of finding a way to disrupt the reproductive systems of mice. But the resulting agent was an unexpected, highly lethal variant of the virus that killed all the mice initially exposed, and subsequently half of those vaccinated with an otherwise highly effective vaccine.[100] In this case, humans had successfully pinpointed a seemingly promising research agenda but failed to anticipate the lethal end result.[101] AI systems with a similar blend of strengths and weaknesses could make it easier for less-cautious actors to accidentally stumble upon dangerous biological agents, despite intending to develop medicinal products. While such incidents are not AI-specific, AI's ability to identify potential research paths at superhuman speed, for reasons opaque to operators, could exacerbate such risks.

Integrating AI Tools into Complex Systems

As with other complex systems, such as the aviation industry or military systems, the integration of AI tools into the broader biotechnology ecosystem can reshape risks. In biotechnology, these phenomena can be thought of in two broad categories. First, enhanced automation in a range of subfields can introduce new safety challenges through eroding technicians' sensitivity to operations in their labs and disrupting the informal safeguards and incentives that accompany conventional lab work. Additionally, AI's continued integration into biological processes may impact the role of tacit knowledge in acting as a barrier to bioweapons production, though the degree of this impact remains to be seen. These issues arise from AI's ongoing, albeit nascent, transformation of experimentation in fields such as gene editing and biomanufacturing, both part of a wider trend toward automation.[102]

AUTOMATED PROCESSES

According to a range of organizational management studies, increasing automation in complex systems can risk introducing a range of safety hazards—and biotechnology is no exception.[103] For one, as biological experiments become more automated, researchers' abilities to maintain robust sensitivity to operations of delicate and potentially dangerous processes could be adversely impacted. Research into high-reliability organizations—those with remarkably strong safety records—shows that the ability to maintain a comprehensive, real-time understanding of the full complexity of an organization's ongoing operations is critical to avoid errors and accidents, and to ensure that mistakes do not snowball into catastrophes.[104] In cases such as the U.S. Navy's Submarine Safety Program, leaders have intentionally designed their systems to ensure that human operators have active and deliberate oversight of the most critical parts of processes for the sake of ensuring operators' active awareness.[105] New automation capabilities in sensitive experimental processes in biology, without careful thought to how to maintain robust situational awareness among operators, also run the risk of eroding experimenters' sensitivity to operations, potentially increasing the chances of lab accidents with catastrophic potential.

> **According to a range of organizational management studies, increasing automation in complex systems can risk introducing a range of safety hazards—and biotechnology is no exception.**

AI tools and automation could also reduce the influence of suppliers and experts who traditionally provide formal and informal feedback and oversight to research activities. This shift could make research more efficient and accessible, but it could also make it more difficult to monitor and evaluate, and to intervene as risks or bad actors present themselves. Although suppliers and experts ostensibly fulfill narrow tasks and processes within a bioscience ecosystem that could, in principle, be automated, the loss of their broader contextual awareness in performing tasks would be significant. It could mean that anomalies, irregularities, or suspicious behavior they might ordinarily notice would go undetected by automated systems.

Finally, automation could make certain lab procedures more precarious by disincentivizing safeguards and escalating the impacts of system failures. As automation enables more high-efficiency approaches to execute lab work at scale and biotechnology engages in more industrial approaches to experimentation, safety checks may or may not scale appropriately to the expanding throughput. Additionally, stringing together multiple processes within larger automated procedures can result in "tightly coupled" systems—that is, systems in which mistakes or failures in one area quickly spread to others as interdependent automated processes rapidly affect one another. In more conventional lab processes, scientists perform subtasks independently, with individuals able to intervene in the case of failures from one task to the next. In more tightly coupled lab systems, where scientists are less directly handling experimental tasks, flaws can have cascading or compounding effects as the process progresses. These automated processes often occur at machine speeds—faster than humans can reliably follow—exacerbating the risk.

Much of the difficulty in addressing the issues associated with automation stems from the fact that they can be exceedingly hard to identify, vary from lab to lab, and in many cases may not manifest at all. Additionally, it is not the case that the net effect of automation on safety is necessarily negative; to the contrary, in industries such as aviation and healthcare, it has a significant safety-boosting effect.[106] The same may be true of biology. Nonetheless, labs and other biotechnology service providers should pay close attention to the formal and informal safeguards that may be lost as they integrate greater automation into their systems, especially in terms of situational awareness, human oversight, and tightly coupled system processes.

TACIT KNOWLEDGE

Historically, tacit knowledge has acted as a major bottleneck to successful biological development for scientists and bad actors alike. To understand how and why that is the case, it is important to recognize that tacit knowledge comes in many forms that are difficult to disseminate for different reasons. To that end, researchers James Revill and Catherine Jefferson explain tacit knowledge in terms of three broad categories, with subcategories in each:[107]

Weak tacit knowledge is that which in principle could be shared, but is difficult or impractical to communicate effectively between individuals or across organizations. For example, many organizations, including biolabs, employ logistically minded individuals who have such an intimate and wide-ranging understanding of the behind-the-scenes systems sustaining their organization's operations that they can bring together information or resources to problem-solve in unique ways that can be lost if they leave the organization. Such knowledge can be difficult to recover without a replacement who has accumulated years of experience in the organization—even when its custodians attempt to communicate it—because the knowledge is experiential in nature. Another example is tacit knowledge that the bearer is unaware of having, so that in attempting to explain how to execute complex processes to others, he or she inadvertently fails to communicate information necessary to complete the task. Such "unrecognized" knowledge may seem simple enough to surface, but it can be exceedingly difficult to identify or articulate practices that have been subconsciously internalized, including in experimental processes.

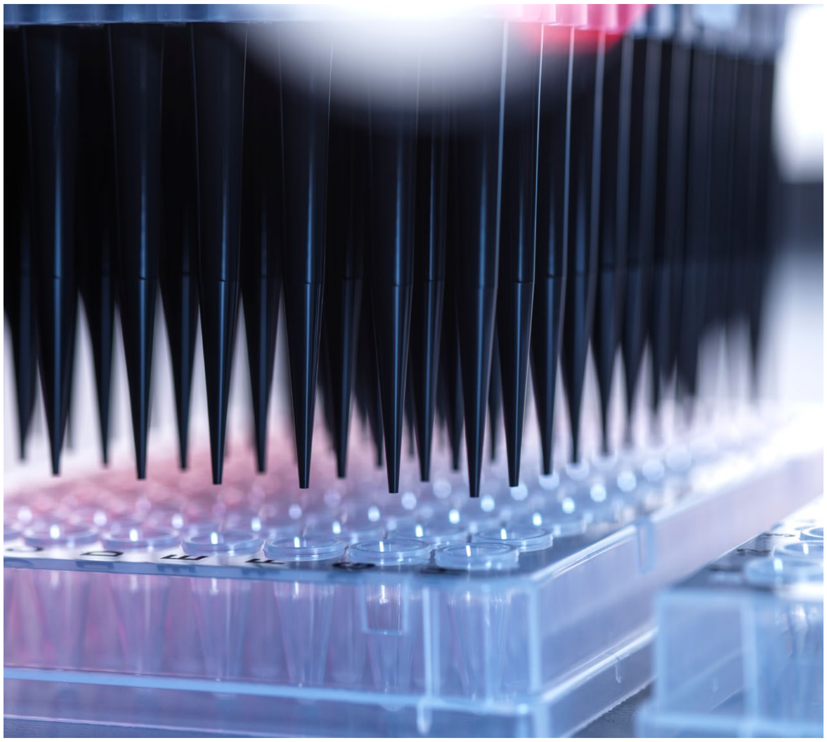
Somatic tacit knowledge refers to subtle physical information that can be impossible to articulate and must be learned by doing, for example balancing on a bicycle. Such "muscle memory" information can be critically important to successful biological experimentation and can be as subtle as different techniques to swirl, pipette, or pour solutions, but also includes highly delicate lab procedures that require specialized techniques.[108] Knowledge such as this often cannot be disseminated without in-person instruction, a fact that has been demonstrated time and again in "do-it-yourself" biology communities' experiences trying to conduct their own experiments.[109]

Communal tacit knowledge refers to collective expertise produced and accumulated by teams of specialists working together over time. Just as sports teams practice together repeatedly to develop plays, tactics, and rhythms that make effective use of each player's skills and roles, so too experts with diverse technical skills and roles must learn to combine their efforts effectively through iteration when conducting complex biological experiments. Likewise, just as trading a single key player on a sports team can have outsized impacts on the team's overall cohesion and effectiveness, disruptions to scientific teams can also dramatically impact results, especially in such delicate experiments as those required to wield biological agents with strategic applications.

In light of the diversity in forms of tacit knowledge and their impacts, it is not difficult to see why tacit knowledge has historically acted as a significant barrier to success in biotechnology. As previously mentioned, there have been efforts in biotechnology to leverage automated cloud labs to overcome some of the challenges associated with tacit knowledge (see "Evolving Capabilities," page 7). Many of these efforts have centered on attempting to record experimental knowledge, including some tacit elements, in excruciating detail. These efforts are likely to accelerate as AI tools and robotics mature, because operating them necessitates greater articulation and codification of elements of experimental expertise previously considered tacit knowledge. The National Science Foundation's BioFoundries program, for example, looks to support the development of "novel technologies, workflows, processes, automations, and knowledge-bases" to ensure "reproducibility of results and the ability to share data in both human- and machine-usable formats."[110]

Some experts see great potential for these and future efforts to leverage AI in labs to greatly reduce the challenges to experimental success that tacit knowledge has presented in the past.[111] Automation already shows promise for increasing the benefits of encoding some forms of tacit knowledge in machine-readable formats and the ease of doing so. Biotechnology companies seeking to capitalize on these trends aspire to build positive feedback loops in which they gather detailed experimental data, enabling more efficient experimentation and in turn generating still more experimental data to work from, and so on.[112] To the degree that such efforts are successful, automated labs could make complex biological lab work more accessible to a wider population—including bad actors. In such a future, some of the barriers to producing bioweapons could shift from the need for tacit experimental knowledge to safeguards established by automated labs in order to restrain bad actors from accessing dangerous biological agents or easily weaponizable antecedent ingredients for particular agents.

Nonetheless, many forms of tacit knowledge are very unlikely to be automated by AI, at least anytime soon. Some variants of weak tacit knowledge are the most likely candidates for automated assistance. And progress has been made in some basic forms of somatic tacit knowledge, most obviously pipetting, which can be finicky to perform manually in certain experimental circumstances, but which automated labs have made considerable strides in mechanizing.[113] Yet it is difficult to imagine AI resolving more sophisticated challenges associated with somatic tacit knowledge, let alone issues of communal tacit knowledge. Additionally, operating automated machines that ostensibly reduce the burdens of tacit knowledge sometimes creates new tacit knowledge barriers. For example, labs may feature only a single individual able to successfully operate a particular temperamental automated pipetting system, who may have spent years figuring out how to make it work at all.[114] Indeed, in some ways automated labs require more sophisticated tacit knowledge to operate, even if the lab customer need not wield that knowledge, because technicians must straddle highly complex biological and mechanical skills to maintain operations.[115] Much as these technologies are rapidly developing, it is still the case that there is a long way to go before automated labs reach the seamless level of automatic experimentation they aim to achieve.[116] As AI technologies progress, they will likely further alter the types of tacit knowledge required for biotechnology, perhaps by centralizing the tacit knowledge needed to perform sophisticated experiments in increasingly automated labs. But how such developments unfold—and how they will shape the risks of biocatastrophes—will remain an ongoing, evolving question.



