The Effect of Machine Learning on Bio-Engineered Pandemic Risk

This is a doc-capsule of the AISC research proposal submitted for the 2024 Winter AISC by Jacob Haimes. It is unchanged from the date of the conversation recorded in Episode 1 of the Into AI Safety podcast, aside from this text.

Summary

In this project we will be quantifying the likelihood of a bio-engineered pandemic causing massive human suffering and death (catastrophe) as a function of the capabilities and accessibility of machine learning (ML) systems.

Keywords: Threat Model; Forecasting; Misuse; Bio-Risk, Criminology

The non-summary

Motivation

Establishing a Cutoff Point

Currently, I believe that one of our utmost priorities in the field of AI safety should be to establish *any* form of restrictive policy regarding the development and/or deployment of ML systems. This is because:

- We cannot trust companies/corporations to take a morally correct action when doing so does not align very closely with taking the financially optimal action.
- In most cases regarding AI, the financially optimal action is to move forward with development and deployment of increasingly advanced models.
- Because of the potential benefits of AI, there will be many opinions regarding what should be allowed, and what shouldn't be allowed.
- Due to a large number of stakeholders, the bureaucracy of governments/standards agencies, and the highly polarized political climate that we are in, agreeing upon and establishing policy regarding AI cannot happen quickly.
- If we wait until there is an imminent threat to begin attempting to pass safety policies, there won't be enough time to establish those rules prior to the threat being realized.

So, how can we increase the likelihood of getting restrictive policies in place?

Lever: Bio-Engineered Pathogens

At this point, most agree that pandemics have the potential to pose an existential risk to humanity, and to some extent understand the impact that one could have (thanks COVID19). Furthermore, biological weapons development and production (among other actions) has

been banned internationally [1], and these agreements have strong support regardless of political perspective or nationality.

By demonstrating the increase in risk associated with certain levels of accessibility[1] to ML systems with specific capabilities[2], we will provide (1) guidance to policy makers regarding what should be restricted, and (2) evidence needed to justify those policies.

Claims

1. It is possible to approximate the probability of a bioengineered pandemic risk in a *baseline*[3] scenario based on historical data.
2. It is possible to define a relationship between access to a dangerous technology, and attacks using that dangerous technology.
3. It is possible to quantify to what extent certain capabilities/services will make bioengineering pandemic threats possible.

If these three claims are true, then it would be possible to develop a quantitative understanding of the increase in likelihood of bioengineered pandemics as a function of ML capabilities, and the access to those capabilities.

Step-by-Step

Rough approximation of project progression - often multiple steps will be happening simultaneously

1. Understand the task
    - Develop an abstract understanding of the steps that would be required to develop and produce a biologically engineered pathogen.
    - *NOTE: The extent to which this is reported is not defined yet, but we will discuss with Subject Matter Experts (SMEs) in the bio-risk space to determine what we should share, and what we shouldn't.*
2. Develop Baseline Model/Understanding
    - In a *baseline* scenario, how likely would it be for (1) individual actors, and (2) rogue organizations to pursue the development and production of bio-engineered pathogens?
        - Look at accessibility of weapons compared to the frequency with which they are used by malicious actors
            - Accessibility of guns in various countries versus the corresponding number of mass shootings, shootings, and gun related deaths in that country
            - Accessibility of various types of weapons versus the number of attacks per year which use it

---

[1] *E.g.,* government, research, companies, general public
[2] *E.g.,* long horizon planning, ability to conduct research, buying and selling of goods
[3] We will define our *baseline* scenario to be one in which advanced machine learning technologies are not present. This means that the state of our current world is already beyond this point, but we must start here to understand the impact of both future and current machine learning capabilities on bio-terrorism.

        - Look at the probability of attacks that target varying numbers of victims within any given year (*e.g.,* how likely is it that an attack which targets at least 10 individuals happens? What about 50? 100? *etc.*).
    - Assuming 2020 technology, what is the probability that (1) individual actors, and (2) rogue organizations would be able to succeed in an endeavor to cause a bio-engineered pathogen related catastrophe?
        - What are the difficult portions of the process?
        - How prohibitive are they?
        - How easily are they circumvented and/or resolved?

