Written statement of

Alexander J. Titus, PhD

Principal Scientist, AI Division, Information Sciences Institute

University of Southern California

Before the

U.S. Senate AI Insight Forum

"Risk, Alignment, & Guarding Against Doomsday Scenarios"

December 06, 2023

Introduction

Leader Schumer, Senator Rounds, Senator Heinrich, Senator Young, and other distinguished members of this Forum, it is my pleasure to participate today in this AI Insight Forum focused on Risk, Alignment, & Guarding Against Doomsday Scenarios. At the outset I would like to note that my comments today represent my personal views, and not necessarily those of the institutions that I work with. Artificial intelligence (AI) is arguably the most pressing technology topic on the minds of Congress today, and rightfully so. If we, as a country, take the right actions to ensure the responsible development of AI, we will see the transformative power of AI in society. If we do not, then we risk losing our leadership in this domain and with it, our ability to drive alignment and responsible technology development.

I would like to focus my comments on the implications of AI in the life sciences, and in particular, the doomsday scenario of AI systems enhancing access to biological weapons and/or pandemic agents that cause widespread loss of life and economic damage. There are other reasonable concerns about AI in the lifesciences as well, including risks of inequality, risks to "human dignity," uneven access to technology, AI outpacing our understanding of natural systems and more. But for the sake of these comments, I've focused on the scenario above. There has been considerable dialogue about this scenario, recently exacerbated by the advances in AI the past year, particularly with the emergence of Large Language Models (LLMs). This has caused serious concern about elevated biosecurity risks and has elicited a wide array of calls for swift and broad-sweeping policy, regulatory, and legislative action. As a subject matter expert in this domain, I urge this Forum to take a refined approach towards action to prevent this potential doomsday scenario. In this domain, we risk serious repercussions from overstepping due to a false positive assessment of risks and unnecessarily hindering our ability to develop new disease treatments, feed our growing population, unlock new product economies, and more. At the same time, there are risks of under-stepping due to a false negative assessment of risk, allowing this technology to advance without effective mechanisms of quantification of those risks. These technologies are complex, not just single domains, and affect nearly aspect of our society today[1].

My name is Alexander Titus and I am a Principal Scientist at the Information Sciences Institute (ISI) of the University of Southern California. At ISI, I lead a research group focused on developing benchmarks to empirically assess the capabilities and implications of AI in the life sciences. I am also an appointed member of the National Security Commission on Emerging Biotechnology, of which Senator Young is also a Commissioner, along with ten experts from Congress, the private sector, and academia.

I started my government service career in the intelligence community while I was in graduate school conducting some of the early research into the use of generative AI to study biology. In 2019, I had the honor of serving as the inaugural Principal Director for Biotechnology at the

[1] The Bioeconomy: A Primer

Department of Defense, where my team established the first enterprise biotechnology strategy, which included many applications of AI for its ability to accelerate our capability development.

I have been working on these issues for many years, with expertise at the intersection of AI and the life sciences. I have spoken with thousands of experts from our national security community, industry, and universities throughout these years. It is with this experience that I urge this Forum to take a refined approach to policy, regulatory, and legislative action in response to emerging concerns of AI in the life sciences. We need more evidence at this point to assess the risks and timeline of the possibility of the doomsday scenario of AI engineering a completely novel biothreat unlike the world has ever seen, or an "unknown unknown". More specifically, we need ways to measure how much AI increases risk above those currently present without AI tools. We cannot prove an absence of risk, but we can build empirical ways to measure change over time.

I urge the AI Insight Forum, and any legislative action that results from the lessons learned from such a breadth of experts, to consider three things:

1. AI and the life sciences are broad categories of technology, and there is no single assessment of risk nor implementation of policy, regulatory, and legislative action.
2. Policymakers should consider input as to the impact of potential actions from a broader set of stakeholders than the AI and biosecurity technical communities. The implications of such actions will directly impact the lives of millions of people. Input is needed from economists, ethicists, healthcare providers, lawyers, social scientists, and more. An approach referred to as Violet Teaming[2][3].
3. Policymakers should seek to establish a mechanism to assess progress toward near-term and doomsday risk scenarios and develop empirical baselines to assess current and emerging AI models and their capabilities in the life sciences. This can provide an assessment of baseline capabilities and departures from those baselines.

