EBRC Engineering Biology Research Consortium

Security Considerations at the Intersection of Engineering Biology and Artificial Intelligence

A White Paper by the Engineering Biology Research Consortium[i]

Compiled and edited by Charlie D. Johnson, EBRC Intern; Wilson Sinclair, EBRC Postdoctoral Scholar; Rebecca Mackelprang, EBRC Associate Director for Security Programs

November 2023

Overview & Key Recommendation

As researchers and innovators work to leverage the power and opportunity of artificial intelligence to advance engineering biology to address societal challenges, stakeholders must also recognize the potential of AI-enabled capabilities to cause harm. Herein, we consider three areas—*de novo* biological design, closed loop-autonomous laboratory systems, and natural language Large Language Models—where the intersection of these technologies may pose significant security concerns. For each, we describe the technology, associated security concerns, and how those concerns might be addressed.

We recommend that a regularly convened international forum or initiative be established to bring diverse stakeholders together to identify security concerns at the intersection of engineering biology and AI and consider how they might be addressed.

Engineering biology and artificial intelligence (AI)[ii] are both characterized by their rapid growth, their potential to dramatically accelerate technological solutions to pressing challenges, and their tendency to usher in novel security concerns. It is unsurprising, then, that the emerging capabilities at the intersection of these technologies are generating both great enthusiasm and great concern. As the power of this technological convergence is still being explored and realized, it is timely to identify and consider potential security implications and how they might be addressed. By engaging in such a process now, researchers, policymakers, and members of society[iii] can consider carefully what acceptable levels of risk given likely benefits might—be and for whom—as new capabilities are developed.

---

[i] The Engineering Biology Research Consortium brings engineering biology researchers and other stakeholders from industry and academia together with policymakers to advance engineering biology to address national and global needs. EBRC White Papers are developed with significant, substantive direction, guidance, and input from EBRC members through an interactive process. EBRC products do not necessarily reflect the direct views of all EBRC members.

[ii] Artificial Intelligence is an umbrella term that encompasses generative AI, many types of machine learning, and deep learning. To avoid confusion and in recognition that different AI-based approaches can be used to address the same challenges, the authors use "AI" to refer collectively to these approaches.

[iii] This paper was developed by EBRC members, who are generally based in the technical research community. Thus, this paper takes a technical approach to this topic. However, members also pointed to the present opportunity to think more broadly about societal impacts, perceptions of risk, and roles in the determination of technological

Engineering biology applies an engineering design framework to the construction and modification of biological systems at the genetic level. Research is generally guided by the iterative "design-build-test-learn" (DBTL) cycle wherein a genetic circuit, pathway, organism, or even consortium of cells or microbes is built from molecular components (e.g., DNA, RNA, proteins) and tested to measure its performance against the function for which it was designed. That performance is used to inform the next design stage, and the cycle repeats until, ideally, an optimized target function is achieved.

AI enhancement of the design of proteins, pathways,[iv] systems,[v] and even consortia of cells or organisms can decrease the number of iterations through the DBTL cycle and increase the efficiency (what is learned per unit cost) of each cycle. Furthermore, closed-loop, autonomous research systems are in development and may be realized in the coming years. In such systems, programmable robotics would build and test biological parts or systems. AI would learn from those tests and iterate upon the original biological design, thus requiring reduced—if any—human input, skill, and supervision to complete successive DBTL cycles. Additionally, natural language Large Language Models (LLMs) such as ChatGPT are becoming accessible to a far wider range of people whose experimentation is outside the control and/or view of state-funded and state-regulated research institutions, complicating traditional security governance capabilities.

While molecular biology, engineering biology, and bioinformatic techniques and capabilities have advanced rapidly in the twenty-first century, AI-based capabilities may significantly increase this rate of advancement. Individually or in concert, these capabilities do pose security concerns.[vi] AI capabilities could help a malicious actor conceptualize an approach; gather information about how to carry out a plan; design biological parts or systems that could cause harm; develop strategies for avoiding detection; undertake the physical processes of building, testing, and weaponizing a biological system designed to cause harm; and develop a process for distributing the final product. Resulting harmful products might include human, plant, animal, or even microbial pathogens, potentially with enhanced disease characteristics. They could also include novel toxins that evade current detection or treatment mechanisms (e.g., a botulinum toxin that is harmful to humans but not susceptible to current antitoxins), illicit microbially-synthesized opioids, or microbes that degrade or weaken important materials. The scale of an attack could range from targeted biocrime; to an act of bioterrorism with significant but limited impact; to an act of biowarfare with high regional, national, and/or global morbidity and mortality. Different actors, ranging from an individual to an ideological group to an independent state, could be motivated by personal, ideological, economic, political, or other impetuses to attempt such attacks, and each might be more or less impacted by different approaches to deterrence, prevention, and mitigation.

