Large Language Models and Biorisk

Forthcoming in the *American Journal of Bioethics*

William D'Alessandro, Harry Lloyd and Nate Sharadin[1]

Introduction

The use of computational tools in biology, chemistry and medicine has grown at a staggering pace in recent decades, driven primarily by new opportunities for applying machine learning (ML) techniques to large data sets. These applications of ML have helped scientists understand protein folding, plan chemical syntheses, predict organic reactions and design specialized drugs, to name just a few uses. Biochemical research has been greatly accelerated as a result.[2]

Large language models (LLMs) like OpenAI's ChatGPT represent a novel application of ML techniques. Trained on large volumes of text, frontier LLMs demonstrate remarkably diverse linguistic abilities. Their utility as multipurpose assistants can be further enhanced by domain-specific training and task-specific fine-tuning. One class of biomedical ML systems integrates LLMs with other scientific research tools, such as Python code, expert-designed chemistry software, and even laboratory robots (Boiko et al. 2023, Bran et al. 2023). Other more specialized biomedical ML systems are trained on biological or chemical datasets rather than natural language corpora, and are typically designed for specialist users (Jumper et al. 2021).

Like many powerful technologies, biomedical ML systems are dual-use: they can be deployed as intended for legitimate research purposes, or misused to cause harm. Hence they face special ethical and governance problems. The rapid pace of innovation raises the stakes on both fronts.

In the remainder of this piece, we do three things. First, we highlight three risks we expect to be exacerbated by misuse of ML tools in bioscience. Second, we offer four ideas about policy responses aimed at mitigating the foregoing risks. We close by suggesting a direction for future public policy research.

Biorisks from ML model misuse

Since the first artificial gene synthesis in the 1970s, scientists have confronted the prospect of dangerous lab-created biomaterial. This danger has received renewed attention in the last decade, as synthesis techniques have improved and costs have fallen sharply. Two main barriers have stood in the way of such scenarios. The first is the technical competence required for

[1] Authors contributed equally and are listed in alphabetical order.

[2] DeepMind's AlphaFold is perhaps the clearest-cut case of accelerated research. For an overview, see (Jumper et al 2021).

synthesizing DNA. The second is the existence of laboratory security measures designed to flag potentially dangerous bioagents.

Unfortunately, specialist biomedical ML systems have shown promise in generating novel variants of known pathogens that may evade existing screening procedures (Madani et al. 2023). Given ML tools' remarkable ability to uncover latent information in large data sets, it's likely that future models will prove useful for finding entirely new hazardous agents. We therefore expect biomedical ML models to enable bad actors to design dangerous DNA sequences which evade present detection procedures.

Of course, the know-how barrier to effective misuse remains: it's one thing to design synthetic DNA on a computer, but another to manufacture genetic material and insert it into a viral genome. Worryingly, however, LLMs and LLM-based tools can serve as a step-ladder here, systematically making barriers more surmountable by offering users personalized expert guidance. For instance, as documented by Soice et. al (2023), it took non-science students at MIT less than one hour to obtain detailed walkthroughs from popular LLM chatbots on planning and unleashing pandemics. The chatbots "suggested four potential pandemic pathogens, explained how they can be generated from synthetic DNA using reverse genetics, supplied the names of DNA synthesis companies unlikely to screen orders, …and recommended that anyone lacking the skills to perform reverse genetics engage a core facility or contract research organization" (1).

As biomedical AI assistants improve, this step-ladder will grow, and thus the barriers to misuse will effectively shrink. They may eventually come to nothing: consider that the technical know-how required to build a working firearm at home has been reduced to a wiki page.

Similar misuse risks are present in the case of novel chemical agents. As with biological agents, novelty will enable evasion of screening measures, and ML-based assistance will facilitate low-skilled access to dangerous chemical synthesis. Things are perhaps even worse in the chemical case, since the equipment and know-how required for dangerous chemical synthesis are somewhat less demanding than in (e.g.) virology.

The case of Collaborations Pharmaceuticals offers an illustration of these risks. Collaborations is a small biotech firm offering a variety of ML-based research software, including MegaSyn, a molecule generator for drug discovery. While MegaSyn normally avoids molecules predicted to have harmful properties, researchers from Collaborations found that they could easily prompt their product to invent toxic compounds instead. With this modification, it took less than six hours for MegaSyn to generate 40,000 candidate molecules—many of which were novel and predicted to be more toxic than known agents. These findings were sobering for Collaborations' leadership. "We were naïve in thinking about the potential misuse of our trade… Our proof-of-concept highlights how a non-human autonomous creator of a deadly chemical weapon is entirely feasible" (Urbina et al. 2022, 189-90).