3. Quantify effect of ML capabilities
    - How much easier (in terms of time, money, information, access) does ML capability X make it to conduct a biological attack?

4. Examine the differences between complete open source (full access for the general public) and other varying amounts of restriction (this informs the number of people who have access to the technology).

5. Finalize paper

Potential Difficulties

Specific tasks that we may struggle to accomplish, which would diminish the value of the idealized project

1. Relevant historical data may be difficult or impossible to find.
    - Especially if we are limiting ourselves to finding non-controversial evidence (I wish I wasn't serious about this).

2. It may be difficult to create a robust argument without providing a dangerous amount of detail into the process for developing and using a bio-engineered pathogen.

3. Making this matter (*i.e.,* getting it in front of policy makers) may not be easy.

4. There may not be a clear relationship between accessibility of a weapons technology and the likelihood that it is used for malicious purposes (doubt it).

5. Attack frequency may not have a clear dependence on the number of individuals with direct access to a given technology (doubt it).

Bounded Outcome

Here we outline two versions of our project to establish upper and lower bounds on our expectations

Conservative Version

In the event that all research difficulties noted above are realized, what does our project look like?

1. Detailed report outlining why specific ML capabilities necessarily increase the likelihood of x-risk level catastrophe (of the bio-engineered pathogen variety) - necessarily uses broad ideas to prevent spreading dangerous ideas.

2. Comparison of resulting increase in biological attacks for multiple capability/access level pairs (*e.g.,* general public access to chatbots, academic access to chatbots, government restricted access to models with chemical compound discovery, corporate access to models with chemical compound discovery).

3. Rough approximation of the difficulty required (bottlenecks and restrictiveness of them) for a (1) individual actor, or (2) rogue organization to carry out a biological attack with the technology and level of access present in 2020.

Ambitious Version

In the event that there are no research difficulties, and we are able to do more than initially anticipated

1. Detailed report outlining why specific ML capabilities necessarily increase the likelihood of x-risk level catastrophe (of the bio-engineered pathogen variety)

2. High precision estimate of the probability of a (1) individual actor, and (2) rogue organization (a) planning, and (b) carrying out a biological attack in our *baseline* scenario. This is likely a mathematical model in some manner.
    - Also obtain high precision estimates of the same values assuming various combinations of capability/access pairs (*e.g.,* general public access to research capable agents, academic access to research capable agents, government restricted access to models that can intentionally deceive, corporate access to models that can intentionally deceive)
    - Visualize these probabilities in a clear and concise manner, potentially creating an infographic

3. Detailed analysis of the difficulty required (bottlenecks and restrictiveness of them) for a (1) individual actor, or (2) rogue organization to carry out a biological attack with the technology and level of access present in our *baseline* scenario. This is likely a mathematical model in some manner.
    - Quantify in what manner these bottlenecks are made less restrictive as a result of ML capabilities/access combinations

Scope

In Scope

Explicitly, what will we be investigating, and what will we include in our analysis?

1. Historical trends in crime
    - Likelihood of actors perpetrating and attempting attacks as a function of target size
    - Likelihood of perpetrated and attempted attacks utilizing a given technology as a function of access to that technology

        - How many catastrophes directly resulted in increased restrictions on a related technology, which made future attempts to do the same thing significantly more difficult.
            - Use guns as an example, compare time/effort spent to get a gun in the US vs. the same time/effort in other countries, compare number of mass shootings and/or gun deaths (probably there will be some correlation)
    - Likelihood of *misfire - i.e.,* instances in which a weapon is constructed without intent to harm others it, but it still does harm others (*e.g.,* accidents, split-second decisions)
    - Likelihood of success in perpetrated attacks as a function of technological improvements
2. In our *baseline* scenario, how substantial are the barriers to bio-terrorism? Is it easy to get around them already?
3. What are potential and already realized capabilities of machine learning systems that will either reduce the barriers to bio-terrorism, or make circumventing those barriers more easy?
    - Based on what these capabilities are, how much they reduce the cost required to enact a bio-terrorist attack, and our understanding of historical trends in crime, how much will specific advances (accompanied by their level of accessibility) increase the likelihood of bio-terrorism

Out of Scope

What research will we be avoiding?