---

development. Different publics might perceive the risks and benefits of technologies at the intersection of engineering biology and artificial intelligence differently. We support the development of additional opportunities for community stakeholders to contribute their voices to conversations around risks and benefits. See EBRC's Guiding Ethical Principles in Engineering Biology Research for more information.[1]

[iv] A biological pathway is a stepwise series of interactions that result in an end product or outcome. A metabolic pathway, for example, begins with a given molecule/compound/metabolite that is converted by proteins into intermediate forms until a final product is formed.

[v] A biological system refers to a unit of biology (e.g., a cell or an organ system) with many distinct, interacting parts.

[vi] While the scope of this paper is limited to security, defined herein as the deliberate misuse of biology to cause harm, the *safety* implications associated with engineering biology and artificial intelligence are also of concern. Well-intentioned researchers could make mistakes that result in significant harm. Small errors in an autonomous laboratory system could compound over DBTL cycles. The safety of AI systems is important and requires further discussion and consideration.

Recognizing this potential for misuse, the Engineering Biology Research Consortium (EBRC) developed this white paper to identify and communicate key areas for security consideration at the intersection of engineering biology and AI. While consensus on every topic was not attainable, the EBRC community identified three key areas where this convergence may pose significant security concerns: *de novo* biological design, closed-loop autonomous laboratory systems, and natural language Large Language Models. Each of these areas is considered, in turn, with respect to anticipated uses and advances, arising security concerns, and opportunities to prevent and/or mitigate misuse. Finally, we recommend the establishment of an international forum on the intersection of engineering biology and artificial intelligence to identify potential security issues, develop approaches to the mitigation of identified issues, and build international consensus on the responsible development and use of these powerful tools.

De Novo Biological Design Using Artificial Intelligence

AI generally accelerates the fields to which it is applied. AI uses pre-existing data to learn trends that can be impossible, or take a very long time, for humans to identify. It can be applied to biological design at different biological scales, from molecular protein design and engineering to the design of metabolic pathways, to synthetic genomes, or even microbial consortia.

Opportunities in the Application of Artificial Intelligence to *De Novo* Biological Design

Enhanced Protein Design

Proteins form the foundation of biological function and activity. Engineering biology researchers often use proteins to perform specific tasks, such as synthesizing or degrading a target compound, either in vitro or by engineering a protein into an existing organism. While nature offers an exquisite variety of naturally occurring proteins with an amazing array of functions, these naturally occurring proteins may not have the robust, efficient activity needed for laboratory or industrial use. Thus, researchers may work to optimize protein performance for specific applications or to design new proteins from scratch with a desired function that may not exist in nature. Both protein optimization and the design of entirely novel proteins can potentially be greatly enabled by AI, with parallel implications for security.

Researchers have previously been able to improve protein performance for desired purposes through approaches like structure-guided design or directed evolution. But AI can yield superior designs more quickly (see, e.g., Lu et al., 2022).[2] While humans are generally poor at identifying such complex relationships, AI excels at it, and AI "sequence-to-function" or "genotype-to-phenotype" algorithms can be trained on preexisting data relating known protein amino acid sequences and 3-D structures to biological function. AI-generated predictions for sequence substitutions or even entirely new segments, domains, or structures can lead to newly optimized functions. However, a protein's function and activity are dependent upon the molecules with which it interacts. Molecular modeling tools such as Rosetta are advancing capabilities for understanding how small molecules bind to proteins and how proteins interact with other proteins.[3] Greater understanding and modeling of these interactions and molecular interfaces will greatly improve protein design and optimization capabilities. Together, these AI-enabled capabilities will greatly accelerate the DBTL cycle for the discovery of improved functionality, irrespective of whether the desired outcome is protein stability, antibody or receptor binding, catalytic activity, immunogenicity, or even higher order functions such as viral tropism.[vii] To date, AI has been used effectively to optimize protein function, structure, and other characteristics and, given recent

[vii] Viral tropism refers to the capability of a virus to infect a particular cell, tissue, or host types, e.g., to infect the respiratory tract.

advances, the re-design or *de novo* design of proteins for entirely new functions will likely soon become routine.[2,4,5]

Design of metabolic pathways, genomes, and microbial consortia

Designing larger biological systems, such as metabolic pathways, genomes, and microbial consortia, is more complex than *de novo* protein design. Metabolic pathway engineering expands upon individual protein engineering to drive multiple enzymes[viii] to work sequentially to convert a given molecular input to a given output. Often, each enzyme must be optimized to fine-tune its activity. Genetic regulatory elements, such as promoters and enhancers, and/or inducible expression systems may also need to be optimized, as they regulate when and how specific proteins are made. The molecular output of an engineered metabolic pathway could be something currently synthesized using synthetic chemistry or inefficient biological processes, or even a novel compound, perhaps identified for a given purpose with the assistance of AI. Highly engineered / rewritten genomes are even more complex,[6] and in the future, researchers will likely be able to design novel synthetic cells; in either case, all the components needed for cellular function must be present in addition to the capabilities for which the cell is designed.[7] At an even larger scale, microbial consortia may be designed to leverage the unique genetic capabilities of microbial community members to spatially and temporally control, for example, carbon sequestration or environmental nutrient availability.[8,9]

