simeon@safer-ai.org

How Can Biosafety Inform AI Safety?

Olivia Jimenez

Summary

Many AI scientists and other notable figures now believe that AI poses a risk of extinction on par with pandemics and nuclear war.[1] A sufficiently powerful AI could self-replicate beyond developers' control. Less powerful AI could also be misused. Given these risks, it is crucial that AI research be held to high standards of caution, trustworthiness, security, and oversight.

To determine what AI research standards should be and how they should be implemented, it may be helpful to consider precedents from other fields conducting dangerous research.

This memo outlines select standards in biosafety, with a focus on how high-risk biological agents are treated in biosafety level (BSL) 3 and 4 labs in the United States. It then considers how similar standards could be applied to high-risk AI research.

1. High-risk research must be conducted in designated labs subject to stringent standards.
2. Personnel must be trained and screened for reliability.
3. Someone at each lab is responsible for safety, and they are empowered to shut projects down if they determine them to be unsafe.
4. Physical and information security are prioritised.
5. Labs record and respond to every incident and plan for emergencies.
6. Labs have extensive oversight from governments and independent auditors.

Biosafety standards

The field of biology has comprehensive standards for maintaining safety while working with potentially dangerous biological agents, such as viruses that could escape the lab and infect many people.

Biosafety standards are applied in a tiered approach, where the riskiest research can only occur in laboratories taking the most stringent precautions. The levels of laboratory precautions range from biosafety level 1 (BSL-1, the lowest) to biosafety level 4 (BSL-4, the highest). Easily-transmissible and lethal viruses, such as the Ebola virus or smallpox viruses, can only be studied in BSL-4 labs.[2] BSL standards also govern the storage and transportation of pathogens, access to labs, and lab protocols.

[1] "Statement on AI Risk", Center for AI Safety

[2] "Facing Down the World's Deadliest Pathogens in a BSL4 Lab", Scientific American

By enforcing tiered standards according to research risk at all labs, biology has been relatively successful at mitigating the **two types of risks it faces:**

- **accident risks**, such as a virus escaping the lab because it was accidentally mishandled, and
- **misuse risk**, such as a virus being intentionally taken out of the lab by someone who intends to cause harm with it.

Selected biosafety standards & potential applications in AI

1. High-risk research must be conducted in designated labs subject to stringent standards.

The USDA maintains a list of biological agents that pose severe public health and safety threats (henceforth select agents). **Any lab wishing to work with a select agent must be licensed** by the Federal Select Agent Program (FSAP). These facilities are continuously monitored for compliance.

Additionally, **each BSL-4 research project must get specific approval** before it can be conducted. Researchers must submit detailed research plans and undergo a thorough review by their institutional committee, and sometimes by government agencies, overseen by a government-approved Responsible Official. The review process assesses the potential risks associated with the select agent being researched, as well as the suitability of the researcher and the lab. Research that poses especially high risk, "restricted experiments", are banned unless individually approved by the Health and Human Services Secretary.[3]

Research that is less risky is subject to correspondingly less severe restrictions, which range from BSL-1 laboratories in grade schools, homes, and some high schools, to high-school and hobbyist BSL-2 biology laboratories with minimal oversight, to BSL-2 or -3 clinical or research laboratories with full-time biosafety officers and institutional biosafety reviews.

Potential application in AI:

A government agency would determine various risk levels for AI research and set out standards for each level. For moderate-risk research, self-governance and independent audits would be sufficient. For high-risk research, research would only be allowed to be conducted in designated labs, using designated compute providers. Governments would provide licences to designated labs and compute providers once they demonstrated compliance with the most restrictive security and research standards, as well as continuously monitor their compliance. For especially large training runs expected to pose the greatest risk, labs would additionally be required to apply for project-specific approval from the government. Governments would block research projects deemed too risky for any lab whatsoever to carry out, until

[3] "Restricted Experiments Guidance", Center for Disease Control and Animal and Plant Health Inspection Service

sufficient precautions are developed. It would be illegal to conduct research outside licensed facilities or without project-specific approval.

2. Personnel must be trained and screened for reliability.

All personnel with access to listed select agents must successfully undergo FBI Security Risk Assessments and be approved by either the CDC or the USDA's select agent program. These resemble or can include DOD security clearances, Department of Energy reliability checks, or Office of Personnel Management "Public Trust" investigations.