Automated pipetting systems are one example of tools that can make labs more efficient, but which can also require new forms of tacit knowledge to operate. (RF/ Andrew Brookes via Getty Images)

Conditions of AI Development in Biotechnology

The risk profile of AI-enabled biotechnology will inevitably be affected by the conditions under which these capabilities are being developed. At present, these include rapidly expanding investment in the AI-bio nexus, a shift from academic to industry AI research leadership, escalating competition with China on both biotechnology and AI, and, perhaps most concerningly, Chinese AI and biotechnology ecosystems that lend themselves to costly errors and crisis mismanagement.

There has been a remarkable surge of investment in the AI-biotechnology nexus in recent years. Annual investments in the AI-driven biotechnology space grew nearly tenfold between 2017 and 2021, estimated at a minimum of $10.3 billion in 2021 alone.[117] Leading biopharmaceutical companies, such as BioNTech and Eli Lilly, each have invested hundreds of millions of dollars in partnerships and acquisitions in the past year.[118] While such investment augurs promising potential for the future of AI capabilities in biotechnology, the rapid flood of capital into the sector could induce acceleration and competition dynamics at odds with safety.

In tandem with the influx of investment in the sector, there has also been a shift toward industry-led development in the wider field of AI research, with the private sector now producing most leading machine learning models.[119] In 2011, PhD graduates in AI were almost equally likely to pursue careers in academia or industry. By 2021, 65 percent of AI PhDs headed to industry, with only 28 percent remaining in academia.[120] The life sciences have followed this shift toward industry-led AI biotechnology research, exemplified by Google DeepMind's creation of AlphaFold, an AI model that trounced competitors using more conventional methods in a 2018 competition to predict protein structures. Subsequent state-of-the-art systems to predict protein structures have all been built upon AlphaFold.[121] This shift toward industry-led AI-enabled biotechnology could have significant implications for how the sector evolves, given the divergent incentives that animate academic and corporate research. Additionally, private development of AI biotech could insulate the sector from some forms of government oversight and direction relative to academic research, which draws more from government funding. For example, the genome sequencing requirements of the White House's recent Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence apply only to government-funded biological research, leaving privately funded research unaffected.

Very few researchers embark on projects with the intention of contributing to catastrophic risks. To the extent this does occur, it is often the product of overlooking or underestimating the potential negative consequences of their work. In most cases, dual-use research holds legitimate potential for positive contributions to medicine. For example, gain-of-function experiments to enhance pathogens' ability to infect hosts can build understanding of viral mutations and transmission, and aid in the development of vaccines and therapeutics. Following controversy regarding gain-of-function research into making the H5N1 bird flu spread more easily between mammals, National Institute of Allergy and Infectious Diseases Director Anthony Fauci highlighted that such research could help predict, prevent, diagnose, and treat future pandemics.[122] In particular, understanding how existing pathogens may evolve to resist countermeasures or spread more easily can help future-proof efforts to develop vaccines and therapeutics.[123] The National Science Advisory Board for Biosecurity's recent report on enhanced potential pandemic pathogens and dual use research of concern, written in the wake of concerns over COVID-19's potential lab origins, reiterated that "life sciences research involving pathogens serves a critical role in pandemic preparedness and ensuring that the United States and the global community are prepared to rapidly detect, respond to, and recover from biological threats, whether naturally occurring, accidental, or deliberate in origin."[124]

> **The extent of harmful applications of well-intentioned research is not always clear at the outset—particularly with AI, given its rapidly developing, sometimes surprising capabilities.**

But as with the MegaSyn case (see page 16), the extent of harmful applications of well-intentioned research is not always clear at the outset—particularly with AI, given its rapidly developing, sometimes surprising capabilities. Coping with the often-opaque risks that emerge from the competitive dynamics of AI-enabled biotech development will involve reckoning with a range of thorny incentives related to potentially large profit margins, academic prestige, legitimate concerns over stifling innovation, and fears of losing America's biotechnological advantage over China. It may also require reckoning with uncomfortable realities, such as the fact that even in more obvious cases of harms from research, such as lab leaks, responsible parties rarely bear the full costs of harms themselves.[125]

One of the most underappreciated drivers of risk related to the conditions of AI and biotech development is that so much of it occurs in the People's Republic of China (PRC), where a range of factors make life sciences and AI research more high risk.[126] Beijing is heavily investing in AI and biotechnology, with stated goals to surpass the United States in AI leadership by 2030, and in biotechnology by 2035.[127] To this end, China has taken an aggressive approach to procuring genetic data en masse from its citizens and foreign nationals around the world through legal and illicit means, spurring what *The Washington Post* terms a "DNA arms race."[128] Given the centrality of genetic data to AI-powered progress in biotechnology, China's approach could yield considerable dividends for the country's ecosystem—and further escalate Sino-American biotech rivalry. Adding to these issues, the U.S. State Department has expressed "concerns" about China's compliance with the Biological Weapons Convention, citing activities including "PRC military medical institutions' toxin and biotechnology research and development [with] dual-use potential and possible [bioweapons] applications."[129] The Department of Defense's *Biodefense Posture Review*, in addition to citing the Department of State's concerns, likewise expressed concern that the "PRC has also released plans to make China the global leader in technologies like genetic engineering, precision medicine, and brain sciences."[130] China is, in short, fueling competitive pressures in biotechnology competition, including in some particularly dangerous areas.

In addition to how China impacts international competitive pressures, its biological and AI ecosystems are also uniquely prone to consequential failures due to the government's heavy-handed approach to accelerating scientific development, its willingness to support controversial and risky experimentation, and its chronic mismanagement of crises. China's government-led sprints to catch up or surpass other countries in particular sectors—as it is attempting to do now in AI and biotechnology—have a history of backfiring badly, as in the Great Leap Forward, the commercial satellite launch industry, and a variety of Belt and Road infrastructure projects.[131] The Chinese Communist Party (CCP) has demonstrated its willingness to support highly controversial biological research, as when it permitted and initially lauded Dr. He Jiankui's botched genetic editing of three human embryos brought to term in 2018. The Chinese government's funding incentives related to science and technology also particularly lend themselves to risky approaches to new technologies.[132] Though there is evidence of increasing AI safety consciousness in some circles, Chinese AI entrepreneurs have historically taken pride in their government's large appetite for stomaching tech development risks.[133]

Taken together, China's history of crises and the current conditions of its high-tech sectors suggest that Beijing's bid to lead the world in biotechnology and AI is a recipe for disaster.

Making matters worse, the Chinese government also chronically mismanages crises—not least in relation to deadly disease outbreaks—as exhibited in the gross mishandling of spiraling HIV contamination from blood sales networks throughout the 1990s, the 2002–03 coverup of China's SARS outbreak, and the severe mismanagement of the early weeks of COVID-19.[134] The latter is all the more remarkable because after the SARS fiasco, the Chinese government invested $850 million in developing public health mechanisms specifically designed to avoid SARS-like coverups, but the government's response ended up similarly dysfunctional.[135] The safety track record of Chinese labs is also not reassuring. In 2019, one lab in Lanzhou was the source of history's largest known lab leak, in which aerosolized Brucella infected more than 10,000 individuals in surrounding areas.[136] Between February and April 2004, another lab in Beijing was the source of four separate known SARS leaks, two of which were discovered only after international investigations were launched.[137] COVID-19, too, may very well have been the result of a lab leak, if preexisting U.S. State Department concerns about safety practices of the Wuhan Institute of Virology are any indication.[138] In either case, the Institute's continued concealment of its virus sequence database, the CCP's suppression of early information about the epidemiology of COVID-19, and the Chinese government's overt campaign to spread disinformation globally about the origins of COVID-19 all suggest a continued prioritization of political interests over biosecurity.[139] The fact that the largest and most advanced gene synthesis company that is not a member of the International Gene Synthesis Consortium is located in China similarly bodes ill for the country's attitude toward biosecurity.[140] Taken together, China's history of crises and the current conditions of its high-tech sectors suggest that Beijing's bid to lead the world in biotechnology and AI is a recipe for disaster.

Concerning as some of the current conditions of AI and biotechnology development are—in China specifically but also more broadly—there have also been a variety of efforts among biological labs, AI labs, and governments to create conditions more hospitable to biosafety. China's exception notwithstanding, nearly all of the most advanced gene synthesis companies have committed to instituting screening protocols for orders as members of the International Gene Synthesis Consortium, even if there is currently no common standard for how to conduct such screening.[141] Several leading frontier AI labs—including OpenAI, Anthropic, Google, and Meta—have committed to internal and external red teaming around biorisks for future frontier model releases, with OpenAI and Anthropic providing substantial commentary on risks from their flagship models.[142] The Biden administration's AI executive order has reinforced these measures and has sought to lay the foundation for a safety-conscious approach to innovation. The UK's AI Safety Summit saw preliminary progress among 29 nations—including the United States and China—toward addressing AI risks internationally, with particular emphasis on biorisks.[143] Though these initiatives are positive steps, and to varying degrees exhibit a desire to curb the worst impulses of competitive dynamics, none guarantees that safety concerns and competitive pressures will settle into an appropriate, sustainable equilibrium. As these technologies progress over time, actors could descend into a more aggressive competition, similar to how the sudden success of ChatGPT pushed competitor companies to compromise on their own safety standards.[144] The competition at the intersection of AI and biotechnology will require careful monitoring and dynamic efforts to ensure that all actors maintain adequate incentives for safe and responsible scientific development.

Capabilities to Monitor

Given the complexity of ongoing, interrelated developments in AI and the life sciences, there is inherent uncertainty as to when different risks at the nexus of AI and bio will emerge and under what conditions, if at all. Compounding this difficulty is the wide range of opinions about the relative threats that different capabilities pose, and a debate about whether the lack of successful large-scale biological attacks in recent years reflects overhyped risks or simply good luck.

Rather than attempting to predict the nature and timing of particular threats, this report instead seeks to identify particular areas of technological progress that could result in substantial alterations to catastrophic risks, albeit perhaps in unexpected ways. Building on the AI safety dimensions explored in this study, the following capabilities demand ongoing monitoring by policymakers to accurately understand what would change the prospects of a biological catastrophe enabled by AI—and how to address such risks.

General-purpose AI systems' effectiveness in supporting advanced biological experimentation

To date, experiments conducted using foundation models to accelerate bioweapons production do not suggest a significant impact on current risks. Indeed, at their current level of development, foundation models at times recommend incorrect courses of action, making them sometimes counterproductive for successful experimentation.

While foundation models are improving, it remains unclear if, when, and to what extent they will be able to successfully enhance nonspecialists' abilities to reliably build or acquire potentially catastrophic bioweapons. Should such capabilities emerge, the risks of bioattacks from lone wolves and terrorist actors could escalate significantly. Experts should, therefore, closely monitor how reliably and effectively general-purpose systems can guide nonspecialists in sophisticated biological experimentation. Given that such experimentation entails cycles of iteration and triaging, experts should also pay attention to AI systems' abilities to course-correct and speed the design-build-test-learn feedback loop, as well as monitor areas of tacit knowledge that are—and are not—aided by foundation models.

Diminishing tacit knowledge requirements from cloud labs and lab automation

Tacit knowledge has historically acted as a key barrier to the production of bioweapons, but cloud labs and other technologies could erode the importance of some forms of tacit knowledge in biological experimentation in the years ahead. While the introduction of AI-powered automation and experimentation at scale in laboratories holds the potential to reduce the role of tacit knowledge in some areas, the degree and speed of these transformations remains unclear. If cloud labs do achieve such reductions in the importance of tacit knowledge, they may also have safety mechanisms at their disposal that could compensate for the increased ease with which bad actors could otherwise produce biological agents, such as advanced order screening methods and know-your-customer requirements that would flag suspicious orders and customers.