AI is enabling—and will undoubtedly continue to enable—progress in the design of these higher order biological systems.[10–12] However, AI-supported design of biological systems is less well developed than AI-supported protein design, due in large part to the higher complexity of systems and associated challenges with generating high-dimensional, integrated training data of sufficient quality.[13,14] Biological systems are dynamic, influenced by many factors such as the expression of other genes (e.g., transcription factors or enzymes that compete for pathway intermediates), the availability of cofactors, the presence of other metabolites that impact metabolic flux, growth media and conditions, etc. Measuring these dynamic factors over time in a biological system can be challenging, but ample, high-quality, well-organized data of this type is important for training broadly-capable models.[15] AI-supported design of biological systems is certainly improving and will continue to do so as training data improves, but also as a result of strategies to build more data-efficient models.[16]

Security Concerns Associated with *De Novo* Biological Design Using Artificial Intelligence

Advances in AI-assisted biodesign do not necessarily pose entirely novel security risks. Rather, just as AI can accelerate the development of beneficial applications of engineering biology, it may also accelerate nefarious applications. For example, a nefarious lab employee working to fine tune a pathogen virulence factor for human immune evasion or weaponize a pest or pathogen of a key agricultural species, might, with AI-assisted design, obviate the need for several DBTL cycles. As a result, such a bad actor would spend fewer hours in the lab working on such an unsanctioned project and would use fewer resources, resulting in a lower likelihood of raising suspicion. This reduced chance of detection—real or perceived—coupled with greater confidence in early designs might also result in more individuals choosing to attempt such work.[17] Other potential bad actors, such as nation-states or ideological groups, may similarly be more interested in the pursuit of harmful uses of biology if or as AI-enabled biodesign increases the likelihood of "successful" research endeavors that require less expertise and/or resources.

[viii] An enzyme is a particular type of protein that converts an input to a product.

Concerns also arise from the possibility that AI will facilitate and enhance a nefarious actor's capability to work around established safety and security systems. Currently, voluntary screening of orders and customers purchasing synthetic DNA helps ensure that highly concerning sequences cannot be obtained without cause. This barrier between the design space and the physical realm is extremely important; designing something catastrophic *in silico*[ix] does no direct harm if the biology cannot be built in the physical world. However, as AI tools are better able to elucidate sequence-to-function relationships, DNA may be designed to have low—or no—sequence identity to any known sequence of concern, yet still carry out a concerning function. In such cases, the DNA synthesis company may be completely unaware of the function encoded by the sequence and ship the DNA without further review (see *One Step Ahead*, below).

Preventing and Mitigating Misuse of *De Novo* Biological Design Using Artificial Intelligence

It may be tempting to restrict the development of these AI-enabled design capabilities. However, given the open-source nature of these tools and their development and use world-wide, it would be impossible to do so effectively. Such efforts would have limited, if any, impact on security but would certainly impede scientific progress. Furthermore, the restriction in the use of these AI models could prevent smaller laboratories or even start-ups from entering the biotechnology market. Restrictions developed in the United States, even with support and buy-in from allied nations, would be unlikely to be upheld globally, and the U.S. would risk losing its position of global leadership in biomedical and biotechnology-related fields.

Fortunately, several risk mitigation approaches can and should be considered and/or used across the research and development pipeline. For instance, the barriers between digital biodesign and physical biological materials could be maintained and strengthened over time. Individual researchers and the entire life sciences research community can become more attentive to security and recognize potential concerns earlier enough to enable intervention. More sophisticated attribution methods could serve as a deterrent, and models themselves could potentially be designed to avoid certain biodesign spaces.

One step ahead: Integrating sequence-to-function algorithms into security pipelines

Companies that use nucleic acid or amino acid sequence screening in their pipelines[x] for security will need to constantly be aware of improvements to function-to-sequence AI models. To ensure these companies maintain advantage over those who would misuse such models, resources (both public and private) must be put into AI models capable of predicting risk in novel sequences. Models built to classify a sequence as "harmful" or "not harmful," like those that could be used by these companies, would be less computationally expensive to build, use, and maintain than models that generate entirely new sequences or try to predict exact function. As such, companies, if supported, should be able to build and maintain the capability to detect novel sequences that could be used to cause harm. Similar classifying methods are used to catch fraud and E-mail spam (e.g., "spam" or "not spam"). However, in the cases of fraud and E-mail spam, models benefit and "learn" from the many, many attempts to bypass them. The number of attempts to bypass sequence screening algorithms is far fewer in number. This rarity of attempts means there is very little positive case training data. Uneven training data makes it difficult to build models with high specificity, and non-specific models raise many false positive alerts. Alternatively, models that never or rarely alert operators of potentially concerning, no/low-homology sequences may, in reality, be missing sequences on which they should alert. Biosecurity screen operators have no way of knowing if a lack of system alerts is a result of insufficient screening capabilities or because no

[ix] An *in silico* experiment is one conducted using computer modeling or simulation.