Security risk assessments check for risk factors for intentional subversion of rules or irresponsible behaviours. This can include whether a candidate has been involved in certain crimes, has used or possessed certain drugs, is affiliated with potentially hostile nations, has been shown themselves to be mentally unstable, or has been discharged from the armed services of the United States under dishonourable conditions.[4]

Personnel assessments check for "emotional stability, capacity for communication and cooperation, integrity, capacity to resist external pressure, acceptance and capacity to follow instructions, active approach to safety and security, mental alertness, mental and emotional stability, trustworthiness, freedom from unstable medical conditions, dependability in accepting responsibilities, effective performance, flexibility in adjusting to changes, good social adjustment, ability to exercise sound judgement in meeting adverse or emergency situations, freedom from drug/alcohol abuse or dependence, compliance with requirements, positive attitude toward PRP [the assessment check itself]" as well as whether the candidate can be "blackmailed, coerced, or otherwise manipulated."[5]

Some assessments are ongoing. Personnel need to self-report, "medical matters... prescription and over-the-counter medications used, alcohol abuse, legal actions, public record court actions (eg, separation, divorce, lawsuit), financial problems or concerns, or any other activity that may influence the staff member's day-to-day reliability."[6] Peers and supervisors also need to report, "any behaviors or events they suspect are affecting an individual's day-to-day reliability."[7]

There is also extensive training required for the use of BSL-3 and -4 laboratories, which cover safety, security, and accident procedures.

Potential application in AI research:

In order to work on a high-risk AI research project, candidates would be required to go through training and screening to ensure they deeply understand and are committed to safety. Training might cover risks

[4] "Biological Safety and Security Program", US Department of Defense
[5] "Implementation of a Personnel Reliability Program", Higgins et al.
[6] "Implementation of a Personnel Reliability Program", Higgins et al.
[7] "Implementation of a Personnel Reliability Program", Higgins et al.

from misalignment and misuse, the limits of current safety and security techniques, and the limits of current understanding of AI systems. Screening might check for a candidate's understanding of known risks and limitations, ability to identify new risks, psychological stability, integrity, commitment to the common good, freedom from conflicts of interest and other potential burdens on decision making, and ability to make good decisions under pressure.

3. Someone at each lab is responsible for safety, and they are empowered to shut projects down if they determine them to be unsafe.

A biosafety officer is required for managing biosafety at laboratories. Their duties include, but are not limited to, "developing the safety policy and procedures...; orientation and training of all laboratory staff in biosafety; providing safety advice; ensuring staff compliance with safety policies and procedures by performing regular safety inspections, and documenting and submitting reports of such inspections to the Laboratory Manager for review and action; [and] investigating laboratory accidents and documenting accident reports".[8]

The biosafety officer requirement is universal across different safety levels, but the work involved differs depending on the laboratory and safety requirements. BSL-4 laboratories will frequently have a group or departments rather than a single individual in charge of this function. BSL-3 and BSL-4 **labs handling select agents are also required to have a Responsible Official, who is not only in charge of ensuring compliance, but is legally required to be on-site, has authority to make decisions on the part of the organisation, and ensures reporting is performed** - including by ensuring that whistleblowers have a secure way to report issues directly to the relevant federal inspector general.[9]

Potential application in AI research:

Every lab and data centre involved in high-risk AI research would be required to have at least one officer responsible for ensuring safety. They would need authority to act on behalf of the organisation, such as to unilaterally shut down projects or require additional procedures. The safety officer would be accountable to the national or international agency or external oversight body. To that body, the safety officer would be required to report all incidents (e.g., "the model demonstrated deceptive and power seeking tendencies" or "a staff member tried to download model weights onto their personal laptop") and anticipated risks (e.g., "the proposed fine-tuning seems too risky given the limitations of current safety techniques" or "our security system may soon be outdated,") and in addition to reporting, would be empowered to intervene to stop or at least pause projects until these issues are addressed.

Professional training of such officers and professional codes of conduct would be needed. These should be developed in concert with other efforts such as model safety reviews and safety research.

[8] "Appoint a Biosafety Officer", World Health Organization
[9] "Responsible Officer Resource Manual", Federal Select Agent Program

4. Physical and information security are prioritised.

BSL-4 levels take extensive physical security precautions to minimise the risk of select agents infecting people and leaving the lab, both by accident and misuse.