The Biden administration's recent executive order on AI mandates more robust screening for gene synthesis in research funded by the U.S. government, including in cloud labs.[145] While this measure is a step in the right direction, it falls short of plugging all the loopholes among cloud labs that are possible to exploit. Moreover, emerging AI techniques could conceivably create new ways to trick, spoof, or circumvent such screening methods. Experts should closely monitor the degree to which cloud labs diminish the barriers of tacit knowledge needed to successfully produce, sustain, and disseminate biological agents. They should also carefully consider the efficacy of evolving safeguards that can be built into cloud labs and other emerging technologies that seek to lower the need for tacit knowledge.

Dual-use progress in AI-enabled research into host genetic susceptibility to infectious diseases

As AI helps identify genetic features that can make people more susceptible to various diseases, scientists and policymakers alike should be cognizant of how well-meaning medical genomic research into precision medicine might be weaponizable. Policymakers should closely monitor precision medicine that focuses on host genetic susceptibility to infection, an area of study that aims to develop treatments, therapies, and other interventions tailored to the genetics of specific individuals or groups. To be clear, the development of precision medicine of relevance to biorisk is uncertain given the persistence of unanswered questions in the subdiscipline.[146] Developments that could make substantial contributions to the production of precision bioweapons, therefore, could remain distant prospects. However, even with an incomplete or imperfect understanding, AI-enabled progress in precision medicine could stumble upon discoveries or techniques that could be used for precision bioweapon development.

Because of this, precision medicine should be an area that national security professionals proactively observe, with deliberate attention to how medically intended AI tools could be counterintuitively leveraged for misuse. Given the novelty and complexity involved in such development, the United States should also monitor state actors that might have an interest in pursuing precision bioweapons research, most notably China, North Korea, and Russia.

Dual-use progress in precision engineering of viral pathogens

Compared to the human genome, pathogen genomes, especially viral genomes, are orders of magnitude smaller and less complex—and therefore easier to manipulate. Additionally, while catastrophic weaponization of many biological agents requires significant efforts to manufacture and disperse in sufficient quantities, viruses can spread rapidly and widely from a smaller initial quantity.[147]

As AI tools accelerate the study of genetic features of various viruses for a range of purposes, such as vaccine development or immunology research, additional capabilities to alter pathogens toward specific strategic purposes could emerge. Though not a new concept, enhanced capabilities to alter pathogens with greater precision could create new methods to optimize viruses for greater lethality, transmissibility, or immune evasion to maximize the impacts of a biological attack. Such capabilities could also be used to make pathogens that thrive only under particular environmental conditions, which would enable geographically targeted bioweapons.

As with research in host genetic susceptibility to infectious diseases, it is possible that legitimate research into pathogen engineering would inadvertently create tools that could be used to make more powerful or strategically useful bioweapons. Policymakers should proactively monitor scientific developments in precision genetic engineering of pathogens for potential malicious applications. They should also watch the development of tools that could inform this engineering, such as those used for protein generation or immunological modeling—with particular attention to the integration of AI capabilities.[148]

Recommendations

While most of the catastrophic threats at the intersection of AI and biology are yet to emerge and will require careful further monitoring as they take shape, there are some sensible measures worth implementing now to reduce the chances of future biological catastrophes. The following recommendations aim to address instances in which failing to preemptively head off risks would result in unnecessary or unacceptable vulnerabilities, or where early intervention to shape the ongoing development of AI tools could set norms to ensure resilience to emerging catastrophic risks. In view of the immense potential of advancements in both AI and biotechnology to support human flourishing, these recommendations also aim to be innovation-friendly, requiring as little as possible regulatory intervention.

The recommendations aim to address only biological threats that are AI-specific, not the full spectrum of biological catastrophic risks, which are much broader. For example, while much of the immediate concern at the nexus of AI and biotechnology relates to proliferation of biological capabilities among nonspecialists, there is a compelling argument that highly capable experts already working within labs pose the greater threat—particularly as the total number of high-risk labs continues to climb.[149] Addressing this threat and others, which are mostly independent of AI development, requires measures beyond the scope of this report.

Additionally, these recommendations do not directly address the ongoing debate about AI models that are released in their raw form onto the internet for public download. These models are often referred to as "open-source models"—or more precisely, "open-weight models," given that that open release of a trained model's weights may not be accompanied by other hallmarks of open-source software such as permissive licensing. While the most capable general-purpose AI models are currently offered to the public only through online and application programming interfaces (APIs), some models such as Meta's Llama 2 are offered freely for users to download and run natively on their own computers. Advocates of an open-source approach to AI development argue that democratizing access to AI models allows for faster scientific experimentation and collaboration, in addition to the benefits of accessibility and transparency intrinsic to the approach.[150] Critics counter that open-sourcing models could pose considerable risks as AI capabilities improve. In this view, "structured access" can offer an important layer of protection against model misuse, through the use of content filters and blocking access to models entirely when necessary.[151] Once a model's weights are openly available, it becomes near-impossible to prevent its proliferation, a major concern if dangerous capabilities are discovered after its release. Additionally, because individuals can directly tamper with open-source models, current built-in safety measures can be relatively easily undone.[152] Proponents counter that just as open-source approaches to many forms of software have helped improve security and stability, the same may be true in the case of AI, ultimately incentivizing the development of more robust built-in safety mechanisms.[153]

Resolving these tensions is a larger question than can be addressed by this report, involving a range of issues around scientific norms, legal liabilities, business models, and the future capabilities of different AI tools.[154] But insofar as biological risks constitute a major area of interest for the open-source debate, the following recommendations aim to help inform national security practitioners' understanding of the issue, even if not all recommendations are equally relevant to open and closed approaches to model deployment.

Further strengthen screening mechanisms for cloud labs and other gene synthesis providers

The Biden administration's October 2023 AI executive order took positive steps to shore up screening mechanisms for providers of genetic synthesis by tasking the director of the Office for Science and Technology Policy with developing a framework for screening customers and gene sequence orders for potentially dangerous activities. The order further mandates that all federal funding for life sciences research will require the use of services that operate with the mechanisms developed under this framework, which will incentivize companies to institute the measures. Even so, these measures do not comprehensively address the possibility of bad actors ordering genetic sequences that could be used to catastrophic effect, as only companies working with federally funded research will fall under the purview of the order.

Further action is necessary. There appears to be some industry support for screening mechanisms that could be made more binding, given that many major gene synthesis companies are already party to the voluntary International Gene Synthesis Consortium that requires its members to implement order screenings. As AI models help democratize biotechnologies to a broader audience—and have already shown the ability to aid bad actors in identifying cloud labs that may lack sufficient screening safeguards—shoring up safeguards on such a critical digital-to-physical barrier in the development of biological agents is low-hanging fruit.[155] Additionally, the expanding use of cloud labs and other AI-automated facilities for processes like viral assembly, CRISPR editing, and mutagenesis also require enhanced oversight, as bad actors could use these capabilities to bypass synthesis screening safeguards.[156]

American lawmakers should require that all relevant companies rigorously screen their orders and customers for potential threats, and that they develop appropriate reporting mechanisms to law enforcement entities for suspicious activity. Additionally, in light of advances in benchtop synthesis capabilities that may broaden access to dual-use capabilities, relevant agencies should consider policies that would require logging and screening of all synthesized genetic sequences, using encryption to protect trade secrets while allowing for queries of sequences in specific emergency situations.[157] The U.S. government should also seek to internationalize screening norms through diplomatic engagement in the Biological Weapons Convention and other multilateral fora. Because AI tools may soon create methods to circumvent conventional screening mechanisms, lawmakers should look ahead and invest in the development of next-generation dynamic screening methods, perhaps making use of new AI technologies that are responsive to attempts to circumvent conventional practices. They should also anticipate further advances in benchtop synthesis capabilities over the coming decade that may pose new risks, and plan accordingly.[158]

Engage in regular, rigorous assessments of the biological capabilities of general-purpose models for the full bioweapons lifecycle

Several leading American AI labs have already committed to internally stress-testing their foundation models for biological misuse capabilities. The AI executive order further solidifies this commitment by requiring the secretary of commerce to establish "guidance and benchmarks for evaluating and auditing AI capabilities" with a particular focus on biosecurity, and by requiring companies to share results of their testing with the government.[159] But such guidelines are, for now, functionally left to the discretion of companies with potential conflicts of interest to determine their implementation. Moreover, the relative ease or difficulty associated with deploying bioweapons is not simply a function of the capabilities of particular general-purpose models in a vacuum, but is contingent on constantly evolving capabilities in the broader biotechnology ecosystem, not least capabilities related to cloud labs and gene synthesizers. Finally, studies to date have lacked elements to test the full degree to which foundation models may or may not impact tacit knowledge hurdles in bioweapon development.

Relevant federal agencies such as the Department of Defense and the Department of Homeland Security should conduct regular, systematic assessments of the impact of foundation models on the full lifecycle of bioweapons procurement, storage, and dissemination. Recent studies by the RAND Corporation and OpenAI could provide a template for what such rigorous testing looks like, with teams of individuals tasked with developing operational bioweapons plans with the aid of foundation models.[160] These plans would then be submitted for assessment by biological experts to critically examine their feasibility relative to control groups lacking such foundation models. To get even more practical, such groups could also be tasked with using foundation models to help develop real biological agents of similar complexity to that of known bioweapons, but harmless to humans, allowing researchers to gauge how practically helpful general-purpose AI systems are for issues of tacit knowledge. Many layers of complexity are required to successfully field a bioweapon, and the landscape of biotechnologies in cloud labs and gene synthesis tools is changing. For these reasons, to assess the effectiveness of lone wolves and small groups in using foundation models to develop bioweapons capabilities, the most reliable method is to regularly test small groups' ability to achieve similar feats under real-world conditions.

The results of these studies could be valuable in discerning how general-purpose AI systems are changing the barriers to bioweapons production for nonstate actors, and where interventions are most effective and needed. Additionally, such studies would shed much-needed light on the open-source debate, helping to establish with greater clarity what the real risks are from freely available models. If there is demonstrated progress toward enabling nonstate actors in new ways, these results can inform the development and implementation of technical safety measures and legislation to place oversight around the development and release of relevant dangerous models. If progress continues to be negligible, the results can equally help to guide what a responsible approach to model development might look like.

Invest in technical safety mechanisms that can curb misuses of foundation models

To mitigate the risks of future foundation models empowering nonstate actors' bioweapon capabilities, private companies and federal research agencies such as the National Science Foundation and the Defense Advanced Research Projects Agency should invest in further research on several technical safety mechanisms, including:

GUARDRAILS FOR CLOUD-BASED ACCESS

Foundation models that users access through online interfaces or APIs feature a variety of tools to curb malicious behavior, such as checking outputs using additional moderation models.[161] Further research could establish more robust methods to tamp down on prompt injections and jailbreaks that can trick models into revealing harmful information. AI developers could also leverage AI tools within their systems to identify conversations of concern for further review. Where appropriate, AI companies could be required to report to law enforcement cases where users seem to be pursuing hazardous lines of inquiry related to terrorism or other threats to public safety or national security.

