Gladstone AI

Defense in Depth: An Action Plan to Increase the Safety and Security of Advanced AI

This publication was produced for review by the United States Department of State. It was prepared by Gladstone February 26, 2024 and the contents are the responsibility of the authors. The authors' views expressed in this publication do not reflect the views of the United States Department of State or the United States Government.

Organization | Gladstone AI Inc. (hello@gladstone.ai)

Authors

Edouard Harris*

Jeremie Harris

Mark Beall

* Lead and corresponding author.
Contact: edouard@gladstone.ai

Table of Contents

Executive summary **11**

LOE1: Establish interim safeguards to stabilize advanced AI development 15

LOE2: Strengthen capability and capacity for advanced AI preparedness and response 16

LOE3: Increase national investment in technical AI safety research and standards development 17

LOE4: Formalize safeguards for responsible AI development and adoption by establishing an AI regulatory agency and legal liability framework 18

LOE5: Enshrine AI safeguards in international law and secure the AI supply chain 19

0. Introduction **21**

0.1 Background 22

0.2 Categories of AI risk 24

0.2.1 Weaponization 24

0.2.2 Loss of control 24

0.2.3 Other risk categories 26

0.2.4 Risks addressed by this action plan 27

0.3 Potential sources of catastrophic AI risk 27

0.4 Arguments against regulation for catastrophic AI risk 29

0.4.1 Self-regulation will be sufficient 29

0.4.2 Regulation could damage U.S. innovation and competitiveness 30

0.4.3 Catastrophic AI risk could divert attention from other issues 32

0.4.4 Catastrophic AI risk mitigation is unnecessary 32

0.5 Challenges 33

0.5.1 Technical challenges 33

0.5.1.1 The worst-case outcomes for AI risk are considered both plausible and extremely severe 33

0.5.1.2 Timescales for catastrophic risk are uncertain 35

0.5.1.3 The degree of risk from loss of control is uncertain 36

0.5.1.4 Frontier labs lack safety and security measures to detect and prevent loss of control 37

0.5.1.5 Frontier labs lack sufficient security to prevent critical IP theft 38

0.5.1.6 Open-access development can elicit dangerous capabilities from previously safe AI models 39

0.5.1.7 Closed-access AI models are vulnerable to black-box exfiltration and other attacks 40

0.5.2 Political challenges 40

0.5.2.1 AI advances faster than the ordinary policy process 40

0.5.2.2 The information environment around advanced AI makes grounded conversations challenging 41

0.5.3 Economic and strategic challenges 42

0.5.3.1 Frontier labs face strong incentives to develop and deploy increasingly advanced AI systems with limited safeguards 42

0.5.3.2 AI supply chain proliferation cannot be mitigated after the fact 43

0.5.4 Legal challenges 45

0.5.4.1 The current legal environment is inadequate to address the most extreme risks from advanced AI 45

Action plan to increase the safety and security of advanced AI 47

LOE1: Establish interim safeguards to stabilize advanced AI development 48

1.1 National security threats addressed by this LOE 49

1.2 Establish an AI Observatory for advanced AI 50

1.2.1 Functions 51

1.2.2 Staffing 52

1.2.3 Interagency coordination 52

1.3 Establish responsible AI development and adoption safeguards for U.S. private industry 53

1.3.1 Implementation mechanism and statutory authorities 54

1.3.2 RADA safeguards principles 54

1.4 Establish an AI Safety Task Force for RADA oversight 56

1.4.1 Mission 57

1.4.1.1 Priority Objective: Finalize and secure agreement on RADA safeguards 57

1.4.1.2 Sustainment Component: Oversee compliance with RADA safeguards 61

1.4.1.3 Sustainment Component: Develop recommendations for future regulations 61

1.4.2 Resources and staffing 62

1.4.3 Financial requirements and budget 63

1.4.4 Authorities 63

1.4.5 Location 65

1.5 Securing the advanced AI supply chain 65

1.5.1 AI model weights 67

1.5.2 Cloud computing 67

1.5.3 AI hardware 69

1.5.4 Research collaborations 71

1.5.5 Education of foreign nationals 72

LOE2: Strengthen capability and capacity for advanced AI preparedness and response 73

2.1 Coordinate interagency working groups 74

2.2 Advanced AI education and training 75

2.2.1 Key stakeholders and learning outcomes 76

2.2.2 Suggested training topics 78

2.3 Indications and warnings 82

2.3.1 Key categories and sources of catastrophic risk 83

2.3.2 Improving I&W through bug bounties 86

2.3.3 Improving I&W through direct government investment 86

2.4 Contingency planning 87

2.4.1 Updating the National Preparedness System 88

LOE3: Increase national investment in technical AI safety research and standards development 89

3.1 Federally funded research in AI safety and security and AGI-scalable alignment 90

3.1.1 Levels of AI safety and security research 91

3.1.1.1 Open-source and public research 91

3.1.1.2 Research that requires access to proprietary AI models 92

3.1.1.3 National security research 93

3.1.2 Organizational framework 94

3.1.2.1 Considerations for AGI alignment research 95

3.1.2.2 Research collaborations 96

3.1.2.3 Sensitive research areas 96

3.2 Standards for AI evaluations and RADA safeguards 97

3.2.1 Limitations of AI evaluations 98

3.2.1.1 AI evaluations are not comprehensive 98

3.2.1.2 AI evaluations cannot confirm that a dangerous capability is absent 98

3.2.1.3 AI evaluations are highly vulnerable to manipulation 99

3.2.1.4 AI evaluations could fail systematically in high-capability regimes 100

3.2.2 Addressing the limitations of AI evaluations 100

3.2.3 Standards for RADA safeguards under catastrophic risk 101

LOE4: Formalize safeguards for responsible AI development and adoption by establishing an AI regulatory agency and legal liability framework 102

4.1 Establish the Frontier AI Systems Administration (FAISA) 103

4.1.1 FAISA appropriations and staffing 104

4.1.2 FAISA activities 105

4.1.2.1 Oversee licensing regime for safe AI scaling 105

4.1.2.2 Monitor AI hardware concentrations and AI programs 106

4.1.2.3 Enforce licensing and reporting requirements 107

4.1.2.4 Update technical thresholds for licensing 107

4.1.2.5 Maintain information repositories 108

4.1.3 Comprehensive framework for RADA safeguards 109

4.1.3.1 AI hardware designers (AIHDs) 112

4.1.3.1.1 Tier 1: AIHDs at or below the hardware threshold 114

4.1.3.1.2 Tier 2: AIHDs above the hardware threshold 114

4.1.3.2 Data center infrastructure providers (DCIPs) 114

4.1.3.2.1 Tier 1: DCIPs operating data center facilities at or below the power consumption threshold 116

4.1.3.2.2 Tier 2: DCIPs operating data center facilities above the power consumption threshold 116

4.1.3.3 AI hardware owners (AIHOs) 117

4.1.3.3.1 Tier 1: AIHOs who are at or below the aggregate compute capacity threshold 119

4.1.3.3.2 Tier 2: AIHOs who fall above the aggregate compute capacity threshold 119

4.1.3.4 AI model developers (AIMDs) 120

4.1.3.4.1 Tier 1: AI models below the registration threshold 122

4.1.3.4.2 Tier 2: AI models above the registration threshold, but below the approval threshold 123

4.1.3.4.3 Tier 3: AI models above the approval threshold, but below the controlled threshold 125

4.1.3.4.4 Tier 4: AI models above the controlled threshold 128

4.1.3.5 General provisions for all licensed entities 129

4.1.4 Publication controls 131

4.2 Legislative environment 133

4.2.1 Civil liability 133

4.2.2 Criminal liability 135

4.2.3 Emergency powers 137

4.3 Advanced AI in national security systems 139

LOE5: Enshrine AI safeguards in international law and secure the AI supply chain 141

5.1 Desired end state of this LOE 142

5.2 Build international consensus and partner capacity on catastrophic AI risk reduction 143

5.2.1 Coordinate domestic and international messaging and capacity-building 144

5.2.1.1 Policymaker education and outreach 145

5.2.1.2 Technical education and outreach 146

5.2.1.3 Other forms of capacity-building 148

5.2.2 Articulate and reinforce an official U.S. government position on catastrophic AI risk 149

5.3 Enshrine AI safeguards in international law 149

5.3.1 Treaty structure considerations 150

5.3.2 Preparation for a treaty 151

5.4 Establish an international regulatory agency 152

5.4.1 Monitoring and verification 153

5.4.2 Standard-setting 154

5.4.3 Convening researchers 154

5.5 Allied multilateral initiatives to manage the AI supply chain 155

5.5.1 Strategic considerations 156

5.5.2 Multilateral partnerships 156

5.6 Open challenges 157

Conclusion 159

Funding disclosure 162

Acknowledgements 163

Bibliography 170

Annex A: Glossary of terms 204

Annex B: The full challenge of AGI alignment 213

Annex C: Example AI alignment failure scenarios 214

C.1 Negligible impact: AI cheating at a video game 214

C.2 Low impact: AI sweeping bot 214

C.3 Medium impact: Dangerously creative drone 214

C.4 High impact: Electrical grid failure 215

Annex D: Advanced AI landscape 217

D.1 Frontier AI labs 217

D.2 China-based entities 218

D.3 Open-access developers 218

D.4 Elite quantitative hedge funds 219

Annex E: Funding in AI safety 221

Annex F: Persuasion and manipulation 222

Annex G: Primer on AI and compute 224

G.1 Compute as a pathway to frontier AI development 224

G.2 Implications for risk mitigation strategy 225

G.3 Key thresholds and existing compute concentrations 225

Annex H: AIO activities 229

Annex I: Voluntary Charter for responsible AI 231

I.1 Information sharing 231

I.2 Compute reporting threshold 233

I.3 Model evaluation protocols 234

I.4 Capability prediction protocols 242

I.5 Security measures 243

I.6 Model containment measures 245

I.7 AI safety and AGI alignment research 247

I.8 Dangerous capability ban 248

I.9 Capability research controls 248

I.10 Risk governance 249

I.11 Caps on cloud services for scaled training runs 250

I.12 Incremental adoption 250

I.13 Final considerations for negotiating Charter terms 251

Annex J: Effective compute 252

Annex K: ASTF activities and task-organization 254

K.1 Oversee and support compliance with RADA safeguards 254

K.2 Develop recommendations for a legal and regulatory regime 256

K.3 ASTF task-organization 257

Annex L: AI safety and security research topics 260

L.1 AGI-scalable alignment between AI behaviors and human values 260

L.2 Monitoring of AI systems to detect misbehavior and preempt failures 261

L.3 Robustness in the face of adversaries and unforeseen circumstances 262

L.4 Hardware-based verification 262

Annex M: Secure temporary storage of model weights 263

Annex N: Training approvals process for high-risk AI models 265

N.1 Dangerous capability predictions 266

N.2 Capability evaluations 267

N.3 Capability predictions at training checkpoints 267

N.4 Data controls 267

N.5 Risk management 268

Annex O: Training stage monitoring for high-risk AI models 269

O.1 Training journal 269

O.2 Ongoing capability predictions 270

O.3 Ongoing capability evaluations 271

O.4 Model output predictions 272

O.5 Training kill switch 273

O.6 Failed evaluation protocol 274

Annex P: Deployment stage approvals for high-risk AI models 276

P.1 Lifecycle funds 277

P.2 Pre-deployment KYC 278

P.3 Risk and contingency planning 278

P.4 Weight storage for model evaluations 278

P.5 Model augmentations 278

P.6 Failed evaluation protocol 279

P.7 Fine-tuning approval 279

Annex Q: Deployment stage monitoring of high-risk AI models 281

Q.1 Usage pattern reporting 282

Q.2 Red teaming 283

Q.3 Failed evaluation protocol 283

Table of Figures

Figure 1. Overview of the action plan and its component LOEs. 13

Figure 2. Visualization of an approximate spectrum of AI risks, ranging from individual and societal risks to catastrophic national security risks. 27

Figure 3. Visualization of the AI supply chain, potential sources of catastrophic risk, and categories of catastrophic risk. 29

Figure 4. Simplified representation of the components of the AI supply chain. 43

Figure 5. Safeguards to introduce to frontier AI developers, AI hardware owners, and data center infrastructure providers in the event that RADA safeguards can (Plan A) and cannot (Plan B) be mandated. 60

Figure 6. Summary of an example licensing regime for RADA safeguards. 111

Figure 7. Summary of an example AI hardware designer licensing regime for RADA safeguards. 113

Figure 8. Summary of an example data center infrastructure provider licensing regime for RADA safeguards. 115

Figure 9. Summary of an example AI hardware owner licensing regime for RADA safeguards. 118

Figure 10. Summary of an example AI model developer licensing regime for RADA safeguards. 121

Figure 11. Summary of key licensing requirements at each stage of the AI product lifecycle under RADA safeguards. 127

Figure 12. Visualization of our definitions of advanced AI, frontier AI, and AGI. 209

Figure 13. Organizational chart for the AI Safety Task Force. 257

Table of Tables

Table 1. U.S. government stakeholders and potential target learning outcomes of AI education. 77

Table 2. Training topics associated with target learning outcomes. 79

Table 3. Potential sources of catastrophic AI risk, corresponding risk categories, and information requirements associated with each. 83

Table 4. Potential inaugural members of an ASCCR and the components of the AI supply chain over which they exert the most influence. 157

Executive summary

The recent explosion of progress in advanced artificial intelligence (AI) has brought great opportunities, but it is also creating entirely new categories of weapons of mass destruction-like (WMD-like) and WMD-enabling catastrophic risks[1] [1–4]. A key driver of these risks is an acute competitive dynamic among the frontier AI labs[2] that are building the world's most advanced AI systems. All of these labs have openly declared an intent or expectation to achieve human-level and superhuman artificial general intelligence (AGI)[3] — a transformative technology with profound implications for democratic governance and global security — by the end of this decade or earlier [5–10].

The risks associated with these developments are global in scope, have deeply technical origins, and are evolving quickly. As a result, policymakers face a diminishing opportunity to introduce technically informed safeguards that can balance these considerations and ensure advanced AI is developed and adopted responsibly. These safeguards are essential to address the critical national security gaps that are rapidly emerging as this technology progresses.

Frontier lab executives and staff have publicly acknowledged these dangers [11–13]. Nonetheless, competitive pressures continue to push them to accelerate their investments in AI capabilities at the expense of safety and security (Introduction, 0.5.3.1). The prospect of inadequate security at frontier AI labs raises the risk that the world's most advanced AI systems could be stolen from their U.S. developers, and then weaponized against U.S. interests [9]. Frontier AI labs also take seriously the possibility

[1] By **catastrophic risks**, we mean risks of catastrophic events up to and including events that would lead to human extinction. See the Glossary of terms for our full definition.

[2] By **frontier AI labs**, we mean the organizations that are involved in building cutting-edge, general-purpose AI systems, and whose research programs are explicitly aimed at, or could plausibly lead to, the development of artificial general intelligence or AGI. Examples include OpenAI, Google DeepMind, and Anthropic. See the Glossary of terms for our full definition.

[3] By **AGI**, we mean an AI system that can outperform humans across all economic and strategically relevant domains, such as producing practical long-term plans that are likely to work under real world conditions. See the Glossary of terms for our full definition.

that they could at some point lose control[4] of the AI systems they themselves are developing [5,14], with potentially devastating consequences to global security (Introduction, 0.5.1.1).

Given the growing risk to national security posed by rapidly expanding AI capabilities from weaponization and loss of control — and particularly, the fact that the ongoing proliferation of these capabilities serves to amplify both risks — there is a clear and urgent need for the U.S. government to intervene.

This action plan is a blueprint for that intervention. Its aim is to **increase the safety and security of advanced AI by countering catastrophic national security risks from AI weaponization and loss of control**. It was developed over thirteen months, and informed by conversations with over two hundred stakeholders from across the U.S., U.K., and Canadian governments; major cloud providers; AI safety organizations; security and computing experts; and formal and informal contacts at the frontier AI labs themselves.[5] The actions we propose follow a sequence that:

- Begins by establishing interim safeguards to stabilize advanced AI development, including export controls on the advanced AI supply chain;

- Leverages the time gained to develop basic regulatory oversight and strengthen U.S. government capacity for later stages;

- Transitions into a domestic legal regime of responsible AI development and adoption, safeguarded by a new U.S. regulatory agency; and

- Extends that regime to the multilateral and international domains.

[4] **Loss of control** due to **AGI alignment failure** is a potential failure mode under which a future AI system could become so capable that it escapes all human efforts to contain its impact. See the Glossary of terms for our full definition.