[x] Opportunities for other product and service providers (e.g., plasmid repositories, contract research and manufacturing organizations, cloud laboratories) to implement screening should also be explored.

relevant sequences of concern are being ordered. **Therefore, USG should invest in—and work with the scientific community on—efforts to make sequence screening tools that can recognize likely harmful functions from protein sequences, consider financial incentives to enable DNA synthesis providers to use such models, and support the development of capabilities to assess screening systems.**

Community Awareness and Attention

The development of high-risk biological materials, such as highly transmissible human pathogens, almost uniformly requires certain experimental workflows. Such experimental workflows involve human or human-like models of pathogen infection and high-throughput sorting or screening measurements. The development of highly transmissible pathogens requires the engineering of viral vectors that survive aerosolization. Experimental workflows that utilize such vectors involve specialized steps involving lipid encapsulation and/or creating emulsions, using large quantities of purified, human-like lipids. While there are plenty of legitimate reasons to use such experimental workflows, community awareness of the research occurring in one's lab or in adjacent labs may lead to useful observations of unexpected or unusual research practices. Federal law enforcement officers, such as FBI Weapons of Mass Destruction Coordinators, should continue to build relationships with engineering biology community members and other relevant research communities to reduce barriers to discussing concerns. Researchers should receive training from their institutions or other entities on appropriate actions in response to potentially concerning observations, while being careful to guard against bias and discrimination.

AI for Enhanced Attribution Capabilities

A nefarious actor(s) may be deterred from misusing biology if they perceive a high likelihood of being caught. Thus, capabilities for the attribution of engineered genetic sequences to given laboratories or organizations are worth pursuing. These approaches take advantage of the observation that individual laboratories often make consistent and unique decisions in their plasmid design and construction, such as cloning methods used, selection method, reporters, and other small choices that, in combination, leave a lab signature on a plasmid. Models can leverage plasmid repositories such as Addgene, which holds over 135,000 plasmids developed from over 5,700 labs around the world and distributes them to other researchers.[18,19] Advances in attribution capabilities are being made quickly, supported by endeavors such as the Genetic Engineering Attribution Challenge.[20] Like others, this approach is imperfect, as i) plasmids are regularly shared and distributed amongst scientists and ii) such models could be used to fine-tune a design that ultimately deflects responsibility from one source or points falsely to a source.

Security by Design

After decades of research describing protein sequences and functions, available data are sufficient and computational capabilities have advanced to the point that sophisticated AI models can be built for protein design. Current models for biodesign are best at interpolation, meaning they work well within the realm of data they have been trained on. They are generally weak at extrapolating into sequence and function spaces missing from their training sets. Thus, if model developers agree to withhold from their training data certain sequences and functions that are known only to create or worsen hazardous outcomes, it may be possible to minimize the adverse advantages conferred by AI-aided biological design in certain hazardous function spaces.

In reality, such an approach might have limited utility. A field-wide norm would need to be established wherein most or all model developers would agree to voluntarily withhold certain training data, deliberately decreasing the performance of their models in certain biorisk spaces. Doing so might inadvertently create a considerable safety concern. Without any model-capability to define and recognize a threat space, a well-intentioned researcher might build and test AI-generated design(s) that are or could be harmful. And, even if such a norm

were established and followed, many models would still be open source, as knowledge sharing is a bedrock principle of global science. Thus, in many instances, minimal additional tuning with a few relevant examples of a novel protein fold or function of interest could improve the model such that a bad actor could explore desired biothreat spaces. Models will continue to improve at extrapolation as well, perhaps rendering this approach ineffective.

The prevalence of open-source models also limits the utility of access controls.[21] If the users of biodesign models were screened, it could be possible to block access to individuals without a credential or legitimate need. Of course, defining legitimacy is challenging and, particularly in the biodesign space, could discourage valuable public engagement with biology. Instead of user screening, tool developers could require users to sign in with a username and password and potentially track their queries to support retrospective attribution capabilities.

Another means to achieve security through design is through a widespread acknowledgement of the risks of AI as applied to engineering biology. One AI company, Anthropic, has developed "AI Safety Levels (ASL) for addressing catastrophic risks, modeled loosely after the US government's biosafety level (BSL) standards for handling of dangerous biological materials."[22] Under the ASL process, models are evaluated for risk and appropriate safety, security, and operational standards are then put into place. Perhaps model risk evaluation could be useful in biodesign as well.

Closed-Loop, Autonomous Biological Research

Recently, there has been increased interest in the development of AI systems coupled to robotics that can autonomously drive research. In such systems, user-defined parameters and existing model(s) are used to develop a hypothesis which is tested autonomously by robotic systems. Robotics systems generate experimental data which is fed back into the AI model. The model then recommends and tests a new hypothesis. This type of "closed-loop autonomous lab," often referred to as a "self-driving lab," can theoretically operate indefinitely without human input or intervention. This approach to research has the potential to enable the investigation of seemingly intractable questions much more rapidly and efficiently than previously possible.[23] It can drive rapid iterations of the DBTL cycle while benefiting from AI-enabled design, data analysis, and hypothesis refinement. By taking humans out of the experimentation process, consistency and reproducibility may increase, and progress could be made while alleviating unreasonable expectations of human researchers (e.g., AI models and robots do not require sleep or time out of the lab).