For perimeter security, labs generally have security guards, metal detectors, fences and barriers that can withstand impact, limited entries and exits, mantraps, and biometric readers and keycards. To control the working environment, labs generally airlock workspaces, ensure airflow in one direction, and sanitise water and air before it exits the facility. To protect staff, labs generally require researchers to wear personal positive pressure suits inside the lab, take a chemical and regular shower before exiting the lab, and avoid using the restroom until outside the lab. While working in the lab, researchers are always watched; scientists outside communicate via radio, operate microscopes remotely, etc. Throughout the lab, sensors are set up to detect select agents in the air to help notice potential release.[10]

Labs also have strict access controls. **The number of people given access to high-risk areas and information is kept to the minimum needed.** Accordingly, information siloing within research institutes is standard. For instance, Boston University's National Emerging Infectious Diseases Laboratories contains BSL-2, BSL-3, and BSL-4 containment spaces. Only a small, designated set of their researchers are allowed to access the BSL-4 space. Even outside the BSL-4 lab, visitors are kept to a minimum and may not be unattended.[11]

There are also strict rules for how information can be accessed and shared, to minimise the risk of research insights being used by outside actors to cause harm.

Potential application in AI research:

Labs would invest drastically more in securing their AI model (e.g., to prevent weights from being leaked or stolen or to prevent AIs getting unintended access to the internet) and secure research insights (to reduce the spread of information that enables others to build high-risk systems). Unlimited access that allows for model-parroting and model-extraction attacks would need to be restricted.

Computers and networks involved in high-risk research would be air-gapped (as is already standard for some military research, and recommended or legally-required for high-security applications like critical infrastructure or cryptocurrency storage by depository institutions). Mechanisms for detecting anomalies in AI models during training and deployment would need to be improved, and continually revised and updated over time. Finally, access to AI models and labs would be limited, and oversight would be increased throughout the research process.

[10] "Threading the NEIDL", This Week In Virology
[11] "Threading the NEIDL", This Week In Virology

5. Labs record and respond to every incident and plan for emergencies.

Labs must perform an initial risk assessment before any work is started. **Risk assessments must also be reviewed whenever a change occurs in factors that may significantly impact risk** such as personnel changes, relocations or renovations, ageing of equipment, new animal species, new biological agents, and new routes of administration of biological agents.[12]

Risk assessments must also be performed after every incident, even if it was a near-miss. The norm is for labs to make concrete changes that will robustly prevent similar accidents in the future each time an accident occurs. This follows best practice for all high reliability organisations, such as nuclear power and aviation.[13]

Labs also prepare thoroughly for emergencies, such as an earthquake compromising physical security. They generally maintain emergency egress routes and backup generators, design buildings and barriers to withstand impact, and have procedures in case of suspected release of a select agent.[14]

Potential application in AI research:

Every lab and data centre involved in high-risk AI could be required to monitor for, report, and address any incidents that suggest heightened research risk. For instance, they could be required to report whenever an AI model develops unexpected or emergent capabilities; shows tendencies towards deception, power seeking, or avoiding shutdown; shows signs of situational awareness, agency, or autonomy; or outputs harmful content, such as text that offends users or information that could assist people in carrying out acts of violence. They could also be required to develop mechanisms in case of security emergencies, such as mechanisms to rapidly shut down compute clusters or destroy weights.

6. Labs have extensive oversight from governments and independent auditors.

Regular audits are conducted by teams of experts from the CDC, other government agencies, and external consultants and inspectors, who are contracted by the CDC or hired directly by the lab. Auditors thoroughly review a lab's procedures, inspect physical infrastructure and equipment, observe the handling of select agents, and interview personnel. At the conclusion of the audit, they provide detailed reports outlining their findings and recommendations. The lab is generally required to address any concerns identified in the report.[15]

Potential application in AI research:

AI labs could be overseen by external bodies that evaluate the reliability of their facilities, personnel, research plans, and AI models. Questions evaluators may ask include: do the AI models have dangerous capabilities, how robustly do the safety techniques being used mitigate risks, are the researchers

[12] "Developing a Biosafety Risk Assessment", Federal Select Agents Program
[13] "Developing a Biosafety Risk Assessment", Federal Select Agents Program
[14] "Threading the NEIDL", This Week In Virology
[15] "Inspection Checklist for BSL-4 Suit Laboratories", Federal Select Agents Program

trustworthy, and are the facilities secure? Government bodies, internal evaluators, and internal governance could all be involved in these assessments.