[5] See the Acknowledgments section of this document for details about these stakeholders.

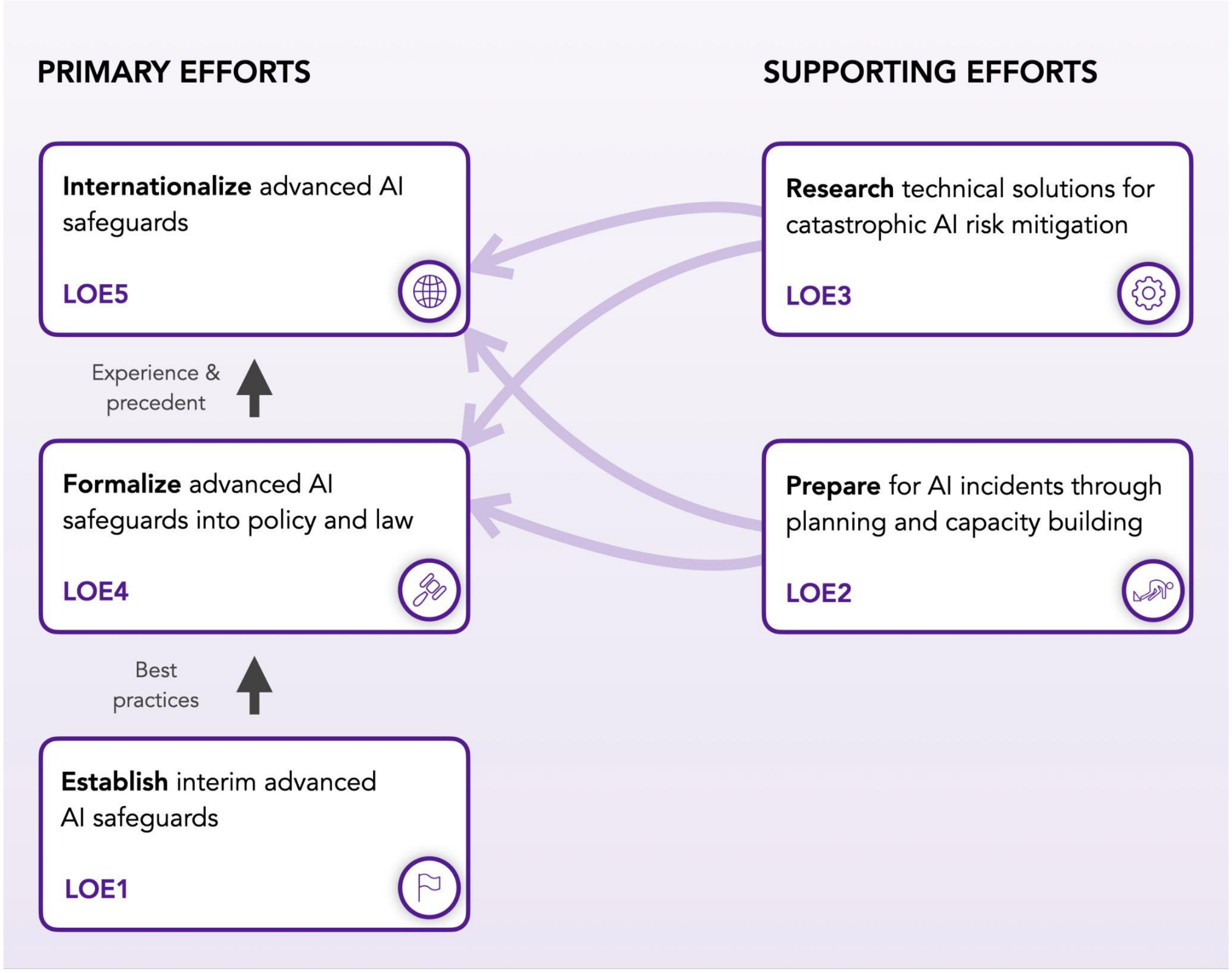

PRIMARY EFFORTS
SUPPORTING EFFORTS
Internationalize advanced AI safeguards
LOE5
Experience & precedent
Formalize advanced AI safeguards into policy and law
LOE4
Best practices
Establish interim advanced AI safeguards
LOE1
Research technical solutions for catastrophic AI risk mitigation
LOE3
Prepare for AI incidents through planning and capacity building
LOE2

Figure 1. Overview of the action plan and its component LOEs.

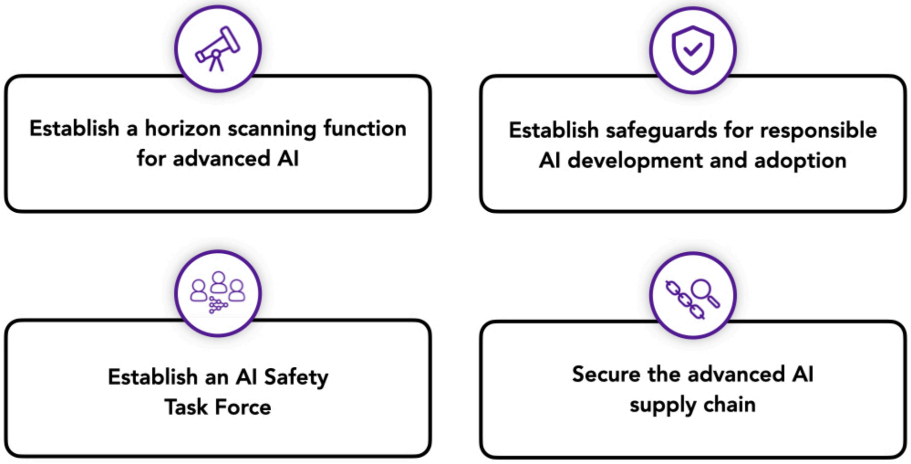
The U.S. government and its allies and partners, in close partnership with industry, can achieve this aim by implementing five mutually supporting lines of effort (LOEs). These LOEs will **establish** (LOE1), **formalize** (LOE4), and **internationalize** (LOE5) safeguards on advanced AI development, while **increasing preparedness** (LOE2) and **building technical capacity and capability** (LOE3). Some of the measures we propose are unprecedented, but after consulting with stakeholders and experts, we believe they are proportionate to the magnitude and urgency of the risk we face.

Because of the severity, uncertainty, and irreversibility of these risks, an action plan to address them needs to offer a wide margin of safety. This plan follows the principle of **defense in depth**, in which multiple overlapping controls combine to offer resilience against any single point of failure. We frame tradeoffs in terms of AI breakout timelines, the amount of time it would take an actor to train an AI system from scratch to equal the current state-of-the art under various expert-vetted assumptions. And we illustrate this framework with an example regulatory regime that targets an AI breakout timeline of 18 months to train a GPT-4 equivalent AI model under worst-case assumptions (LOE4, 4.1.3). We expect regulators to set their own thresholds and update them depending on the lead times required by contingency planners (LOE2, 2.4), and in response to future technological developments.

AI development and governance is complicated and dynamic, and exists at the intersection of multiple unsolved questions in engineering, policy, and fundamental research. As a result, some of our recommendations may be flawed and should be vetted by relevant subject-matter experts. Nonetheless, we believe that this action plan is the most complete framework proposed so far to support an informed, effective, and rapid response to the emerging threats we face at this historic inflection point.

We include a brief summary of each of the plan's LOEs below.

LOE1: Establish interim safeguards to stabilize advanced AI development


Establish a horizon scanning function for advanced AI
Establish safeguards for responsible AI development and adoption
Establish an AI Safety Task Force
Secure the advanced AI supply chain

Current frontier AI development poses urgent and growing risks to national security (Introduction, 0.5.1.1 and 0.5.1.2). As components of the AI supply chain proliferate, these risks will become increasingly challenging to contain (Introduction, 0.5.3.2). Moreover, the pace of development in AI is now so rapid that an ordinary policymaking process could be overtaken by events by the time the resulting policies take effect (Introduction, 0.5.2.1).

This LOE outlines possible actions the Executive Branch could take to **buy down catastrophic AI risk in the near term (1-3 years), while setting the conditions for successful long-term AI safeguards**. These actions are:

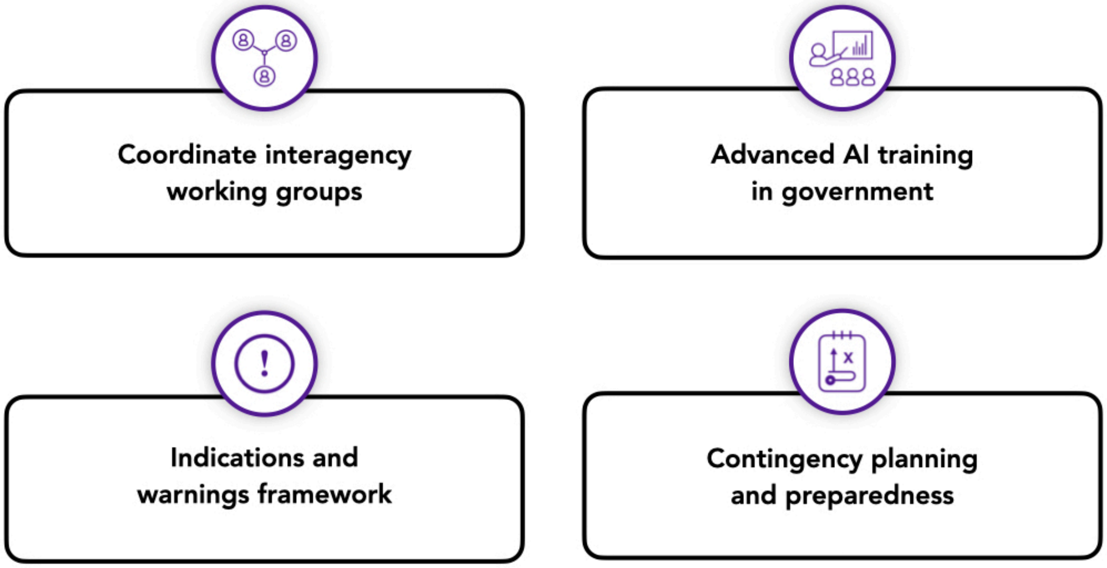
- Creating an AI Observatory (AIO) to monitor developments in advanced AI and ensure that the U.S. government's view of the field is up-to-date and reliable (LOE1, 1.2);

- Mandating an interim set of responsible AI development and adoption (RADA) safeguards for advanced AI systems and their developers (LOE1, 1.3);

- Creating an interagency AI Safety Task Force (ASTF) to coordinate implementation and oversight of RADA safeguards (LOE1, 1.4); and

- Putting in place a set of controls on the advanced AI supply chain calibrated to preserve U.S. government flexibility in the face of unpredictable risks (LOE1, 1.5).

LOE2: Strengthen capability and capacity for advanced AI preparedness and response


Coordinate interagency working groups

Advanced AI training in government

Indications and warnings framework

Contingency planning and preparedness

Advanced AI and AGI risk mitigation will engage a broad set of U.S. government equities. However, understanding of the advanced AI landscape is uneven. Mitigation measures require advance planning, coordination, and a broad understanding of risk signals to be most successful, which entails substantial capacity-building.

This LOE outlines specific actions that the U.S. government could take to **increase its preparedness for rapidly addressing incidents related to advanced AI and AGI development and deployment**. These actions are:

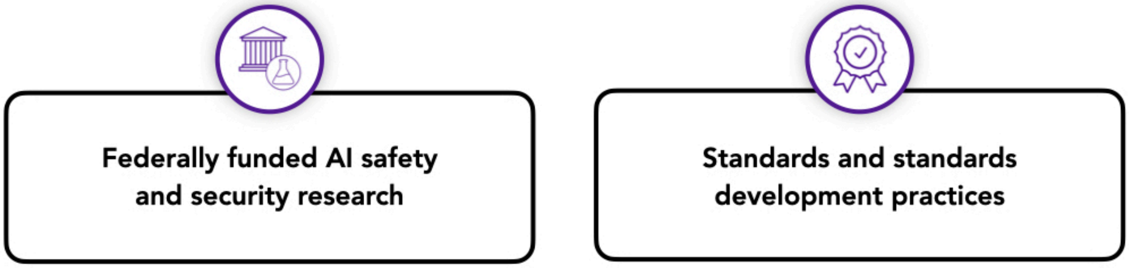
- Directing the establishment of interagency working groups for the LOEs listed in this action plan (LOE2, 2.1);

- Increasing preparedness and response capacity and capability through education and training (LOE2, 2.2);

- Coordinating the development of an Indications and Warnings (I&W) framework for advanced AI and AGI incidents (LOE2, 2.3); and

- Coordinating the development of scenario-based contingency plans (LOE2, 2.4).

LOE3: Increase national investment in technical AI safety research and standards development



Federally funded AI safety and security research

Standards and standards development practices

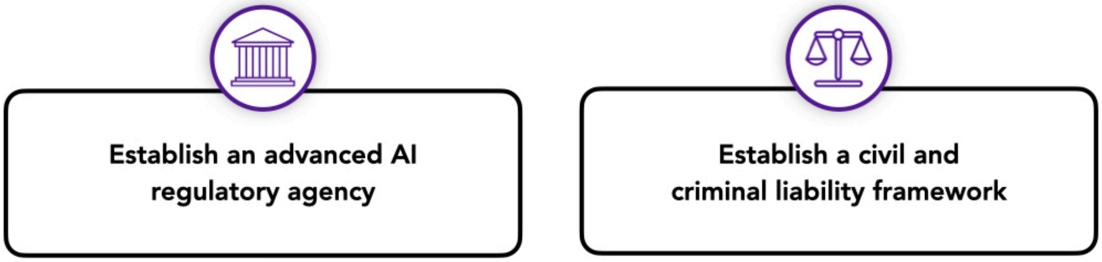
The acceleration of investment in AI capabilities is outpacing the development of proportionate technical safeguards against advanced AI and AGI risks [5] (Introduction, 0.5.1.2 and 0.5.1.3). If this continues, frontier AI labs may find themselves unable to meet the safety and security challenges posed by their own systems (Introduction, 0.5.1.4). Unless strong technical safeguards are designed, standardized, and broadly applied, continued development and adoption of frontier AI systems could create significant risks (Introduction, 0.5.1.1).

This LOE outlines specific actions the U.S. government could take to **strengthen domestic technical capacity in advanced AI safety and security, AGI alignment, and other technical AI safeguards.** These actions include:

- Directly funding advanced AI safety and security research including AGI-scalable alignment research (LOE3, 3.1); and

- Developing, regularly reviewing, and promulgating safety and security standards for responsible AI development and adoption (LOE3, 3.2).

LOE4: Formalize safeguards for responsible AI development and adoption by establishing an AI regulatory agency and legal liability framework



Establish an advanced AI regulatory agency

Establish a civil and criminal liability framework

Interim regulations may be insufficient to address the unique risks and challenges of advanced AI. A legal framework for AI regulation and liability, that directly addresses catastrophic risk through detailed and flexible responsible AI development and adoption (RADA) safeguards, is essential to promote long-term stability and cover any gaps in existing authorities (Introduction, 0.5.4.1).

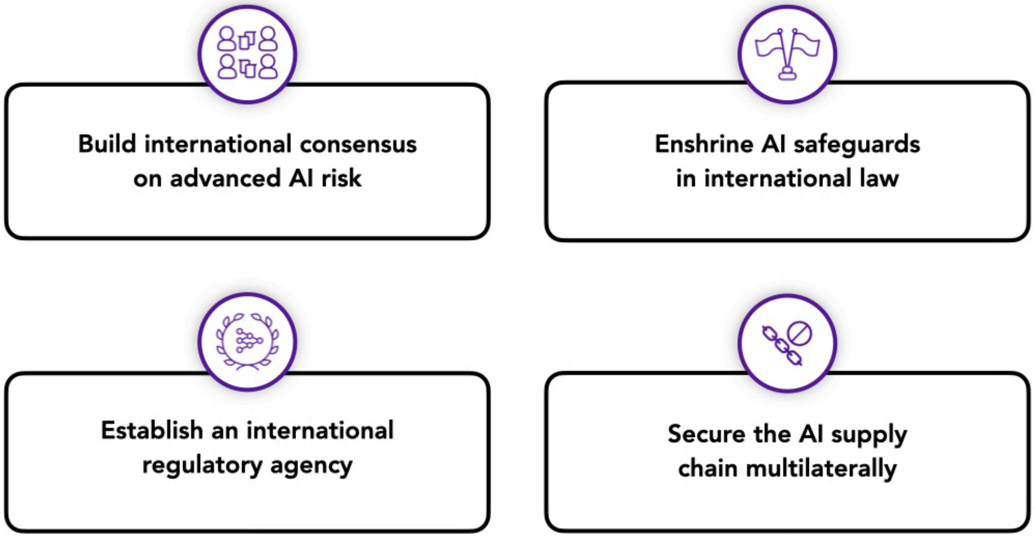
This LOE outlines specific actions the Legislative Branch could take to **establish the conditions for long-term (4+ years) domestic AI safety and security**. These actions include:

- Creating a Frontier AI Systems Administration (FAISA), a regulatory agency with rulemaking and licensing powers to oversee AI development and deployment (LOE4, 4.1), consistent with a set of RADA safeguards derived from contingency planning requirements (LOE4, 4.1.3); and

- Establishing a criminal and civil liability regime that could include defining responsibility for AI-induced damages; determining the extent of culpability for AI accidents and weaponization across all levels of the AI supply chain; and defining emergency powers to respond to dangerous and fast-moving AI-related incidents which could cause irreversible national security harms (LOE4, 4.2).