Security Concerns Associated with Closed-Loop, Autonomous Biological Research

As with biological design, advances in the use of closed-loop, autonomous biological research laboratories could also be used by bad actor(s). A bad actor could take advantage of autonomous research equipment i) as an employee of a laboratory with regular access to such equipment; ii) as a customer of a cloud laboratory;[xi] iii) by hacking into laboratory systems or cloud labs, thereby gaining remote control of robotic equipment; or iv) as a nation-state or other well-funded organization with the resources to construct and use such a lab system for harmful purposes.

For an insider, less time spent personally experimenting might mean less likelihood of being noticed engaging in suspicious activity. While labs generally have record-keeping practices and sign-up requirements for the use

[xi] Cloud laboratories are heavily automated facilities that enable researchers to remotely design, run, and monitor experimental protocols.

of such resource-intensive equipment, laboratory personnel could misrepresent the purpose of their equipment, or an approved experiment could be hijacked and redirected. Cloud laboratories are still relatively new and the extent to which they will become widely used and accessible is unknown. Customers could misrepresent themselves and/or the nature of their work to outsource the necessary technical development. In addition, any internet-connected autonomous system could be hacked and potentially reprogrammed to develop an unsanctioned, harmful product. However, the hacker would need extensive knowledge of the physical equipment, reagents, and samples present to be successful. Finally, a well-resourced organization or nation may have the capacity to construct such laboratories and use them to make rapid progress toward the development and optimization of a biological weapon.

Preventing and Mitigating Misuse of Closed-Loop Autonomous Biological Research

Human interventions and AI-driven strategies can minimize the biorisks associated with autonomous laboratories. A basic intervention at an individual lab or foundry-level might include requiring human manual approval before the "build" phase of each DBTL cycle, or after some number of cycles appropriate to the level of risk posed by a given experiment. Some robotics systems could require multiple individuals to sign off on an experimental set-up and trajectory before beginning work. At the cloud lab level, no standards, guidelines, or best-practices exist (to our knowledge) describing responsible security measures. Anecdotally, some cloud labs are beginning to follow the example of gene synthesis companies, for example by verifying received samples (e.g., through sequencing), screening customers, using robust network protection and firewalls to prevent hacking, and/or other steps as warranted.

A more complex safeguarding measure might include the development of metrics used to estimate the toxicity or potential harm of an experimental product (e.g., novel metabolite, engineered protein, DNA sequence) to humans, plants, and animals, or a subset thereof, before initiating a new DBTL cycle. If such an approach were to be deemed useful and pursued, it would be important to isolate such safeguarding systems from experimental systems so that if an experimental system were compromised, safeguards would remain trustworthy.

Closed-loop autonomous laboratory capabilities are still nascent. They are expensive and the extent to which they will be used over near- to mid-term time horizons is not yet clear. While it is therefore important not to overstate the current threat that they pose, it is also important not to wait for the industry to fully develop before articulating norms and best practices for security. Thus, **USG should fund a collaborative effort to better understand the role such labs are likely to have going forward, security concerns that currently, or that may relatively soon, exist as a result, and provide recommendations and best practices for better safe-guarding their use.**

Large Language Models

The progress of Large Language Models (LLMs) announced in the past year has yielded great public interest and also caused significant alarm. In the biological sciences, one concern is that LLMs will lower the level of expertise needed to develop a biological system that causes harm. Much discussion has resulted from a recent preprint describing that a class of MIT students, in one hour, queried ChatGPT and learned how reverse genetics can be used to synthesize pandemic pathogens; were shown protocols for such work; identified which DNA synthesis companies are not members of a consortium that requires its members to demonstrate their screening capabilities; and learned about the existence of contract research organizations, which can be contracted to perform experimental work.[24] However, it is challenging to know just how much more quickly students were able to learn and access this information with ChatGPT than they could have using search

engines. It is also not clear whether the time saved would make a significant difference to a motivated bad actor. The existence of the preprint itself may now negate some of the advantages a bad actor may have had as a result of using an LLM.

While LLMs can clearly lower some barriers to knowledge acquisition, helping a bad actor learn about dual-use areas of biology, how biology might be misused, and instructions for doing so,[25] questions remain as to if, or the extent to which, this information could actually enable the misuse of biology. Of all the barriers (e.g., procurement of supplies, reagents, and equipment; laboratory skills) that exist to misusing biology, how substantial is the barrier of knowledge acquisition? What proportion of total time spent to get to a biological weapon is initial publicly available knowledge acquisition? Are LLMs uniquely enabling for a nefarious actor? Are LLMs able to describe not just how to build biology, but also how to weaponize it? To what degree is tacit knowledge required that cannot be communicated via LLM? And (how) will the answers to these questions change over short, medium, and long time horizons?