1. Details of bio-engineering pathogens
2. Policy design
    - *E.g.,* with intent to restrict the proliferation of AI systems, or slow down capabilities research
3. State-level actors (government biological weapons programs)
    - The effect of machine learning on the likelihood and/or success of bioweapons research by governments
    - Repercussions of biological warfare

Output

In an ideal scenario, I think the output would be an academic paper, ideally submitted for either a conference, although I am unsure where to submit as of right now. Most importantly, we would need a very well written executive summary so that we can grab policymaker attention (who will then get their aids to read the rest of the paper).

As part of our stretch goals, we would want to aim for an infographic which consolidates our information in a visually appealing manner, and a blog post which can serve as a shorter version of our paper (as this will increase the number of people that we can get our ideas to).

Potential Harmful Externalities and Risks

Addressing potential negative externalities

1. **Our project increases bio-risk in some way, by:**
    - **Spreading the idea in general**
    - **Demonstrating what would have to be done**
    - **Providing examples of how advanced ML systems could be used**

    I think that it is highly unlikely that our project will actually increase bio-risk, as we will be very cognizant of this potential - I will make sure that we are able to have oversight from others who are more experienced in this domain before sharing anything.

2. **We can't share our work due to safety concerns.**

    This is a more significant concern. I plan on investigating if, in the case that we feel we cannot provide public access due to safety concerns, we can still provide our work as a resource to some smaller group of people (most likely a government agency or group).

3. **We end up polarizing the issue of AI safety, making future policy action more difficult.**

    When I began writing this proposal, this issue was not even on my radar. Upon reflection, however, I realized that the parallels between gun violence and gun accessibility in the United States and the claims I am making may not be the best angle to take. I am not sure what to do about this as of right now.

4. **The analysis presented is significantly wrong, and important decisions are made poorly as a result.**

    By avoiding grandiose claims and grounding all of our assumptions in reality via historically relevant data, this risk is significantly reduced.

5. **Regardless of the quality of the output, it doesn't actually influence policymakers, or provide those who were already convinced of this issue with enough evidence to sway others.**

    While I do see this as a legitimate concern, I think we can leverage connections with the members of the AI safety community who are also members of congressional staff or in the Tech Policy Fellowship.

Acknowledgements

Linda Linsefors	Initial problem specification, resource recommendations, understanding what makes a good AISC project
Abby Hoskin	Connections!

Tomek Korbak	Understanding problem specification, resource recommendations
David Manheim	Mitigating Goodhart's Law, value add
Darryl Wright	Brainstorming, promising angles
Michael Middleton	Paths, problem definition
Chris Lonsberry	Reasonability, value add

Citations

[1] United Nations. "Biological Weapons Convention – UNODA." *UNODA*, https://disarmament.unoda.org/biological-weapons/.

Team

Team size
Ideally I would want a team of at least 2 others, but I think that 3 or 4 would be great, and I would be more than happy to work with more if enough people are interested.

Research Lead
Jacob Haimes

Email: jacob.d.haimes@gmail.com
LinkedIn: https://www.linkedin.com/in/jacob-haimes/

Feel free to reach out with any questions or concerns.

About me:

- I have co-authored two academic papers, so I have some experience with research in general, and lots of experience with LaTeX (both of these papers, along with some of my other work, can be found on my portfolio)
- MS in computational modeling, with a focus in optimization
- AI safety relevant courses
    - BlueDot Impact's AI Alignment Course (independent)
    - BlueDot Impact's AI Governance Course (cohort)
    - Center for AI Safety's AI Safety Sprint Course (cohort)
- Actively working on mechanistic interpretability research with a collaborator

I commit to working a minimum of 20 hours a week, every week, with the intent to work more.

Team Coordinator

I would greatly appreciate another member taking on the responsibility of TC. That said, if others are not willing/don't have the time to do this, I can.

Skill requirements

Requirements:

- You think that getting our first restrictive rules established regarding AI is significantly important
- You have had some exposure to AI safety, and you understand the core concepts
- You are excited to work in a collaborative and encouraging group

Ideal Experiences (*i.e., if you have experience in any of these areas, please apply*):

- Threat modeling
- Forecasting
- Bio-risk
- Criminology
- Bio-engineering
- Tech policy
- Writing for policymakers
- Technical writing