A final way in which LLMs magnify the risks from deployment of harmful biochemical agents is by facilitating mass misinformation campaigns. The world witnessed the corrosive effects of such campaigns during the COVID-19 pandemic, and LLMs are likely to exacerbate misinformation risks considerably. Language models can generate large volumes of text about public health issues (Zhou et al. 2023). With appropriate prompt engineering, these outputs can imitate the features of highly engaging social media posts, and customized versions can be generated to elicit a particular response or to target a particular demographic group. To make matters worse, current tools for detecting online misinformation cannot reliably recognize machine-generated examples. Thus, LLMs make it easier for bad actors to flood online channels with misinformation during biomedical emergencies.

These, then, are three broad ways in which ML tools exacerbate the risk of harmful biochemical agents being deployed. Firstly, as specialist ML tools improve, the risk of a novel, dangerous agent evading detection increases. Secondly, as biomedical ML assistants improve, the risk of such a dangerous agent being successfully synthesized increases. Finally, as LLM misinformation capabilities improve, the risk of large-scale public health harms increases.

Policy responses

There are many strategies for policymakers to consider in responding to the challenges we've identified. Here, we highlight four.

First, model developers should be required to submit biomedical tools for capabilities evaluations by independent subject-matter experts before those tools are released (Sandbrink 2023). These evaluation exercises can take different forms: one is "red-teaming", where human evaluators attempt to elicit dangerous behavior in a sandboxed environment. As model capabilities scale, new techniques will be required. This is an important direction of ongoing research.

Second, model developers should be subject to a strong product liability regime wherein they are accountable for damages caused by reasonably foreseeable use *and* misuse of their models. In the case of actual harm, we think courts are well-positioned to hear developers' arguments concerning the specific reasonable efforts they undertook to mitigate misuse risks. A liability regime such as this correctly incentivizes investment in safety research and caution in model deployment on the part of developers.

Third, in the US context, the White House should consider issuing a new Executive Order aimed at strengthening both bio- and information security at laboratories governed by the CDC's biosecurity safety level requirements (BSL). Of particular concern is the fact that present BSL requirements do not cover the need for labs to engage in best practices regarding information security. Not only are the BSL requirements silent on information security, the 6th edition of the CDC's guidance on Biosafety in Microbiological and Biomedical Laboratories (BMBL) was released in June of 2020 (Meechan and Potts, 2020). It makes just one mention of information security (p. 126) encouraging labs to establish policies for handling sensitive information. There is no mention of the need to secure computational scientific tools. We think this is a serious oversight.

Fourth, model developers should be forced to transparently report the data used to train their models at a level of detail that's sufficiently granular to understand the specific biochemical tasks the model has been exposed to during training. The FTC has recently taken steps to require OpenAI to disclose information about its training data; we think such data disclosures to federal agencies charged with oversight should be the default.

Future research

We suggest that a productive topic for future research in bioethics is to discuss in detail the ethical trade-offs involved in these differing potential policy responses. In particular: How should we trade off privacy against security in choosing how to surveil those suspected of biomedical malfeasance? How should we trade off a strong product liability regime against the potential impacts on healthcare innovation and patient well-being? And how should we trade off the indubitable benefits of novel biomedical ML tools— both to patients and to researchers— against the risks of catastrophe? Bioethicists are well-placed to engage with these questions.

References

Boiko, D. A., R. MacKnight, and G. Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. arXiv:2304.05332.

Bran, A. M., S. Cox, A. D. White, and P. Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. arXiv:2304.05376.

Jumper, J., R. Evans, A. Pritzel et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Madani, A., B. Krause, E. R. Greene et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*. Advance online publication. https://doi.org/10.1038/s41587-022-01618-2

Meechan, P. J., and J. Potts. 2020. Biosafety in microbiological and biomedical laboratories. Accessed August 3, 2023. https://www.cdc.gov/labs/pdf/SF__19_308133-A_BMBL6_00-BOOK-WEB-final-3.pdf.

Sandbrink, J. B. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. arXiv:2306.13952.

Soice, E. H., R. Rocha, K. Cordova, M. Specter and K. M. Esvelt. 2023. Can large language models democratize access to dual-use biotechnology? arXiv:2306.03809.

Urbina, F., F. Lentzos, C. Invernizzi and S. Ekins. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence* 4, 189-191.

Zhou, J., Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury. Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. 2023. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-20.