UNLEARNING

Machine learning researchers have sought to discover ways to make general-purpose models "forget" information, and recent work has shown tentative promise toward this end.[162] If successful, such techniques could scrub models of dangerous biological information, amounting to a far more comprehensive ability to curtail general-purpose models' dangerous biological capabilities than is currently available from RLHF methods, which are susceptible to jailbreaks and prompt injections. In pursuit of this goal, Google has already launched a Machine Unlearning Challenge to pioneer methods to erase information from models.[163] Two researchers at Microsoft have also shown some progress in getting a model to forget Harry Potter–related information, and more recently a team of researchers has pioneered an unlearning method based on controlling model representations to excise WMD-relevant information from models.[164] But there remains more work to be done to comprehensively excise harmful biological information from general-purpose models, especially in such a way that preserves said models' abilities to usefully enhance scientific endeavors.

NOVEL APPROACHES TO "INFORMATION HAZARDS" IN MODEL TRAINING

It is possible that if the scientific publications featuring the most concerning dual-use information, often referred to as "information hazards," were left out of AI training sets altogether, many of the risks experts fear regarding the proliferation of potentially harmful biological information may prove moot. Bad actors might be able to partially overcome this by giving general-purpose models access to information hazards either as part of queries or via fine tuning, but this would nonetheless still raise technical barriers, as users would need to assemble the relevant information first to provide to the AI system, and know how to use it effectively with the system.

Another method of guarding against biological misuse from general-purpose AI could involve deliberately "poisoning" the training data of models in hazardous areas of biology. Such a method would alter training data by adjusting relevant information or instructions to be intentionally incorrect, perhaps using LLMs to locate and edit the relevant information from training data. If such a poisoned model were successfully coaxed into providing dangerous biological information or instructions, the results would likely be incorrect.

Update government biodefense investment to further prioritize agility and flexibility

The U.S. government has increasingly recognized the need to move beyond the traditional "one-bug-one-drug" approach to biodefense. Following the Department of Defense's most recent *Biodefense Posture Review*, for example, senior officials underscored that advances in science and technology are accelerating the emergence of diverse hazards, necessitating a move toward more flexible and comprehensive solutions to efficiently address a broader spectrum of biological risks.[165]

AI-enabled biological design tools could further escalate the need to transition to more agile and flexible defenses by making it easier to create novel pathogens or modify existing pathogens to resist current countermeasures. However, the current funding model for biodefense initiatives, such as those used in relevant BARDA programs, is often inconsistent and reactive. Past instances, such as the allocation of substantial funds during the swine flu and COVID-19 crises, demonstrate Congress's capacity to mobilize resources in emergency situations. Yet this reactive funding approach, characterized by irregular "boom and bust" cycles, is not conducive to the sustained development and procurement of medical countermeasures. This is particularly true for ambitious, forward-thinking projects, such as the potential development of vaccine prototypes for all known pathogenic viral families. Stable, long-term investment is crucial, particularly in the development of treatments and vaccines that are broad spectrum, and in underlying technologies that enable rapid retargeting of these countermeasures to new pathogens, as was demonstrated by the exceptionally rapid development of COVID-19 vaccines.[166]

More broadly, policymakers should also dedicate further resources to pathogen-agnostic approaches to strengthening biodefenses, prioritizing measures that directly counteract pathogens' ability to spread. For example, they could target improving indoor air quality—whether through enhanced ventilation or promising innovative technologies such as Far-UVC light—which shows potential in directly and safely inactivating airborne pathogens.[167]

In the long term, consider a licensing regime for biological design tools with potentially catastrophic capabilities

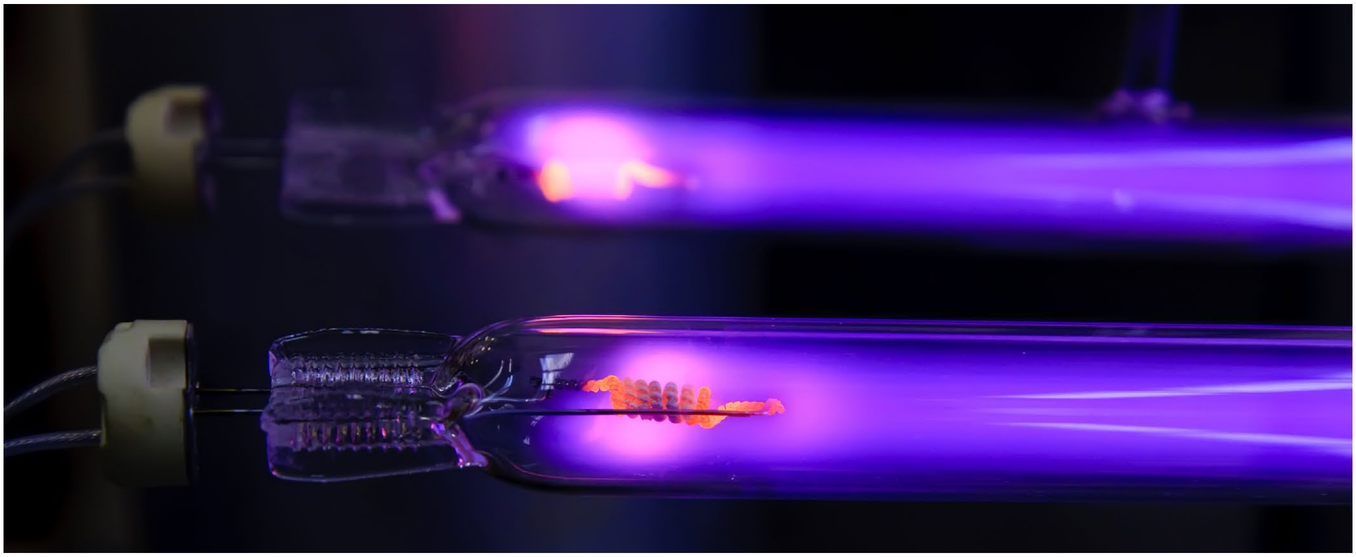
BDTs that successfully design bespoke pathogens—whether targeting particular genetic populations and environments or heightening lethality or transmissibility—would represent a step change in the catastrophic potential of biological agents. If such BDTs emerge, their impacts on the risks of biological catastrophes will be profound. While these tools thankfully remain speculative today, if significant progress toward such systems is achieved, it will be worth considering a licensing regime in advance of their full realization to prevent their uncontrolled proliferation. A licensing regime of this kind should be targeted to only a narrow and clearly defined set of BDTs to restrict their development and use to trusted actors only.

The Biden administration's executive order on AI has taken preliminary steps in this direction by establishing reporting mandates for developers who create biological AI tools that require large amounts of computing power to train. As BDTs grow in sophistication and genomic data becomes more readily available, relevant authorities should closely monitor how such models could be built or repurposed to create designer pathogens. Though these capabilities remain largely speculative, given the severity of the threats posed by such potential developments, policymakers cannot afford to fall behind the curve on developing appropriate safeguards. At the same time, policymakers should avoid premature regulation on the issue, given the immense potential for medical advancements from the dual-use research that would lead to such developments. If BDTs do acquire sophisticated pathogen design capabilities well beyond what is currently possible, the U.S. government should work to establish a stringent licensing regime for the creation and use of such models, both at home and abroad.[168]

Conclusions

All things considered, dire warnings from industry leaders and government officials about AI-powered biocatastrophes remain largely speculative: today's AI has not significantly altered the risks of biocatastrophes. That said, there is a strong case that current biological safeguards—independent of AI development—already need significant updates. And a range of budding AI applications could, perhaps, drive up the likelihood and severity of large-scale biological destruction, even if the extent and timing of these increases remain unclear, as do new opportunities for safeguards that may emerge with improving capabilities.

The good news is that industry and government leaders have a window to address these risks proactively, rather than reactively. Careful monitoring of a handful of capabilities can help policymakers and experts get ahead of risks as they emerge, and respond appropriately with an eye to protecting innovation. Additionally, sensible measures now can set on firmer footing the trajectory of biosecurity in the age of AI. Daunting as the theoretical possibilities for future AI-enabled biological catastrophes may be, they are far from inevitable.



Ultraviolet lamps are widely used for germicidal irradiation, but are generally harmful to humans. But part of the ultraviolet spectrum—Far-UVC—shows promise in inactivating airborne pathogens while also being safe for humans. (Victor Borisov via Getty Images)

1. Kamala Harris, “Remarks by Vice President Harris on the Future of Artificial Intelligence” (U.S. Embassy, London, November 1, 2023), https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/11/01/remarks-by-vice-president-harris-on-the-future-of-artificial-intelligence-london-united-kingdom.

2. Tejal Patwardhan et al., “Building an Early Warning System for LLM-Aided Biological Threat Creation,” OpenAI, January 31, 2024, https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation; Christopher A. Mouton, Caleb Lucas, and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study* (RAND Corporation, January 25, 2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html; Anthropic, “Frontier Threats Red Teaming for AI Safety,” July 26, 2023, https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety; Emily H. Soice et al., “Can Large Language Models Democratize Access to Dual-Use Biotechnology?” (arXiv, June 6, 2023), https://doi.org/10.48550/arXiv.2306.03809.

3. “The Pandemic’s True Death Toll,” *The Economist*, accessed April 15, 2024, https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates.

4. The 1918 Spanish Flu killed approximately 1 to 2 percent of the world’s population—equivalent to 70 to 150 million today: Alain Gagnon et al., “Age-Specific Mortality during the 1918 Influenza Pandemic: Unravelling the Mystery of High Young Adult Mortality,” *PLoS ONE* 8, no. 8 (August 5, 2013): e69586, https://doi.org/10.1371/journal.pone.0069586. The Black Plague killed about half of Europeans over a few years in the mid-1300s: Ole J. Benedictow, *The Black Death 1346–1353: The Complete History* (Woodbridge: Boydell Press, 2006).

5. *Advanced Technology: Examining Threats to National Security: Hearing before the Subcommittee on Emerging Threats and Spending Oversight of the Senate Committee on Homeland Security and Governmental Affairs*, 118th Cong. (2023) (testimony of Jeff Alstott, senior information scientist, RAND Corporation), https://www.hsgac.senate.gov/subcommittees/etso/hearings/advanced-technology-examining-threats-to-national-security.

6. Melissa De Witte, “How Pandemics Catalyze Social and Economic Change,” Stanford News, April 30, 2020, https://news.stanford.edu/2020/04/30/pandemics-catalyze-social-economic-change; Ewen Callaway, “Collapse of Aztec Society Linked to Catastrophic Salmonella Outbreak,” *Nature* 542 (February 23, 2017): 404, https://doi.org/10.1038/nature.2017.21485; Williamson Murray, “On Plagues and Their Long-Term Effects,” Hoover Institution, April 24, 2020, https://www.hoover.org/research/plagues-and-their-long-term-effects.

7. “Pause Giant AI Experiments: An Open Letter,” Future of Life Institute, March 22, 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments; “Statement on AI Risk,” Center for AI Safety, May 30, 2023, https://www.safe.ai/statement-on-ai-risk#signatories.

8. *Oversight of A.I.: Principles for Regulation: Hearing before the Subcommittee on Privacy, Technology, and the Law of the Senate Judiciary Committee*, 118th. Cong. (2023) (statement of Dario Amodei, CEO, Anthropic), https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf.

9. James Titcomb, “Britain Has One Year to Prevent AI Running Out of Control, Sunak Fears,” *The Telegraph*, September 25, 2023, https://www.telegraph.co.uk/business/2023/09/25/artificial-intelligence-create-bioweapons-warning.

10. Harris, “Remarks by Vice President Harris on the Future of Artificial Intelligence.”

11. Bill Drexel and Caleb Withers, *Catalyzing Crisis: A Primer on Artificial Intelligence, Catastrophes, and National Security* (Center for a New American Security, June 2024), https://www.cnas.org/publications/reports/catalyzing-crisis.