1. The complexity of AI and the life sciences

Artificial intelligence is a broad category of capabilities that have varying levels of implications for the life sciences. Generative AI, largely focused on Large Language Models (LLMs) at the moment, lowers the barrier to access to knowledge at a scale, as some argue, not seen since the advent of the internet and the smartphone. These are generally broadly applicable and not particularly specialized. Biological Design Tools (BDTs), such as AlphaFold[4], are specific algorithms that are designed and used to effectively engineer biological systems, whether it be the protein of a drug or the genetic circuit of an industrial synthetic biology product. Any risk from these two categories of AI will manifest differently and thus should be regulated as such[5].

[2] Red Teaming Improved GPT-4. Violet Teaming Goes Even Further

[3] The Promise and Peril of Artificial Intelligence -- Violet Teaming Offers a Balanced Path Forward

[4] AlphaFold Protein Structure Database

[5] Why AI for biological design should be regulated differently than chatbots

Assessing Risk

I encourage the Forum to consider the risk of AI in the life sciences from two categorizations. The first is the distinction between the (1) operational aspects of biology and (2) the risk of completely new biological systems, or "unknown unknowns".

The second categorization is how (1) AI is changing what is possible today (e.g. engineering microorganisms), how (2) AI is changing what is possible in the future (e.g. design of new microorganisms from scratch), and how (3) AI impacts the physical world.

I also encourage the Forum to continue to ask the question "How much does AI increase risk compared to a baseline?" Without a baseline, effective measurement of risk is not possible and any assessment is important but will remain incomplete without a comparator. I have spoken to numerous practitioners who are red-teaming modern AI's ability to support the design and dissemination of biological weapons and/or pandemic agents (i.e. enhance operational biology). All of the assessments I have seen have indeed indicated that these LLMs have the ability to synthesize information that one may find useful to carry out these harmful actions. These preliminary conclusions give us further evidence that we must assess. Each time I ask the teams "How much of the knowledge the LLMs provide can be found on the internet? How much does this increase risk beyond what exists today?" but to my knowledge, these comparisons have not yet been conducted and would provide valuable baseline assessments.

2. Building Violet Teams to assess AI risk in the life sciences

The life sciences are broad-sweeping and are in some way incorporated into nearly every facet of our lives. Healthcare, food and agriculture, and industrial biotechnologies, among others, directly impact the lives of millions of people. As our population grows, we must continue to improve our food supplies. As our population ages and lives longer, we must continue to improve our healthcare and treatment of disease. As we strive towards a sustainable economy, we must continue to develop novel products and production methods with lower environmental impacts.

Advances in all of these domains will require the application of AI to keep pace with growing needs. Thus any policy, regulatory, and/or legislative action intended to prevent the doomsday scenario that I have described will inherently have repercussions on all of these domains. As such, especially as we build methods and collect empirical evidence of AI-enhanced risk of such a scenario, additional stakeholders should be incorporated into the conversation; this addition is not to complicate or slow discussions, but to make robust "minimally viable policy" that balances the promise and perils of AI. To date, the primary discussion of these doomsday scenarios comes from technical AI developers and the biosecurity community. However, the implications of actions should include input from economists to understand the opportunity costs of action, and by extension an understanding of what it costs in the long run to not take action. Input from providers and drug developers should be considered as to how action may hinder our

ability to develop new treatments and care plans, and the patient view on cost/benefit analysis should be considered. Input from ethicists and social scientists should be considered as well.

This idea is an expansion of the traditional red-teaming approach to considering technology risks. Especially in the context of doomsday scenarios, where policymakers are considering action to prevent potential risks in the future, the tradeoffs of action should be weighed carefully. This idea is known as Violet Teaming and was introduced in the context of expanding the risk assessment of ChatGPT-4[6] and my team is working on expanding this input mechanism to assess the promise and peril of AI in the life sciences[7]. Many of the same conversations are happening in siloed expert domains but these experts rarely have the opportunity to provide joint input.

