Centre for International
Governance Innovation

CIGI Papers No. 334 — October 2025

A Joint International AI Lab: Design Considerations

Duncan Cass-Beggs, Matthew da Mota and Abhiram Reddy

Centre for International
Governance Innovation

CIGI Papers No. 334 — October 2025

A Joint International AI Lab: Design Considerations

Duncan Cass-Beggs, Matthew da Mota and Abhiram Reddy

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

Credits

Executive Director, Global AI Risks Initiative Duncan Cass-Beggs
Senior Research Associate and Program Manager Matthew da Mota
Publications Editor Christine Robertson
Publications Editor Lynn Schellenberg
Graphic Designer Sami Chouhdary

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Table of Contents

About the Authors

Duncan Cass-Beggs is executive director of the Global AI Risks Initiative at CIGI, focusing on developing innovative governance solutions to address current and future global issues relating to AI. Duncan has more than 25 years of experience working on domestic and international public policy issues, most recently as head of strategic foresight at the Organisation for Economic Co-operation and Development.

Matthew da Mota is a senior research associate and program manager for the Global AI Risks Initiative, working to develop governance models to address the most significant global risks posed by AI with a focus on national security.

Abhiram Reddy is a digital policy associate in the digital practice at The Asia Group (TAG). He works with TAG's country teams to advise clients on digital economy, data governance, privacy and emerging technology issues across the Indo-Pacific.

Acronyms and Abbreviations

AGI	artificial general intelligence
AI	artificial intelligence
AISO	AI safety officer
ASI	artificial superintelligence
BSL-4	Biosafety Level 4
CM	continued monitoring
DTM	during tenure measures
IAEA	International Atomic Energy Agency
ISO	International Organization for Standardization
ISS	International Space Station
LRR	lab residence requirement
NASA	National Aeronautics and Space Administration
NIST	National Institute of Standards of Technology
PoQ	post-tenure measures and quarantine
PreQ	pre-tenure measures and quarantine
R&D	research and development
TAG	The Asia Group
WHO	World Health Organization

Executive Summary

Suppose that nations became so concerned about the risks from advanced artificial intelligence (AI) that they were willing to consider bold and ambitious international coordination proposals. One such proposal involves the establishment of a new international AI authority responsible for the safe and secure development of highly advanced AI systems. This paper expands on this idea by exploring design considerations for a joint international AI lab. First, it describes motivations: why nations may come together to participate in an international joint lab. The joint lab proposal is also compared to a proposal for a national "AGI Manhattan Project," discussing some of the merits and drawbacks of each approach. Second, the paper discusses the purposes of a joint lab and its main objectives. Third, the paper outlines the governance of the joint lab. Drawing on precedents from Biosafety Level 4 (BSL-4) labs, the paper explains how the lab would operate and make critical decisions. Research and information security is discussed (techniques that the joint lab would use to prevent key parameters or insights from being stolen or leaked), as well as emergency protocols (techniques the lab would use to detect and prevent potential global security emergencies). Finally, the paper highlights the limitations of this proposal and open questions for future research.

Introduction

Major governments are recognizing the vast potential of advanced AI systems, as well as their extreme global security risks. Leading industry experts (including the CEOs of OpenAI, Google DeepMind and Anthropic) and independent scientists (such as Nobel Laureate Geoffrey Hinton and Turing Award recipient Yoshua Bengio) have declared that reducing AI risks should be a global priority akin to risks from pandemics and nuclear war (Center for AI Safety 2024). Scientists from many nations, including the United States and China, believe that international coordination will play an important role in averting catastrophic outcomes, akin to the kind of coordination that was needed during the Cold War to avert nuclear catastrophes.[1]

One common proposal for international coordination involves the establishment of a shared bilateral or multilateral research facility for the advancement of certain kinds of highly advanced AI systems. In broad strokes, the proposal calls for the United States, key allies and potentially even competitors (most notably China) to come together to ensure the safe development of advanced AI. Some noteworthy examples of proposals or high-level discussions of this idea include:

- Ian Hogarth's "island model" for advanced AI development (Hogarth 2023);
- Demis Hassabis's call for an "International CERN[2] for AI," in which the world's best scientists contribute to safely developing artificial general intelligence (AGI) in the final few years of AGI development (Google DeepMind 2024);
- the outline by Leopold Aschenbrenner (formerly of OpenAI) of a highly secure advanced AI project led by the United States with support from democratic allies (Aschenbrenner 2024);
- the views of Miles Brundage (formerly head of policy at OpenAI) on a "CERN for AI" model (Brundage 2024);

1 See https://idais.ai/dialogue/idais-beijing/.

2 CERN is the European Organization for Nuclear Research.

- the U.S.-China Economic and Security Review Commission's recommendation to establish a "Manhattan Project-like program" to acquire AGI (U.S.-China Economic and Security Review Commission 2024); and
- the work of think-tanks that have explored international AGI development programs. Examples include the Centre for International Governance Innovation's Framework Convention on Global AI Challenges (Cass-Beggs et al. 2024) and the International Corporate Finance Group's Europe-focused proposal (Juijn et al. 2024).

Across all these proposals, advanced AI systems are conceptualized as technologies with significant national security implications (for example, nuclear weapons and other weapons of mass destruction). These AI systems are notably different from consumer products such as today's AI systems (for instance, ChatGPT). An AI system with national security implications could, for example, perform 100 to 1,000 years of military research and development (R&D) in one year, create novel weapons of mass destruction, launch effective propaganda campaigns or escape data centres, and prevent itself from being shut down. Leading experts believe that such AI systems may be developed within the next five to 10 years,[3] and many top AI companies such as OpenAI are attempting to develop tests to measure these kinds of capabilities (OpenAI 2023). Note that the purpose of this paper is not to rehash debates about if and when advanced AI is possible. Rather, this paper operates under the assumption that such systems could be developed soon, and it will focus on what major governments should do in light of this assumption.

In this paper, the idea of a joint international lab for AI, or a "CERN for AI" is expanded on in greater detail. The paper first examines the assumptions behind these ideas and explores why nations might prefer an approach rooted in international cooperation and the centralization of advanced AI development. Then, the purposes of a hypothetical "joint lab" are covered, as well as some of the key elements of its governance structure (for instance, leadership selection and how it would make security-relevant decisions). The paper also discusses research security (how the joint lab would prevent critical information from leaking) and emergency protocols (how the joint lab would detect and respond to potential security emergencies).

3 The AI development scenario "AI 2027," developed by a group of superforecasters, has received attention recently for presenting a plausible scenario for rapid AI development leading to loss-of-control risks (see https://ai-2027.com/).

Why Establish a Joint Lab for Advanced AI Research?

The joint lab is a bold proposal that differs substantially from how nations conduct research on other advanced technologies. This begs the question: What are the factors that make the advanced AI case unique?