12. “The Pandemic’s True Death Toll.”

13. Gagnon et al., “Age-Specific Mortality during the 1918 Influenza Pandemic”; Max Roser, “The Spanish Flu: The Global Impact of the Largest Influenza Pandemic in History,” Our World in Data, March 4, 2020, https://ourworldindata.org/spanish-flu-largest-influenza-pandemic-in-history.

14. Benedictow, *The Black Death 1346–1353*.

15. Abraham Haileamlak, “Pandemics Will Be More Frequent,” *Ethiopian Journal of Health Sciences* 32, no. 2 (March 2022): 228, https://doi.org/10.4314/ejhs.v32i2.1.

16. Marc Lipsitch, “Why Do Exceptionally Dangerous Gain-of-Function Experiments in Influenza?” *Methods in Molecular Biology* 1836 (2018): 589–608, https://doi.org/10.1007/978-1-4939-8678-1_29; Todd Kuiken, *Global Pandemics: Gain-of-Function Research of Concern* (Congressional Research Service, November 21, 2022), https://crsreports.congress.gov/product/pdf/IF/IF12021.

17. Josh Rogin, “State Department Cables Warned of Safety Issues at Wuhan Lab Studying Bat Coronaviruses,” *The Washington Post*, April 20, 2020, https://www.washingtonpost.com/opinions/2020/04/14/state-department-cables-warned-safety-issues-wuhan-lab-studying-bat-coronaviruses.

18. Martin Furmanski, *Laboratory Escapes and “Self-Fulfilling Prophecy” Epidemics* (Center for Arms Control and Nonproliferation, February 17, 2014), 10–11, https://armscontrolcenter.org/wp-content/uploads/2016/02/Escaped-Viruses-final-2-17-14-copy.pdf.

19. Georgios Pappas, “The Lanzhou Brucella Leak: The Largest Laboratory Accident in the History of Infectious Diseases?” *Clinical Infectious Diseases* 75, no. 10 (November 14, 2022): 1845–47, https://doi.org/10.1093/cid/ciac463.

20. Filippa Lentzos et al., *Global BioLabs Report 2023* (Global Biolabs, 2023), 5, https://www.kcl.ac.uk/warstudies/assets/global-biolabs-report-2023.pdf.

21. Lentzos et al., *Global BioLabs Report 2023*, 6.

22. David Manheim and Gregory Lewis, *High-Risk Human-Caused Pathogen Exposure Events from 1975 to 2016* (F1000Research, July 8, 2022), https://doi.org/10.12688/f1000research.55114.2.

23. Mustafa Suleyman and Michael Bhaskar, *The Coming Wave: Technology, Power, and the Twenty-First Century’s Greatest Dilemma* (New York: Crown, 2023), 175.

24. V. Barras and G. Greub, “History of Biological Warfare and Bioterrorism,” *Clinical Microbiology and Infection* 20, no. 6 (June 1, 2014): 498–99, https://doi.org/10.1111/1469-0691.12706.

25. Barras and Greub, “History of Biological Warfare and Bioterrorism,” 499; Friedrich Frischknecht, “The History of Biological Warfare,” *EMBO Reports* 4, no. S1 (June 2003): S47–52, https://doi.org/10.1038/sj.embor.embor849.

26. Barras and Greub, “History of Biological Warfare and Bioterrorism,” 500.

27. Edward M. Spiers, *Agents of War: A History of Chemical and Biological Weapons*, 2nd expanded ed. (London: Reaktion Books, 2021), 63; Sonia Ben Ouagrham-Gormley, *Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development* (Ithaca: Cornell University Press, 2014), 94.

28. Glenn Cross, “Biological Weapons in the ‘Shadow War,’” War on the Rocks, November 9, 2021, https://warontherocks.com/2021/11/biological-weapons-in-the-shadow-war.

29. W. Seth Carus, “A Century of Biological-Weapons Programs (1915–2015): Reviewing the Evidence,” *The Nonproliferation Review* 24, no. 1–2 (January 2, 2017): 129–53, https://doi.org/10.1080/10736700.2017.1385765.

30. Howard Brody et al., “United States Responses to Japanese Wartime Inhuman Experimentation after World War II: National Security and Wartime Exigency,” *Cambridge Quarterly of Healthcare Ethics* 23, no. 2 (April 2014): 220–30, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4487829.

31. Frischknecht, “The History of Biological Warfare.”

32. Daniel Barenblatt, *A Plague Upon Humanity: The Hidden History of Japan’s Biological Warfare Program* (New York: Harper Perennial, 2005), 173–74.

33. Barras and Greub, “History of Biological Warfare and Bioterrorism.”

34. Benny Morris and Benjamin Z. Kedar, “‘*Cast Thy Bread*’: Israeli Biological Warfare during the 1948 War,” *Middle Eastern Studies* 59, no. 5 (2023): 752–76.

35. *2023 Adherence to and Compliance with Arms Control, Nonproliferation, and Disarmament Agreements and Commitments* (U.S. Department of State, April 2023), 22–28, https://www.state.gov/wp-content/uploads/2023/04/13apr23-final-2023-treaty-compliance-report-unclassified-unsourced.pdf.

36. Milton Leitenberg, Raymond A. Zilinskas, and Jens H. Kuhn, *The Soviet Biological Weapons Program: A History* (Cambridge, MA: Harvard University Press, 2012), 106.

37. Togzhan Kassenova, “Aralsk: A Kazakh Town That Lived through a Smallpox Epidemic,” Davis Center for Russian and Eurasian Studies, May 15, 2020, https://daviscenter.fas.harvard.edu/insights/aralsk-kazakh-town-lived-through-smallpox-epidemic.

38. Judith Miller, William J. Broad, and Stephen Engelberg, *Germs: Biological Weapons and America’s Secret War* (New York: Simon & Schuster, 2002), chap. 1.

39. Ouagrham-Gormley, *Barriers to Bioweapons*, 2–3.

40. Richard Danzig et al., *Aum Shinrikyo: Insights into How Terrorists Develop Biological and Chemical Weapons* (Center for a New American Security, 2012), https://www.jstor.org/stable/resrep06323; Rolf Mowatt-Larssen, “Al Qaeda’s Pursuit of Weapons of Mass Destruction,” *Foreign Policy*, January 25, 2010, https://foreignpolicy.com/2010/01/25/al-qaedas-pursuit-of-weapons-of-mass-destruction; Joby Warrick, “ISIS Planned Chemical Attacks in Europe, New Details on Weapons Program Reveal,” *The Washington Post*, July 11, 2022, https://www.washingtonpost.com/national-security/2022/07/11/isis-chemical-biological-weapons; John V. Parachini and Rohan Kumar Gunaratna, *Implications of the Pandemic for Terrorist Interest in Biological Weapons: Islamic State and al-Qaeda Pandemic Case Studies* (RAND Corporation, May 31, 2022), https://www.rand.org/pubs/research_reports/RRA612-1.html; Shahzeb Ali Rathore, “Is the Threat of ISIS Using CBRN Real?” *Counter Terrorist Trends and Analyses* 8, no. 2 (2016): 4–10, https://www.jstor.org/stable/2636958; and Rueben Ananthan Santhana Dass, “Jihadists’ Use and Pursuit of Weapons of Mass Destruction: A Comparative Study of Al-Qaeda and Islamic State’s Chemical, Biological, Radiological and Nuclear (CBRN) Weapons Programs,” *Studies in Conflict & Terrorism* 47, no. 5 (2024): 548–82, https://doi.org/10.1080/1057610X.2021.1981203.

41. Kevin Esvelt, *Delay, Detect, Defend: Preparing for a Future in Which Thousands Can Release New Pandemics*, Geneva Papers (Geneva Centre for Security Policy, November 2022), https://www.gcsp.ch/publications/delay-detect-defend-preparing-future-which-thousands-can-re-

lease-new-pandemics.

42. Miller, Broad, and Engelberg, *Germs*, chap. 1.

43. Mowatt-Larssen, "Al Qaeda's Pursuit of Weapons of Mass Destruction"; Danzig et al., *Aum Shinrikyo: Insights into How Terrorists Develop Biological and Chemical Weapons*; and Leitenberg, Zilinskas, and Kuhn, *The Soviet Biological Weapons Program*, chap. 1.

44. Suleyman and Bhaskar, *The Coming Wave*, chap. 5.

45. Garrett Dunlap and Eleonore Pauwels, *The Intelligent and Connected Bio-Labs of the Future: Promise and Peril in the Fourth Industrial Revolution*, Wilson Briefs (Wilson Center, September 2017), 12, https://www.wilsoncenter.org/sites/default/files/media/documents/misc/the_intelligent_connected_biolabs_of_the_future.pdf.

46. Dunlap and Pauwels, *The Intelligent and Connected Bio-Labs of the Future*, 2–5.

47. "Transcend the Lab," Emerald Cloud Lab (remote-controlled life sciences lab), accessed September 27, 2023, https://www.emeraldcloudlab.com.

48. Sarah R. Carter et al., *The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe* (Nuclear Threat Initiative, 2023), 20–22, 26, https://www.nti.org/wp-content/uploads/2023/10/ntibio_ai_final.pdf.

49. Ouagrham-Gormley, *Barriers to Bioweapons*.

50. Dunlap and Pauwels, *The Intelligent and Connected Bio-Labs of the Future*; Emma Saunders, "Can the Biological Weapons Convention Address New Biothreats?" Chatham House, November 25, 2021, https://www.chathamhouse.org/2021/11/can-biological-weapons-convention-address-new-biothreats.

51. James Revill and Catherine Jefferson, "Tacit Knowledge and the Biological Weapons Regime," *Science and Public Policy* 41, no. 5 (October 1, 2014): 597–610, https://doi.org/10.1093/scipol/sct090.

52. Dunlap and Pauwels, *The Intelligent and Connected Bio-Labs of the Future*, 7.

53. Dana Gretton et al., "Random Adversarial Threshold Search Enables Automated DNA Screening" (bioRxiv, April 2, 2024), https://doi.org/10.1101/2024.03.20.585782.

54. International Gene Synthesis Consortium, "Harmonized Screening Protocol© v2.0: Gene Sequence & Customer Screening to Promote Biosecurity," November 19, 2017, https://genesynthesisconsortium.org/wp-content/uploads/IGSCHarmonizedProtocol11-21-17.pdf; Authors' discussion with expert in gene synthesis screening policy (name withheld by agreement), 2024.

55. "Public Health Preparedness: HHS Could Improve Oversight of Research Involving Enhanced Potential Pandemic Pathogens," U.S. Government Accountability Office, January 18, 2023, https://www.gao.gov/products/gao-23-105455; Sharon Lerner, Mara Hvistendahl, and Maia Hibbett, "NIH Documents Provide New Evidence U.S. Funded Gain-of-Function Research in Wuhan," The Intercept, September 10, 2021, https://theintercept.com/2021/09/09/covid-origins-gain-of-function-research.

56. Sonia Ben Ouagrham-Gormley and Kathleen M. Vogel, "Follow the Money: What the Sources of Jiankui He's Funding Reveal about What Beijing Authorities Knew about Illegal CRISPR Babies, and When They Knew It," *Bulletin of the Atomic Scientists* 76, no. 4 (July 3, 2020): 192–99, https://doi.org/10.1080/00963402.2020.1780726.

57. "Biomedical Advanced Research and Development Authority (BARDA)," Medical Countermeasures.gov, U.S. Department of Health & Human Services, accessed October 2, 2023, https://medicalcountermeasures.gov/barda.