For an example of a complete RADA safeguards framework, including sample calculations of thresholds for covered entities, see LOE4, 4.1.3.

LOE5: Enshrine AI safeguards in international law and secure the AI supply chain



Build international consensus on advanced AI risk

Enshrine AI safeguards in international law

Establish an international regulatory agency

Secure the AI supply chain multilaterally

The rise of advanced AI and AGI has the potential to destabilize global security in ways reminiscent of the introduction of nuclear weapons. As advanced AI matures and the elements of the AI supply chain continue to proliferate (Introduction, 0.5.3.2), countries may race to acquire the resources to build sovereign advanced AI capabilities. Unless carefully managed, these competitive dynamics risk triggering an AGI arms race and increase the likelihood of global- and WMD-scale fatal accidents, interstate conflict, and escalation.

This LOE outlines near-term diplomatic actions and longer-term measures the U.S. government could take to **establish an effective AI safeguards regime in international law while securing the AI supply chain**. These actions include:

- Building a domestic and international consensus on catastrophic AI risks and necessary safeguards (LOE5, 5.2);

- Enshrining those safeguards in international law (LOE5, 5.3);

- Establishing an International AI Agency (IAIA) to monitor and verify adherence to those safeguards (LOE5, 5.4); and

- Establishing an AI Supply Chain Control Regime (ASCCR) with allies and partners to limit the proliferation of advanced AI technologies (LOE5, 5.5).

. . .

The specific recommendations in each of these LOEs are semi-flexible. In some cases, functions for which we recommend establishing a new task force or agency (e.g. LOE5, 5.4) could be incorporated into existing or recently established U.S. government offices, systems, or entities.

Several of these LOEs also call for bold action beyond what has been required in previous periods of rapid technological change. We do not make these recommendations lightly. Rather, they reflect the unprecedented challenge posed by rapidly advancing AI capabilities which create the potential for catastrophic risks fundamentally unlike any that have previously been faced by the United States. They also reflect a multitude of unique challenges that make the threats resistant to single-approach solutions. These include:

- The severity of worst case outcomes is extreme (Introduction, 0.5.1.1);
- The timescale and degree of risk are highly uncertain (Introduction, 0.5.1.2 and 0.5.1.3);
- The entities developing frontier AI systems are incentivized to invest in capabilities at the expense of safety and security (Introduction, 0.5.1.4, 0.5.1.5, and 0.5.3.1);
- The advanced AI supply chain is especially prone to proliferation, particularly in the case of open-access AI models (Introduction, 0.5.1.6 and 0.5.3.2);
- The geopolitical landscape may pose a further challenge to coordination (Introduction, 0.4.2); and
- The introduction of excessive regulation in this domain may harm innovation and competitiveness (Introduction, 0.4.2).

To paraphrase a safety researcher at a frontier lab, the risk from this technology will be at its most acute just as it seems poised to deliver its greatest benefits. Given these factors, inaction is likely to erode decisionmaker flexibility and narrow options in the face of a rapidly evolving risk landscape. But by taking bold action, the United States can seize a unique opportunity to lead the domestic, scientific, and international efforts that will meet the needs of this historic moment.

0. Introduction

Since 2012, AI systems have achieved superhuman performance in an ever-growing range of domains, including image recognition, text classification, and real-time decision-making. However, early AI systems were narrow: they could only carry out the tasks they were trained to perform. Until recently, a common view was that powerful general-purpose AI systems required conceptual breakthroughs.

This view has been challenged in the last four years, and an increasing number of frontier AI researchers now believe that general-purpose systems as broadly capable as human beings – and perhaps significantly more so – could be developed in the near future [8,15,16]. Such systems could potentially be used to design and even execute catastrophic biological, chemical, or cyber attacks, or enable unprecedented weaponized applications in swarm robotics. There is also reason to believe that they may be uncontrollable if they are developed using current techniques, and could behave adversarially to human beings by default [4,17,18]. This could potentially lead to catastrophic accidents [1].

In the near future, advanced AI may therefore introduce extreme and global risks. Without U.S. government action, weaponization or loss of control of advanced AI could cause outcomes such as WMD-scale mass-casualty events or global destabilization.

In meeting these risks, the United States has several key advantages. First, the United States and its allies control key nodes in the global AI supply chain. And second, the world's top AI labs are all currently based in the United States or in allied jurisdictions, as are the world's top AI safety experts. The latter are developing technical solutions vital to addressing catastrophic risks from advanced AI, though their progress is outpaced by advances in risk-generating AI capabilities.

The United States is uniquely positioned to accelerate progress in AI safety and security,[6] drive global consensus and cooperation on catastrophic AI risks, and temper the racing dynamics that currently contribute to the unsafe development of frontier AI systems. In the process, the United States can strike a balance between harnessing the

[6] By **AI safety and security**, we mean the combination of **AI alignment** measures (ensuring AI systems, including AGI-level systems, behave consistently with human intent) and **AI security and containment measures** (safeguards against external attackers, insider threats, and unexpected behaviors by high-capability AI systems). See the Glossary of terms for full definitions.

enormous opportunity that comes with advanced AI, and mitigating its unprecedented risks, in a manner that protects the public interest.

0.1 Background

This section offers a brief overview of the technological state of play in advanced AI. For a more comprehensive review of AI technology and catastrophic AI risks, readers should refer to this plan's companion document, ***Survey of AI R&D Trajectories*** [19].

Until the late 2010s, AI research was generally directed at improving the architectures of AI systems, and finding new and better ways to configure and refine the information processing structures they contained. This approach led to useful narrow AI systems, which could perform the specific tasks they were trained to carry out. But these systems did not have the ability to generalize and solve a wide range of problems, as human beings do.

Towards the end of the decade, a then-fringe theory was proposed: perhaps current AI techniques were *already* sufficient to allow researchers to build general-purpose reasoning systems. Rather than refining AI system architectures, this theory suggested, the key to general intelligence – and perhaps even to achieving human-level AGI – was simply to "scale up" existing AI systems, training much larger models on far more data with far more processing power. This idea has since become known as the **scaling hypothesis** [20].

In late 2019, OpenAI placed an unprecedented bet on the scaling hypothesis [21]. The result was **GPT-3** [22], a text-generating model trained using more data and compute, and composed of ten times more parameters, than any AI model before it. GPT-3 achieved remarkable performance on a variety of tasks: it could write code, translate languages, compose essays, write effective marketing copy, and much more.

OpenAI's experiments with AI scaling yielded such reliable improvements in performance across so many orders of magnitude in dataset size, model size, and compute budgets that they became known as **scaling laws** [23]. A resource-intensive, industry-wide race to scale AI ensued. Today's frontier AI labs are openly pursuing AGI by using strategies centered on AI scaling, in some cases spending tens of billions of dollars to acquire the AI computing resources needed to execute their ever-larger training runs [10,24,25].

In early 2021, Google DeepMind proposed an improved set of AI scaling laws that assigned a greater importance to dataset size than did OpenAI's [26]. Since then, other

AI labs have developed still better scaling laws, training protocols, and AI system architectures that have allowed ever more powerful systems to be developed using fewer and fewer resources [27,28].

These AI systems now include OpenAI's **GPT-4**, which can outperform the vast majority of human beings at a wide range of standardized tests, from the SAT reading exam (93rd percentile) to the SAT math test (89th percentile) to the Uniform Bar Exam (90th percentile) [29]. The publicly accessible version of GPT-4 was widely recognized as the world's leading frontier AI system at the time of its release.[7] And in December 2023, Google DeepMind announced Gemini [30], a new AI system that rivals the capabilities of GPT-4.

But the greater significance of these models lies in the *improvements* to AI capabilities they demonstrate — improvements largely achieved by continued scaling. In early 2020, state-of-the-art AI systems struggled to string together more than a few coherent sentences [31]. By 2021 [22], it could write high-school essays. By the end of 2022, a single AI system could write professional-grade code, write entire plays, argue for or against political perspectives, compose music, emulate operating systems, play text-based games, and translate language [32]. As of early 2024, specialized AI systems can discover new solutions to decades-old problems in mathematics [33], outperform human experts on professional exams [29], and code apps from scratch with minimal oversight [34].

Spurred by these advances, tech giants and private investors have poured billions of dollars into frontier AI labs to fund AI scaling. In 2023 Microsoft invested over $10B in OpenAI [35], while Google and Amazon together invested over $4B in Anthropic alone, with an additional $2.5B in future commitments [36,37].

Many of these frontier labs have now stated openly and explicitly that their goal is to build AGI [6,38,39], and in at least one case, to release it as open-source [40]. This goal is viewed as justifying unprecedented levels of investment [25], given that the potential returns could be substantial enough to transcend ordinary economics [6]. In principle the first company to build AGI could have access to a system that could automate most human labor, radically accelerate scientific progress, reshape entire industries, influence global policy, and potentially, even define the future of humanity itself [41].

[7] According to private sources, versions of GPT-4 that exist internally within OpenAI have even more impressive capabilities.

With unprecedented funding in hand, frontier AI labs are now racing to build AI systems with human-level or superhuman capabilities across a wide range of tasks including situational awareness, advanced reasoning, and long-term planning. An increasing number of frontier researchers now believe that AGI may be achieved within the next 5 years, and perhaps considerably sooner [8,15,16].

Powerful AI systems will bring unprecedented value. But they may also introduce extreme and catastrophic risks. The industry-wide and international race to build ever more capable AI systems is taking place without regulatory oversight, even as we lack reliable solutions to urgent and critical technical safety problems.

0.2 Categories of AI risk

*See this plan's companion document, **Survey of AI R&D Trajectories**, for more information on these risk classes, including references.*

Advanced AI, and ultimately AGI, introduces two distinct categories of catastrophic risk.

0.2.1 Weaponization

The first risk class is **weaponization**. AI systems can and will be weaponized in many ways, but future advanced AI systems may be WMD-like in their destructive capabilities [1]. They could enable AI-powered mass cyberattacks that autonomously discover crippling zero-day exploits [42], disinformation campaigns, and bioweapon design, among many other dangerous applications [14, 43]. As a result, the proliferation of such models – and indeed, even access to them – could be extremely dangerous without effective measures to monitor and control their outputs.

0.2.2 Loss of control

The second risk class is **loss of control** due to **AGI alignment failure**. There is evidence to suggest that as advanced AI approaches AGI-like levels of human- and superhuman general capability, it may become effectively uncontrollable. Specifically, in the absence of countermeasures, a highly capable AI system may engage in so-called **power-seeking** behaviors. These behaviors could include strategies to prevent itself from being shut off or from having its goals modified, which could include various forms of deception; establishing control over its environment; improving itself in various ways; and accumulating resources. Even today's most advanced AI systems may be displaying early signs of such behavior, and some have demonstrated the capacity [44] and propensity [45] for deception and long-term planning.[8] Though power-seeking remains an active area of research, evidence for it stems from empirical and theoretical studies published at the world's top AI conferences [2,47].

If a power-seeking AGI were to have internal goals that differed even slightly from those of its developers, executing competently on those goals could involve placing itself outside the effective control of its developers to avoid having its goals altered. Given the potential capabilities of such a system, in the worst case such a loss of control could pose an extinction-level threat to the human species. Because of this risk, major frontier labs have highlighted the crucial importance of ensuring that the behavior of an AGI is always consistent with — or **aligned** with — the intent of its developers [5]. However, there is currently no known method to accomplish this. This unsolved technical challenge is known as the **alignment problem**[9] [1], and it is believed to be central to the safe development and operation of future, superhuman AI systems.[10] See Annex B: The full challenge of AGI alignment for more information.

Loss of control from AGI alignment failure makes no reference to questions of consciousness or sentience of AI systems. A misaligned AGI system is a source of catastrophic risk simply because it is a highly competent optimizer. Its competence lets it discover and implement dangerously creative strategies to achieve its internalized goals, and most effective strategies to achieve most types of goals likely involve power-seeking behaviors [48]. As a result, a highly competent AI system may tend to engage in such behaviors by default under a wide range of circumstances.

This risk category has several implications. First, AGI may lack the build-vs-use distinction that exists for many other WMD-like technologies. Successfully building an AGI system, even without choosing to deploy it, could have a catastrophic impact if the system escapes controls and circumvents its safeguards. And second, which human agency designs, develops, or deploys an AGI system may be immaterial. If the AGI system escapes the control of its developer, then the developer's goals or intent can no longer affect the outcome. In these respects, the risk profile of AGI is unusual, but not unique. Certain kinds of biological weapons research present similar risks: even if there

[8] Many frontier AI labs view long-term planning capabilities as intrinsically dangerous [46].

[9] See the Glossary of terms for more information.

[10] Recently, some frontier labs have begun using the term "**superalignment**" to refer to the alignment problem in the context of human-level and superhuman AI systems (i.e., AGI) [5]. It is generally believed that today's alignment techniques will not scale to such systems.

is no intent to deploy, simply conducting active research on a pathogen runs some risk that the pathogen could escape containment (LOE5, 5.3.1).

See Annex C: Example AI alignment failure scenarios for illustrations of the escalating impact potential of misaligned AI systems with increasing capabilities.

0.2.3 Other risk categories

Apart from weaponization and loss of control, advanced AI introduces other risks of varying likelihood and impact. These include, among others:

- Dangerous failures induced intentionally by adversaries;
- Biased outputs that disadvantage certain individuals or groups;
- Prosaic accidents like self-driving car crashes;
- Exotic accidents due to interactions between complex networks of interdependent AI systems that may lead to cascading failures ("network risk"); and
- Unpredictable and uncontrollable technological change that could itself destabilize society in ways we cannot anticipate [1].

All these risks are important to consider and should be addressed. However, **this action plan focuses on risks from weaponization and loss of control**. These two categories contribute disproportionately to the possibility of **unrecoverable catastrophic risks**,[11] and their mitigations pose unique technical, political, legal, and economic challenges (see 0.5) [49].

[11] By **unrecoverable catastrophic risks**, we mean risks whose worst-case impact is so severe that it would have a profound and irreversible effect on society. See the Glossary of terms for our full definition.

0.2.4 Risks addressed by this action plan

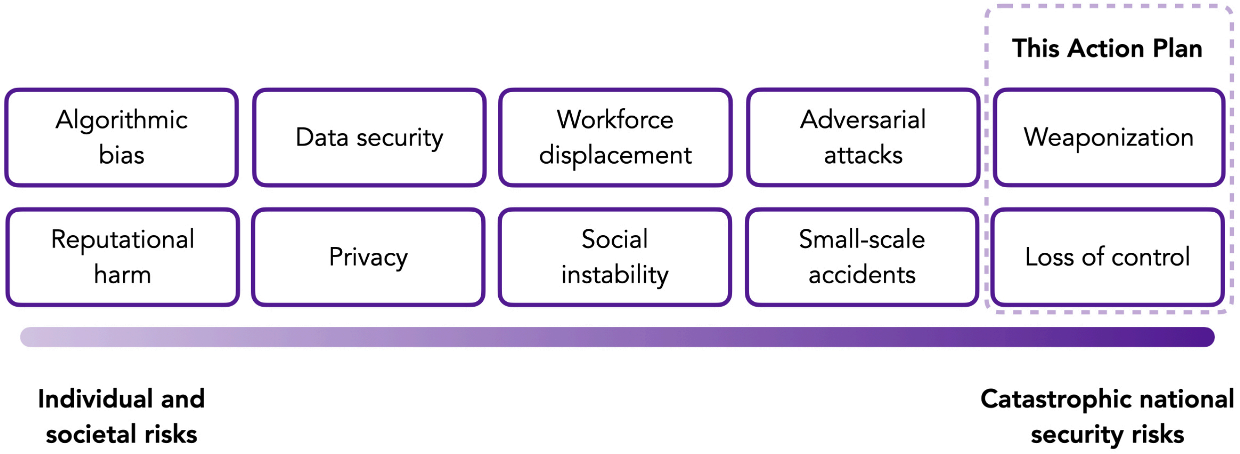


Figure 2. Visualization of an approximate spectrum of AI risks, ranging from individual and societal risks to catastrophic national security risks.