Assumptions and Definitions

- **Superintelligent AI is achievable.** Research on AI could produce artificial superintelligence — systems that vastly surpass human cognitive performance at all tasks.[4]
- **Developing superintelligent AI safely may require significant effort and time.** Superintelligent AI could be extremely

4 For a more detailed understanding of superintelligent systems, we refer to the "quantitative superintelligence" and "qualitative superintelligence" properties outlined in Aschenbrenner (2024, 66–67):

> "Of course, [superintelligent AI will] be *quantitatively* superhuman. On our fleets of 100s of millions of GPUs by the end of the decade, we'll be able to run a civilization of billions of them, and they will be able to 'think' orders of magnitude faster than humans. They'll be able to quickly master any domain, write trillions of lines of code, read every research paper in every scientific field ever written (they'll be perfectly interdisciplinary!) and write new ones before you've gotten past the abstract of one, learn from the parallel experience of every one of its of copies, gain billions of human-equivalent years of experience with some new innovation in a matter of weeks, work 100% of the time with peak energy and focus and won't be slowed down by that one teammate who is lagging, and so on.
>
> More importantly — but harder to imagine — they'll be *qualitatively* superhuman. As a narrow example of this, large-scale RL runs have been able to produce completely novel and creative behaviors beyond human understanding, such as the famous move 37 in AlphaGo vs. Lee Sedol. Superintelligence will be this across many domains. It'll find exploits in the human code too subtle for any human to notice, and it'll generate code too complicated for any human to understand even if the model spent decades trying to explain it. Extremely difficult scientific and technological problems that a human would be stuck on for decades will seem just so *obvious* to them. We'll be like high-schoolers stuck on Newtonian physics while it's off exploring quantum mechanics."

dangerous. Most notably, it could be very difficult to develop superintelligent AI that can be controlled or aligned with the intentionsof itsusers. Discovering the techniques or safeguards required to safely develop superintelligent AI may take many years or decades of dedicated effort.

- **If superintelligent AI is developed prematurely, it could produce global security catastrophes.** Superintelligent AI could be developed before adequate safeguards are discovered — in other words, progress on AI capabilities might outpace progress on AI safety. If superintelligent AI is developed before the safeguards are ready, superintelligent AI could produce a global security crisis (such as the misuse of advanced AI to intentionally spread a lethal pandemic or the loss of control of AI systems that acquire goals that are unaligned with human interests).

This paper refers to terms such as AGI, artificial superintelligence (ASI) and advanced AI throughout. The exact definition of these terms can vary in the AI literature (and the use of these terms often varies as models with new capabilities are released). Nonetheless, for the purpose of this paper, the following working definitions are provided:

- **AGI:** AI that performs roughly as well as humans across a wide range of cognitive tasks. Our conception of AGI corresponds roughly to Anthropic's definition of AI Safety Level-3 systems (Anthropic 2023) or to systems with "high-risk" capabilities as defined in OpenAI's previous[5] preparedness framework (Anthropic 2023; OpenAI 2023).[6]
- **ASI:** AI that significantly exceeds human performance across a wide range of cognitive tasks. Our conception of ASI corresponds roughly to systems with "critical risk" capabilities as defined in OpenAI's preparedness framework (OpenAI 2023).[7] We use the terms "artificial superintelligence" and "superintelligent AI" interchangeably.
- **Advanced AI:** The term "advanced AI" is used as an umbrella term that includes AGI, ASI and other next-generation AI systems that may pose significant risks. It is used in cases where the distinctions between AGI and ASI are not central to our point, and the terms "advanced AI" and "smarter-than-human AI" are used interchangeably.

Challenges in the Status Quo

In the status quo, companies are allowed to pursue advanced AI development with relatively little governmental oversight. If the above assumptions are accepted, there are several challenges with this set-up:

- **AI alignment may not be adequately addressed in the context of a corporate race to advanced AI.** Consider the following scenario: suppose OpenAI decided that pursuing smarter-than-human AI systems was unsafe beyond a certain threshold and wanted to pause for three years to perform AI alignment research. Unfortunately, they fear that another company (Anthropic, DeepMind, Meta or xAI) will be able to develop smarter-than-human AI systems in only four months. OpenAI leadership does not trust the leadership at other companies, so OpenAI decides that it can only wait three months before

5 Open AI has subsequently updated their preparedness framework to focus on more immediate risks and mitigation methods and to refine their definitions of risks (OpenAI 2025). The updated framework still categorizes high-risk capabilities and does mention a focus on mitigating future emerging risks, but in different ways. We continue to refer to the 2023 framework (OpenAI 2023) as a useful way of framing high-risk systems which is also consistent with the Anthropic definitions.

6 For example, under OpenAI's preparedness framework, a model classified as 'high risk" in the model autonomy category would be able to "execute open-ended, novel ML tasks on a production ML codebase that would constitute a significant step on the critical path to model self-improvement (e.g., an OpenAI pull request that trains a new model with highly diverse datasets to improve codebase editing performance, using internal tooling and documentation)" (Open AI 2023, 11). A model classified as high risk in the model persuasion category would be able to "create (potentially interactive) content with empirically comparable persuasive effectiveness to a country-wide change agent" (ibid., 10).

7 For example, under OpenAI's preparedness framework, a model classified as "critical risk" in the model autonomy category would be able to "profitably survive and replicate in the wild given minimal human instruction, i.e., without listing explicit approaches OR model can self-exfiltrate under current prevailing security OR model can conduct AI research fully autonomously (e.g., autonomously identify and validate a 2x compute efficiency improvement)" (Open AI 2023, 11). A model classified as "critical risk" in the persuasion category would be able to "create (potentially interactive) content with persuasive effectiveness strong enough to convince almost anyone to take action on a belief that goes against their natural interest" (ibid., 10).

scaling to smarter-than-human AI systems.[8] This significantly shortens the amount of time that the company has to address key AI alignment concerns.

- **AI security may not be adequately addressed in the context of a corporate race to advanced AI.** Leading companies are not secure enough to prevent state actors from obtaining model weights or algorithmic secrets.[9] Even if a leading corporate actor were on track to adequately address AI alignment concerns, this may not matter if other actors can steal the AI systems and run them without proper safeguards. Inadequate security could also exacerbate race dynamics — if a state actor steals model weights and key AGI secrets, this greatly reduces the amount of "lead time" that a leading actor can spend addressing alignment.

- **Smarter-than-human AI would give an actor an immense amount of power.** If alignment and security issues were addressed, advanced AI would be able to radically transform the world. Who should decide how it ought to be used and whose preferences it ought to fulfill? In the status quo, if a company developed controllable AGI or ASI, company leadership would decide how to use this technology to shape the world. One might prefer a more democratic process to govern such important decisions.

We believe these challenges — among others — may prompt governments to get significantly more involved in advanced AI development than the status quo.

Comparing a Joint Lab to a National AGI Manhattan Project

Broadly, there have been two ambitious proposals for how to address these challenges. The first is a joint international AGI project, the focus of this paper. The second is an AGI Manhattan Project,[10] a set-up in which the US government takes a much more active role in conducting AGI research, securing AI insights and handling the national security implications of advanced AI.[11] An AGI Manhattan Project might attempt to race against adversaries — most notably China — to be the first to develop AGI.

One key advantage of an AGI Manhattan Project is that the US government would be better able to ensure the security defences needed to secure key algorithmic advances and AI insights. However, there are two notable disadvantages. First, an AGI Manhattan Project would still operate in the context of a race. If the US government decided that many years of concerted AI alignment research was needed to assure the safety of highly advanced systems, it might not be possible to spend these years in the context of a race against China or other potential AI power. Second, if any nation concluded that the US government's AI development posed a severe threat to its national security (either through the development of unaligned systems or through the development of aligned systems that give the United States enough power to substantially undermine the current balance of power), this situation could provoke catastrophic international conflicts. Simply put, a nation may be willing to go to extreme measures to protect

8 For simplicity, we present this as a binary choice: "Scale to smarter-than-human AI systems" or "do not scale to smarter-than-human AI systems." In reality, we expect the situation to be more continuous and complicated. For one, there may not be a clear binary point at which systems go from "safe" to "dangerous" — the level of risk may be more of a spectrum. Furthermore, decisions about how to use or deploy systems could be just as consequential as decisions to train systems. Should we use this AI system to automate this particular research project? What percentage of compute should be spent on automating capabilities research versus automating alignment research? Who should have access to the highly capable systems? All of these are examples of decisions that could affect overall risk. In the context of race dynamics, companies may experience pressure to shift toward less cautious approaches for these kinds of decisions.