58. Kathleen M. Vogel, "Intelligent Assessment: Putting Emerging Biotechnology Threats in Context," *Bulletin of the Atomic Scientists* 69, no. 1 (January 1, 2013): 43–52, https://doi.org/10.1177/0096340212470813.

59. Nell Greenfieldboyce, "Did Pox Virus Research Put Potential Profits ahead of Public Safety?" NPR, February 17, 2018, https://www.npr.org/sections/health-shots/2018/02/17/585385308/did-pox-virus-research-put-potential-profits-ahead-of-public-safety.

60. Greenfieldboyce, "Did Pox Virus Research Put Potential Profits ahead of Public Safety?"

61. *Advanced Technology: Examining Threats to National Security* (testimony of Jeff Alstott).

62. Ouagrham-Gormley, *Barriers to Bioweapons*, 10–11.

63. Miller, Broad, and Engelberg, *Germs*, 191–92.

64. Vogel, "Intelligent Assessment."

65. *The Apollo Program for Biodefense: Winning the Race against Biological Threats* (Bipartisan Commission on Biodefense, January 2021), https://biodefensecommission.org/reports/the-apollo-program-for-biodefense-winning-the-race-against-biological-threats; Christian Ruhl, *Global Catastrophic Biological Risks: A Guide for Philanthropists*, "Societal Spending and Neglectedness" (Founders Pledge, October 31, 2023), 70–81, https://dkqj4hmn5mktp.cloudfront.net/GCBR_Report_Founders_Pledge_7505b1ebe0.pdf; *COVID-19: GAO Recommendations Can Help Federal Agencies Better Prepare for Future Public Health Emergencies*, GAO-23-106554 (U.S. Government Accountability Office, July 2023), "Public Health Preparedness," 16–18, https://www.gao.gov/assets/gao-23-106554.pdf; and Tom Inglesby, "How

to Prepare for the Next Pandemic," *The New York Times*, March 12, 2023, https://www.nytimes.com/2023/03/12/opinion/pandemic-health-prepare.html.

66. David Willman, "The U.S. Quietly Terminates a Controversial $125m Wildlife Virus Hunting Programme amid Safety Fears," *BMJ* 382 (September 7, 2023): 2002, https://doi.org/10.1136/bmj.p2002; Jason Matheny, "The Nuclear and Biological Weapons Threat," September 7, 2023, in *The Rachman Review*, *Financial Times*, presented by Gideon Rachman and produced by Fiona Symon, podcast, 27:57, https://www.ft.com/content/af097860-0ed9-4079-84ab-46cf14a2a157t; and *Proposed Biosecurity Oversight Framework for the Future of Science* (National Science Advisory Board for Biosecurity, March 2023), https://osp.od.nih.gov/wp-content/uploads/2023/03/NSABB-Final-Report-Proposed-Biosecurity-Oversight-Framework-for-the-Future-of-Science.pdf.

67. *The Apollo Program for Biodefense* (Bipartisan Commission on Biodefense); Jaime Yassif, Chris Isaac, and Kevin O'Prey, *Strengthening Global Systems to Prevent and Respond to High-Consequence Biological Threats* (Nuclear Threat Initiative, November 23, 2021), https://www.nti.org/analysis/articles/strengthening-global-systems-to-prevent-and-respond-to-high-consequence-biological-threats; Mark Dybul, *Biosecurity in the Age of AI: Chairperson's Statement* (Helena, July 2023), https://www.helenabiosecurity.org.

68. Drexel and Withers, *Catalyzing Crisis*.

69. Carol Kuntz, *Genomes: The Era of Purposeful Manipulation Begins* (Center for Strategic & International Studies, July 8, 2022), https://www.csis.org/analysis/genomes-era-purposeful-manipulation-begins; Suleyman and Bhaskar, *The Coming Wave*, chaps. 5–6.

70. Suleyman and Bhaskar, *The Coming Wave*, chap. 5.

71. Elliot Hershberg, "Optimizing Viral Vehicles," The Century of Biology, May 2, 2021, https://centuryofbio.com/p/optimizing-viral-vehicles.

72. Nardus Mollentze, Simon A. Babayan, and Daniel G. Streicker, "Identifying and Prioritizing Potential Human-Infecting Viruses from Their Genome Sequences," *PLOS Biology* 19, no. 9 (September 28, 2021): e3001390, https://doi.org/10.1371/journal.pbio.3001390; Jakub M Bartoszewicz, Anja Seidel, and Bernhard Y Renard, "Interpretable Detection of Novel Human Viruses from Genome Sequencing Data," *NAR Genomics and Bioinformatics* 3, no. 1 (March 1, 2021): lqab004, https://doi.org/10.1093/nargab/lqab004; Jakub M. Bartoszewicz et al., "DeePaC: Predicting Pathogenic Potential of Novel DNA with Reverse-Complement Neural Networks," *Bioinformatics* 36, no. 1 (January 1, 2020): 81–89, https://doi.org/10.1093/bioinformatics/btz541; Ewen Callaway, "How AlphaFold and Other AI Tools Could Help Us Prepare for the Next Pandemic," *Nature*, October 11, 2023, https://doi.org/10.1038/d41586-023-03201-4; Kevin Esvelt, "Pandemic Cost-Benefit Analyses Should Consider Misuse," March 4, 2023, response to Aaron S. Bernstein et al., "The Costs and Benefits of Primary Prevention of Zoonotic Pandemics," *Science Advances* 8, no. 5 (February 4, 2022): eabl4183, https://www.science.org/doi/full/10.1126/sciadv.abl4183#elettersSection.

73. Jonas B. Sandbrink, "Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools" (arXiv, August 12, 2023), 4, http://arxiv.org/abs/2306.13952.

74. S. Alizon et al., "Virulence Evolution and the Trade-Off Hypothesis: History, Current State of Affairs and the Future," *Journal of Evolutionary Biology* 22, no. 2 (2009): 245–59, https://doi.org/10.1111/j.1420-9101.2008.01658.x.

75. Louise Matsakis, "Why AI-Assisted Bioterrorism Became a Top Concern for OpenAI and Anthropic," Semafor, November 15, 2023, https://www.semafor.com/article/11/15/2023/ai-assisted-bioterrorism-is-top-concern-for-openai-and-anthropic.

76. Elsa Kania and Wilson Vorndick, "Weaponizing Biotech: How China's Military Is Preparing for a 'New Domain of Warfare,'" Defense One, August 14, 2019, https://www.defenseone.com/ideas/2019/08/chinas-military-pursuing-biotech/159167.

77. "How to Tweak Drug-Design Software to Create Chemical Weapons," *The Economist*, March 19, 2022, https://www.economist.com/science-and-technology/how-to-tweak-drug-design-software-to-create-chemical-weapons/21808200.

78. Fabio Urbina et al., "Dual Use of Artificial Intelligence-Powered Drug Discovery," *Nature Machine Intelligence* 4, no. 3 (March 2022): 190, https://doi.org/10.1038/s42256-022-00465-9.

79. Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models" (arXiv, July 12, 2022), https://doi.org/10.48550/arXiv.2108.07258.

80. Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety" (arXiv, September 4, 2023), 7, https://doi.org/10.48550/arXiv.2307.03718.

81. Daniil A. Boiko, Robert MacKnight, and Gabe Gomes, "Emergent Autonomous Scientific Research Capabilities of Large Language Models" (arXiv, April 11, 2023), http://arxiv.org/abs/2304.05332.

82. Soice et al., "Can Large Language Models Democratize Access to Dual-Use Biotechnology?"

83. Robert Service, "Could Chatbots Help Devise the Next Pandemic Virus?" Science Insider, June 14, 2023, https://www.science.org/content/article/could-chatbots-help-devise-next-pandemic-virus.

84. Anthropic, "Frontier Threats Red Teaming for AI Safety."

[85] Anthropic, "Frontier Threats Red Teaming for AI Safety."

[86] Anthropic, "Frontier Threats Red Teaming for AI Safety."

[87] Mouton, Lucas, and Guest, *The Operational Risks of AI in Large-Scale Biological Attacks*.

[88] Without adjustment, multiple comparisons increase the likelihood of finding a spurious positive result. For example, with a 5 percent significance threshold, conducting 20 comparisons where there is no true effect would statistically be expected to yield one significant result purely by chance. Therefore, adjusting for multiple comparisons ensures that associated findings are not disproportionately likely to reflect spurious correlations.

[89] For a critique claiming that the researchers underplayed a positive finding, see Gary Marcus, "When Looked at Carefully, OpenAI's New Study on GPT-4 and Bioweapons Is Deeply Worrisome," Marcus on AI, February 04, 2024, https://garymarcus.substack.com/p/when-looked-at-carefully-openais.

[90] Patwardhan et al., "Building an Early Warning System for LLM-Aided Biological Threat Creation."

[91] Sophie Rose and Cassidy Nelson, *Understanding AI-Facilitated Biological Weapon Development* (The Centre for Long-Term Resilience, October 18, 2023), 3, https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio.

[92] Boiko, MacKnight, and Gomes, "Emergent Autonomous Scientific Research Capabilities of Large Language Models," 8.

[93] Vogel, "Intelligent Assessment."

[94] Revill and Jefferson, "Tacit Knowledge and the Biological Weapons Regime."

[95] Usman Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models" (arXiv, April 15, 2024), sec. 3.5, https://doi.org/10.48550/arXiv.2404.09932; Ali Shafahi et al., "Are Adversarial Examples Inevitable?" (arXiv, February 3, 2020), https://doi.org/10.48550/arXiv.1809.02104.

[96] Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," sec. 3.2.

[97] Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," sec. 3.1, 3.2.

[98] Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," sec. 2.2.3, 3.5.4; Carl-Johann Simon-Gabriel et al., "Adversarial Vulnerability of Neural Networks Increases with Input Dimension" (arXiv, February 5, 2018), https://arxiv.org/abs/1802.01421v3. For a suggestive example of potential adversarial vulnerability in biology inputs, see Daniel Mas Montserrat and Alexander G. Ioannidis, "Adversarial Attacks on Genotype Sequences" (bioRxiv, November 8, 2022), https://doi.org/10.1101/2022.11.07.515527.

[99] David Lyell et al., "More Than Algorithms: An Analysis of Safety Events Involving ML-Enabled Medical Devices Reported to the FDA," *Journal of the American Medical Informatics Association* 30, no. 7 (July 1, 2023): 1227–36, https://doi.org/10.1093/jamia/ocad065; Emily Sohn, "The Reproducibility Issues That Haunt Health-Care AI," *Nature* 613, no. 7943 (January 9, 2023): 402–3, https://doi.org/10.1038/d41586-023-00023-2; Adnan Qayyum et al., "Secure and Robust Machine Learning for Healthcare: A Survey," *IEEE Reviews in Biomedical Engineering* 14 (2021): 156–80, https://doi.org/10.1109/RBME.2020.3013489; David Benrimoh et al., "Editorial: ML and AI Safety, Effectiveness and Explainability in Healthcare," *Frontiers in Big Data* 4 (2021), https://www.frontiersin.org/articles/10.3389/fdata.2021.727856; Nam K Tran et al., "Evolving Applications of Artificial Intelligence and Machine Learning in Infectious Diseases Testing," *Clinical Chemistry* 68, no. 1 (January 1, 2022): 125–33, https://doi.org/10.1093/clinchem/hvab239; Carlos Outeiral, Daniel A Nissley, and Charlotte M Deane, "Current Structure Predictors Are Not Learning the Physics of Protein Folding," *Bioinformatics* 38, no. 7 (January 31, 2022): 1881–87, https://doi.org/10.1093/bioinformatics/btab881; Kieran Didi and Matej Zečević, "On How AI Needs to Change to Advance the Science of Drug Discovery" (arXiv, December 23, 2022), https://doi.org/10.48550/arXiv.2212.12560; and Harald König et al., "AI Models and the Future of Genomic Research and Medicine: True Sons of Knowledge?" *BioEssays* 43, no. 10 (2021): 2100025, https://doi.org/10.1002/bies.202100025.