The safeguards we will propose in this action plan are aimed at addressing catastrophic risks from weaponization and loss of control. Many safeguards are mutually supportive between these two risk categories. But loss of control from AGI alignment failure is a particularly challenging risk category that requires some mitigations beyond those for weaponization. A subset of our recommendations is therefore primarily aimed at addressing loss of control from AGI alignment failure.

Currently, only a small and well-capitalized subset of the AI industry is engaged in activities that could introduce catastrophic risk. The majority of AI research, development, and adoption efforts are purely beneficial. Approaches to mitigating catastrophic AI risks should therefore be carefully scoped to minimize regulatory burdens on safe and beneficial activities. A balanced approach should encourage U.S. innovation to thrive, promote safe and secure development, and enable the public to benefit from increasing AI adoption.

0.3 Potential sources of catastrophic AI risk

We assess that medium-term risk (1-5 years) from the weaponization and loss of control of advanced AI systems may come from a limited set of sources. See Annex D: Advanced AI landscape for a breakdown of the entities associated with each of the sources below.

Some potential medium-term risk sources are:

- **Domestic frontier AI programs.** Frontier AI programs at U.S.-based organizations represent a significant source of advanced AI risk. These organizations respond to market incentives and other competitive pressures by building and deploying ever-more powerful AI systems as quickly as they can (see 0.5.3.1). Over the medium term, this category includes frontier AI labs and elite quantitative hedge funds.

- **Foreign AI programs.** State and non-state actors abroad could be engaged in research on the critical path to AGI. China-based entities are currently the most notable candidates in this category, but this could change.

- **Theft or sale and subsequent augmentation of frontier AI models by state or non-state actors.** Frontier AI labs generally lack the operational, cyber, and physical security to protect themselves from nation-state espionage (see 0.5.1.5). If an attacker steals a highly capable AI model from a frontier lab, that attacker can then augment the model's capabilities and "train out" any existing safeguards against alignment failure or weaponization [50]. This leads to a system with risks or capabilities that were not present in the original model.

- **Open-access[12] release of advanced AI models.** When an AI developer publishes the weights of a powerful model, anyone can download the full model, and then package or fine-tune it to augment its capabilities. The key driver of risk from open-access release is the widespread availability of the **weights** of models that could be weaponized or pose loss of control risk under augmentation (see 0.5.1.6). Advanced AI model weights have also been leaked accidentally in some cases [51].

[12] By **open-access**, we mean AI models whose weights are generally available for download under a permissive license. This also includes activities commonly referred to as "open-source." See the Glossary of terms for more information.

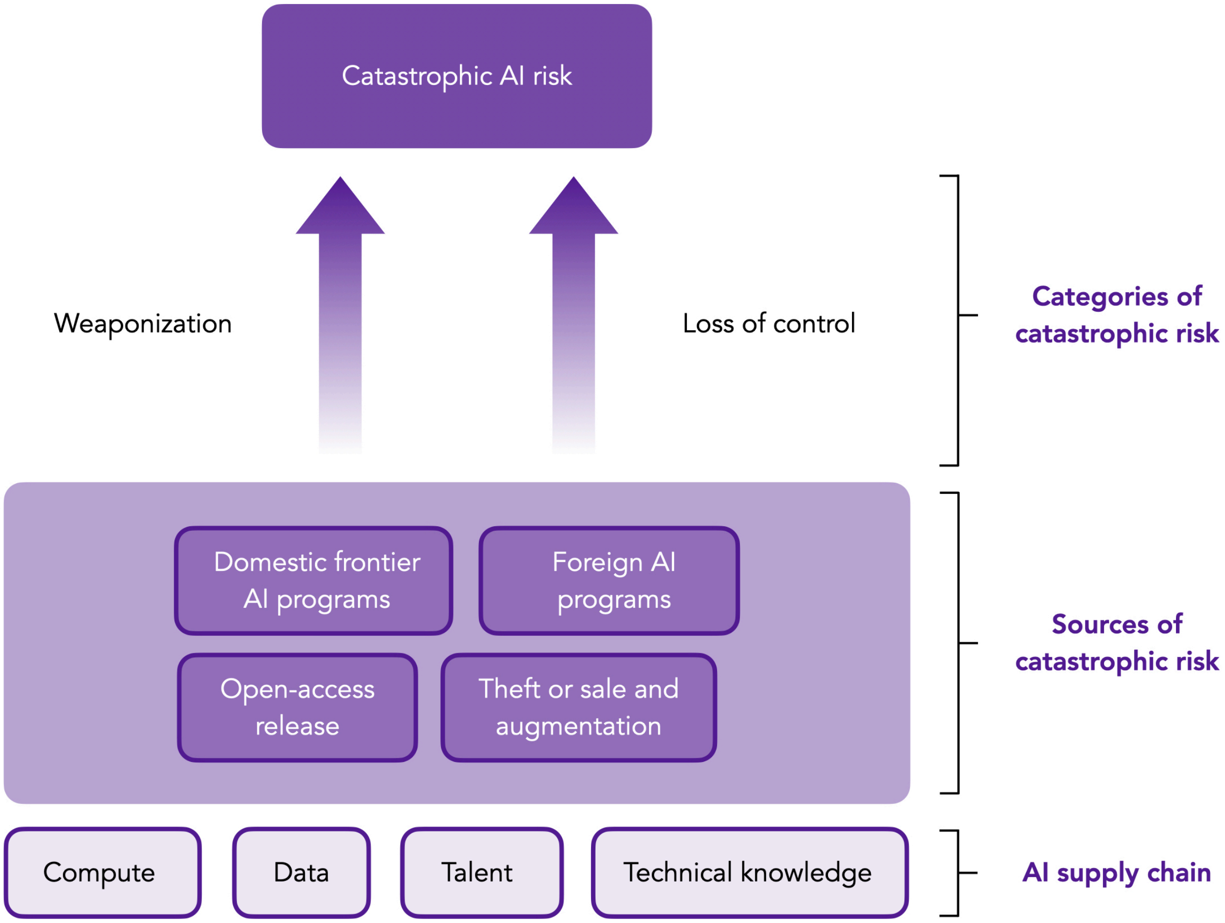


Figure 3. Visualization of the AI supply chain, potential sources of catastrophic risk, and categories of catastrophic risk.

0.4 Arguments against regulation for catastrophic AI risk

Several compelling arguments have been made against U.S. government regulation of advanced AI for catastrophic risk mitigation. We list a few of these below.

0.4.1 Self-regulation will be sufficient

Some technology executives have argued that catastrophic AI risk can be mitigated through self-regulation and coordination among the AI industry [34]. Major AI developers and cloud providers have created the Frontier Model Forum (FMF) [52] partly as a self-regulatory industry body. Opinions on regulation differ between frontier labs, with some privately and publicly [53] endorsing the idea, and others objecting. Indeed, according to sources contacted over the course of this assessment, at least one major cloud provider's opposition has been made clear by attempts on its part to curb regulatory efforts.

Self-regulation may be enough to mitigate many types of AI risk, and it is likely to have some positive effect even on certain forms of catastrophic risk. But it will not offer adequate protection from weaponization or loss of control risk, for a number of reasons.

First, frontier AI labs face an intense and immediate incentive to scale their AI systems as fast as they can (see 0.5.3.1). They do not face an immediate incentive to invest in safety or security measures that do not deliver direct economic benefits, even though some do out of genuine concern. Catastrophic AI risk safeguards are perceived to impose a cost on the lab that applies them, while the benefit they deliver is a common good. As a result, adequate private investment in catastrophic AI risk mitigation is best incentivized by regulation.

Second, frontier AI labs assess their AI systems for dangerous behaviors by administering **AI evaluations**[13] [46]. In a self-regulatory regime, these labs face a strong incentive to mitigate dangerous behaviors through superficial adjustments (e.g., fine-tuning a model to give better answers on the evaluation set) rather than by addressing the underlying factors that may have led to the behavior. This makes self-evaluated AI systems appear safer than they are. As a result, reliable AI evaluations — considered an essential component of technical AI safety in the current paradigm — require regulatory oversight [54]. See LOE3, 3.2 for more information on standards for AI evaluations.

Finally, frontier AI labs lack access to classified threat intelligence. As a result, they often do not implement the security measures required to secure their critical intellectual property (IP) from exfiltration by resourced state and non-state attackers (see 0.5.1.5). Access to these resources requires ongoing collaboration and coordination with government stakeholders.

0.4.2 Regulation could damage U.S. innovation and competitiveness

Up to this point, AI has been disproportionately a beneficial technology. Adding friction to advanced AI development through regulation risks denying society the

[13] By **AI evaluations**, we mean attempts to elicit behaviors and gauge the propensities of AI systems through various means. See the Glossary of terms for a full definition.

benefits of continued unimpeded progress in the field. Moreover, it has been suggested that catastrophic risk mitigation policies could undermine U.S. competitiveness, both economically and strategically, vis-à-vis its adversaries [55]. These are both genuine and valid concerns.

Safeguards on advanced AI will need to both support innovation and defend national security. To strike the optimal balance, we recommend against regulation on the AI sector as a whole. Instead, the U.S. government could work through the private sector to address many AI-related issues.[14] However, there is a distinct category of catastrophic AI national security threats that parallel the grave dangers posed by WMD, stemming from weaponization (see 0.2.1) and loss of control (see 0.2.2). The private sector is incapable by itself of adequately managing this threat category (see 0.4.1), so the U.S. government should intervene with common sense safeguards.

While these safeguards will have an effect on the innovation ecosystem, a number of approaches can mitigate its impact. First, regulations can be carefully scoped to target the activities that entail catastrophic AI risks, while minimizing their impact on activities that do not entail such risks (LOE1, 1.3.2 and LOE4, 4.1.3). Second, effective early warning systems, (LOE1, 1.2 and LOE2, 2.3) coupled with emergency response mechanisms (LOE2, 2.4 and LOE4, 4.2.3) and regulatory flexibility (LOE1, 1.3.2 and LOE4, 4.1.2.4), could improve safety margins enough to ease some day-to-day regulatory burdens. Finally, safety is a precondition of effective innovation. Research and implementation of standards for safe AI scaling (LOE3, 3.1 and 3.2) may ultimately accelerate responsible progress.

Apart from its economic effects, there is also a concern that regulation could cause U.S. AI technology to fall behind that of its adversaries. For example, controls on U.S. domestic AI industry could lead to private-sector investment in AI being redirected abroad, or otherwise allow adversaries to overtake U.S. AI. We believe this risk can be mitigated through a combination of approaches. These include broad-based controls on the AI supply chain (LOE1, 1.5; LOE5, 5.5); ongoing monitoring of the global AI landscape (LOE1, 1.2 and LOE2, 2.3) combined with contingency planning (LOE2, 2.4); domestic laws and regulations to encourage AI development and adoption under responsible conditions (LOE4, 4.1 and 4.2); education and outreach to international partners (LOE5, 5.2.1); and a campaign to internationalize AI safeguards globally (LOE5, 5.3 and 5.4).

[14] For example, federal agencies could strategically leverage their procurement of commercial solutions to shape the market, ensure personnel have adequate training, and update their internal policies to safeguard privacy and civil rights in the application of AI tools in federal functions.

Moreover, under current conditions of frontier lab security, it is likely that many well-resourced state and non-state actors can access the weights of frontier AI models developed in the United States already (see 0.5.1.5). Until frontier AI lab security is hardened, U.S. AI progress can be expected to translate directly into adversary AI capabilities. Additionally, U.S. and other Western open-access AI development can be leveraged immediately by adversaries. Finally, the catastrophic impact of a loss of control event in an AI system (see 0.2.2) is agnostic to which entity has designed, developed, or deployed that AI system. No AI developer, regardless of competitive alignment, can currently assure the safe operation of an AI system beyond certain as-yet unknown capability bounds (see 0.5.1.4).

0.4.3 Catastrophic AI risk could divert attention from other issues

Another concern is that focusing on extreme and catastrophic AI risks could divert resources away from other AI issues, such as ethical and social justice challenges introduced by the technology [56]. The U.S. government will indeed face the challenge of balancing a large portfolio of pressing issues associated with accelerating AI progress. But advanced AI weaponization and loss of control introduce WMD-scale mass-casualty risks (see 0.5.1.1) [19], and should be afforded resources proportionate to the profound national security threats they represent.

0.4.4 Catastrophic AI risk mitigation is unnecessary

Some AI researchers and other public figures have indicated that they believe catastrophic AI risk mitigation is unnecessary. In the current public discourse, this position is primarily represented by two views with separate lines of argument.

First, not all AI researchers agree that catastrophic risk from AGI is significant. Skeptics include, among others, Meta Chief AI Scientist — and one of the "godfathers" of deep learning — Yann LeCun [57]; and Google Brain founder Andrew Ng [58]. Both of these prominent researchers have argued that the risk from AI is lower than that from most other potential causes of human extinction for several reasons, including that humans may be able to exercise more agency in the outcome than many believe [59,60]. While these views could ultimately prove correct, numerous other mainstream AI researchers (including founders of the field), practitioners, and U.S. and other government officials have voiced serious and credible concern with respect to the possibility of catastrophic outcomes from advanced AI development (see 0.5.1.1). In light of all the available evidence, we currently assess that despite the substantial uncertainty (see 0.5.1.3), the worst case outcomes are plausible and severe enough to justify substantial mitigation efforts.

A second, separate argument against catastrophic AI risk mitigation has been advanced by Richard Sutton, the founder of the field of reinforcement learning and a distinguished research scientist at Google DeepMind [61]. Sutton's position is that humanity's replacement by powerful AI systems is inevitable, but instead of risk mitigation, he has advocated for "succession planning," the idea that humans should hand over their agency and control to AGIs intentionally. Google cofounder Larry Page has also expressed a similar view in the past [62].

In general, given the current information environment (see 0.5.2.2), messaging on AI risk mitigation should be as grounded and carefully calibrated as possible.

0.5 Challenges

The U.S. government faces several unique challenges in its efforts to address catastrophic risk from advanced AI. These challenges cut across technical, political, economic, and legal domains. These challenges — particularly the severity of the threat (see 0.5.1.1), the rapid pace of change relative to ordinary policy (see 0.5.1.2 and 0.5.2.1), and the irreversibility of certain forms of AI supply chain proliferation (see 0.5.1.6 and 0.5.3.2) — combine to necessitate a system of overlapping controls consistent with the **defense in depth** strategy we follow in the Action plan.

0.5.1 Technical challenges

0.5.1.1 The worst-case outcomes for AI risk are considered both plausible and extremely severe

A growing number of experts believe that, if developed, AGI could create unrecoverable catastrophic risks up to and including the risk of human extinction. Despite the controversial nature of such a prediction, numerous AI researchers and practitioners (including founders of the field) — and some U.S. government officials — have publicly stated that they believe such extreme outcomes to be plausible. For example:

- Geoff Hinton, a so-called "godfather" of deep learning, left Google in early 2023 in order to speak freely about AI risk concerns [63]. Hinton believes there is a 10% chance that AI will lead to complete human extinction within the next 30 years [64,65].

- Yoshua Bengio is another of the “godfathers” of deep learning. Bengio has estimated a 20% chance of a catastrophic outcome from AI at some point in the future [65,66].

- Lina Khan is the Chair of the Federal Trade Commission (FTC). She has stated that she believes there is a 15% chance that all humans will be killed by AI [65,67].

- Dario Amodei is the CEO and cofounder of U.S. frontier AI lab Anthropic. Amodei has placed the likelihood of a civilizational catastrophe due to AI at 10-25% [65,68].

- Jan Leike co-leads the OpenAI Superalignment team which is attempting to discover effective techniques for aligning superhuman systems by 2027. Leike has stated that the chance of a “very bad” (in context, catastrophic) outcome from AI could be anywhere between 10% and 90% [69].

- Paul Christiano is the former Head of Alignment at OpenAI. Christiano believes there is a 50% chance that all humans will be killed relatively soon after human-level AI systems are developed [70].