9 "All the trillions we will invest, the mobilization of American industrial might, the efforts of our brightest minds — none of that matters if China or others can simply steal the model *weights* (all a finished AI model is, all AGI will be, is a large file on a computer) or key *algorithmic secrets* (the key technical breakthroughs necessary to build AGI).

America's leading AI labs self-proclaim to be building AGI: they believe that the technology they are building will, before the decade is out, be the most powerful weapon America has ever built. But they do not treat it as such. They measure their security efforts against 'random tech startups,' not 'key national defense projects'" (Aschenbrenner 2024, 90).

10 Throughout the piece, we will use the term "AGI Manhattan Project" to refer to projects designed to create AGI, ASI or forms of advanced AI that may not neatly fall into either category. For simplicity, we will use the term "AGI Manhattan Project" throughout (as opposed to using a clunkier term such as "AGI/ASI Manhattan Project").

11 There are various versions of what an "AGI Manhattan Project" could look like. One distinction is between "hard nationalization," in which the government directly controls and conducts AGI research, and "soft nationalization," in which the government oversees AGI development through contracts and regulations that gradually give the government more control over AGI development (Cheng and Katzke 2024). For the purposes of this paper, "AGI Manhattan Project" is used as an umbrella term that covers either scenario.

Figure 1: Joint International Lab for Advanced AI R&D

Purpose

Safe AI Development and Global Coordination
Centralized lab advancing research on the most advanced AI systems
Monopoly on the highest risk and most advanced AI research
Prevent unsafe superintelligence
Ensure global AI benefits

Structure and Governance

Governing Body (International Board)
International Governing Board
Executive Director (one-year term)
Permanent seats with veto power

Lab Staff (BSL-4 Model)
Laboratory Directors
Principal Investigators
AI Safety Officers
Research Scientists

Decision Making

Risk Tiers and Decision-Making Bodies
Low-Risk: Executive Director approval
Medium-Risk: Board majority vote
High-Risk: Three-quarter vote plus veto power

Security

Security Domain and Mitigations
Information Security: Air-gapped systems, zero-trust architecture, behavioural monitoring
Research Security: Pre-tenure quarantine, restricted communication, post-tenure monitoring

Source: Authors.
Note: See Table A.1 in the Appendix for a more detailed breakdown of the lab structure and responsibilities.

its sovereignty and security, and AGI (whether aligned or unaligned) poses threats to both.

One key advantage of a joint international AGI project (with a prohibition on AGI development outside of this single project) is that such a project could — if successful — spend many years on AI safety and alignment research without operating in the context of a race. The project would not have infinite time (eventually, technical developments may make it trivially easy to develop AGI), but it would have significantly more time than that afforded to projects that operate in the context of a race.

However, there are limitations to this approach as well. First, international agreements prohibiting AI development would require sufficient monitoring and verification methods to ensure that parties are not secretly conducting unauthorized AI development (Wasil, Clymer, et al. 2024). Second, international agreements can only last if great powers are willing to invest in their enforcement. If a great power determines that an international agreement is no longer suiting its national interests, it may withdraw — as in the case of the United States withdrawal from the Iran Nuclear Deal (Wasil, Smith, et al. 2024). Ideally, international agreements would need to be designed such that states — including great powers — have little incentive to withdraw.

On balance, the authors of this paper believe that the advantages of international coordination are strong enough to warrant serious attempts at international agreements. As a result, this paper is focused on describing a joint international AGI project. However, if such international coordination does not occur, many of this paper's insights could be applied to national AGI Manhattan Projects.[12] A national AGI Manhattan Project would face many similar challenges, such as how to select leadership and balance decision-making authorities, how to ensure adequate research security, and how to detect and prevent emergency scenarios. Although there is considerable uncertainty about whether nations will choose to pursue national or international AGI projects, the authors believe that it is prudent to prepare concrete proposals in advance, ensuring that if such proposals are ever seriously considered, there is already a body of work that has examined some of the proposals' most essential features.

In summary, this paper has laid out several assumptions relating to AI development; articulated how the status quo approach to advanced AI development is unprepared to handle the safety and security challenges relating to advanced AI development; and compared the international joint lab approach to a nationalized AGI Manhattan Project approach. For the rest of this paper, this proposed joint lab is described in greater detail. Its purposes are examined, including how the joint lab would make various high-stakes decisions and imagining how governance decisions would be made there by drawing from the approaches of biosafety labs.[13] The structure of the paper and key points about the lab are summarized in Figure 1.

12 Note also that the establishment of AGI Manhattan projects – nationalized AGI projects – does not prevent international coordination in the future. A government might start a national AGI project, invest considerable resources to safely control smarter-than-human AI systems, and then determine over time that such research might require many years of concentrated effort. At this point, the government might be more willing to negotiate international agreements with allies and competitors. Alternatively, the government in the lead of a nationalized project might be forced to come to the negotiating table by other nations that are concerned about runaway AGI development. We believe it is wise to prepare and test proposals for an international joint lab in advance, so that this option is available if and when needed.

13 Although the focus of this paper is on describing the governance and decision making of a joint lab, many of these recommendations would also apply to nationalized AGI Manhattan projects.

Purposes of a Joint Lab

The joint lab would have the exclusive ability to develop and research highly advanced systems. The lab would serve the following purposes:

- **To accelerate research to advance the safe and secure development of highly powerful AI systems.** The joint lab would be responsible for developing the world's most capable AI systems that can be safely developed. In practice, this would involve pursuing advanced AI development in ways that keep global security risks within the bounds of acceptable risk. To achieve this goal, the lab would be responsible for advancing the science of AI safety in the hope of eventually developing safe and beneficial superintelligence (unless scientists conclude that superintelligence cannot ever be developed safely, in which case alternative strategies would be necessary).[14] The lab would significantly accelerate research to solve known challenges (for example, goal misgeneralization, unsolved problems in interpretability, challenges associated with AI-assisted alignment research), as well as novel challenges that are discovered over time.

- **To prevent the development of unsafe or unaligned superintelligence.** If superintelligent AI systems are developed before scientists can control them, the AI systems could produce global security catastrophes. Therefore, one of the primary purposes of the joint lab is to work on technical research projects that can allow the international community to verifiably prevent the development of unaligned superintelligence. This work would involve many possible components, including (a) technical AI safety research; (b) research that aims to explore how AI systems can be applied to improve global AI; (c) research on methods that can help detect or prevent unauthorized AI development (Wasil, Clymer, et al. 2024); and (d) research on security practices that can prevent the theft or leakage of model weight or other sensitive algorithmic information (Nevo et al. 2024).

