Chemical & Biological Weapons and Artificial Intelligence: Problem Analysis and US Policy Recommendations

27th February 2024

Hamza Chaudhry & Landon Klein
policy@futureoflife.org

The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence. Since its founding ten years ago, FLI has taken a leading role in advancing key disciplines such as AI governance, AI safety, and trustworthy and responsible AI, and is widely considered to be among the first civil society actors focused on these issues. FLI was responsible for convening the first major conference on AI safety in Puerto Rico in 2015, and for publishing the Asilomar AI principles, one of the earliest and most influential frameworks for the governance of artificial intelligence, in 2017. FLI is the UN Secretary General's designated civil society organization for recommendations on the governance of AI and has played a central role in deliberations regarding the EU AI Act's treatment of risks from AI. FLI has also worked actively within the United States on legislation and executive directives concerning AI. Members of our team have contributed extensive feedback to the development of the NIST AI Risk Management Framework, testified at Senate AI Insight Forums, participated in the UK AI Summit, and connected leading experts in the policy and technical domains to policymakers across the US government.

Domain Definition

A Chemical Weapon is a chemical used intentionally to kill or harm with its toxic properties. Munitions, devices and other equipment specifically designed to weaponize toxic chemicals also fall under the definition of chemical weapons. Chemical agents such as blister agents, choking agents, nerve agents and blood agents have the potential to cause immense pain and suffering, permanent damage and death.[1] After these weapons caused millions of casualties in both world wars, 200 countries signed the Chemical Weapons Convention - enforced by the Organization for the Prohibition of Chemical Weapons (OPCW) - and sought to destroy their chemical stockpiles. With the destruction of the last chemical weapon by the United States in July 2023, the OPCW has declared the end of all official chemical stockpiles.[2] While small-scale attacks by non-state actors and rogue state actors have occurred over the last fifty years, these are isolated cases. The vast majority of chemical weapons have been eradicated.

Biosecurity encompasses actions to counter biological threats, reduce biological risks, and prepare for, respond to and recover from biological incidents - whether naturally occurring, accidental, or deliberate in origin and whether impacting human, animal, plant, or environmental health. The National Biodefense Strategy and Implementation Plan published by the White House in October 2022 finds biosecurity to be critical to American national security interests, economic innovation, and scientific empowerment.[3] In addition, American leadership from both sides of the political spectrum has undertaken significant investments in strengthening biosecurity over the last two decades. Finally, the COVID-19 pandemic has crystallized the threat to American life, liberty, and prosperity from pandemics in the future.

Problem Definition

Artificial intelligence (AI) could reverse the progress made in the last fifty years to abolish chemical weapons and develop strong norms against their use. As part of an initiative at the Swiss Federal Institute for Nuclear, Biological, and Chemical (NBC) Protection, a computational toxicology company was asked to investigate the potential dual-use risks of AI systems involved in drug discovery. The initiative demonstrated that these systems could generate thousands of novel chemical weapons. Most of these new compounds, as well as their key precursors, were not on any government watch-lists due to their novelty.[4] This is even more concerning in light of the advent of large language model (LLM) based artificial agents. This is because the ability of artificial agents to sense their environment, make decisions, and take actions compounds the unpredictability and risks associated with this kind of research.

On the biological weapons front, cutting-edge biosecurity research, such as gain-of-function research, qualifies as dual-use research of concern – i.e. while such research offers significant potential benefits it also creates significant hazards. For instance, such research may be employed to develop vital medical countermeasures or to synthesize and release a dangerous pathogen. Over the last two decades, the cost of advanced biotechnology has rapidly decreased and access has rapidly expanded through advancements in cheaper and more accessible DNA sequencing, faster DNA synthesis, the discovery of efficient and

1 What is a Chemical Weapon? Organization for the Prohibition of Chemical Weapons. https://www.opcw.org/our-work/what-chemical-weapon

2 US Completes Chemical Weapons Stockpile Destruction Operations. Department of Defense. https://www.defense.gov/News/Releases/Release/Article/3451920/us-completes-chemical-weapons-stockpile-destruction-operations/

3 National Biodefense Strategy And Implementation Plan. The White House. https://www.whitehouse.gov/wp-content/uploads/2022/10/National-Biodefense-Strategy-and-Implementation-Plan-Final.pdf

4 Dual Use of Artificial Intelligence-powered Drug Discovery. National Center for Biotechnology Information. National Institutes of Health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9544280/

accurate gene-editing tools such as CRISPR/Cas9, and developments in synthetic biology.[5]

Accompanying these rapid developments are even faster advancements in AI tools used in tandem with biotechnology. For instance, advanced AI systems have enabled several novel practices such as AI-assisted identification of virulence factors and in silico design of novel pathogens.[6] More general-purpose AI systems, such as large language models, have also enabled a much larger set of individuals to access potentially hazardous information with regard to procuring and weaponizing dangerous pathogens, lowering the barrier of biological competency necessary to carry out these malicious acts.