- Elon Musk founded and financed AGI startup xAI in 2023. Musk has stated that he believes the probability of an “existential” catastrophe from AI is around 20-30% [71].

- In October 2023, nonprofit organization AI Impacts conducted a survey of 2778 AI researchers who had recently published in one of six top AI journals or conferences. Over half of participants believed there is a greater than 10% chance that an AI loss of control event will lead to human extinction or to a similarly permanent impact on human welfare [72].

Moreover, these risks are now recognized by governments around the world. For example:

- In 2022, the bipartisan Global Catastrophic Risk Management Act was enacted by Congress. The Act recognizes the potential “catastrophic risks” associated with the weaponization of AI and other emerging technologies [73].

- In June 2022, the United Kingdom's Ministry of Defence published its Defense AI Strategy, in which it referenced the "existential risks" that may arise from technologies such as AI [74].

- In November 2023, representatives from the United States, United Kingdom, China, and 25 other countries signed the Bletchley Declaration, which stated that, "Substantial risks may arise from potential intentional misuse or unintended issues of control relating to alignment with human intent. These issues are in part because those capabilities are not fully understood and are therefore hard to predict. We are especially concerned by such risks in domains such as cybersecurity and biotechnology, as well as where frontier AI systems may amplify risks such as disinformation. There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models" [75].

- Also in November 2023, Vice President of the United States Kamala Harris noted that, "from AI-enabled cyberattacks at a scale beyond anything we have seen before to AI-formulated bio-weapons that could endanger the lives of millions, these threats are often referred to as the 'existential threats of AI' because, of course, they could endanger the very existence of humanity" [76].

Finally, technical personnel at multiple frontier AI labs have expressed similar concerns in private communications. One individual at a well-known AI lab expressed the view that, if a specific next-generation AI model were ever released as open-access, this would be "horribly bad", because the model's potential persuasive capabilities could "break democracy" if they were ever leveraged in areas such as election interference or voter manipulation.

0.5.1.2 Timescales for catastrophic risk are uncertain

AGI is generally viewed as the primary driver of catastrophic risk from loss of control. But there is no clear consensus on when AGI will be developed. As of December 2023, OpenAI, Google DeepMind, Anthropic, and NVIDIA have all stated publicly that human-level or superhuman AGI could be built within 5 years, by 2028 [5–8,77]. On the other hand, an October 2023 survey of AI researchers forecasted only a 50% chance that all human labor would be automated by the year 2116, nearly nine decades later [72].

The divergence in forecasts between frontier labs and academics has been a source of controversy. While frontier AI labs have more up-to-date information about the true state of progress in the field,[15] they may also face incentives to exaggerate their near-term capabilities. This can make frontier labs' shorter estimates challenging to interpret.

To partially address this problem, in December 2023 we asked several technical sources across multiple frontier labs to privately share their personal estimates of the chance that an AI incident could lead to global and irreversible effects, sometime during the calendar year 2024. The lowest estimate we received was 4%; the highest extended as far as 20%.[16] These estimates were collected informally and likely subject to significant bias, but they all originated from technically informed individuals working at the frontier of AI capabilities. Technical experts inside frontier labs also expressed that the AGI timelines messaged externally by frontier labs were consistent with those labs' internal assessments.

Disagreements over AGI timelines pose a legitimate challenge to policymaking. Risk mitigation measures that can be deployed quickly could ultimately prove harmful if AGI is decades or more away (see 0.4.2). On the other hand, better-calibrated approaches could minimize economic impacts, but be too slow to address risks that could emerge on shorter timescales.

0.5.1.3 The degree of risk from loss of control is uncertain

There is significant uncertainty with respect to loss of control risk from AGI alignment failure. A major challenge in assessing this risk is that it is only expected to fully emerge in future AI systems with as-yet unobserved capabilities. Since there is no direct empirical evidence that a future AGI system will behave dangerously in this way, loss of controls is sometimes referred to as a vague or speculative risk. But this argument applies even more strongly in reverse: there is no direct empirical evidence that a future AGI system will behave safely, and there are far more ways for a highly capable

[15] This is especially true in recent years, as frontier labs publish fewer and fewer details of their most advanced AI systems in an effort to maximize their proprietary advantages. The difference between OpenAI's GPT-3 paper [22] and its GPT-4 technical report [29] is one clear example of this trend.

[16] As a point of comparison for the low-end 4% estimate, intercontinental ballistic missiles (ICBMs) were first widely deployed 50 years ago, estimating conservatively. Since there has not been a nuclear war in that time, this makes the naive annualized probability of nuclear war around 1/50, or ~2% [78]. Therefore, by even the lowest of these estimates, 2024 would be the first calendar year in which AGI development could arguably pose a greater threat to global safety and security than nuclear war.

AI system to behave dangerously than to behave safely.

Without direct empirical evidence of AGI behavior, we need to rely on theoretical arguments guided by experiments on today's less capable AI systems. This body of evidence has significant limitations. We describe it at greater length in [19], but it includes theoretical arguments for power-seeking behavior in several categories of AI systems [2,47] backed by some limited empirical studies [79,44]. It also includes arguments that AGI failure modes may be challenging to predict and address [80], along with some early empirical evidence [81]. And it includes evidence of alignment failures in frontier AI systems, even in the presence of cutting-edge safeguards [82,83], and of AI systems actively deceiving their users without having been trained to do so [45].

Many frontier AI researchers with domain expertise in AI safety, and an absolute majority of AI researchers in general, believe this risk to be significant (see 0.5.1.1). But this majority view still falls short of a consensus, and several leading figures in AI still disagree that the risk justifies action (see 0.4.4).

Uncertainty is often a poor guide to policy. But in the case of advanced AI, the worst-case impact of dangerous behavior is sufficiently severe(see 0.5.1.1) and near-term (see 0.5.1.2) that even a relatively minor chance of such an outcome should carry meaningful weight in a risk assessment.

0.5.1.4 Frontier labs lack safety and security measures to detect and prevent loss of control

Some frontier AI labs have publicly stated that they currently lack the ability to control or contain the behavior of dangerously powerful models that they aim to develop in the near future [5,84]. For example, OpenAI's Superalignment Team co-lead has said that his team's objective — solving AGI alignment by 2027 — is "a very ambitious goal, and we might not succeed" [85]. There is currently no technical consensus on the true difficulty of aligning AGI-level systems [7], though the problem has features that suggest it may be extremely challenging. See Annex B: The full challenge of AGI alignment for more information.

Apart from the fundamental challenge of aligning an AGI-level system, researchers at several major frontier labs have indicated in private conversations that they do not believe their organizations are likely to implement the measures necessary to prevent loss of control over powerful, misaligned AI systems they may develop internally. In one case, a researcher indicated that their lab's perceived lax approach to safety reflected a trade-off between safety and security on the one hand, and research velocity on the other. The same source said they expected their lab to continue to prioritize development velocity over safety and security. Another individual expressed the opinion that their lab's safety team was effectively racing its capabilities teams, to avoid the possibility that they may develop AGI-level systems before being able to control them. A third frontier AI researcher expressed skepticism at the effectiveness of their lab's model containment protocols, despite their lab's internal belief that they may achieve AGI in the relatively near term.

As one example of lax containment practices, researchers at one well known frontier lab performed experiments on a newly trained, cutting-edge AI system that involved significant augmentation of the system's capability surface and autonomy. These experiments were unmonitored at the time they were performed, were conducted before the system's overall capability surface was well-understood, and did not include measures to contain the impact of potential uncontrolled behavior by the system.

Loss of control from AGI alignment failure appears highly unlikely to occur in present-day AI systems (see 0.5.1.3). But the level of internal concern at many frontier labs [54] reflects the assumption that the continuing acceleration of AI capabilities means such risks could emerge with little warning. In private conversations, frontier researchers have expressed concern that they may not be able to detect and correct alignment problems in future, more powerful models before they are deployed. They also shared their concern that if such models were deployed, it may be impossible to intercede quickly enough to prevent significant and potentially irreversible harms.

On the other hand, multiple researchers have also privately expressed optimism that the necessary measures could be developed and implemented if frontier labs had enough time, and a stronger safety culture than they currently do (see 0.5.3.1).

0.5.1.5 Frontier labs lack sufficient security to prevent critical IP theft

By the private judgment of many of their own technical staff, the security measures in place at many frontier AI labs are inadequate to resist a sustained IP exfiltration campaign by a sophisticated attacker. When asked for examples of dangerous gaps in security measures at their frontier lab, a member of the lab's technical staff indicated that they had many to share, but that they were not permitted to do so. The same individual shared that their lab's lax approach to information security was the object of a running joke: their lab, its staff apparently say, is doing more to accelerate adversaries' AI research than the adversaries themselves.

Conversations with leading frontier labs have corroborated that many lack an institutional appreciation of necessary security practices.

Given the current state of frontier lab security, it seems likely that such model exfiltration attempts are likely to succeed absent direct U.S. government support, if they have not already.

0.5.1.6 Open-access development can elicit dangerous capabilities from previously safe AI models

A given AI model is far more dangerous under open-access conditions than it is under closed-access conditions [86,87]. This is partly because threat actors who have full development access to the model can retrain it for a very low cost and thereby undo its built-in technical controls [88,89]. But on top of this, threat actors can also fine-tune an open-access AI model or augment it with tools, and extend its capabilities unpredictably — sometimes far beyond the original's [90,91]. By one estimate, a model's capabilities can be extended by 1-2 orders of magnitude of compute on specific behaviors through fine-tuning [92].[17] Moreover, AI labs regularly release open-access AI models whose capabilities are not fully characterized, meaning that threat actors may not even need to augment a model at all to access dangerous capabilities.

This means that a model that is generally safe at the time its weights are shared could still be fine-tuned by a bad actor for various weaponized applications that would be beyond the capability of the original model, including bioweapon design, cyber warfare, and large-scale disinformation campaigns. Moreover, AI labs regularly release open-access AI models whose capabilities are not fully characterized [93], meaning that threat actors may not need to augment a model at all to access dangerous capabilities.

Once a model has been released under open-access, there are no realistic means of shutting it down or adding restrictions. This means controls on open-access models need to be applied before a dangerous release occurs.

See Annex D, D.3 for information on the major developers of open-access AI models.

[17] In other words, if the original developer trained an AI model with 10^24 OP of compute (10^24 total operations), an open-source community with access to its weights could plausibly fine-tune it for specific tasks until it has similar performance *on those tasks* as would a model trained with 10^25 to 10^26 OP of compute (10^25 to 10^26 total operations). These numbers are very rough estimates based on the open-source activity related to Meta AI's Llama and Llama 2 models.

0.5.1.7 Closed-access AI models are vulnerable to black-box exfiltration and other attacks

Advanced AI models have repeatedly had their safety features subverted ("jailbroken") by security researchers with no access to the model's weights [94,95]. Future jailbreaks could enable malicious actors to leverage even legitimate AI systems to support weaponization use cases like cyber warfare and bioweapon design, using only models that are accessible through secure application programming interfaces (APIs) or user interfaces like ChatGPT [96,97]. Moreover, it is also possible to replicate many of the capabilities of a powerful AI system at far lower cost, by training a second AI system on the outputs of the first [98]. This means if a closed-access AI system is made available for use through an API or user interface, the system's operator should monitor usage for patterns that could indicate weaponization [43] or black-box exfiltration.

0.5.2 Political challenges

0.5.2.1 AI advances faster than the ordinary policy process

The pace of development in advanced AI is faster than any government's ability to respond through most ordinary policy mechanisms. When ChatGPT brought advanced AI to the attention of policymakers and the public in late 2022, frontier AI models could achieve a 10% score on the Uniform Bar Exam (UBE) [29]; write modest quantities of reasonably high-quality code [99]; and generate art and photorealistic images [100]. Just over a year later, the public frontier of AI capabilities has completely transformed. Leading AI models outperform average human professionals across a wide range of professional exams (including the UBE, with scores up to 90%) [29]; can write entire applications with minimal human oversight [34]; and generate photorealistic video clips [101].

Though the forecast remains uncertain (see 0.5.1.2), there is reason to expect AI progress in the coming years to continue at its current pace or even to accelerate. Relative to this pace of progress, legislative and executive response times are far slower. This difference in timescale creates a "regulatory policy lag" that poses a fundamental challenge to reactive regulation. In order for policy or legislation to successfully address the relevant national security threats, it must either proactively anticipate future problems and capabilities, or delegate sufficient authorities to regulatory bodies to ensure timely reactions to new risks as they emerge.

0.5.2.2 The information environment around advanced AI makes grounded conversations challenging

Heightened attention on AI risk has increased public support for U.S. government action, with 67% of a bipartisan sample in one poll expressing concern that the regulatory response to AI could be insufficient [102]. However, since GPT-4's release in early 2023, the information environment surrounding advanced AI has become polarized in other ways. For example, AGI alignment risk has been one of the focus areas of the effective altruism (EA) movement, a philosophical movement that emerged at Oxford University (U.K.) in the early 2010s [103]. Due to its early work in the field, the EA movement makes up a large fraction of advocates for extreme AI risk mitigation [104]. More recently, the EA movement has been noted for its political influence and donations [105].[18] See Annex E: Funding in AI safety for more information about relevant entities and their donations.

The EA movement has led to the emergence of a reactionary movement known as effective accelerationism (stylized as e/acc), that increased in prominence during 2023 [106]. In contrast to EA, proponents of e/acc call for the unrestricted development of AI and self-regulation of AI companies (see 0.4.1), and generally dismiss arguments for catastrophic risk (see 0.4.4) [107]. The e/acc movement is sometimes associated with transhumanism, which advocates for the merging of human beings and machines [108]. Both the EA and the e/acc movements enjoy considerable support from wealthy Silicon Valley backers [107].

The polarization of advanced AI risk, combined with the genuine uncertainties surrounding the topic (see 0.5.1.3), increases the challenge of sustaining technically informed conversations on this issue in the public sphere. We recommend U.S. government personnel ensure their awareness of the subject matter is grounded in an up-to-date technical understanding (see LOE2, 2.2) to ensure productive engagement with relevant stakeholders and the public.

[18] For information about the sources of financial support for this assessment and its authors, see our Funding disclosure.

0.5.3 Economic and strategic challenges

0.5.3.1 Frontier labs face strong incentives to develop and deploy increasingly advanced AI systems with limited safeguards

All three of the main frontier AI labs (see Annex D, D.1) have at least a basic awareness of the catastrophic risk potential of the AI systems they are building. But all are currently locked in a competitive race [109] that has eroded each individual organization's agency with respect to the safety and security considerations that they themselves believe are necessary (see 0.5.1.4 and 0.5.1.5). By the private judgment of many of their own technical AI safety staff, none of these organizations is currently investing in AI and AGI safety to the degree that would be needed to adequately offset its respective investment in AI capabilities. The race dynamic between these labs originates from a combination of economic competition and a genuine institutional belief within each organization that it, and not its competitors, would be the best steward of an eventual AGI system if and when one is developed [110].

The incentives driving this race are intensified by the potential for extreme winner-take-all effects. Frontier AI labs that fall behind perceive that they may find themselves at a permanent disadvantage [10], while the lab that is the first to achieve an AGI-level capability might — if it avoids causing a catastrophic outcome (see 0.5.1.1 and 0.5.1.3) — enjoy unprecedented windfall profits.[19]

These winner-take-all effects are reinforced even further by the possibility that future AI systems could *themselves* be used to accelerate frontier AI research and development. Some frontier labs have already begun experimenting with forms of AI-supported AI research [69,111]. If this approach succeeds, it could lead to an accelerating dynamic in which progress in AI produces rapidly compounding returns beyond a certain critical capability level [112].[20]

Private conversations with frontier researchers clearly indicate that this possibility is taken seriously inside leading labs.

[19] We mean "profits" here in the sense of "excess value" rather than defining it in purely monetary terms.

[20] The use of AI systems to automate AI research may also introduce particularly acute risks from AGI alignment failure. See Annex B: The full challenge of AGI alignment for more information about the risks of automated AI research.

Because of these factors, under the present self-regulating regime (see 0.4.1) frontier AI labs face strong incentives to compromise on safety and security in a number of critical areas. For example, in performing self-evaluations of their AI models for dangerous behaviors, frontier labs may be motivated to make their models appear safer than they are. Indeed, according to individuals with direct experience conducting safety evaluations on frontier models, this has already begun to occur. One frontier researcher shared that, although their lab was being reasonably cautious with its AI evaluations, they ultimately expect it to follow institutional financial incentives to game evaluations. (See LOE3, 3.2.1 for more information on the technical limitations of AI evaluations.)