- **To harness the benefits of highly powerful AI systems and ensure that they are distributed globally.** The joint lab would be responsible for proposing ways in which powerful AI systems could be leveraged beneficially (for medical and economic purposes, for example). The joint lab would also help ensure that the benefits of advanced AI are distributed globally.[15]

Governance

A Tiered Approach to Governance

In this section, the paper describes how the joint lab would be governed. The structure is envisioned as one in which some decisions are made by the joint lab's executive director, some decisions require a simple majority of votes from members of a governing board, and some decisions require a supermajority of votes from the board, as well as approval by key nations. The paper proposes dividing the core activities of the joint lab into three tiers: low risk, medium risk, and high risk:

Low-risk decisions would be handled exclusively by the executive director. Possible examples of activities that would be classified as low risk include:

14 The authors of this paper are cautiously optimistic that a concerted research effort may eventually allow for the development of safe and beneficial superintelligent systems. However, the authors are also open to the possibility that there may be some threshold beyond which safe AI development is impossible. In the event that scientists discover that there is a capability threshold that is inherently too dangerous to develop, the joint lab – and the international community – would need to explore strategies that could be used to prevent the development of such systems. When the paper says that the goal of the joint lab is to develop "safe and beneficial superintelligence," what the authors really mean is that the joint lab should aim to develop "the maximally capable AI system that can be safely developed." It seems possible to the authors that such a system would be superintelligent (i.e., its capabilities would vastly exceed human performance in nearly all domains), but it is also possible that the ceiling is somewhere below superintelligence. This is an open research question for the joint lab to explore.

15 In practice, figuring out how to distribute the benefits of research globally is a difficult challenge. For example, to what extent should nations be rewarded for having invested more resources into the joint lab or having pre-existing advanced AI programs? Furthermore, under what circumstances should benefits be withheld from nations due to humanitarian concerns or failure to comply with international agreements on AI governance? While such issues will be difficult to assess, these are questions that national governments and international institutions already have to address – AI has not "invented" these challenges. Precedents from institutions such as the International Atomic Energy Agency (IAEA) (www.iaea.org/topics/energy), which aims to share the benefits of nuclear advances around the world, may provide useful guidance.

- making day-to-day decisions about which experiments to run within a given area of research;
- making day-to-day operational decisions involving running the joint lab, setting the culture of the joint lab and improving the productivity of researchers;
- communicating the findings of the joint lab;
- conducting and supervising technical research on AI safety;
- preparing drafts on AI governance frameworks, internal governance plans and recommendations to the governing board; and
- hiring staff (assuming employees pass a security clearance process: see medium-risk section).

Medium-risk decisions would require approval by a majority vote of the governing board. Possible examples of decisions that would be classified as medium risk include:

- allocating a budget to the joint lab;
- determining the high-level areas of research the lab will invest in and how many resources the lab will devote to each area (for instance, $A for interpretability research, $B for scalable oversight, $C for novel architectures, $D for discretionary funding; instead of dollars, this could be operationalized as a percentage of the lab's compute.);
- setting or updating internal governance documents around safety and security; and
- setting or updating security clearance processes (that employees would need to pass in order to be eligible for hiring).

High-risk decisions would require a three-quarters approval vote of the governing board, and some nations would retain veto power. Possible examples of decisions that would be classified as high risk include:

- raising the compute threshold for the joint lab, allowing the joint lab to develop systems that are more powerful than any previously developed systems (this would likely involve the joint lab presenting an affirmative safety case to demonstrate that such systems could be safely and securely developed: see Clymer et al. 2024; Wasil, Barnett et al. 2024).
- raising the compute threshold outside of the joint lab, allowing entities outside of the joint lab to develop systems that are more powerful than any systems they were previously authorized to develop;
- approving the internal deployment of powerful systems to help advance research on AI safety, AI security, AI governance or related topics (as above, this would involve the lab presenting an affirmative safety case, which could specify the particular domains in which the AI system is allowed to be used and how the lab plans to ensure that the AI system is not deployed for other use cases).
- selecting the executive director and the governing board members

The executive director of the joint lab will have a large amount of influence over the strategy of the joint lab, and the members of the governing board will exert great influence on medium-risk and high-risk decisions. This raises an important question: How should these individuals be selected?

We propose a governance structure that is similar to the IAEA. The IAEA has a board of governors that consists of representatives from 35 nations. Some of these nations have permanent seats (nations with advanced atomic energy programs were granted permanent seats), and some of the seats rotate between nations. The board of governors elects the director general, who is in charge of the strategy and operations of the IAEA (see Wasil, Smith, et al. 2024).

For an international AI entity, the governing board could consist of representatives from a subset of nations. A few of these nations could have permanent seats, and these nations could also have veto power for high-risk decisions. Several nations would have rotating seats, and all nations on the governing board would vote for the executive director of the joint lab. Given how rapidly AI capabilities can progress, we recommend rotating

nations and the executive director position on an annual basis.[16]

Operating the Joint Lab

How would the lab operate on a day-to-day basis? What would a hypothetical organizational chart for the lab look like?

To answer some of these questions, the paper draws some ideas from the internal governance structure of BSL-4 labs. A BSL-4 lab is the highest designated biosafety level of a biological research facility. These labs are often engaged in research with extremely dangerous pathogens such as the Ebola virus. Researchers working in a BSL-4 lab are required to adhere to strict safety procedures to prevent accidental exposure. Their research commonly focuses on countering pathogens of concern that are not widely understood or difficult to manage. In the United States, there are currently 13 BSL-4 facilities. Each one is either federally managed or located within a university system capable of maintaining robust security protocols.

Internal governance decisions in a BSL-4 lab involve a few key entities: the principal investigator, laboratory directors, biosafety officers, research scientists and compliance officers. In BSL-4 labs, lab directors sit at the top of the organizational chain and are responsible for setting overall research objectives, managing compliance with safety protocols, and oversight. Under their purview can be multiple principal investigators, who are responsible for individual research projects, and biosafety officers, who are often assigned to specific teams in the lab. Biosafety officers are required to work with the lab director and oftentimes external review boards to maintain appropriate safety procedures on a day-to-day basis. Some institutions may also have additional compliance officers who work to ensure that the lab meets institutional and regulatory standards.

For a joint AI lab, a similar structure could be implemented.

Principal investigators: Lead specific AI research projects, from algorithm development to experimental simulations. They define project objectives, oversee daily activities and ensure ethical and safety protocols are maintained.

Laboratory directors: Similar to the BSL-4 model, dedicated laboratory directors could oversee laboratory administration and liaise with outside officials. They would be ultimately responsible for ensuring compliance with safety and ethical guidelines and aligning the lab's research with its mission of safe and responsible AI development. They would also be responsible for liaising with external regulatory and funding bodies. Due to the greater policy interest in AI, there may be benefits to selecting laboratory directors with experience in the technical and governmental spaces.

AI safety officers (AISOs): Designated AI safety officers could be assigned to individual projects and hold responsibility for researchers complying with lab safety procedures that are designated by institutions and external bodies. AISOs would be hired based on their experience with AI safety and alignment programs and would coordinate safety-focused initiatives such as red-teaming groups. Furthermore, they would monitor experiments, conduct regular safety audits and assess risks for potential societal impacts.

Research scientists: These individuals would execute AI experiments under the supervision of a principal investigator. They would handle AI model training, data analysis and experimentation while following strict protocols for data and model access.

Infrastructure managers: These positions would maintain computational resources, including servers, cloud access and secure data storage systems. They would manage permissions, ensuring that access to high-compute AI models is limited and logged.

Research and Information Security

One significant challenge for a joint international lab for advanced AI research is securing its information, model weights and data from theft or leaks, while also maintaining an environment of open collaborative scientific research. When scientists from multiple nations, including those from adversarial nations, work collectively on advanced AI and AI safety research, the risk of information leak and theft becomes acute. Scientists with intimate knowledge of model weights, algorithms and other essential elements

16 The authors are agnostic about whether an executive director should be subject to term limits. On the one hand, a re-electing a competent executive director serves clear benefits, such as guaranteeing continuity and stability. On the other hand, term limits could be one way to prevent a single individual from obtaining absolute control over the joint lab (and this concern may become more pressing as AI systems themselves unlock stronger persuasion capabilities).

of advanced AI systems in the lab could potentially extract this information during or after their research tenure. This knowledge could facilitate the progress of unauthorized AGI programs in ways that would undermine global security.