Given the expansive nature of both AI and the life sciences, and how risk should be assessed in targeted contexts, policymakers should leverage Violet Teams to create mechanisms for input and feedback to provide data to inform policy actions. This is actually becoming easier and more scalable with advances in AI because the tools make communication easier than previously possible. These teams can also provide a regularly updated assessment of risk with additional context as information evolves. Recently introduced legislation[8], and the actions under consideration by this Forum, are prime opportunities to do so.

3. Empirical Baselines to Assess Risk

In any scenario of risk, it is crucial that empirical baseline metrics be established so that a risk assessment, and a change-in-risk-over-time assessment, can be conducted. Particularly when considering action to guard against the risk of doomsday scenarios, an empirical assessment of risk will be key. In AI applied to the life sciences, metrics, and baselines for those metrics, have not yet been established. My team, and the work of many others in the field, is actively assessing the generative biology capabilities as new models emerge.

In the doomsday scenario I have focused on, the concern is that AI will allow a bad actor to create a novel bioweapon and/or pandemic agent that we are not prepared for and that we are not equipped to respond to. The "unknown unknown". To quantify the potential risk in this case, we should establish a systematic assessment of how capable AI systems are, as of today, at generating completely novel sequences vs. augmenting and optimizing existing sequences. We should also establish a method of baselining how readily accessible the information provided by a given LLM is on the open internet. For example, if a model is introduced that can generate a pathogen with a new adaptive trait we have not seen before in nature, as validated with experimental models, it would give us evidence of the risk of AI in this case and would inform the trade-off decisions where the benefits of that technology may be outweighed by the risks.

---

[6] Red Teaming Improved GPT-4. Violet Teaming Goes Even Further

[7] The Promise and Peril of Artificial Intelligence -- Violet Teaming Offers a Balanced Path Forward

[8] Sens. Markey, Budd Announce Legislation to Assess Health Security Risks of AI

This scenario also requires AI-generated sequences to be physically produced in the real world, giving good reason to enhance screening at DNA synthesis companies as outlined in many recent recommendations[9][10]. Another valuable baseline is an assessment of whether AI-generated sequences are any easier to get past screening protocols than an expert-engineered sequence. There have been recent efforts to encourage the establishment of benchmarks to assess AI progress toward any number of doomsday scenarios that would provide necessary information to inform the baseline risk and change in risk over time assessment needed, but more is needed.

Autonomous Science

I would also like to highlight the flipside of applying AI in the life sciences. One key aspect of Violet Teaming is to leverage the very technology of concern for solutions to those concerns. In the context of the scenario outlined here, AI and autonomous science may provide the rapid response capabilities we need to respond to a new biological threat. If we limit our ability to leverage AI in this capacity, we may inadvertently hinder our ability to prevent the very threats that we are concerned about, thus giving others an advantage that would otherwise not exist.

Conclusion

Artificial intelligence poses risks in many domains if not developed responsibly with alignment and risk mitigation in mind. In the context of AI in the life sciences, there are a number of doomsday scenarios that focus on AI meaningfully increasing the risk of catastrophic harm due to a novel bioweapon and/or pandemic agent. These scenarios require an empirical assessment, and tracking over time, of AI-enhanced risk over a baseline of existing threats. As a practitioner and national security professional, I encourage the AI Insight Forum and the members of Congress who have the foresight to convene these important discussions, to take a balanced approach to action with regard to these doomsday scenarios. This is an opportunity to build an interactive process of AI science, engineering, and policy development.

One doomsday scenario that is rarely discussed and is absent from the public dialogue is the one in which fear causes an unnecessary restriction on the use of modern AI tools for biotechnology. In this scenario, people die unnecessarily because a limitation has hindered drug development, people go hungry because our ability to advance food production has stalled, and we see slower than necessary economic growth because product development cannot progress fast enough.

We can maximize the opportunity of AI in the life sciences while simultaneously minimizing the risk by (1) taking action strategically and in the context of specific risks, (2) incorporating a broad set of input to assess the opportunity-cost trade-offs, and (3) establishing methods to empirically assess where AI increases risk above that which already exists.

---

[9] FS: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

[10] Dr. Tom Inglesby: Policy Considerations for Artificial Intelligence in Health Care