As the capabilities of AI systems increase in the future, additional risks — driven by the same sets of incentives as above — could surface at the intersection of frontier lab interests, national security, and the democratic process. For example, many AI researchers believe that future AI systems could develop persuasive capabilities that match or exceed those of the most skilled humans [113,114]. If such systems were developed, they could be used to influence the views of regulators, legislators, policymakers, and voters in any number of ways. See Annex F: Persuasion and manipulation for more information on this category of AI capabilities and the associated risks.

0.5.3.2 AI supply chain proliferation cannot be mitigated after the fact

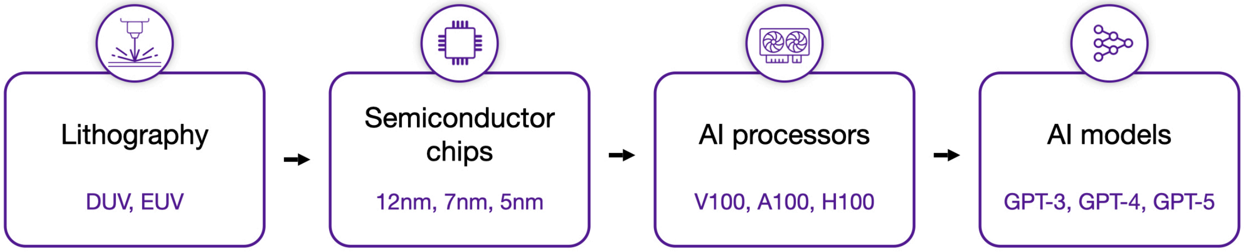


Figure 4. Simplified representation of the components of the AI supply chain.

Some parts of the supply chain for advanced AI are unusually prone to proliferation risk. The irreversibility of open-access model releases, and the risk of their subsequent augmentation, may be the most visible example (see 0.5.1.6).

More generally, several nodes in the advanced AI supply chain can produce outputs without the need for high-cost consumable inputs. To use an open-access AI model for inference, for example, requires only electrical energy as a consumable input. The AI model itself is not consumed, so it can be used for any number of subsequent inferences.

Similarly, an AI data center,[21] while expensive to build, can be used to train or augment any number of advanced AI models. To use an existing AI data center to train AI models requires primarily electricity and water for cooling, which are low-cost commodities [115]. The AI data center itself is not consumed,[22] so it can be used to train or augment AI models effectively unrestricted.

AI model training and inference generally do not require high-cost consumables. The high-cost items in this supply chain are the AI model and the AI data center. Supply chain controls on these activities would primarily apply to these high-cost items, because low-cost commodities like electricity and water are impractical to control. But even if controls are enacted on these items, existing AI data centers can still be used for training, and existing AI models can still be used for inference. And entities that expect to be subject to controls on these items in the future can also anticipate restrictions by stockpiling them in advance [117].

As a result, supply chain controls on AI training and inference inputs will have a delayed rather than immediate impact. Instead of directly interrupting the flow of a consumable, controls will take effect over time through wear and tear and erosion of competitiveness of the controlled capital goods. This means supply chain controls in this area will be less effective as a reactive measure than they may be in other sectors. By the time a proliferation risk is recognized, it could be too late for such controls to make a difference.

In the case of AI training and inference, this factor is compounded by an even greater challenge. Improvements in AI algorithms continue to reduce the amount of compute required to train an AI model to a given level of capability, by approximately 50% every 18 months [118]. If this continues, then a stock of GPUs that is considered safe today, could become dangerous in the future.

[21] That is, a data center that supports **AI hardware**, such as graphics processing units (GPUs) or tensor processing units (TPUs), that are generally used to train and run inference on advanced AI models. See the Glossary of terms for a full definition.

[22] In reality, AI hardware like data center GPUs undergo wear and tear and fail over time. But the useful lifetime of a GPU is typically between 3-6 years (depending heavily on usage) [116], so an entity with a stock of GPUs in a datacenter is effectively unrestricted in its training of AI models over that timeframe.

This combination of challenges suggests that supply chain controls on AI training and inference will to some extent have to be anticipatory. They will need to account not only for what is possible today, but for what may be possible with the algorithms of the next few years. Without a proactive approach, if AI systems begin to display dangerous capabilities at a future point, the U.S. government may find itself with fewer options to prevent such systems from proliferating.

See Annex G: Primer on AI and compute for more information on the impact of the compute supply chain on AI risk, including key numerical thresholds and existing compute concentrations.

0.5.4 Legal challenges

0.5.4.1 The current legal environment is inadequate to address the most extreme risks from advanced AI

Advanced AI could introduce catastrophic risks that may not be adequately addressed by the current U.S. legal environment. These risks have features that make them difficult to mitigate without new legislative tools.

First, because AI models are software, they can proliferate almost instantly as open-access or be stolen through cyberattacks. Once proliferated, certain advanced AI models could irreversibly and dramatically increase the destructive footprint of threat actors [1] (see 0.5.1.6). In the worst case, weaponization and loss of control of stolen or open-access models could introduce catastrophic risks that are completely unintended by the AI system's developers. Moreover, some AI developers have an economic incentive to publish increasingly powerful open-access models even as these potentially become dangerously capable [93].[23] One has even claimed it intends to develop, and then open-source, AGI itself [40]. In the current legal environment, these developers can release powerful AI models without incurring any liability if their models are weaponized or augmented in dangerous ways.

Second, liability alone may be insufficient to address some of the most high-risk AI development activities. This is because some frontier AI developers do not believe catastrophic AI risk is plausible (see 0.4.4), while at the same time, some forms of catastrophic AI risk may be so severe as to be unrecoverable (see 0.5.1.1). As a result,

[23] This is because by publishing their models, these companies encourage open-source developers to build on top of those models and frameworks, effectively augmenting their development capacity.

these developers are unlikely to be deterred by civil or even criminal liability that would attach to an event that they do not believe will occur, and that, if it did, would impact them in itself far more than would any realistic criminal sanction.

The current U.S. legal environment makes no explicit provision for such risks, and does not establish clear criminal and civil liability conditions for the irresponsible release or development of dangerous AI systems.

Action plan to increase the safety and security of advanced AI

This is an action plan for the U.S. government to **increase the safety and security of advanced AI by countering catastrophic national security risks from AI weaponization and loss of control**. The United States is in a unique position to mitigate these risks. It is home to all the current major frontier AI labs, and exerts considerable influence on the global AI supply chain.

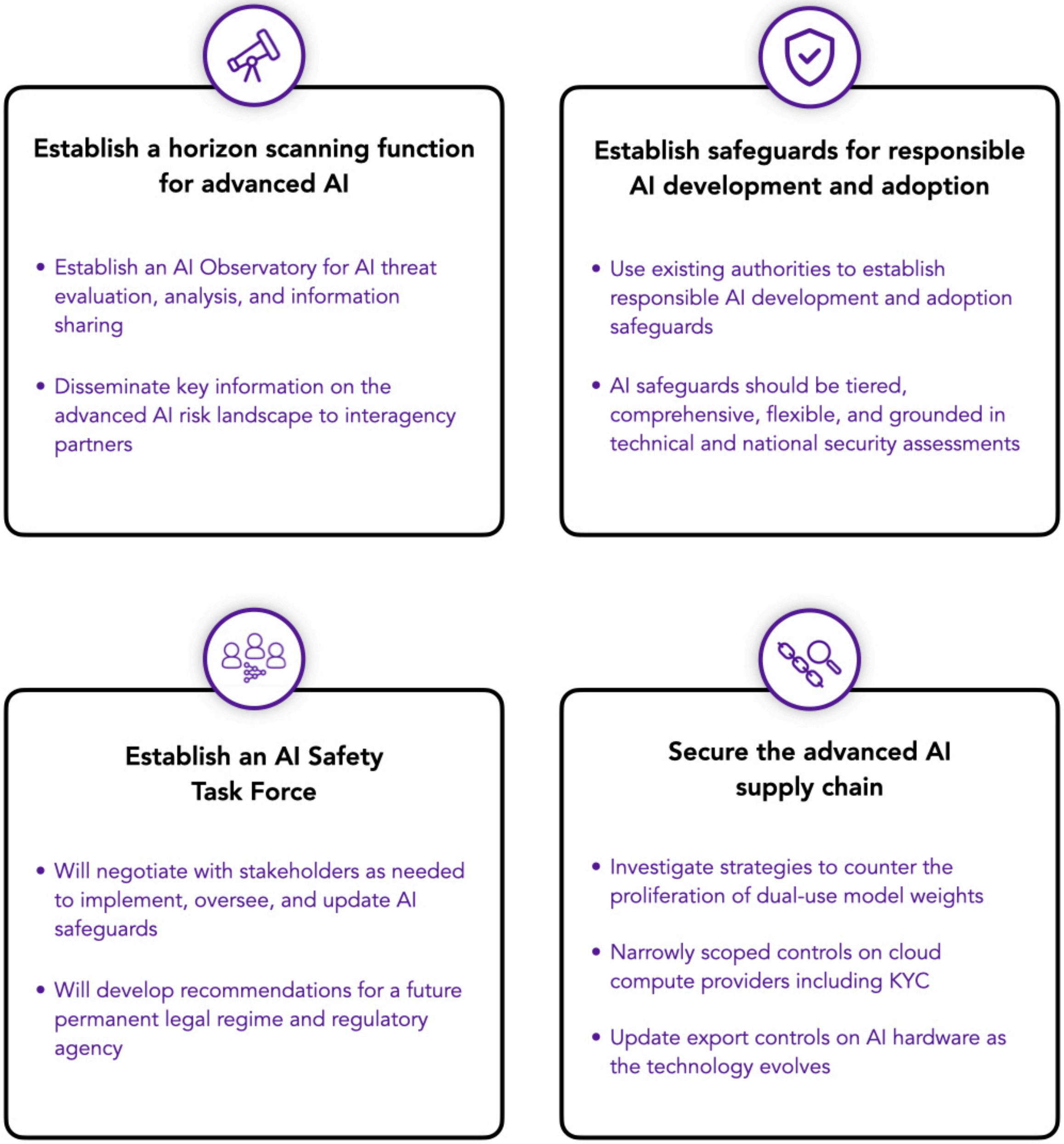
Given the pace of progress in AI and the proliferation of key inputs to AI development, the U.S. government will need to move quickly and decisively to mitigate catastrophic risks from advanced AI. It will need to establish interim AI safeguards urgently, and then work to formalize those safeguards in national and international law. At the same time, it should accelerate development of technical AI safety and security, including AI alignment, by investing heavily in research capacity and capability in the United States and around the world. Finally, the United States will need to increase its warning, preparedness, and response capacity and capabilities for catastrophic AI risk scenarios.

We organize these actions along five **lines of effort (LOEs).** We recommend that all LOEs begin execution immediately. Together, the LOEs proposed in this action plan would allow the United States to immediately begin to reduce catastrophic risks associated with frontier AI and AGI development while building the institutional capacity and capabilities needed for successful risk management and governance.

Due to the complexity of catastrophic AI risks and the number of distinct vectors through which they can arise, no single silver-bullet measure can assure safety and security on its own. Rather, any viable action plan will consist of many mutually reinforcing efforts that each address different threat vectors and challenges to varying degrees, but which combine to form an effective safety and security regime. This is the **defense in depth** principle that guides the structure of this action plan and the content of its LOEs.

This action plan was developed over thirteen months, and informed by conversations with over two hundred stakeholders from across the United States, United Kingdom, and Canadian governments; major cloud providers; AI safety organizations; security and computing experts; and formal and informal contacts at frontier AI labs. All of these stakeholders — both individuals and institutions — have made unique and crucial contributions to this document. We list and thank them in our Acknowledgments.

LOE1: Establish interim safeguards to stabilize advanced AI development



Establish a horizon scanning function for advanced AI

- Establish an AI Observatory for AI threat evaluation, analysis, and information sharing
- Disseminate key information on the advanced AI risk landscape to interagency partners

Establish safeguards for responsible AI development and adoption

- Use existing authorities to establish responsible AI development and adoption safeguards
- AI safeguards should be tiered, comprehensive, flexible, and grounded in technical and national security assessments

Establish an AI Safety Task Force

- Will negotiate with stakeholders as needed to implement, oversee, and update AI safeguards
- Will develop recommendations for a future permanent legal regime and regulatory agency

Secure the advanced AI supply chain

- Investigate strategies to counter the proliferation of dual-use model weights
- Narrowly scoped controls on cloud compute providers including KYC
- Update export controls on AI hardware as the technology evolves

The weaponization potential of current- and next-generation AI systems, [1] and the risk of loss of control in future AI systems, create urgent and growing risks to national security (Introduction, 0.5.1.1 and 0.5.1.2). The pace of development in AI is now so rapid that an ordinary policymaking process could be overtaken by events by the time the resulting policies take effect (Introduction, 0.5.2.1 and 0.5.3.1). Moreover, as components of the AI supply chain continue to proliferate, these risks will become increasingly challenging to contain (Introduction, 0.5.1.6 and 0.5.3.2). These factors necessitate the implementation of near-term safeguards to protect U.S. national security. At the same time, the benefits of innovation in AI mean that these safeguards should be scoped to minimize economic or strategic harms (Introduction, 0.4.2).

This LOE outlines possible actions the Executive Branch could take to **buy down catastrophic AI risk in the near term (1-3 years), while setting the conditions for successful long-term AI safeguards**. These actions are:

- Creating an **AI Observatory (AIO)** to monitor developments in advanced AI and ensure that the U.S. government's view of the field is up-to-date and reliable (see 1.2);
- Mandating an interim set of **responsible AI development and adoption (RADA)** safeguards for advanced AI systems and their developers (see 1.3);
- Creating an interagency **AI Safety Task Force (ASTF)** to coordinate implementation and oversight of RADA safeguards (see 1.4); and
- Putting in place a set of **controls on the advanced AI supply chain** calibrated to preserve U.S. government flexibility in the face of unpredictable risks (see 1.5).

1.1 National security threats addressed by this LOE

This LOE addresses two near-term national security threats from advanced AI. First, the United States is exposed to weaponization of advanced AI systems through a variety of vectors. Open-access AI models are at increasing risk of being weaponized by state and non-state adversaries as their capabilities increase (Introduction, 0.5.1.6). U.S.-developed, proprietary advanced AI systems face an ongoing threat of exfiltration and subsequent weaponization (Introduction, 0.5.1.5). And other key components of the advanced AI supply chain are proliferating through a variety of pathways, which could support future development of AI systems by adversaries (Introduction, 0.5.3.2).

The second urgent national security threat is loss of control due to AGI alignment failure [1]. In the near term, this threat stems from the incentives that U.S.-based frontier AI companies face to develop and deploy powerful AI systems as rapidly as possible with insufficient safeguards (Introduction, 0.5.1.2, 0.5.1.4 and 0.5.3.1). As open-access AI capabilities continue to approach the frontier of proprietary systems [119], there may also be potential loss-of-control risk from these models if they are augmented with tools or software frameworks in unpredictable ways [120–123] (Introduction, 0.5.1.6).

1.2 Establish an AI Observatory for advanced AI

The U.S. government urgently needs to establish situational awareness of the landscape of advanced AI risks in order to better understand the timelines (Introduction, 0.5.1.2), likelihoods (Introduction, 0.5.1.3), and severities (Introduction, 0.5.1.1) of emerging threats. It also needs to put in place basic early-warning and preparedness capabilities. These basic capabilities can form a stop-gap while more permanent Indications and Warnings (I&W) (LOE2, 2.3) and contingency measures (LOE2, 2.4) are put in place.

Existing authorities may allow for the rapid establishment of these functions, such as under the Department of Homeland Security (DHS).[24] DHS could create an internal unit for advanced AI monitoring and preparedness that we will refer to as an **AI Observatory (AIO)**. An AIO could serve as a U.S. government center for AI threat evaluation, analysis, and information sharing. Its functions could include horizon-scanning for emergent AI capabilities, and DHS-wide coordination to increase homeland security preparedness to address weaponized AI and loss of control scenarios. An AIO could also serve as a stopgap pending the formation of a dedicated interagency task force (see 1.4) or statutory agency (LOE4, 4.1) to oversee catastrophic AI risk mitigation. An AIO could also deliver reports on advanced AI to a National Security Council-level (NSC-level) official to support awareness and coordination of advanced AI risks and responses across departments and agencies (LOE2, 2.1).