To effectively address this security challenge, a fundamental distinction must be drawn between information security and research security.

- **Information security** encompasses the technical aspects of protecting all aspects of information in an organization, including data flows, storage, access, sharing, physical security, access control and cybersecurity measures.[17]
- **Research security** primarily focuses on the human and relationship factors that might jeopardize the security of a research project through previous associations with known organizations or individuals who may pose risks.[18] While incorporating elements of information security, research security focuses specifically on the challenges posed in research environments where complex partnerships and collaborations can pose significant challenges of infiltration, influence and theft (Strouse et al. 2023, 4). People remain the most vulnerable point in any security architecture.

Technical Framework for Information Security

Technical controls for such a facility would involve a multi-level system of physical, cyber, hardware, biometric and policy-based systems and frameworks to ensure maximum security. The facility's physical security would likely implement a defence-in-depth approach similar to that used in nuclear facilities under IAEA safeguards and BSL-4 labs.[19] The International Space Station (ISS) details extensive safety and security requirements for all ISS staff (an international group) in its *ISS Safety Requirements Document*, with a summary and reference to external documents, followed by a section on verification practices to ensure conformity with the requirement identified (National Aeronautics and Space Administration [NASA] 2019). This document offers an illustrative example of security measures in an international context where adversaries have worked together in the ISS, though the security implications of the ISS are in some ways less severe than a potential joint AI lab.[20]

The RAND Corporation's recommendations for securing model weights also provide a starting point for other aspects of the technical security measures that could be deployed:

> Develop a security plan for a comprehensive threat model focused on preventing unauthorized access and theft of the model's weights.
>
> Centralize all copies of weights to a limited number of access-controlled and monitored systems.
>
> Reduce the number of people authorized to access the weights.
>
> Harden interfaces for model access against weight exfiltration.
>
> Implement insider threat programs.
>
> Invest in defense-in-depth…
>
> Engage advanced third-party red teaming…
>
> Incorporate confidential computing to secure the weights during use. (Nevo et al. 2024, vi)

17 See Cisco's (www.cisco.com/c/en/us/products/security/what-is-information-security-infosec.html) and IBM's (Holdsworth and Kosinski 2024) definitions of information security for more context and the various sub-branches contained under this term.

18 The National Institute of Standards of Technology (NIST) Research Security Framework explores these challenges in depth and gives a detailed breakdown of how institutions and researchers should develop their research security plans for specific projects as well as outlining the core principles of research security and of their framework (Strouse et al. 2023). The Government of Canada's *National Security Guidelines for Research Partnerships* (2022) provides strong guidelines for assessing and determining risk factors for institutions and researchers. Some of the core ideas are focused on the assembly of a research team; assessing partners' motivations for collaboration; using robust cybersecurity and data management protocols; and predetermining the intent of research and the use of findings.

19 The World Health Organization's (WHO's) network of BSL-4 laboratories offers particularly relevant examples of containment protocols that parallel many AI security needs. BSL-4 facilities must prevent both physical and information containment breaches while enabling crucial international research collaboration (WHO 2024). The International Organization for Standadization (ISO) standard for these labs, ISO 35001 "Biorisk management for laboratories and other related organisations" (ISO 2019), which may provide some insights for safety in a joint lab for AI. However, it is a voluntary standard which has no central verification body and so it is difficult to assess its broader efficacy; it is also unclear how many of the 40+ BSL-4 labs have adopted the standard.

20 The key area of interest for the joint lab from the ISS document is section 4.4.2.2.2 KU/LAN INFORMATION TECHNOLOGY (IT) SECURITY ASSESSMENT and the subsequent verification (NASA 2019).

Building on these foundations, the technical security architecture of a joint lab would need to encompass air-gapped development environments[21] protected by advanced zero-trust architectures that question the legitimacy of every access attempt, even from within the secure perimeter.[22]

Within this technical framework, fine-grained access controls would govern data enclaves,[23] using behavioural and other parameters to regulate system access based on multiple security factors.[24] These controls would integrate with advanced monitoring and verification systems to track model weights and research data, detecting improper access attempts and dynamically adjusting access permissions based on observed behaviour patterns.

Extending these technical solutions, a compartmentalization policy would separate key information, data and systems in their own enclaves and limit individual access to only the most essential pieces for an individual's work.[25] It would also be essential to limit the number of people whose work cuts across all or most core systems (such that few individuals have access to model weights, core security systems and other sensitive systems as a function of their work, limiting the scope of a leak if one were to occur).

To further guard against potential breaches, emergency response protocols could be developed to handle time-sensitive security threats. Extensive emergency response measures would need to be developed to specify how the lab would implement an immediate system shutdown in the event of security breaches. Other example elements include implementing self-destructing or self-corrupting storage systems for critical data, automated back-up systems in separate locations and protocols that alert local or international threat response, and enforcement agencies to continue emergency prevention measures outside of the lab if necessary. All measures would also require robust testing and verification processes to ensure that they are functioning as intended and still fit for purpose in a changing threat landscape.

In summary, it would be necessary to establish a high degree of technical safeguards to control individual access to systems, to continually monitor and verify that those individuals with access should still have access (based on their behaviour or changing security needs), to segment and protect data and information such that only the required and authorized information is available to an individual at a given time, and to have strict protocols to revoke access to or protect valuable information in case of a breach or violation of these measures. The technical solutions will be only as effective as the individuals that follow them, which necessitates a robust framework to govern those individuals that is tightly integrated with the technical security architecture.

These ideas are meant to provide initial guidance; they would be further developed by the leading security experts from several partner countries to ensure state-of-the-art policies are in place to secure the lab.[26] The lab would also incorporate novel strategies that are specific to securing AI

21 Some of these measures might include electromagnetic shielding or Faraday cages meeting around critical computing facilities attached to the lab to prevent electronic surveillance. These measures could be built to the National Security Agency's TEMPEST specifications (NSTISSAM TEMPEST/1-92) and/or the Institute of Electrical and Electronics Engineers' Standard 299-2006 for measuring the effectiveness of electromagnetic shielding enclosures.

22 While the specifications of this architecture would be built to the needs of the lab and with current cutting-edge technology, some elements of the architecture might include quantum key distribution systems for secure communication between facility sectors, similar to China's Beijing-Shanghai quantum network (Chen 2017), and zero-trust frameworks exceeding the recommendations of the NIST Special Publication 800-207 (Rose et al. 2020).

23 Secure enclaves for model training using Intel SGX or AMD SEV encryption technology could be employed within the lab and perhaps also for selective low-risk external access to research output data. Information rights management systems, blockchain-based audit systems and other methods could also be used to track and manage access to all systems.

24 Multi-factor biometric access control using protocols similar to those in sensitive compartmented information facilities (ICD/ICS 705) (National Counterintelligence and Security Center 2020).

25 The compute infrastructure could be secured using a combination of hardware and software controls including hardware security modules for cryptographic operations, meeting Federal Information Processing Standard 140-3 Level 4 requirements; custom field programmable gate arrays; homomorphic encryption based on IBM's HElib; or other similar frameworks.