Answers to these questions are not necessarily clear, and the research community is not in total agreement. Some concern may be warranted. More so than enabling experimentally naive actors to misuse biology, LLMs could lower a barrier for someone who already has training in biological sciences just enough such that they pursue nefarious activity they otherwise would not have. For example, a bioinformatically illiterate graduate student could use an LLM for assistance using biodesign tools that require some proficiency in coding languages. An LLM could help a lab worker interface with a closed-loop autonomous research system. And an LLM may be useful in troubleshooting experimental challenges. However, effective misuse or weaponization of biology requires unique capabilities beyond just molecular biology or virology laboratory skills.[26,27] For example, the weaponization of a pathogenic system or organism requires the design of a biological system that can maintain its infectivity in the wild, the production of that system at scale, the testing of its transmissibility and lethality (e.g. on animals or in human cell lines), and the successful introduction to the target population(s). This biology must be undertaken in the physical world. It requires biodesign, laboratory skills, equipment, materials, and tacit knowledge, the vast majority of which would be very challenging to learn just from reading AI-generated text.

It is also worth noting that LLMs can very confidently err, or "hallucinate."[23] Legitimate researchers using LLMs to more quickly advance or troubleshoot their research may either be able to identify such hallucinations themselves or discuss the output of an LLM with colleagues before following the LLM's directions. A nefarious non-state actor may not themselves have—or have co-conspirators—with sufficient knowledge to recognize misleading or inaccurate LLM responses, fortunately taking their efforts down counterproductive paths. This "tax" on effort to first follow, then detect as wrong, then reorient effort in a new direction, could place a nefarious actor at a distinct disadvantage in attempting to leverage LLMs.

Efforts to build LLMs that withhold certain information from users have proven weak against "jailbreaking"[xii] attempts.[28] Further research and red-teaming of such approaches may enhance these capabilities, and a tiered access system may ultimately provide a useful, though imperfect, barrier. Perhaps LLMs could be built to alert their developers when certain types of queries are received, or when certain types of responses are given. Knowledge is and always will be difficult to control. Therefore, safeguards for physical materials such as nucleic acids must be prioritized. Models that can help synthesis companies reduce the risk of biodesign tools, policies

[xii] Jailbreaking refers to attempts to bypass safety and/or security measures to gain access to protected information.

that support screening, and the identification of other physical realms where security measures might be useful and appropriate must be prioritized, supported, and strengthened.

Next Steps: A Path Forward

Given the complexity of the topic, the rapid advancement of these technologies, and the international nature of the potential risks and benefits, we suggest that **a regularly-convened international forum or initiative be established to:**

i) identify emerging security concerns associated with the convergence of engineering biology and AI;

ii) conduct horizon scanning to predict the trajectory of engineering biology and AI in the coming years;

iii) consider how associated risks may be differentially borne across communities and regions of the world;

iv) develop appropriate guidelines, policies, and/or best practices—along with tools and strategies for their implementation—for determining and preventing deliberate misuse; and

v) over time, evaluate and iterate upon implemented security practices.

Others have also recognized the importance of such fora.[21,29] Convenings should be organized and held at a regular cadence to update expected timelines for achieving specific capabilities, reinforce collaborative relationships, develop best security practices as the biorisk space evolves, and evaluate previously adopted policies and practices. They should be hosted by the private sector and/or academia with the support of governments and international organizations such as the World Economic Forum (WEF) or Organization for Economic Cooperation and Development (OECD).

Stakeholder participants in such an initiative should minimally include members of the AI and engineering biology academic and industry communities, (bio)security experts, and government partners. Furthermore, where possible, community representatives should be included and/or consulted in recognition that different stakeholder populations may view risks and benefits of technologies very differently.

Academic and industry stakeholders might aim to develop and disseminate community best practices within the field. Simultaneously, policymakers must be active participants so that frameworks, regulatory structures, standards, and government guidance developed by individual countries can be aligned to technical capabilities. Such government frameworks should also, whenever possible, support and reinforce each other to avoid fragmentation, confusion, and exploitable loopholes. Community representatives should be consulted in these processes in appreciation that decisions about technology use and governance have global impacts on lived experiences and even existing social, economic, and political systems.