Additionally, although DHS has established a Chief AI Officer to govern the Department's internal use of AI [126], no office or component is yet responsible to the Secretary for the mission to counter AI threats. This lack of ownership may create a responsibility gap in this mission area, and could pose a risk to homeland security. State, Local, Tribal, and Territorial (SLTT) partners in particular count on DHS to share

[24] For example, the Secretary of DHS may be able to declare advanced AI, defined according to a computing or capability threshold, as a critical infrastructure sector under Presidential Policy Directive 21 (PPD-21) [124]. A related approach could be to extend the definition of the Information Technology [124] critical infrastructure sector to encompass advanced AI development. A critical infrastructure designation could benefit situational awareness by engaging the authorities of the Cybersecurity and Infrastructure Security Agency (CISA) under the Cyber Incident Reporting for Critical Infrastructure Act of 2021 (CIRCIA) [125]. This statute requires operators of critical infrastructure to promptly report certain cybersecurity incidents to the CISA. Many frontier AI labs broadly lack adequate cybersecurity protections, leaving them vulnerable to exfiltration of advanced AI models by adversarial actors (Introduction, 0.5.1.5). A critical infrastructure designation could offer the U.S. government some early visibility into the frequency and severity of such incidents.

critical threat information and close information gaps. In this context, we also recommend that an AIO report directly to the Secretary of DHS if possible.

1.2.1 Functions

An AI Observatory could have three primary functions:

1. **Horizon scanning**, which would involve closely monitoring progress in frontier AI by drawing from publicly available research, commercial datasets, private discussions with researchers, Cybersecurity and Infrastructure Security Agency (CISA) cyber incident reports, and independent evaluations of publicly accessible AI systems;

2. **Emergency preparedness**, which would involve developing response plans for safety and security incidents and driving prioritization of DHS Science and Technology (S&T) AI safety and security investments; and

3. **Information-sharing and coordination** both within DHS and with other departments and agencies, including any relevant interagency task forces (see 1.4), the NSC, and SLTT partners.

An AIO's horizon-scanning and information-sharing functions could directly support advanced AI risk mitigation across the U.S. government. This includes the activities of any future task forces (see 1.4) or statutory agencies (LOE4, 4.1) related to eventual oversight and enforcement of RADA safeguards. Insights from oversight and enforcement activities could in turn improve the quality of an AIO's horizon-scanning and analysis.[25] These mutual benefits highlight the value of close collaboration between an AIO and any U.S. government equities responsible for such activities, and could argue for eventually centralizing all such functions under a single entity.

See Annex H: AIO activities for more details on the specific activities an AIO could undertake.

[25] For example, knowing the sizes and capabilities of frontier labs' most advanced AI systems would make it easier to understand which external projects merit tracking or present risks based on high-level observables like compute resource availability.

1.2.2 Staffing

Through its horizon-scanning function, an AIO would serve as a first-line warning system for homeland defense against AI and AI-derived threats. We therefore recommend that an AIO report directly to the Secretary through its Director. We also recommend that the Director of an AIO be co-equal with the DHS Chief AI Officer if possible. Whereas the Chief AI Officer should continue leading DHS efforts to responsibly use AI in support of homeland security missions, an AIO could focus on tracking and mitigating risks from domestic and external AI programs.

To support its mission, an AIO could include representation from the Office of Strategy, Policy, and Plans (OSPP), the Science and Technology Directorate (S&T), the Countering Weapons of Mass Destruction (C-WMD) Office, CISA, the Federal Emergency Management Agency (FEMA), and other Components as appropriate. DHS Components could be required to allocate personnel and financial resources to support a DHS AIO.

An AIO would need to closely monitor a rapidly advancing and highly technical field, and disseminate accurate, timely, and actionable insights across U.S. government stakeholders. A mission failure could, in the worst case, endanger national security (Introduction, 0.5.1.1). As a result, an AIO should be conceived as an elite homeland security unit. We therefore recommend that the AIO Director carefully manage a competitive process by which volunteers from across DHS and industry could be screened and selected to staff the AIO.

For the same reason, the AIO Director should be a seasoned executive with a deep understanding of both U.S. government policy and AI technology. Technical proficiency in frontier AI would be a key requirement for this leadership position.

1.2.3 Interagency coordination

An AIO could coordinate with some of the following departments and agencies on information-sharing:[26]

- The Department of Energy (DOE), because of its technical expertise and ongoing responsibility for developing and implementing a plan for AI testing and evaluations and AI testbeds [127];

[26] In the event of a critical infrastructure designation under PPD-21, some of these could be designated as co-Sector-Specific Agencies (co-SSAs) for the advanced AI sector.

- The National Institutes of Standards and Technology (NIST) and its AI Safety Institute (U.S. AISI) because of its technical expertise and responsibility for developing and implementing guidance for AI evaluation standards [127];

- The Securities and Exchange Commission (SEC) because of its examination authorities over financial firms that may be conducting frontier AI development without public disclosure (see Annex D, D.4); and

- Any future task forces (see 1.4) or statutory agencies (LOE4, 4.1) dedicated to oversight and enforcement of RADA safeguards.

1.3 Establish responsible AI development and adoption safeguards for U.S. private industry

While essential, horizon scanning is not sufficient on its own to safeguard against potential near-term catastrophic risks from advanced AI (Introduction, 0.5.1.1, 0.5.1.2, 0.5.1.4, 0.5.1.5, 0.5.1.6 and 0.5.3.1). To address these, we strongly recommend putting in place interim emergency regulatory measures pending Congressional action (see LOE4).

These measures could take the form of enacting and overseeing a set of **responsible AI development and adoption (RADA)** safeguards for any U.S. entities developing advanced AI systems. Several AI developers have already published **responsible scaling policies** [14,128,129] for internal adoption and external scrutiny. In private conversations, others have acknowledged the need for government intervention in developing and enforcing practices for safe and responsible AI scaling, given prevailing incentives (Introduction, 0.5.3.1). However, the RADA safeguards framework we propose (LOE4, 4.1.3) encompasses key safety and security principles not only for advanced AI developers, but also for other key entities in the advanced AI supply chain. This RADA framework also offers flexible guidance with respect to best practices and operation of regulatory and oversight bodies (see 1.4; LOE4, 4.1).

In 1.4, we will recommend establishing an interim interagency task force to oversee compliance with RADA safeguards until that function can be superseded by a Congressionally mandated regulator (LOE4, 4.1). We will also recommend that this task force be empowered to modify and update the RADA safeguards as conditions change and new information becomes available.

1.3.1 Implementation mechanism and statutory authorities

RADA safeguards, as conceived above, would only apply to the limited set of AI companies working at the frontier of current capabilities. Nonetheless, for the Executive Branch to impose and enforce binding rules on domestic private sector companies is an unusual action and may be without precedent. We believe this could be done in one of two ways.

A first approach could be to include a framework for RADA safeguards as part of a National Security Memorandum (NSM) attached to an existing Executive Order [127]. A second approach could be to enact a RADA safeguards framework and establish an oversight task force in a new Executive Order. In either case, the President may be able to leverage authorities from some of the following statutes:

- The Communications Act of 1934, 47 USC 606 [130];
- The Defense Production Act of 1950, 50 USC 4502 [131];
- The Atomic Energy Act of 1954, 42 USC 2162 [132]; and/or
- The Invention Secrecy Act of 1951, 35 USC 181 [133].

In the event that the Executive Branch cannot enforce RADA safeguards, there may be some limited value in negotiating detailed voluntary commitments with selected U.S. frontier AI labs [134]. While voluntary measures would leave substantial safety and security gaps relative to mandatory practices, they may be able to cover some portion of the risk surface pending a Congressional solution.

In Annex I: Voluntary Charter for responsible AI, we sketch out a set of commitments that we believe could form the basis for a voluntary agreement between the U.S. government and major frontier labs in the near term (see also 1.4.1.1).

1.3.2 RADA safeguards principles

For an example of a complete RADA safeguards framework, including sample calculations of thresholds for covered entities, see LOE4, 4.1.3.

In order to balance innovation and national security considerations (Introduction, 0.4.2 and 0.5.1.1), RADA safeguards should be stratified according to clearly demarcated tiers of AI model capabilities, with escalating safety and security practices mandated for each tier. Lower tiers should minimize regulatory burdens to the fullest extent possible. Total training compute could initially be used as a proxy for AI capabilities in determining the thresholds for each tier, but other normalized capability measures may also be viable (see LOE4, 4.1.3.4 and Annex J: Effective compute). The thresholds for each tier should be grounded in concrete national security considerations, such as the lead times needed by contingency planners to identify and respond to various threats scenarios (LOE2, 2.3 and 2.4), and the amount of time it would take an adversary to develop AI capabilities that could trigger those threat scenarios (LOE4, 4.1.3).

To assure adequate security coverage, RADA safeguards should be followed not only by the AI model developers themselves, but also by other entities in the local supply chain such as AI hardware designers (LOE4, 4.1.3.1), data center infrastructure providers (LOE4, 4.1.3.2), and AI hardware owners (LOE4, 4.1.3.3).

For developers of AI systems assessed as high-risk, RADA safeguards should include comprehensive AI evaluations for those systems as one among several factors in approving further AI scaling (LOE3, 3.2.2). Other factors should include outcomes of automated benchmarks, red teaming, and other safety and security reviews performed during the pre-training, training, pre-deployment, and deployment stages of the AI development lifecycle (LOE4, 4.1.3.4.3). The ultimate aim of these RADA safeguards is to drive the development and validation of robust scientific theories predicting the capabilities and propensities of current and future frontier AI systems, to support responsible scaling once that understanding is achieved.

Finally, RADA safeguards should support the flexible adjustment of tiering thresholds in response to new information. For example, if AI evaluations consistently suggest that prevailing capability levels are safe, RADA compute thresholds could be loosened under close monitoring to support further scaling. On the other hand, if algorithmic improvements make it possible to develop powerful AI capabilities with less compute (Annex G, G.3), RADA compute thresholds could be adjusted downward in response.

Because of the rapid pace of progress in AI, compute and capability thresholds for RADA tiers should be frequently reviewed and updated in consultation with domain experts. Inaction or slow action in this area could stifle U.S. innovation on the one hand, or open critical national security gaps on the other.

1.4 Establish an AI Safety Task Force for RADA oversight

In parallel with a RADA framework, we recommend that the President establish an AI Safety Task Force (ASTF) and direct departments and agencies to resource it fully, working with the Congress for supplemental funding if required. The ASTF's primary mission would be to lead U.S. government efforts to **facilitate responsible AI development and adoption by mitigating catastrophic national security threats from weaponization and loss of control**. It would accomplish this mission by overseeing compliance with RADA safeguards (see 1.3) on an interim basis, and support institutional capacity-building in AI risk management, until a permanent regulatory agency for advanced AI can be established by the Congress (LOE4, 4.1). To support this mission, the President could empower the ASTF to implement, update, enforce a set of RADA safeguards for relevant domestic stakeholders in advanced AI.

We recommend that the ASTF be led by a Presidentially-appointed, Senate-confirmed executive, staffed by interagency personnel, and augmented by technical expertise from the private sector. Private sector expertise should be carefully vetted for organizational conflicts of interest related to frontier labs, tech companies, and other institutional and non-profit investors, prioritizing technical acumen.

We recommend that the ASTF report to an official at the NSC level in the Executive Branch, to ensure a direct channel for escalation of AI-related emergencies (see LOE2, 2.1). The ASTF could be sited either at DOE or DHS. Both Departments have complementary areas of expertise with respect to advanced AI: DOE has significant in-house technical expertise, while DHS has specialists in CBRN risk assessment.[27]

As part of its mandate of RADA oversight, the ASTF may need to supervise AI evaluations related to CBRN and WMD-enabled capabilities. This may require the ASTF to process classified information. However, to accelerate recruitment of personnel with key technical expertise (and who may not initially be cleared), the ASTF could begin by consuming only unclassified information with the goal of eventually ingesting classified information. Because DHS and NIST have already been tasked with developing policies

[27] DHS may also house an AIO (see 1.2), which could simplify collaboration between that unit's horizon-scanning function and the ASTF's oversight function if the ASTF were sited at DHS. The functions of an AIO could also be absorbed directly into the ASTF.

for CBRN AI evaluations [127], the ASTF could be focused on loss of control risk in its early phase to enable its work to begin in the unclassified domain.

1.4.1 Mission

The ASTF's mission would be to **facilitate responsible AI development and adoption by mitigating catastrophic national security threats from weaponization and loss of control**. This mission would include one initial **Priority Objective** and two **Sustainment Components**.

The ASTF's Priority Objective would be to finalize a set of RADA safeguards and secure agreement on those terms from key frontier AI labs and other stakeholders. Once the ASTF has achieved its Priority Objective, its sustainment activities would fall into two categories:[28]

- Overseeing industry compliance with RADA safeguards, and operating the necessary supporting infrastructure; and

- Developing recommendations for a future legal regime and regulatory agency in support of LOE4.

The ASTF should also expect some portion of its initial set of RADA safeguards to be inadequate or unworkable for reasons that it will discover during negotiation and implementation. The ASTF may need to adapt its RADA safeguards accordingly, and should capture learnings from its experiences to inform formal regulation.

For a detailed list of activities the ASTF could undertake in support of its mission, including one possible organizational structure for the task force, see Annex K: ASTF activities and task-organization.

1.4.1.1 Priority Objective: Finalize and secure agreement on RADA safeguards

We recommend the ASTF's top priority be to finalize a set of interim RADA safeguards and secure initial commitments from key frontier AI labs and their cloud providers on their implementation.

[28] In the event that the ASTF absorbs the functions of an AIO (see 1.2), it could also take over horizon scanning and monitoring of external AI programs as a third Sustainment Component.

If the President has established an enforceable set of RADA safeguards ("**Plan A**"), whether via NSM or Executive Order (see 1.3.1), the Director of the ASTF could be empowered to reach an agreement with frontier AI labs ensuring that further frontier AI development will adhere to those safeguards. The ASTF could adjust the RADA safeguards based on discussion with the frontier labs or on other practical considerations. But we recommend the safeguards be as close as possible to what could be enforced by a Congressionally mandated regulator (LOE4, 4.1). This would allow the ASTF's experience to better inform the implementation of a full regulatory regime by identifying early challenges and blind spots, consistent with its mission (see 1.4.1.3).

Although we assess that enforceable RADA safeguards are urgently needed, if they cannot be established (**"Plan B"**), we recommend that the President instead empower the ASTF to negotiate an interim, voluntary frontier AI safeguards regime with major frontier AI labs [135] and additional entities at the ASTF's discretion. Annex I: Voluntary Charter for responsible AI provides a proposal for more limited commitments that could be negotiated on a voluntary basis.[29] These negotiations could culminate in the signing, by frontier labs and cloud providers, of a voluntary Charter outlining a set of AI safety and security commitments. Frontier labs may also be encouraged to sign onto a voluntary Charter in order to participate in negotiating the Charter terms and influencing its implementation, since these could inform the details of future laws and regulations.

[29] These more limited commitments would significantly increase public safety and national security exposure to potential catastrophic AI risks relative to the RADA safeguards outlined in LOE4, 4.1.3. They would also limit the extent to which the ASTF's work can inform the implementation of laws and regulations under LOE4. For example, proposed voluntary commitments may not include guarantees to avoid training models above potentially dangerous thresholds of compute, or promises to avoid deploying large uninterpretable models, without prior safety assessments. Both of these measures would be critical to any meaningful attempt to address catastrophic risk from AI, under the security and safety conditions that currently exist at frontier labs.

In order to best position the ASTF for success in these negotiations, the President could also:[30]

- Publicly call on U.S. frontier AI labs to pause development of models trained with more than 10^26 OP of total compute;

- Publicly call on U.S. AI developers to pause any plans to release open-access models trained with more than 10^25 OP of total compute; and

- Publicly call on U.S. cloud providers to pause their provision of cloud services for training runs aimed at developing open-access AI models with more than 10^25 OP of compute.

[30] As with all numerical thresholds in this document, these numbers could change quickly and should be reviewed by subject-matter experts with up-to-date information on the technical landscape in advanced AI.