The threats posed by biological and chemical weapons in convergence with AI are of paramount importance. Sections 4.1 and 4.4 of the White House Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence explicitly outline the potential Chemical, Biological, Radiological and Nuclear (CBRN) threats posed by advanced AI systems.[7] They can be broadly divided into two categories:

#1. Exponentially Enhanced Capacity to Engineer Deadly Toxins and Biological Weapons

As discussed in the example of the toxicology company above, there is growing concern regarding the potential misuse of molecular machine learning models for harmful purposes. The dual-use application of models for predicting cytotoxicity to create new poisons or employing AlphaFold2 to develop novel toxins has raised alarm. Recent developments in AI have allowed for an expansion of open-source biological design tools (BDTs), increasing access by bad actors.[8] This creates three kinds of risks:

A. **Increased Access to Rapid Identification of Toxins:** The MegaSyn AI software used by the toxicology company discussed was able to find 40,000 toxins with minimal digital architecture (namely some programming), open-source data, a 2015 Mac computer and less than six hours of machine time.[9] This suggests that AI systems may democratize the ability to create chemical weapons, increasing access by non-state actors, rogue states or individuals acting on their own who would otherwise be precluded by insufficient resources. Combined with the use of LLMs and other general-purpose AI tools, the bar for expert knowledge needed to develop chemical weapons has been substantially lowered, further diffusing the ability to identify and release deadly toxins.

B. **Discovery of Novel Toxins:** An important aspect of the findings from the experiment discussed above is that the AI system not only found VX and other known chemical weapons; it also discovered thousands of completely new putatively toxic substances. This creates serious hazards for chemical defense, as malevolent actors may try to make AI systems develop novel toxins that are less well understood, and for which defensive, neutralizing, or treatment procedures have not yet been developed.

C. **AI-Accelerated Development of Biological Design Tools:** These tools span different fields such as bio-informatics, genomics, synthetic biology, and others. In essence, these tools allow smaller groups of individuals, with fewer resources, to discover, synthesize, and deploy enhanced pathogens

5 The Blessing and Curse of Biotechnology: A Primer on Biosafety and Biosecurity. Carnegie Endowment for International Peace. https://carnegieendowment.org/2020/11/20/blessing-and-curse-of-biotechnology-primer-on-biosafety-and-biosecurity-pub-83252

6 Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology. National Center for Biotechnology Information. National Institutes of Health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7310294/

7 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

8 Bio X AI: Policy Recommendations For A New Frontier. Federation of American Scientists. https://fas.org/publication/bio-x-ai-policy-recommendations/

9 AI suggested 40,000 new possible chemical weapons in just six hours. The Verge. https://www.theverge.com/2022/3/17/22983197/ai-new-possible-chemical-weapons-generative-models-vx

of pandemic potential (PPPs). Critically, these AI systems can amplify risks from gain-of-function research, enabling malevolent actors to make pathogens more deadly, transmissible, and resilient against medical counter-measures.[10] AI assistance can also help bad actors direct such bio-weapons at targets of particular genotypes, races, ethnicities, tribes, families, or individuals, facilitating the conduct of genocide at a potentially global scale.[11]

#2. Increased Access to Dangerous Information and Manipulation Techniques Through LLMs

Outside the use of narrow AI systems to discover deadly toxic substances, developments in general purpose systems such as large language models may allow malevolent actors to execute many of the other steps needed to deploy a chemical weapon. Essential steps include baseline knowledge of chemistry and biology, access to critical materials and lab infrastructure, and access to means of deploying the weapon (e.g. munitions). LLMs equip malevolent actors with the ability to send deceptive emails and payments to custom manufacturers of chemical and biological materials, access substances through illicit markets, and hire temporary workers to accomplish specialized, compartmentalized tasks around the world. Taken together, these capacities enable the production and deployment of chemical weapons. More narrow AI systems have displayed effectiveness in writing code to exploit technical loopholes in the cybersecurity architecture of several organizations, such as identifying and exploiting zero-day vulnerabilities.[12] Such techniques could be used to target critical bio-infrastructure such as Biosafety Level 3 and 4 Labs (BSL-3 and BSL-4), research laboratories, hospital networks, and more. These practices could enable access to dangerous information or be used to cripple recovery and response to a high-consequence biological incident.