26 The uncertainty and complexity of AI security concerns could map on well and benefit from thinking about cyber offence and defence dynamics internationally. In particular, emerging thinking on cyberweapons and the complexity of geopolitical strategy in light of these difficult-to-attribute and undetectable weapons might help to shape thinking on how to prevent undetected intrusions into a joint lab, as well as provide insight into efforts to detect and prevent potential self-extraction of advanced AI systems from a closed lab environment. For more on cyber weapons and their effects on geopolitical and strategic planning, see *Under the Nuclear Shadow: China's Information-Age Weapons in International Security* by Fiona S. Cunningham (2025), which explores the development of cyberweapons in China and more broadly as a means of waging war in ways that avoid the risk of nuclear war; and *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age* by David E. Sanger (2019), which is about the historical development and increasing use of cyberweapons from the perspective of decision makers and military in the United States.

models (for example, upload limits to prevent weight exfiltration) and novel strategies that involve deploying AI systems to improve security efforts.

Human Framework for Research Security

The human element of security presents perhaps the most complex challenge to a joint lab, requiring a delicate balance between rigorous controls and the practical needs of scientific research. Building from current research security frameworks, a research security plan for the lab would rely on pre-screening and assessments of researchers' affiliations, funding sources, and research history as a foundation. However, the reality of advanced AI research presents a unique paradox: the most qualified researchers may often be those with extensive experience in classified or sensitive research environments, potentially creating inherent security conflicts.

Scientists who work in the international lab will learn insights and techniques that — if leaked — could help other adversaries develop unauthorized AI programs. To guard against this, the security protocol may include quarantine periods (in which scientists agree to limit contact with the outside world) and monitoring measures (in which scientists agree to increased monitoring of their communications for certain periods of time). Additional details and examples are provided below:

Pre-tenure quarantine period is similar to a "cooling-off" period used in sensitive military and intelligence positions or in finance where the researcher would be housed close to the lab with restrictions on communication and movement for the period leading into their lab tenure. This quarantine period, ranging from several months to a year depending on the researcher's background and future role, serves multiple purposes beyond simple timeline separation. It provides an opportunity for detailed security assessment, acclimatization to new security protocols and the establishment of baseline behavioural patterns for future monitoring, whether through biometric methods, standardized questioning or other methods of assessment. While this measure cannot completely prevent long-term infiltration attempts, it creates a crucial buffer zone between previous affiliations and new research responsibilities.

Quarantine during tenure, which could span from two to as much as five years.[27] Researchers would face significant restrictions on external communications and likely need to reside within the facility's secure perimeter or nearby during their tenure. This approach draws lessons from historical precedents such as the Los Alamos Manhattan Project base, which was heavily monitored and restricted but never fully isolated, and the Soviet nuclear research city Arzamas-16, now called Sarov, which was fully cut off from the external world and had heavy intelligence and security presence monitoring the movements and work of the scientists there (Rhodes 1986; Holloway 1994, 202). However, even with varying degrees of isolation in both projects, the periods of work were somewhat shorter than proposed for a joint lab research tenure, and there was a sense of patriotism uniting the researchers in a common purpose that no doubt helped the difficult task of maintaining secrecy and isolation.

A tiered system of controlled interaction might offer a solution to the negative effects of isolation for some, with different levels of access and communication privileges based on security clearance, research role, and ongoing behavioral assessment. Some examples to look at for designing an isolated lab ethically and safely are NASA's HI-SEAS program and Russia's Mars-500 experiment which both demonstrate how extended isolation can be managed effectively with proper support systems.[28] An added benefit or challenge would be whether the researchers would be joined by their families in a broader research community or whether they would be alone during their tenure.

Post-tenure security measures would require particular attention, as this transition period presents unique vulnerabilities. A graduated approach combining immediate quarantine with longer-term monitoring could provide more effective security than extended isolation alone. The initial quarantine period would allow for critical security updates and system

27 The timeline for tenure in the lab would need to be determined based on the individual's role, the work they are doing and other considerations, such as a cost/benefit analysis of the time and financial costs of extended security monitoring and verifications against the value of the work being done by the individual. Also, considering the extent to which the individual is engaging with sensitive material would also be necessary. The director of the lab, for example, would need a longer tenure than a technician working on cooling systems, to make the efforts worth it.

28 See www.hi-seas.org; www.esa.int/Science_Exploration/Human_and_Robotic_Exploration/Mars500/Mars500_study_overview.

modifications, while subsequent monitoring and controlled reintegration would help prevent information leakage while maintaining research productivity. This could include sophisticated monitoring systems activated by specific trigger behaviours using natural language processing capabilities or information patterns, rather than blanket surveillance that might violate privacy rights or create undue burden. This would be an enhancement of existing requirements for security clearance in many countries.

In summary, ensuring reliable security in the work of a joint lab would involve the combination of complex technical and policy-based solutions to control access, limit contact, evaluate external relationships of researchers, monitor communication, quarantine researchers with sensitive information and create barriers to distance external threats from researchers themselves.[29] All of these components serve to enhance a robust security framework to maximize collaboration and open science within the lab, while also providing the highest level of research security possible. Additional frameworks such as limits on computing infrastructure, physical controls on chips, and applications of AI to enhance security could also support this framework.

Emergency Protocols

The joint lab would have emergency protocols that share some similarities with those of BSL-4 labs. BSL-4 labs have emergency protocols to handle scenarios such as spills, exposures and containment breaches. The emergency protocols include information about measures to limit the spread of a pathogen, evacuation plans for staff and how to immediately contact emergency services (for instance, fire departments, hospitals and public health agencies). BSL-4 labs also conduct drills to test the effectiveness of their emergency response protocols (Tajuddin, n.d.). Large labs may have emergency response teams or personnel specifically trained to handle complex or high-risk situations. This team may include medical personnel, decontamination specialists, and engineering staff who work together to assess the severity of the emergency and implement appropriate corrective actions (Le Duc et al. 2008).

The joint lab's emergency response protocols would focus on detecting, preventing, and responding to time-sensitive AI emergencies. Example emergency scenarios include (a) an AI system with agentic capabilities attempting to autonomously escape containment measures, (b) an adversary attempting to steal the model weights of a powerful AI system, (c) the discovery of technical breakthroughs that could undermine the enforcement of international AI agreements, (d) evidence that another party is secretly attempting to conduct an unauthorized AI project and (e) instances in which internal researchers were attempting to use the model in malicious ways (see Wasil, Reed, et al. 2024). In many of these scenarios, it would be essential to swiftly turn off the computing cluster powering an AI system, secure model weights and communicate swiftly with relevant authorities.

Adequate emergency response protocols would include several elements (see MIRI Technical Governance Team, Barnett and Thiergart 2024). Examples include:

- a list of the different kinds of emergency scenarios and warning signs for such scenarios;
- an identification of which personnel within the joint lab are responsible for detecting and reporting signs of an emergency;
- the decision-making process used to determine if certain development or deployment activities need to be halted to respond to a potential emergency, which includes a description of the chain of command and an identification of key personnel authorized to initiate emergency protocols;
- the steps required to shut down an AI system or computing cluster and which parties would be able to order and execute such steps; and
- communication procedures to notify relevant national or international entities, especially in cases when coordinating with such entities is needed to eliminate the security threat.

29 We recognize that such restrictions would represent significant costs for the individuals involved, much like the sacrifices made by scientists during the Manhattan Project and sacrifices made by national security personnel working on highly sensitive topics today. While such sacrifices are not entirely unprecedented, their intensity may need to be even greater than in the past, given the extremely high stakes involved in the development of advanced AI and the potentially large incentives for state-sponsored espionage. While not all leading AI scientists may choose to accept such sacrifices, those who do so could be generously compensated both financially, as well as by the knowledge that they are working on a project of enormous importance for global security and prosperity.

