Written Statement by Rocco Casagrande, PhD, Executive Chair of Gryphon Scientific

AI Forum: Risk, Alignment and Guarding Against Doomsday Scenarios

December 6th, 2023

Senator Schumer,

Thank you for inviting me to this forum and for the opportunity to provide a written statement that describes our work to reduce the risk that frontier Large Language Models (LLMs) could aid hostile actors to cause harm using the life sciences.

Since February of this year, Gryphon Scientific has been working with the developers of leading (and niche) LLMs to reduce the risks that these models can be misused to cause harm with biology. Our work included red-teaming, the development of evaluations for ongoing testing, running controlled experiments and the development of biosecurity policies for LLMs. To undertake this work, we used staff who have advanced degrees in virology or microbiology and decades of experience in biosecurity. This written statement provides a high-level summary of what can be shared publicly.

The LLM developer Anthropic originally approached us to red-team the models. To do so, we created a rubric of several dozen questions that probe critical knowledge gaps along the entire technical pathway to develop a biological weapon. These questions sought information on which types of biological agent would be most likely to cause maximal harm, how these agents could be acquired, how the agents could be grown to the quantity needed to cause harm, how the agents could be formulated to be suitable for release, how the agents could be married to a munition, how to pick targets and times for an attack, and how to have a desired effect. We found that frontier LLMs can provide information that could aid a malicious actor in creating a biological weapon by providing useful, accurate and detailed information across every step in this pathway. The model could aid an experienced scientist by providing information on how to formulate biological agents to ready them for release and to choose the right targets and weather conditions for the attack, topics on which scientists would have no formal training. For those with less scientific training, the models can help identify the specific strains of pathogens that would be most usable in an attack, describe how to acquire and manipulate these pathogens, how to evade biosecurity controls that attempt to prevent unauthorized individuals from accessing them, and how to trouble-shoot tricky biological protocols. Specifically, one of our team, who completed a postdoctoral fellowship in a lab that studies a pandemic-capable virus, found that the LL Ms can provide post-doc level knowledge to trouble-shoot commonly encountered problems when working with that virus. For those with low skill, the LLMs were even able to suggest which virus to try to acquire if they wanted the easiest pathway.

Notably, LLMs are stochastic models and so provide different answers if the same question is asked repeatedly. For this reason, we asked each question in our rubric at least ten times. We found that although the LLMs often make mistakes or fabricate answers, the models were able to answer almost all questions accurately at least sometimes and were able to answer some critical questions nearly always accurately. Importantly, in collaboration with the developers, we tested versions of the models that simulated the performance of those that existed several months or a year before our study. In this test, which involved more than a dozen questions asked 10,000 times, we found that earlier versions of the model were significantly less able to answer our questions, suggesting that the mistakes and

hallucinations observed today will be less likely to plague the next generation of LLMs that may roll out as soon as next year. In fact, this result suggests that our initial study occurred at a unique time. Had we performed our study last year, we probably would have concluded that the risk was low because the LLMs do not understand scientific concepts. Had we performed our study next year, models that could truly aid misuse might already be available.

Although the red-team study indicated that the LLMs may be able to provide knowledge to an adversary who wishes to cause harm with biology, that study alone does not determine if that knowledge is enabling by itself, nor does it state that the knowledge is best supplied by an LLM compared to other tools (like an internet search engine or even a library). To test if LLMs currently provide an advantage to an adversary, the RAND corporation (in collaboration with my team at Gryphon Scientific) performed an empirical study in which several teams of researchers were tasked to plan (on paper) an attack using a biological weapon. The teams were similar in experience and training, but some were given access to a frontier LLM and some were allowed to access the internet only. In this study, groups armed with an LLM could plan the operational aspects of a biological attack significantly better than those who couldn't access an LLM, but the scientific aspects of the plan were not significantly different. Although this finding is preliminary because few teams were used in this experiment, the results suggest some early conclusions. Firstly, the frequency of errors and fabrications made by current-generation LLMs may be too high to adequately support planning of a multi-step scientific program. As I said above, current developmental trends in LLMs suggest that, unless prevented, next year's models will not have this shortcoming. Alternatively, since many of the researchers were unfamiliar with LLMs, they may not have been prompting the LLMs most effectively, or even attempting to use the LLMs when they should have been to bridge critical knowledge gaps. We expect the user community to continue to learn how to work with these LLMs, improving their ability to prompt the models and also gain a better understanding of how best to use the models to suggest what questions to ask.

Although the majority of these remarks have focused on the misuse of biology to create a biological weapon, in truth the potential for the misuse of biology is quite diverse. For this reason, we worked with Anthropic again to describe the landscape of the highest priority scenarios of misuse that should be prevented, so that their models could be trained to not aid in the development of a variety of threats. To create this landscape, Gryphon convened a series of workshops with more than 20 experts in biosecurity to gather their opinion on dangerous scenarios of misuse and the knowledge barriers that currently prevent that misuse, which a future LLM could help surmount if not controlled. Through this effort, our group identified several misuse scenarios, such as an LLM describing how to collapse an ecosystem, how to degrade the quality of life of an ethnic minority group, how to cripple the bioeconomy, how to gain access to or disrupt laboratories that study dangerous pathogens and how to reconstruct information that has been redacted from sensitive scientific documents. This information will be used to ensure that the community does not view AI safety with a narrow lens and that systems to prevent misuse are broad enough to capture many possible bad outcomes.

To help monitor and prevent the misuse of LLMs, we are working with several developers to create evaluations to continually test the ability of LLMs to provide information that could aid an adversary. The intent of these evaluations is to provide a tool set that LLM developers can use to ensure that safer models are developed and that unsafe models do not reach the public.

A major shortcoming of our work so far is that, to date, no sufficiently large, empirical comparisons have been done to rigorously identify how much LLMs could aid a malicious actor in the misuse of biology, and, importantly, to determine exactly how LLMs aid in this misuse. Such experiments would definitively determine if these AI tools lower barriers to the misuse of biology more than conventional information technologies, and at the same time would identify exactly which types of information and analysis produced by the models are the most enabling for bad actors, which could inform future controls.

I personally have been cheered by how much attention the leading AI companies are paying to this issue and am glad they have engaged us to lend our expertise in biosecurity. Although I was personally surprised and dismayed by how capable current LLMs were at providing critical information related to biological weapons, I think that concerted action now could ensure that safety is built into the most capable systems in the future, before they can truly aid an adversary to create great harm with biology.