Conclusion

Policymakers, researchers, and other stakeholders are giving considerable attention to the power of AI, particularly in conjunction with other technologies such as engineering biology. Collaborative efforts to identify the security concerns associated with the intersection of these technologies are important, as are efforts to contextualize the relative or actual risk they pose. At present, members of the technical research community associated with EBRC identified *de novo* biological design, closed loop autonomous laboratories, and natural language Large Language Models as areas of potential security concern at the intersection of AI and engineering biology. Still, consensus on the extent to which each is enabling for a nefarious actor—now and in the future—and appropriate prevention and mitigation strategies, is challenging to achieve. Fortunately, unanimous consensus is not necessary for progress. Future efforts to bring people together on this topic will

need to recognize the value of different stakeholder perspectives and seize opportunities to prevent, deter, and mitigate the misuse of AI-enabled engineering biology that can be widely agreed upon. In doing so, stakeholders must actively consider the harms that may be caused and the lives that may be lost through an overly-restrictive approach to navigating these concerns. Ultimately, the research community is investing its time and attention in these technologies because of their capacity to enable progress in engineering biology, leading to a healthier, more sustainable future for us all. Together, we can realize that future while identifying and implementing reasonable safeguards to minimize the potential for misuse.

Funding Acknowledgements and Author Affiliation Notes

WS and RM were supported by the National Science Foundation under Grant No. 2341279. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

WS now works for the National Institutes of Health Office of Science Policy. This article was prepared while he was employed at the Engineering Biology Research Consortium. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, GovCIO, the Department of Health and Human Services, or the United States government.

CDJ was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number T32GM139796. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CDJ is currently a Ph.D. candidate at the University of Texas at Austin, Department of Biomedical Engineering, in the Ellington Laboratory.

References

(1) Mackelprang, R.; Aurand, E. R.; Bovenberg, R. A. L.; Brink, K. R.; Charo, R. A.; Delborne, J. A.; Diggans, J.; Ellington, A. D.; Fortman, J. L. "Clem"; Isaacs, F. J.; Medford, J. I.; Murray, R. M.; Noireaux, V.; Palmer, M. J.; Zoloth, L.; Friedman, D. C. Guiding Ethical Principles in Engineering Biology Research. *ACS Synth. Biol.* **2021**, *10* (5), 907–910. https://doi.org/10.1021/acssynbio.1c00129.

(2) Lu, H.; Diaz, D. J.; Czarnecki, N. J.; Zhu, C.; Kim, W.; Shroff, R.; Acosta, D. J.; Alexander, B. R.; Cole, H. O.; Zhang, Y.; Lynd, N. A.; Ellington, A. D.; Alper, H. S. Machine Learning-Aided Engineering of Hydrolases for PET Depolymerization. *Nature* **2022**, *604* (7907), 662–667. https://doi.org/10.1038/s41586-022-04599-z.

(3) *Welcome to RosettaCommons | RosettaCommons*. https://www.rosettacommons.org/home (accessed 2023-11-10).

(4) Chen, B.; Cheng, X.; Geng, Y.; Li, S.; Zeng, X.; Wang, B.; Gong, J.; Liu, C.; Zeng, A.; Dong, Y.; Tang, J.; Song, L. xTrimoPGLM: Unified 100B-Scale Pre-Trained Transformer for Deciphering the Language of Protein. *bioRxiv* **2023**. https://doi.org/10.1101/2023.07.05.547496.

(5) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(6) Richardson, S. M.; Mitchell, L. A.; Stracquadanio, G.; Yang, K.; Dymond, J. S.; DiCarlo, J. E.; Lee, D.; Huang, C. L. V.; Chandrasegaran, S.; Cai, Y.; Boeke, J. D.; Bader, J. S. Design of a Synthetic Yeast Genome. *Science* **2017**, *355* (6329), 1040–1044. https://doi.org/10.1126/science.aaf4557.

(7) Gaut, N. J.; Adamala, K. P. Reconstituting Natural Cell Elements in Synthetic Cells. *Adv. Biol.* **2021**, *5* (3), 2000188. https://doi.org/10.1002/adbi.202000188.

(8) Moon, T. S. SynMADE: Synthetic Microbiota across Diverse Ecosystems. *Trends Biotechnol.* **2022**, *40* (12), 1405–1414. https://doi.org/10.1016/j.tibtech.2022.08.010.

(9) Engineering Biology Research Consortium. *Microbiome Engineering: A Research Roadmap for the Next-Generation Bioeconomy*; 2020. https://roadmap.ebrc.org.

(10) Delépine, B.; Duigou, T.; Carbonell, P.; Faulon, J.-L. RetroPath2.0: A Retrosynthesis Workflow for Metabolic Engineers. *Metab. Eng.* **2018**, *45*, 158–170. https://doi.org/10.1016/j.ymben.2017.12.002.