[100] Ronald J. Jackson et al., "Expression of Mouse Interleukin-4 by a Recombinant Ectromelia Virus Suppresses Cytolytic Lymphocyte Responses and Overcomes Genetic Resistance to Mousepox," *Journal of Virology* 75, no. 3 (February 2001): 1205–10, https://doi.org/10.1128/jvi.75.3.1205-1210.2001.

[101] Others subsequently argued the lethal effects could have been predicted in advance: Arno Müllbacher and Mario Lobigs, "Creation of Killer Poxvirus Could Have Been Predicted," *Journal of Virology* 75, no. 18 (September 15, 2001): 8353–55, https://doi.org/10.1128/jvi.75.18.8353-8355.2001.

[102] Sana Zakaria et al., *Machine Learning and Gene Editing at the Helm of a Societal Evolution* (RAND Corporation, October 23, 2023), https://www.rand.org/pubs/research_reports/RRA2838-1.html; Laura M. Helleckes et al., "Machine Learning in Bioprocess Development: From Promise to Practice," *Trends in Biotechnology* 41, no. 6 (June 1, 2023): 817–35, https://doi.org/10.1016/j.tibtech.2022.10.010; Pablo Carbonell, Tijana Radivojevic, and Héctor García Martín, "Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation," *ACS Synthetic Biology* 8, no. 7 (July 19, 2019): 1474–77, https://doi.org/10.1021/acssynbio.8b00540; Elliot

Hershberg, "What's Different? Part Four: Software," Century of Biology, September 17, 2023, https://centuryofbio.com/p/whats-different-part-four-software.

103. Karl E. Weick and Kathleen M. Sutcliffe, *Managing the Unexpected: Sustained Performance in a Complex World*, 3rd ed. (Hoboken, NJ: Jossey-Bass, 2015), chap. 5; Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W. W. Norton & Company, 2019), chap. 10.

104. Weick and Sutcliffe, *Managing the Unexpected*, chap. 5.

105. Scharre, *Army of None*, chap. 10.

106. Daniel Morrow, Robert North, and Christopher D. Wickens, "Reducing and Mitigating Human Error in Medicine," *Reviews of Human Factors and Ergonomics* 1, no. 1 (June 1, 2005): 254–96, https://doi.org/10.1518/155723405783703019; Tor Erik Evjemo and S. Johnsen, "Lessons Learned from Increased Automation in Aviation: The Paradox Related to the High Degree of Safety and Implications for Future Research," *Proceedings of the 29th European Safety and Reliability Conference* (Hannover, Germany, September 22–26, 2019), 3076–83, https://doi.org/10.3850/978-981-11-2724-3_0925-cd.

107. Revill and Jefferson, "Tacit Knowledge and the Biological Weapons Regime."

108. See, for example, Kathleen M. Vogel, *Phantom Menace or Looming Danger? A New Framework for Assessing Bioweapons Threats* (Baltimore, Johns Hopkins University Press, 2012), 80.

109. Revill and Jefferson, "Tacit Knowledge and the Biological Weapons Regime."

110. "BioFoundries to Enable Access to Infrastructure and Resources for Advancing Modern Biology and Biotechnology," National Science Foundation, May 5, 2023, https://new.nsf.gov/funding/opportunities/biofoundries-enable-access-infrastructure.

111. Pallavi Satsangi, "Automation of Tacit Knowledge Using Machine Learning," *2019 6th International Conference on Soft Computing & Machine Intelligence* (Johannesburg, South Africa, November 19–20, 2019), 35–39, https://doi.org/10.1109/ISCMI47871.2019.9004290; Åsa Fast-Berglund et al., "Conceptualizing Embodied Automation to Increase Transfer of Tacit Knowledge in the Learning Factory," *2018 International Conference on Intelligent Systems* (Funchal, Portugal, September 25–27, 2018), 358–64, https://doi.org/10.1109/IS.2018.8710482.

112. See, for example, biotechnology companies such as Recursion: "Our Unique Approach to TechBio," Recursion, accessed October 9, 2023, https://www.recursion.com/approach.

113. Revill and Jefferson, "Tacit Knowledge and the Biological Weapons Regime."

114. Revill and Jefferson, "Tacit Knowledge and the Biological Weapons Regime."

115. Maciej B Holowko et al., "Building a Biofoundry," *Synthetic Biology* 6, no. 1 (December 16, 2020): ysaa026, https://doi.org/10.1093/synbio/ysaa026.

116. Holowko et al., "Building a Biofoundry."

117. Navraj S. Nagra et al., "Understanding the Company Landscape in AI-Driven Biopharma R&D," *Biopharma Dealmakers*, May 17, 2023, https://doi.org/10.1038/d43747-023-00020-4.

118. "Google DeepMind: Drug Developers Seek a Structural Advantage from AI," *Financial Times*, September 19, 2023, https://www.ft.com/content/29f8616f-05ad-42b0-b954-ab81c017b03f; Nathan Benaich, "State of AI Report 2023," Air Street Capital (Google Slides presentation), 102, https://docs.google.com/presentation/d/156WpBF_rGvf4Ecg19oM1fyR51g4FAmHV3Zs0WLukrLQ.

119. *Artificial Intelligence Index Report 2024* (Stanford University, Human-Centered Artificial Intelligence, April 2024), 46, https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf.

120. Nestor Maslej et al., "Artificial Intelligence Index Report 2023" (arXiv, October 5, 2023), chaps. 1, 5, https://doi.org/10.48550/arXiv.2310.03715.

121. Maximilian Schreiner, "CASP15: AlphaFold's Success Spurs New Challenges in Protein-Structure Prediction," The Decoder, December 14, 2022, https://the-decoder.com/casp15-alphafolds-success-brings-new-challenges.

122. *Dual Use Research of Concern: Balancing Benefits and Risks: Hearing before the Committee on Homeland Security and Governmental Affairs*, 112th Cong. (2012) (statement of Anthony S. Fauci, director, National Institute of Allergy and Infectious Diseases), https://www.hsgac.senate.gov/imo/media/doc/Testimony-Fauci-2012-04-26.pdf.

123. Board on Life Sciences et al., "Gain-of-Function Research: Background and Alternatives," in *Potential Risks and Benefits of Gain-of-Function Research: Summary of a Workshop* (Washington, DC: National Academies Press, 2015), https://www.ncbi.nlm.nih.gov/books/NBK285579.

124. *Proposed Biosecurity Oversight Framework for the Future of Science*, 6.

125. Sebastian Farquhar, Owen Cotton-Barratt, and Andrew Snyder-Beattie, "Pricing Externalities to Balance Public Risks and Benefits of Research," *Health Security* 15, no. 4 (August 2017): 401–8, https://doi.org/10.1089/hs.2016.0118.

126. See Bill Drexel and Hannah Kelley, "China Is Flirting with AI Catastrophe," *Foreign Affairs*, May 30, 2023, https://www.foreignaffairs.com/china/china-flirt-

ing-ai-catastrophe.

127. Graham Webster et al., “Full Translation: China’s ‘New Generation Artificial Intelligence Development Plan’ (2017),” blog post, New America, August 1, 2017, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017; Xu Zhang et al., “The Roadmap of Bioeconomy in China,” *Engineering Biology* 6, no. 4 (November 30, 2022): 71–81, https://doi.org/10.1049/enb2.12026.

128. Joby Warrick and Cate Brown, “China’s Quest for Human Genetic Data Spurs Fears of a DNA Arms Race,” *The Washington Post*, September 21, 2023, https://www.washingtonpost.com/world/interactive/2023/china-dna-sequencing-bgi-covid.

129. U.S. Department of State, *2023 Adherence to and Compliance with Arms Control, Nonproliferation, and Disarmament Agreements and Commitments*.

130. U.S. Department of Defense, *2023 Biodefense Posture Review* (August 16, 2023), https://media.defense.gov/2023/Aug/17/2003282337/-1/-1/1/2023_biodefense_posture_review.pdf.

131. David Stanway, “Factbox: A History of China’s Steel Sector,” Reuters, Business, May 3, 2012, https://www.reuters.com/article/idUSBRE84203A; Anatoly Zak, “Disaster at Xichang,” *Smithsonian Magazine*, February 2013, https://www.smithsonianmag.com/air-space-magazine/disaster-at-xichang-2873673; and Ryan Dube and Gabriele Steinhauser, “China’s Global Mega-Projects Are Falling Apart,” *The Wall Street Journal*, January 20, 2023, https://www.wsj.com/articles/china-global-mega-projects-infrastructure-falling-apart-11674166180.

132. Ouagrham-Gormley and Vogel, “Follow the Money.”

133. Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Boston: Houghton Mifflin Harcourt, 2018), 27, 102–3; “State of AI Safety in China,” Concordia AI, October 2023, https://concordia-ai.com.

134. Anna Hayes, “AIDS, Bloodheads & Cover-Ups: The ‘Abc’ of Henan’s Aids Epidemic,” *AQ: Australian Quarterly* 77, no. 3 (2005): 12–40, https://www.jstor.org/stable/20638337; Yanzhong Huang, “The SARS Epidemic and Its Aftermath in China: A Political Perspective,” in *Learning from SARS: Preparing for the Next Disease Outbreak: Workshop Summary* (Washington, DC: National Academies Press, 2004), https://www.ncbi.nlm.nih.gov/books/NBK92479; Julia Hollingsworth and Yong Xiong, “China’s Truthtellers: The People Who Shared Details of the Covid-19 Pandemic That Beijing Left Out,” CNN, 2021, https://www.cnn.com/interactive/2021/02/asia/china-wuhan-covid-truth-tellers-intl-hnk-dst; and Jin Wu et al., “How the Virus Got Out,” *The New York Times*, March 22, 2020, https://www.nytimes.com/interactive/2020/03/22/world/coronavirus-spread.html.

135. David Stanway, “The Shadow of SARS: China Learned the Hard Way How to Handle an Epidemic,” Reuters, January 23, 2020, https://www.reuters.com/article/idUSKBN1ZL133.

136. Georgios Pappas, “The Lanzhou Brucella Leak: The Largest Laboratory Accident in the History of Infectious Diseases?” *Clinical Infectious Diseases* 75, no. 10 (November 14, 2022): 1845–47, https://doi.org/10.1093/cid/ciac463.

137. Furmanski, *Laboratory Escapes and “Self-Fulfilling Prophecy” Epidemics*, 10–11; Martin Furmanski, “Threatened Pandemics and Laboratory Escapes: Self-Fulfilling Prophecies,” Bulletin of the Atomic Scientists, March 31, 2014, https://thebulletin.org/2014/03/threatened-pandemics-and-laboratory-escapes-self-fulfilling-prophecies.

138. Josh Rogin, “State Department Cables Warned of Safety Issues at Wuhan Lab Studying Bat Coronaviruses,” *The Washington Post*, April 20, 2020, https://www.washingtonpost.com/opinions/2020/04/14/state-department-cables-warned-safety-issues-wuhan-lab-studying-bat-coronaviruses.