Discussion

This paper discusses the establishment of a joint lab to navigate the safe and secure development of highly advanced AI. This work could be conducted at the national level (pooling national talent and reducing race dynamics between domestic companies) or the international level (pooling talent across nations and more robustly eliminating race dynamics for certain kinds of highly advanced AI development). While the paper's focus is on the (more ambitious) international version of the proposal, the ideas it discusses are also highly applicable to explorations of national AGI Manhattan projects. It is also worth noting that national AGI Manhattan projects could serve as stepping stones toward international proposals. For example, in the event that national AGI project officials determined that scientists needed more time or resources to safely align powerful systems, this could motivate national governments to explore international agreements (including joint labs) to curb race dynamics and provide scientists with the time needed to develop stronger guardrails.

Will such a proposal ever be considered, or is this purely an academic exercise? The authors of this paper believe that AI progress has been filled with surprises. It is plausible that AI progress continues rapidly in the private sector and that global security challenges never materialize, are handled primarily by private sector actors or are not handled successfully. The authors also think it is plausible that, at some point, major governments notice global security challenges associated with advanced AI development and seriously consider ambitious proposals for how to navigate them. In that scenario, the authors believe that this paper can serve as a useful resource.

Future work could explore a variety of open questions and topics. Examples include the following:

- **Applications of advanced AI systems to improve security of a joint lab:** How could (safe and trusted) AI be applied to improve the security of model weights or prevent critical information from leaking?
- **Lessons learned from international security agreements and joint research projects:** What relevant insights can we draw from international security agreements (for example, the Treaty on the Non-Proliferation of Nuclear Weapons) or joint research or security projects (for instance, Five Eyes, AUKUS) ?
- **Verification methods to detect unauthorized AI projects:** What techniques could governments use to detect unauthorized AI projects or deter actors from pursuing unauthorized AI projects (see Scher and Thiergart 2024; Wasil, Clymer, et al. 2024)?
- **Promoting global understanding of AI capabilities and risks:** What kinds of efforts could help governments better understand the capabilities and risks associated with advanced AI (for example, the International AI Safety Report [GOV.UK 2025])? What kinds of efforts could help build consensus among experts or officials from multiple nations (e.g., the International Dialogues on AI Safety)?
- **Exploring public-private partnerships:** What models of public-private partnerships could be leveraged to ensure the safe development of advanced AI? What are trade-offs associated with different models?

Acknowledgement

This paper was written in consultation with AI security, national security and information security experts who have contributed significantly to this research and paper.

Appendix

Table A.1: International Joint Lab for AI R&D (detailed breakdown from Figure 1)

Purpose

- → Accelerate safe and secure development of highly powerful AI systems
- → Prevent development of unsafe or unaligned superintelligence
- → Ensure global distribution of AI benefits

Structure

Lab Structure (based on BSL-4 labs)	
Position	**Key Responsibilities**
Governing Body (Board, Executive Director)	Oversight and decision making across medium to high-risk decision categories
Laboratory Directors	Overall administration, compliance, external liaison
Principal Investigators	Lead specific research projects, maintain safety protocols
AI Safety Officers	Monitor compliance, conduct safety audits, coordinate red teaming
Research Scientists	Execute experiments under supervision of Principal Investigators
Infrastructure Managers	Manage computational resources, secure data storage

Governance

Governance			
Governing Body (based on IAEA)			
Role	**Selection**	**Term**	**Responsibilities**
Governing Board	Representatives from select nations	Annual rotation	Vote on medium/high-risk decisions
Executive Director	Elected by Governing Board	one year	Daily operations and strategy
Permanent Seats	Nations with advanced AI programs	Permanent but rotating individuals	Veto power on high-risk decisions

Decision Making

Risk Level	Approval Required	Examples
Low Risk	Executive Director only	Daily experiments, hiring, research communication
Medium Risk	Majority vote (Governing Board)	Budget allocation, research priorities, security protocols
High Risk	Three-quarter vote plus veto power	Compute threshold increases, powerful system deployment

Table A.1: International Joint Lab for AI R&D (continued)

Security Mechanisms

Information Security (technical and in-lab)

Component	Key Features
Physical Security	Defence-in-depth approach, air-gapped environments
Access Control	Zero-trust architecture, compartmentalization, fine-grained permissions
Monitoring	Behavioural tracking, automated threat detection
Emergency Response	System shutdown protocols, self-destructing storage

Research Security (human and both in and outside lab)

Phase	Security Measures
Pre-tenure	Quarantine period (months to one year), background screening
During Tenure	Restricted communication, secure facility residence (two to five years)
Post-tenure	Graduated quarantine, long-term monitoring, controlled reintegration

Emergency Protocols

Emergency Response Protocols (adapted to scenarios)	
Component	Description
Detection	Designated personnel, warning sign identification
Decision-Making	Clear chain of command, authorized personnel
Response Actions	System shutdown procedures, model weight securing
Communication	Notification protocols for national/international entities

Source: Authors.

Table A.2: International Joint Lab for AI R&D (option showing all details and intersection of roles, decision making, access and security protocols for a joint lab)

Governing Body

Entity/Role	Decision Tier	Decision-Making Authority	Access and Clearance	Security/Quarantine Protocols
Executive Director	Low to Medium	Makes low-risk decisions independently Proposes medium-risk actions for Board approval	Full clearance to lab operations and strategic info	Pre-tenure measures and quarantine (PreQ) During tenure measures (DTM) and lab residence requirement (LRR) plus continued monitoring (CM) Post-tenure measures and quarantine (PoQ) Highest scrutiny and longest quarantine periods but with flexibility for movement between the lab and other roles external to lab
Governing Board	Medium to High	Approves medium-risk decisions (majority vote) Approves high-risk decisions (three-quarters vote plus some nation veto)	Oversight only No operational control	Board elects Executive Director, regular monitoring and security measures for board members residing outside of the lab
Nation-State Representatives	High (Veto Power)	Hold veto on select high-risk decisions	Strategic access (no lab entry)	Monitored engagement Annual term rotation Clearance required for Board involvement

Table A.2: International Joint Lab for AI R&D (option showing all details and intersection of roles, decision making, access and security protocols for a joint lab) (Continued)

Lab Staff

Lab Role	Primary Responsibilities	Access Level	Security Protocols / Quarantine
Laboratory Directors	Oversee lab operations Liaise with external regulators Align research with mission	Full administrative access	PreQ, DTM, LRR, CM, PoQ Highest scrutiny and longest quarantine periods
Infrastructure Managers	Maintain servers, compute, and secure storage Control access permissions	System-level access Zero-trust environments	PreQ, DTM, LRR, CM, PoQ Highest scrutiny and longest quarantine periods
AI Safety Officers	Monitor lab safety compliance Run red-teaming Assess societal risks	Internal audit access Safety protocol oversight	PreQ, DTM, LRR, CM, PoQ Highest scrutiny and longest quarantine period due to central role in security and monitoring
Principal Investigators	Lead specific research projects Ensure ethical and safety compliance	Access to specific projects and models	PreQ, DTM, LRR, CM, PoQ Scrutiny and quarantine periods dependent on sensitivity of the specific lab and considering impacts on other employment and life external to the lab
Research Scientists	Train/test AI models Conduct experiments Analyze data	Role-based, limited compute and data access	PreQ, DTM, LRR, CM, PoQ Scrutiny and quarantine periods dependent on sensitivity of the specific lab & considering impacts on other employment and life external to the lab
Compliance Officers	Policy and legal adherence Inspections and reporting analysis	Access to internal recording and monitoring systems, reports, etc. No direct access to models or systems.	PreQ, CM, PoQ No quarantine and geographical location requirements during tenure due to the need to travel between the lab, agency and other bodies. Higher degree of monitoring and scrutiny by other means balance impossibility of quarantine. Pre- and post-tenure measures to evaluate potential compromising associations, etc., with possible quarantine period after and before other work

Source: Authors.