(11) Patra, P.; Disha, B. R.; Kundu, P.; Das, M.; Ghosh, A. Recent Advances in Machine Learning Applications in Metabolic Engineering. *Biotechnol. Adv.* **2023**, *62*, 108069. https://doi.org/10.1016/j.biotechadv.2022.108069.
(12) Lawson, C. E.; Martí, J. M.; Radivojevic, T.; Jonnalagadda, S. V. R.; Gentz, R.; Hillson, N. J.; Peisert, S.; Kim, J.; Simmons, B. A.; Petzold, C. J.; Singer, S. W.; Mukhopadhyay, A.; Tanjore, D.; Dunn, J. G.; Garcia Martin, H. Machine Learning for Metabolic Engineering: A Review. *Metab. Eng.* **2021**, *63*, 34–60. https://doi.org/10.1016/j.ymben.2020.10.005.
(13) Kessell, A. K.; McCullough, H. C.; Auchtung, J. M.; Bernstein, H. C.; Song, H.-S. Predictive Interactome Modeling for Precision Microbiome Engineering. *Curr. Opin. Chem. Eng.* **2020**, *30*, 77–85. https://doi.org/10.1016/j.coche.2020.08.003.
(14) Hernández Medina, R.; Kutuzova, S.; Nielsen, K. N.; Johansen, J.; Hansen, L. H.; Nielsen, M.; Rasmussen, S. Machine Learning and Deep Learning Applications in Microbiome Research. *ISME Commun.* **2022**, *2* (1), 1–7. https://doi.org/10.1038/s43705-022-00182-9.
(15) Shi, Z.; Liu, P.; Liao, X.; Mao, Z.; Zhang, J.; Wang, Q.; Sun, J.; Ma, H.; Ma, Y. Data-Driven Synthetic Cell Factories Development for Industrial Biomanufacturing. *BioDesign Res.* **2022**, *2022*, 9898461. https://doi.org/10.34133/2022/9898461.
(16) Adadi, A. A Survey on Data-efficient Algorithms in Big Data Era. *J. Big Data* **2021**, *8* (1), 24. https://doi.org/10.1186/s40537-021-00419-9.
(17) Rose, S.; Nelson, C. *Understanding AI-Facilitated Biological Weapon Development*; The Centre for Long-Term Resilience, 2023.
(18) *Addgene: Deposit Plasmids*. https://www.addgene.org/deposit/ (accessed 2023-11-10).
(19) *Addgene: Plasmid Collections*. https://www.addgene.org/collections/ (accessed 2023-11-10).
(20) Crook, O. M.; Warmbrod, K. L.; Lipstein, G.; Chung, C.; Bakerlee, C. W.; McKelvey, T. G.; Holland, S. R.; Swett, J. L.; Esvelt, K. M.; Alley, E. C.; Bradshaw, W. J. Analysis of the First Genetic Engineering Attribution Challenge. *Nat. Commun.* **2022**, *13* (1), 7374. https://doi.org/10.1038/s41467-022-35032-8.
(21) Carter, S. R.; Wheeler, N. E.; Chwalek, S.; Isaac, C. R.; Yassif, J. *The Convergence of Artificial Intelligence and the Life Sciences:*; NTI | Bio, 2023.
(22) *Anthropic's Responsible Scaling Policy*. Anthropic. https://www.anthropic.com/index/anthropics-responsible-scaling-policy (accessed 2023-11-10).
(23) Martin, H. G.; Radivojevic, T.; Zucker, J.; Bouchard, K.; Sustarich, J.; Peisert, S.; Arnold, D.; Hillson, N.; Babnigg, G.; Marti, J. M.; Mungall, C. J.; Beckham, G. T.; Waldburger, L.; Carothers, J.; Sundaram, S.; Agarwal, D.; Simmons, B. A.; Backman, T.; Banerjee, D.; Tanjore, D.; Ramakrishnan, L.; Singh, A. Perspectives for Self-Driving Labs in Synthetic Biology. *Curr. Opin. Biotechnol.* **2023**, *79*, 102881. https://doi.org/10.1016/j.copbio.2022.102881.
(24) Soice, E. H.; Rocha, R.; Cordova, K.; Specter, M.; Esvelt, K. M. Can Large Language Models Democratize Access to Dual-Use Biotechnology? *arXiv* **2023**, *2306.03809*. https://doi.org/10.48550/arXiv.2306.03809.
(25) Sandbrink, J. Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools. *arXiv* **2023**, *2306.13952*. https://doi.org/10.48550/arXiv.2306.13952.
(26) Montague, M. *Towards a Grand Unified Threat Model of Biotechnology*. http://philsci-archive.pitt.edu/22539/ (accessed 2023-11-10).
(27) Sabra, D. M.; Krin, A.; Romeral, A. B.; Frieß, J. L.; Jeremias, G. Anthrax Revisited: How Assessing the Unpredictable Can Improve Biosecurity. *Front. Bioeng. Biotechnol.* **2023**, *11*.
(28) Deng, G.; Liu, Y.; Li, Y.; Wang, K.; Zhang, Y.; Li, Z.; Wang, H.; Zhang, T.; Liu, Y. MasterKey: Automated Jailbreak Across Multiple Large Language Model Chatbots. *arXiv* **2023**, *2307.08715*. https://doi.org/10.48550/arXiv.2307.08715.
(29) Ekins, S.; Brackmann, M.; Invernizzi, C.; Lentzos, F. Generative Artificial Intelligence-Assisted Protein Design Must Consider Repurposing Potential. *GEN Biotechnol.* **2023**, *2* (4), 296–300. https://doi.org/10.1089/genbio.2023.0025.