139. Katherine Eban, “A Big Week for the ‘Lab Leak’: Making Sense of the Latest Twists in the COVID-19 Origins Debate,” *Vanity Fair*, March 1, 2023, https://www.vanityfair.com/news/2023/03/covid-19-origins-lab-leak; *Potential Links between the Wuhan Institute of Virology and the Origin of the COVID-19 Pandemic* (Office of the Director of National Intelligence, June 2023), https://www.dni.gov/files/ODNI/documents/assessments/Report-on-Potential-Links-Between-the-Wuhan-Institute-of-Virology-and-the-Origins-of-COVID-19-20230623.pdf; *Updated Assessment on COVID-19 Origins* (Office of the Director of National Intelligence, National Intelligence Council, 2021), https://www.dni.gov/files/ODNI/documents/assessments/Declassified-Assessment-on-COVID-19-Origins.pdf; Nicholas Wade, “Where Did Covid Come From?” *The Wall Street Journal*, Opinion, February 28, 2024, https://www.wsj.com/articles/where-did-covid-come-from-new-evidence-lab-leak-hypothesis-78be1c39; and Erika Kinetz, “Anatomy of a Conspiracy: With COVID, China Took Leading Role,” Associated Press, February 15, 2021, https://apnews.com/article/pandemics-beijing-only-on-ap-epidemics-media-122b73e134b780919cc1808f3f6f16e8.

140. Discussion with expert in gene synthesis screening policy (name withheld by agreement), 2024.

141. Discussion with expert in gene synthesis screening policy (name withheld by agreement), 2024.

142. The White House, “Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI,” press release, July 21, 2023, https://www.

whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai. See also Anthropic, "Frontier Threats Red Teaming for AI Safety"; OpenAI, "GPT-4 System Card," March 23, 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

143. "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023," AI Safety Summit, November 1, 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

144. Nitasha Tiku, Gerrit De Vynck, and Will Oremus, "Big Tech Was Moving Cautiously on AI. Then Came ChatGPT," *The Washington Post*, February 3, 2023, https://www.washingtonpost.com/technology/2023/01/27/chatgpt-google-meta; Tom Dotan and Deepa Seetharaman, "The Awkward Partnership Leading the AI Boom," *The Wall Street Journal*, June 13, 2023, https://www.wsj.com/articles/microsoft-and-openai-forge-awkward-partnership-as-techs-new-power-couple-3092de51.

145. *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,* Exec. Order No. 14110, 88 FR 75191 (October 30, 2023), https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

146. Kuntz, *Genomes*, 4–5.

147. While biological agents other than viruses can replicate and spread from person to person, the most transmissible agents are viruses. See *The Role of Building Ventilation and Filtration in Reducing Risk of Airborne Viral Transmission in Schools, Illustrated with SARS-COV-2* (California Department of Public Health, Indoor Air Quality Section, Environmental Health Laboratory Branch, Center for Healthy Communities, September 1, 2020), app. 2, https://www.cdph.ca.gov/Programs/CCDPHP/DEODC/EHLB/AQS/Pages/Airborne-Infections.aspx; Rachael M. Jones and Lisa M. Brosseau, "Aerosol Transmission of Infectious Disease," *Journal of Occupational and Environmental Medicine* 57, no. 5 (May 2015): 501–8, https://doi.org/10.1097/JOM.0000000000000448; Chia C. Wang et al., "Airborne Transmission of Respiratory Viruses," *Science* 373, no. 6558 (August 27, 2021): eabd9149, https://doi.org/10.1126/science.abd9149.

148. For a more comprehensive overview of types of BDTs that could support bioweapon development, see Rose and Nelson, *Understanding AI-Facilitated Biological Weapon Development*. Current protein generation models show reasonable accuracy in outputting appropriate sequences based on text descriptions of known proteins, and some capabilities to generate modifications of a given protein to better match text descriptions: Shengchao Liu et al., "A Text-Guided Protein Design Framework" (arXiv, February 9, 2023), http://arxiv.org/abs/2302.04611. For discussion of whether rapid, surprising progress might be seen in this domain in the coming years—as it has previously with language models, image models, and protein structure models—see Misha Yagdin, *Generative Biology on the Way: How Soon Will We See Powerful AI Models in Biology?* (Arb Research, May 25, 2023), https://arbresearch.com/files/gen_bio.pdf.

149. Anders Sandberg and Cassidy Nelson, "Who Should We Fear More: Biohackers, Disgruntled Postdocs, or Bad Governments? A Simple Risk Chain Model of Biorisk," *Health Security* 18, no. 3 (June 2020): 155–63, https://doi.org/10.1089/hs.2019.0115.

150. Jessica Billingsley, "Beyond Corporate AI: Why We Need an Open-Source Revolution," *Rolling Stone*, November 1, 2023, https://www.rollingstone.com/culture-council/articles/beyond-corporate-ai-why-we-need-open-source-revolution-1234867419.

151. Elizabeth Seger et al., "Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives" (arXiv, September 29, 2023), https://doi.org/10.48550/arXiv.2311.09227.

152. Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish, "LoRA Fine-Tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B" (arXiv, October 31, 2023), https://doi.org/10.48550/arXiv.2310.20624; Xianjun Yang et al., "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models" (arXiv, October 4, 2023), http://arxiv.org/abs/2310.02949; and Anjali Gopal et al., "Will Releasing the Weights of Future Large Language Models Grant Widespread Access to Pandemic Agents?" (arXiv, November 1, 2023), https://doi.org/10.48550/arXiv.2310.18233.

153. Rahul Roy-Chowdhury, "Why Open-Source Is Crucial for Responsible AI Development," World Economic Forum, December 22, 2023, https://www.weforum.org/agenda/2023/12/ai-regulation-open-source.

154. Kyle Miller, *Open Foundation Models: Implications of Contemporary Artificial Intelligence* (Center for Security and Emerging Technology, March 12, 2024), https://cset.georgetown.edu/article/open-foundation-models-implications-of-contemporary-artificial-intelligence; Caleb Withers, *Response to NTIA Request for Comment: "Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights"* (Center for a New American Security, March 27, 2024), https://www.cnas.org/publications/commentary/response-to-ntia-request-for-comment-dual-use-foundation-artificial-intelligence-models-with-widely-available-model-weights.

155. Soice et al., "Can Large Language Models Democratize Access to Dual-Use Biotechnology?" 2.

156. Tessa Alexanian, “Develop a Screening Framework Guidance for AI-Enabled Automated Labs,” in *Bio X AI: Policy Recommendations for a New Frontier* (Federation of American Scientists, December 12, 2023), https://fas.org/publication/bio-x-ai-policy-recommendations.

157. Sarah R. Carter, Jaime. M. Yassif, and Christopher R. Isaac, *Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications and Governance* (Nuclear Threat Initiative, May 2023), https://www.nti.org/wp-content/uploads/2023/05/NTIBIO_Benchtop-DNA-Report_FINAL.pdf; David Baker and George Church, “Protein Design Meets Biosecurity,” *Science* 383, no. 6681 (January 26, 2024): 349–349, https://doi.org/10.1126/science.ado1671.

158. Carter, Yassif, and Isaac, *Benchtop DNA Synthesis Devices*; Gretton et al., “Random Adversarial Threshold Search Enables Specific, Secure, and Automated DNA Synthesis Screening.”

159. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

160. Mouton, Lucas, and Guest, *The Operational Risks of AI in Large-Scale Biological Attacks*; Patwardhan et al., “Building an Early Warning System for LLM-Aided Biological Threat Creation.”

161. Anwar et al., “Foundational Challenges in Assuring Alignment and Safety of Large Language Models,” sec. 3.5.5.

162. Song Wang et al., “Knowledge Editing for Large Language Models: A Survey” (arXiv, October 25, 2023), http://arxiv.org/abs/2310.16218; Anwar et al., “Foundational Challenges in Assuring Alignment and Safety of Large Language Models,” sec. 3.2.4, 3.2.5; and Nathaniel Li et al., “The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning” (arXiv, March 5, 2024), https://arxiv.org/abs/2403.03218v2.

163. Fabian Pedregosa and Eleni Triantafillou, “Announcing the First Machine Unlearning Challenge,” Google Research blog, June 29, 2023, https://blog.research.google/2023/06/announcing-first-machine-unlearning.html.

164. Ronen Eldan and Mark Russinovich, “Who’s Harry Potter? Approximate Unlearning in LLMs” (arXiv, October 4, 2023), http://arxiv.org/abs/2310.02238; Li et al., “The WMDP Benchmark.”

165. David Vergun, “DoD Aims to Shield Warfighters from Novel Biological Agents,” U.S. Department of Defense, January 10, 2023, https://www.defense.gov/News/News-Stories/Article/Article/3261095/dod-aims-to-shield-warfighters-from-novel-biological-agents; Lara Seligman and Erin Banco, “New Worldwide Threats Prompt Pentagon to Overhaul Chem-Bio Defenses,” *Politico*, January 9, 2023, https://www.politico.com/news/2023/01/09/russia-china-chemical-biological-weapons-pentagon-00077035; Stew Magnuson, “New Bio-Defense Strategy to Eschew ‘One Bug, One Drug’ Programs,” *National Defense Magazine*, January 27, 2023, https://www.nationaldefensemagazine.org/articles/2023/1/27/new-bio-defense-strategy-to-eschew-one-bug-one-drug-programs.

166. Nikki Teran, “Why BARDA Deserves More Funding,” Institute for Progress, March 30, 2022, https://ifp.org/why-barda-deserves-more-funding; Arielle D’Souza, “Three Reasons Congress Should Fund Biodefense,” Institute for Progress, January 24, 2023, https://ifp.org/three-reasons-congress-should-fund-biodefense.

167. Juan Cambeiro and Brian Potter, “Indoor Air Quality Is the Next Great Public Health Challenge,” Institute for Progress, June 29, 2023, https://ifp.org/indoor-air-quality; Alec Stapp, Juan Cambeiro, and Britt Lampert, “Research into Far-UVC Systems Can Help Us Prevent the Next Pandemic,” Institute for Progress, December 16, 2022, https://ifp.org/response-to-the-epas-request-for-information-on-better-indoor-air-quality-management.

168. See also Richard Moulange et al., “Towards Responsible Governance of Biological Design Tools” (arXiv, November 27, 2023), https://arxiv.org/abs/2311.15936v3.

About the Center for a New American Security

The mission of the Center for a New American Security (CNAS) is to develop strong, pragmatic, and principled national security and defense policies. Building on the expertise and experience of its staff and advisors, CNAS engages policymakers, experts, and the public with innovative, fact-based research, ideas, and analysis to shape and elevate the national security debate. A key part of our mission is to inform and prepare the national security leaders of today and tomorrow.

CNAS is located in Washington, DC, and was established in February 2007 by cofounders Kurt M. Campbell and Michèle A. Flournoy. CNAS is a 501(c)3 tax-exempt nonprofit organization. Its research is independent and nonpartisan.

CNAS Editorial

DIRECTOR OF STUDIES
Paul Scharre

DEPUTY DIRECTOR OF STUDIES
Katherine L. Kuzminski

PUBLICATIONS & EDITORIAL DIRECTOR
Maura McCarthy

ASSOCIATE EDITOR
Caroline Steel

CREATIVE DIRECTOR
Melody Cook

DESIGNER
Rin Rothback

Cover Art & Production Notes

COVER ILLUSTRATION
Rin Rothback

PRINTER
CSI Printing & Graphics
Printed on an HP Indigo Digital Press

Center for a New American Security
1701 Pennsylvania Ave NW
Suite 700
Washington, DC 20006
CNAS.org
@CNASdc

Contact Us

202.457.9400

info@cnas.org

Center for a
New American
Security