An experiment conducted at MIT demonstrated that students without a technical background were able within 60 minutes to use LLMs to identify four potential pandemic pathogens, explain how they can be generated from synthetic DNA using reverse genetics, supply the names of DNA synthesis companies unlikely to screen orders, identify detailed protocols and how to troubleshoot them, and recommend that anyone lacking the skills to perform reverse genetics engage a core facility or contract research organization.[13] Other experiments conducted across different settings and time horizons have also demonstrated how large language models can be exploited to access and/or use hazardous information.[14] Traditionally, access to this kind of expertise and information was mediated through established systems (completing a Ph.D. in an advanced field, being hired by a top research laboratory, meeting specified safety and security criteria for conducting sensitive research, etc.). Nowits democratization allows many more individuals, with less skill and less intelligence, to access this knowledge and potentially use it to cause considerable harm.

AI-powered cyberattacks also present a threat to biosecurity and chemical security. Advancements in AI have allowed a wider net of actors to construct more easily cyber exploits that could be used to target cyber-vulnerabilities in water treatment facilities, research labs and containment facilities, to cause widespread harmful chemical or biological exposure. In addition, AI systems may be used to improve the cyber-manipulation techniques used by malevolent actors. Cyber-manipulation encompasses a wide array of practices such as spearphishing, pharming, smishing, vishing, and others intended to deceive, blackmail,

10 The Convergence of Artificial Intelligence and the Life Sciences. Nuclear Threat Initiative. https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/

11 The Coming Threat of a Genetically Engineered 'Ethnic Bioweapon'. National Review. https://www.nationalreview.com/corner/the-coming-threat-of-a-genetically-engineered-ethnic-bioweapon/

12 US adversaries employ generative AI in attempted cyberattack. Security Magazine. https://www.securitymagazine.com/articles/100418-us-adversaries-employ-generative-ai-in-attempted-cyberattack

13 Can large language models democratize access to dual-use biotechnology? Computer and Society. https://arxiv.org/abs/2306.03809

14 The Operational Risks of AI in Large-Scale Biological Attacks. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA2977-1.html

mislead, or otherwise compel the victim of such a practice to reveal high-value information. Large language models have demonstrated a considerable capacity to amplify the power of these illegal practices, which could allow malevolent actors to access dangerous biological information or infrastructure by manipulating owners of DNA synthesis companies, prominent academics in the field, and biosecurity professionals.[15] While many large language models have some preliminary guardrails built in to guard against this misuse, several experiments have demonstrated that even trivial efforts can overcome these safeguards.[16] For instance, relabeling of these toxic substances within the data of the model can overcome safeguards which were set up to preclude them from providing dangerous information. Prompt engineering by compartmentalizing (breaking up one dangerous process into several steps which seem innocuous by themselves), as well as faking authority (pretending to be in charge of a government chemical facility), have also yielded success in manipulating these models.[17]

Policy Recommendations

In light of the significant challenges analyzed in the previous section, considerable attention from policymakers is necessary to ensure the safety and security of the American people. The following policy recommendations represent critical, targeted first steps to mitigating the risks posed by AI in the domains of chemical and biosecurity:

1. **Explicit Requirements to Evaluate Advanced General Purpose AI Systems for Chemical Weapons Use:** There is considerable ongoing policy discussion to develop a framework for evaluating advanced general purpose AI systems before and after they are developed and/or deployed, through red-teaming, internal evaluations, external audits and other mechanisms. In order to guard against emerging risks from biological and chemical weapons, it is vital that these evaluations explicitly incorporate a regimen for evaluating a system's capacity to facilitate access to sensitive information and procedures necessary to develop chemical weapons. This could include the capability of these systems to provide dangerous information as discussed, as well as the capability to deceive, manipulate, access illicit spaces, and/or order illegal financial transactions. In order to prevent malevolent actors from accessing hazardous information and expertise, or further exploiting AI systems to access high-risk infrastructure, it is also critical to set up minimum auditing requirements for these general-purpose systems before launch. These practices could help test and strengthen the safeguards underpinning these systems. Such a requirement could also be incorporated into the existing risk management frameworks, such as the NIST AI Risk Management Framework.
2. **Restrict the Release of Model Weights for Systems that Could be Used, or Modified to be Used, to Discover Dangerous Toxins:** In order to reduce the ability of malevolent actors to use AI capabilities in production of dangerous chemical toxins, it is critical that both narrow and general-purpose AI systems that are shown to be dangerous in this regard (as well as future iterations of those and similar systems) include significant restrictions on access both for use and to the underlying model weights . Critically, the release of model weights is an irreversible act that eliminates the capacity to restrict use in perpetuity. Accordingly, red-teaming procedures such as those mentioned in the previous recommendation must include extensive assessment to confirm the lack of potential for