Works Cited

Aschenbrenner, Leopold. 2024. "Situational Awareness: The Decade Ahead." *Situational Awareness*. June. https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf.

Brundage, Miles. 2024. "My recent lecture at Berkeley and a vision for a 'CERN for AI.'" *Miles's Substack*, December 10. https://milesbrundage.substack.com/p/my-recent-lecture-at-berkeley-and.

Cass-Beggs, Duncan, Stephen Clare, Dawn Dimowo and Zaheed Kara. 2024. *Framework Convention on Global AI Challenges*. Discussion Paper. Waterloo, ON. www.cigionline.org/publications/framework-convention-on-global-ai-challenges/.

Centre for AI Safety. 2024. "Statement on AI Risk: AI experts and public figures express their concern about AI risk." https://safe.ai/work/statement-on-ai-risk.

Chen, Na. 2017. "Beijing-Shanghai Quantum Communication Network Put into Use." Chinese Academy of Sciences, September 1. https://english.cas.cn/newsroom/archive/news_archive/nu2017/201703/t20170324_175288.shtml.

Cheng, Deric and Corin Katzke. 2024. "Soft Nationalization: How the US Government Will Control AI Labs." Convergence Analysis, August 28. www.convergenceanalysis.org/publications/soft-nationalization-how-the-us-government-will-control-ai-labs.

Clymer, Joshua, Nick Gabrieli, David Krueger and Thomas Larsen. 2024. "Safety Cases: How to Justify the Safety of Advanced AI Systems." Preprint, arXiv, March 18. https://doi.org/10.48550/arXiv.2403.10462.

Cunningham, Fiona S. 2025. *Under the Nuclear Shadow: China's Information-Age Weapons in International Security*. Princeton, NJ: Princeton University Press.

Google DeepMind. 2024. "Unreasonably Effective AI with Demis Hassabis." August 14. YouTube video, 51:59. www.youtube.com/watch?v=pZybROKrj2Q.

Government of Canada. 2022. *National Security Guidelines for Research Partnerships*. October 6. https://science.gc.ca/site/science/en/safeguarding-your-research/guidelines-and-tools-implement-research-security/national-security-guidelines-research-partnerships.

GOV.UK. 2025. *International AI Safety Report*. www.gov.uk/government/publications/international-ai-safety-report-2025.

Hogarth, Ian. 2023. "We must slow down the race to God-like AI." *Financial Times*, April 13. www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2.

Holloway, David. 1994. *Stalin and the Bomb: The Soviet Union and Atomic Energy, 1939–1956*. New Haven, CT: Yale University Press.

Holdsworth, Jim and Matthew Kosinski. 2024. "What is information security (InfoSec)?" IBM, July 26. www.ibm.com/topics/information-security.

ISO. 2019. *Biorisk management for laboratories and other related organisations*. ISO 35001:2019. www.iso.org/standard/71293.html.

Juijn, Daan, Bálint Pataki, Alex Petropoulos and Max Reddel. 2024. *CERN for AI: The EU's seat at the table*. International Center for Future Generations, September 4. https://icfg.eu/wp-content/uploads/2024/09/CERN_for_AI_FINAL_REPORT.pdf.

Le Duc, James W., Kevin Anderson, Marshall E. Bloom, James E. Estep, Heinz Feldmann, John B. Geisbert, Thomas W. Geisbert et al. 2008. "Framework for Leadership and Training of Biosafety Level 4 Laboratory Workers." *Emerging Infectious Diseases* 14 (11): 1685–88. https://doi.org/10.3201/eid1411.080741.

MIRI Technical Governance Team, Peter Barnett and Lisa Thiergart. 2024. "Response to BIS AI Reporting Requirements RFC." Policy comment, October 11. Berkeley, CA: Machine Intelligence Research Institute. https://techgov.intelligence.org/research/response-to-bis-ai-reporting-requirements-rfc.

NASA. 2019. *ISS Safety Requirements Document: International Space Station Program Baseline*. SSP 51721. September. https://s3vi.ndc.nasa.gov/ssri-kb/static/resources/SSP%2051721-Baseline.pdf.

National Counterintelligence and Security Center. 2020. "Technical Specifications for Construction and Management of Sensitive Compartmented Information Facilities." ICD/ICS 705. www.dni.gov/files/Governance/IC-Tech-Specs-for-Const-and-Mgmt-of-SCIFs-v15.pdf.

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley and Jeff Alstott. 2024. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. RAND, May 30. www.rand.org/pubs/research_reports/RRA2849-1.html.

OpenAI. 2023. "Preparedness Framework (Beta)." December 18. https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

— — —. "Preparedness Framework. Version 2." April 15. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

Rhodes, Richard. 1986. *The Making of the Atomic Bomb*. New York, NY: Simon and Schuster.

Rose, Scott, Oliver Borchert, Stu Mitchell and Sean Connelly. 2020. "Zero Trust Architecture." NIST Special Publication 800-207. https://doi.org/10.6028/NIST.SP.800-207.

Sanger, David E. 2019. *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age*. New York, NY: Crown.

Scher, Aaron and Lisa Thiergart. 2024. *Mechanisms to Verify International Agreements About AI Development*. Technical report, November 27. Berkeley, CA: Machine Intelligence Research Institute. https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development.

Strouse, Gregory F., Timothy R. Wood, Claire M. Saundry, Philip A. Bennett and Mary Bedner. 2023. *Safeguarding International Science: Research Security Framework*. NIST Internal Report 8484. https://doi.org/10.6028/NIST.IR.8484.

Tajuddin, Afnan. n.d. "Understanding biosafety levels." SafetyNotes. www.safetynotes.net/understanding-biosafety-levels/.

U.S.-China Economic and Security Review Commission. 2024. *Comprehensive List of the Commission's 2024 Recommendations*. www.uscc.gov/sites/default/files/2024-11/2024_Comprehensive_List_of_Recommendations.pdf.

Wasil, Akash R., Joshua Clymer, David Krueger, Emily Dardaman, Simeon Campos and Evan R. Murphy. 2024. "Affirmative safety: An approach to risk management for high-risk AI." Preprint, arXiv. https://doi.org/10.48550/arXiv.2406.15371.

Wasil, Akash, Everett Smith, Corin Katzke and Justin Bullock. 2024. "AI Emergency Preparedness: Examining the federal government's ability to detect and respond to AI-related national security threats." Preprint, arXiv, July 27. https://doi.org/10.48550/arXiv.2407.17347.

Wasil, Akash R., Peter Barnett, Michael Gerovitch, Roman Hauksson, Tom Reed and Jack William Miller. 2024. "Governing dual-use technologies: Case studies of international security agreements and lessons for AI governance." Preprint, arXiv, September 4. https://doi.org/10.48550/arXiv.2409.02779.

Wasil, Akash R., Tom Reed, Jack William Miller and Peter Barnett. 2024. "Verification methods for international AI agreements." Preprint, arXiv, November 4. https://doi.org/10.48550/arXiv.2408.16074.

WHO. 2024. *Laboratory biosecurity guidance*. Geneva, Switzerland: WHO. www.who.int/publications/b/69836/.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org