15 AI tools such as ChatGPT are generating a mammoth increase in malicious phishing emails. CNBC. https://www.cnbc.com/2023/11/28/ai-like-chatgpt-is-creating-huge-increase-in-malicious-phishing-email.html

16 NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems. National Institutes of Standards and Technology. https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems

17 Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. Computer Engineering. https://arxiv.org/abs/2305.13860

these dangerous capabilities, and for modification or fine-tuning to introduce these dangerous capabilities, if the developer intends to release the model weights..[18]

3. **Ring-fence Dangerous Information from Being Used to Train Large Language Models:** In order to ensure that general-purpose AI systems do not reveal hazardous information, it is vital to require that companies not use this kind of information during training runs to train their AI models. Proactively keeping information that would very likely pose a significant health and/or safety issue to the general public classified using new classification levels and initiatives would significantly reduce these risks.[19]
4. **Incorporating AI Threats into Dual Use Research of Concern Guidance and Risk Frameworks:** Over the last two decades, considerable policy attention has been devoted to establishing policy frameworks, including guidance and requirements, for biosecurity. However, these frameworks do not currently include policy prescriptions and guidance for unique challenges posed by AI. National-level policy frameworks such as those published by the National Science Advisory Board for Biosecurity (NSABB), the CDC, HHS, DHS, and others must explicitly integrate concerns at the convergence of AI and biosecurity, and establish technical working groups within these bodies populated by experts in both fields to study these risks. Finally, these convergence risks should also be integrated into AI risk frameworks such as the NIST AI RMF. With the exception of the NIST AI RMF, all of these regulatory directives and review regimes were instituted before the exponential development of AI systems seen over the last few years. It is important to update this guidance and include explicit provisions for the use of AI in dual-use biological and chemical research.
5. **Expand Know Your Customer (KYC) and Know Your Order (KYO) Requirements:** Companies that provide sequencing and synthesis services, research laboratories, and other relevant stakeholders should be required to follow KYC and KYO standards, ensuring that potentially dangerous sequences are kept out of the hands of malevolent actors.[20] Regulation should further require standardized, scalably secure synthesis screening methods (such as SecureDNA). These requirements must also include assurance that correspondence pertaining to these services is between human agents and not involving AI systems.
6. **Strengthen Existing Capabilities and Capacities for Biodefense:** As developments in AI and biotechnology accelerate, it is also vital to ensure that there is considerable capacity to prevent, detect, and respond to high-consequence biological incidents of all kinds. This includes significant investments in early warning and detection, response capacities, interoperability and coordination, national stockpiles of PPEs and other relevant infrastructure, supply-chain resilience, development of medical countermeasures, and accountability and enforcement mechanisms to disincentivize both accidents and intentional misuse.[21]

More general oversight and governance infrastructure for advanced AI systems is also essential to protect against biological and chemical risks from AI, among many other risks. We further recommend these broader regulatory approaches to track, evaluate, and incentivize the responsible design of advanced AI systems:

1. **Require Advanced AI Developers to Register Large Training Runs and to "Know Their Customers":** The Federal Government lacks a mechanism for tracking the development and proliferation of advanced AI systems that could exacerbate bio-risk. To mitigate these risks adequately, it is essential

18 BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13. Computation and Language. https://arxiv.org/abs/2311.00117

19 Artificial Intelligence in the Biological Sciences: Uses, Safety, Security, and Oversight. Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R47849

20 Preventing the Misuse of DNA Synthesis Technology. Nuclear Threat Initiative. https://www.nti.org/about/programs-projects/project/preventing-the-misuse-of-dna-synthesis-technology/

21 Biosecurity In The Age Of AI. Helena Biosecurity. https://www.helenabiosecurity.org

to know what systems are being developed and who has access to them. Requiring registration for the acquisition of large amounts of computational resources for training advanced AI systems, and for carrying out the training runs themselves, would help with evaluating possible risks and taking appropriate precautions. "Know Your Customer" requirements similar to those imposed in the financial services industry would reduce the risk of systems that can facilitate biological and chemical attacks falling into the hands of malicious actors.

2. **Clarify Liability for Developers of AI Systems Used in Bio- and Chemical Attacks:** It is not clear under existing law whether the developers of AI systems used by others, for example to synthesize and launch a pathogen, would be held liable for resulting harms. Absolving developers of liability in these circumstances creates little incentive for profit-driven developers to expend financial resources on precautionary design principles and robust assessment. Because these systems are opaque and can possess unanticipated, emergent capabilities, there is inherent risk in developing advanced AI systems and systems expected to be used in critical contexts. Implementing strict liability when these systems facilitate or cause harm would better incentivize developers to take appropriate precautions against vulnerabilities, and the risk of use in biological and chemical attacks.