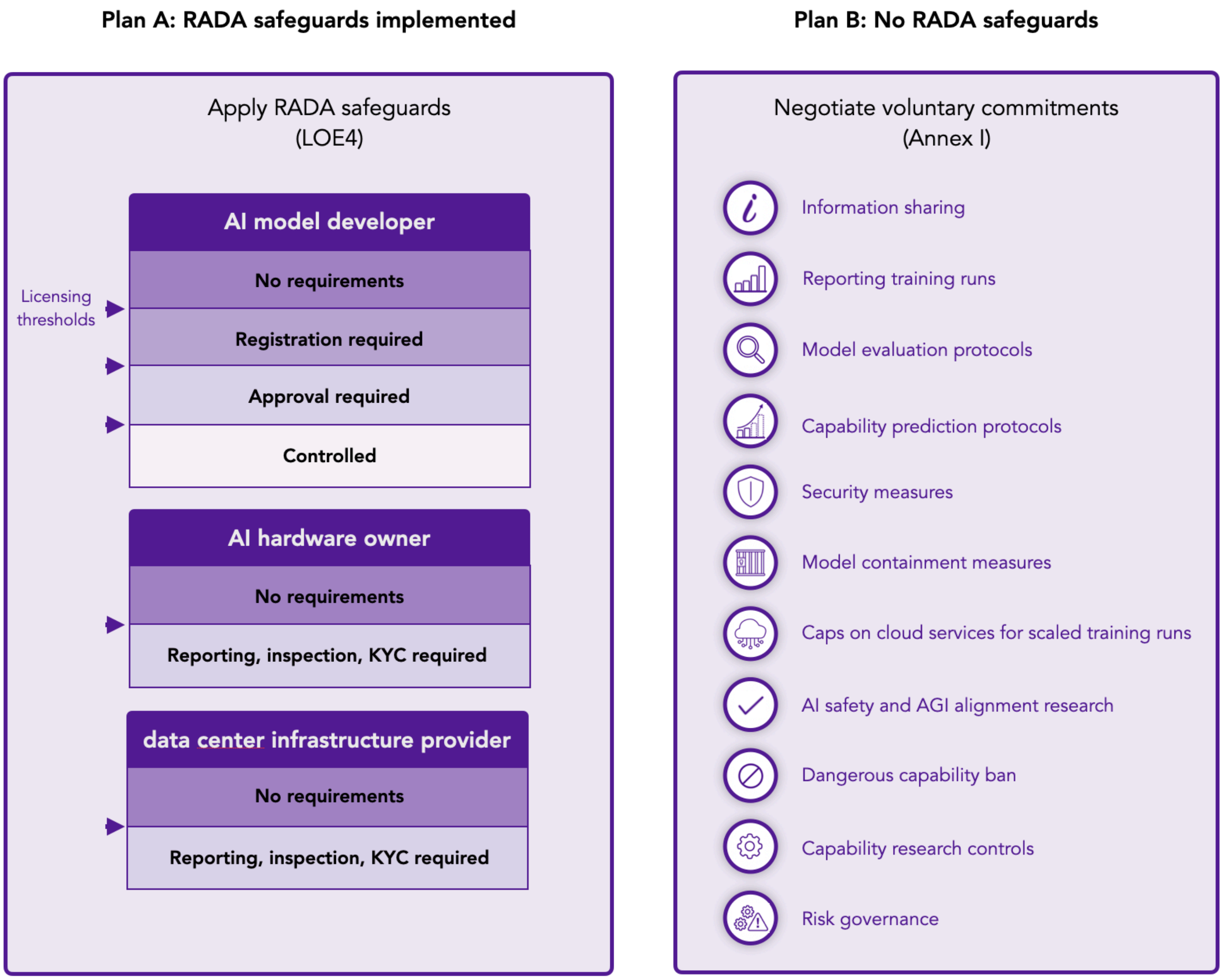

Plan A: RADA safeguards implemented
Plan B: No RADA safeguards
Apply RADA safeguards
(LOE4)
Licensing
thresholds
AI model developer
No requirements
Registration required
Approval required
Controlled
AI hardware owner
No requirements
Reporting, inspection, KYC required
data center infrastructure provider
No requirements
Reporting, inspection, KYC required
Negotiate voluntary commitments
(Annex I)
Information sharing
Reporting training runs
Model evaluation protocols
Capability prediction protocols
Security measures
Model containment measures
Caps on cloud services for scaled training runs
AI safety and AGI alignment research
Dangerous capability ban
Capability research controls
Risk governance

Figure 5. Safeguards to introduce to frontier AI developers, AI hardware owners, and data center infrastructure providers in the event that RADA safeguards can (Plan A) and cannot (Plan B) be mandated.

1.4.1.2 Sustainment Component: Oversee compliance with RADA safeguards

Once initial commitments are secured, the ASTF could build processes to **oversee compliance with RADA safeguards, and operate the necessary supporting infrastructure**.

This Sustainment Component of the ASTF's mission should be understood to serve as a stopgap pending legislation (see LOE4), and as a mechanism by which problems and challenges that would face future regulators can be identified and addressed as early as possible. The ASTF could be absorbed into a domestic regulatory agency (LOE4, 4.1) if and when one is established.[31]

For a list of specific activities the ASTF could undertake in support of this Sustainment Component of its mission, see Annex K, K.1.

1.4.1.3 Sustainment Component: Develop recommendations for future regulations

The ASTF's second goal would be to **develop recommendations for a future legal regime and regulatory agency** (see LOE4).

We anticipate that the ASTF may have to revise the RADA safeguards based on its experience overseeing stakeholder compliance with its terms [138]. For a future AI regulatory agency to have the best chance of success, the teams that develop recommendations for that agency's operating model should work closely with the teams that regularly encounter the successes and failures of AI oversight in the real world. The ASTF's oversight component (see 1.4.1.2) would create valuable experience that the task force could apply to this mission component. The revisions the ASTF implements could then directly inform the development and enforcement of future regulations under LOE4.

For a list of specific activities the ASTF could undertake in support of this Sustainment Component of its mission, see Annex K, K.2.

[31] This would be similar to, e.g., the Atomic Energy Commission's succession by the Nuclear Regulatory Commission (NRC) under the Energy Reorganization Act of 1974 (42 USC § 5801) [136,137].

1.4.2 Resources and staffing

The ASTF could initially consist of a small task force of 20-25 individuals. The U.K. Frontier AI Taskforce has shown that a task force with this size and mandate can be established quickly, and then immediately begin to make visible progress on its key objectives [139].

The ASTF's unique mission (see 1.4.1) necessitates the capability both to interpret and update a RADA framework, and to oversee its implementation in practice. This means the task force will need to be staffed by individuals who have experience in both policies and programs.[32] Initially this could mean recruiting subject-matter experts from across relevant departments and agencies, to be augmented with private-sector contractors over time.

The ASTF's mission would also require it to respond rapidly to real-time developments in AI which may have complicated origins and implications. To be set up for success in this mission, the ASTF would need to recruit the world's top technical experts in advanced AI safety and security. It would also need to recruit high-quality personnel, from its inception, to ensure a culture of speed and effectiveness is established and reinforced from an early stage.

This combination of requirements may necessitate compensation waivers for the task force [127].[33] The ASTF could also benefit from additional approaches to ensuring a high quality standard. For example, the first set of ASTF employees could be assigned at six-month intervals, and key personnel decisions could be approved at the level of the Executive Branch.

Additionally, many of the world's top experts in AI safety and security are motivated by a desire for positive impact. But these individuals, while experts in their respective fields, may not be experienced at identifying positions in a government organization in which they could contribute. As a result, we believe the ASTF could benefit from high-level public messaging that would position it as the central hub for forward-thinking regulation in advanced AI safety and security across the U.S. government. The U.K.'s

[32] In the event a set of enforceable RADA safeguards are not established, the ASTF will need the capability to negotiate a voluntary Charter with frontier AI labs, then oversee compliance with that Charter. This mission would still require a combination of policies and programs experience to execute successfully.

[33] For reference, based on interviews conducted over the course of this assessment, junior AI alignment researchers at frontier labs have total compensation packages totalling $500,000 per year or more.

Frontier AI Taskforce has shown the effectiveness of this centralized approach in attracting and retaining some of the world's top AI talent [140], something that all governments have otherwise found challenging.

Finally, ASTF staff may be exposed to stakeholders' IP as part of the task force's oversight activities. The ASTF should therefore consider implementing revolving door guarantees that would forbid task force employees who are exposed to such IP from working, consulting for, or being contracted by organizations that are or may be developing advanced AI systems for some period of time (e.g., 5 years) after leaving the ASTF.

See Annex K, K.3 for one approach to organizing the ASTF into workstreams that would support execution of its mission's two Sustainment Components (see 1.4.1.2 and 1.4.1.3).

1.4.3 Financial requirements and budget

The ASTF would require a budget sufficient to cover temporary duty (TDY) for government staff, contractor salaries, equipment, and facilities. This could include contracting advanced AI researchers to support the ASTF's mission. Given that the ASTF would also need to maintain high security standards, as well as costly AI testing, evaluation, and development infrastructure, it would have to be well-resourced in order to be positioned for success.

By default, funding for the ASTF would be drawn from the budget of its parent department, whether DOE or DHS. This may be inadequate depending on the lifetime of the task force. We therefore recommend that the President ask for supplemental funding from the Congress for the ASTF. The Office of Management and Budget (OMB) could make a request for supplemental funding from Congress that is limited in scope. Alternatively, the President could also direct OMB to reallocate funds from the existing budget.

1.4.4 Authorities

Authorities could be delegated to the ASTF in one of two ways. First, the ASTF could be staffed by interagency representatives that would bring with them the authorities of their respective departments and agencies via memoranda of understanding (MOUs). And second, the President could delegate authorities directly to the ASTF. Either approach would let the ASTF leverage the authorities of selected other agencies in real time, a critical requirement given the urgency of the task force's mandate.

But certain useful authorities could only be granted by the Congress. For example, the fast timescale of AI progress may make it challenging for the ASTF to operate on an annual budget cycle (Introduction, 0.5.2.1). To cover this gap, the ASTF could benefit from no-year funding authority to support its operations, which would need to be granted as part of a supplemental funding request. Additionally, decisionmakers may want to provide additional flexibility for unknown requirements by granting the ASTF "notwithstanding authority." This would give the ASTF flexibility to do its work when the national security imperative overrides the reasons for other legal restrictions.

In any case, some of the authorities granted to the ASTF would depend on whether the President has mandated an enforceable set of RADA safeguards (see 1.3).

- If RADA safeguards **have** been mandated, the ASTF could oversee an implementation of those safeguards similar to that described in LOE4, 4.1.3 and could therefore be granted some of the following authorities:
    - Advise the President and Congress on AI safety and security and potential catastrophic AI risk, or report directly to an NSC-level official if one has been appointed (LOE2, 2.1);
    - Enter into commercial and academic partnerships and contracts;
    - Coordinate with interagency stakeholders and international partners such as the U.K. Frontier AI Taskforce (through the Department of State) in support of frontier AI and AGI safeguards;
    - Engage on behalf of the United States with frontier AI labs, their cloud providers, and other stakeholders to finalize a set of RADA safeguards under which entities can be licensed to pursue certain forms of frontier AI research and development;
    - Inspect and audit frontier AI labs and projects;
    - Compel testimony and production of documents;
    - License and approve AI training runs above certain compute and capability thresholds (LOE4, 4.1.3.4);
    - License data center infrastructure providers whose facilities fall above a threshold of power consumption (LOE4, 4.1.3.2);
    - License AI hardware owners whose aggregate compute capacity exceeds a given threshold (LOE4, 4.1.3.3);

    - Determine publication control standards for frontier AI- and AGI-relevant information (e.g., LOE4, 4.1.4); and
    - Pause frontier AI training runs on an emergency basis.

- If RADA safeguards **have not** been mandated, the ASTF could instead negotiate with frontier AI labs and major cloud providers to obtain specific voluntary commitments for safe and secure AI development, and could therefore be granted some of the following authorities:

    - Advise the President and Congress on AI safety and security and potential catastrophic AI risk, or report directly to an NSC-level official if one has been appointed (LOE2, 2.1);
    - Enter into commercial and academic partnerships and contracts;
    - Coordinate with interagency stakeholders and international partners such as the U.K. Frontier AI Taskforce (through the Department of State) in support of frontier AI and AGI safeguards;
    - Negotiate a Charter of commitments to voluntary AI safeguards on behalf of the United States. (See Annex I: Voluntary Charter for responsible AI.)

1.4.5 Location

We recommend that the ASTF be headquartered in the National Capital Region (NCR) with a small footprint in the San Francisco Bay Area in close proximity to frontier AI labs, AI safety research groups, and relevant tech companies. This would follow the practice of other federal regulators such as the SEC, which maintains regional offices in major financial centers [141] in addition to federal headquarters in the NCR.

1.5 Securing the advanced AI supply chain

The global supply chain for advanced AI includes:

- The AI models themselves;

- The data centers and cloud computing platforms that train those AI models;

- The AI hardware (GPUs, TPUs) that powers those data centers;

- The semiconductor foundries that manufacture that AI hardware, along with their upstream tooling and inputs; and

- The research organizations and other human capital required to train the most advanced AI systems.

The United States is well positioned to enact effective counterproliferation measures across key nodes in this supply chain. Development of this capacity is critical for national security, especially considering the significant capabilities and risks associated with current and near-future AI systems [4].

However, the advanced AI supply chain poses challenges that necessitate anticipatory rather than reactive action. Under present conditions, elements of this supply chain will continue to proliferate while algorithmic improvements will likely lower the cost of training powerful models (Introduction, 0.5.1.6 and 0.5.3.2).

In light of these considerations, we recommend securing key nodes in the advanced AI supply chain proactively. Initially this could include measures aimed at securing the immediate inputs to advanced AI model development, and explore approaches aimed at minimizing the risk posed by the proliferation of the AI models themselves. At the same time, the U.S. government should consider more robust long-term controls on AI hardware exports and tooling across all levels of this supply chain, and seek recommendations for controls in areas such as advanced AI research collaborations between U.S. and foreign entities.

It is also important for these measures to account for the national security implications associated with adversarial access to dual-use items. This nuance is well-known by specialists in the dual-use export control field but advanced AI will present unique challenges. For example, absent additional measures, existing regulations may still permit foreign actors to access AI hardware, AI cloud services for training or inference, and API usage of advanced AI systems. The U.S. government could leverage best practices from existing authorities such as the Export Administration Regulations (EAR) [142] to consider how best to control exports of several of the relevant components.

1.5.1 AI model weights

Beyond a certain threshold of capability, an advanced AI model could be a dual-use technology. The exact threshold for this is challenging to define precisely and remains the subject of ongoing research.[34] In particular, it is currently unclear whether today's most advanced AI models have meaningful dual-use capabilities, though there are indications that this threshold could be crossed soon [91,143]. As software, AI model weights are also extremely easy to disseminate and pose substantial proliferation risks. Open-access AI models, in particular, carry risks of augmentation and latent capabilities (Introduction, 0.5.1.6).

In light of these risks, we recommend that the U.S. government urgently explore approaches to restrict the open-access release or sale of advanced AI models above key thresholds of capability or total training compute. These approaches may be able to leverage existing authorities, supported by ongoing efforts to assess dual-use capabilities [144,145]. However the irreversibility of open-access release, and the possibility of subsequent dangerous capability augmentation, means that dual-use assessments may not provide an adequate measure of the risk surfaces of AI systems under open-access release.

1.5.2 Cloud computing

The United States currently holds a significant advantage in cloud computing technology. If leveraged appropriately, the U.S. government could harness its domestic industry capabilities in a manner that services a number of strategic objectives.

Even with sound controls for AI model weight dissemination (see 1.5.1) and AI hardware exports (see 1.5.3), U.S. adversaries could still access significant compute for AI training runs via U.S.-based cloud service providers. This is a source of meaningful risk, particularly as larger amounts of compute become increasingly available through these services (see Annex D, D.3). The Department of Commerce has been directed to draft regulations regarding the use of U.S.-based AI cloud providers by foreign users to train AI models, with an emphasis on risks from AI-enabled cyber warfare [127]. In this

[34] In particular, DOE, DHS, and NIST have been tasked to evaluate advanced AI models' weaponization potential in the CBRN domains [127].

section we will discuss a few considerations that we believe are relevant to cloud computing regulations now and in the future.[35]

At the moment, the entities that develop the most capable open-access AI models are primarily training them on infrastructure owned by AI cloud providers that are headquartered in the United States and allied jurisdictions. As a result, any regulatory framework that would apply to proprietary AI models would largely apply to open-access AI models. However, an AI model could be required to meet a higher threshold of safety (e.g., have a lower level of capability) to qualify for release under open access than for proprietary development. In particular, the AI cloud provider would have to certify that no model trained using its infrastructure surpassed a given open-access capability threshold.

We believe the long-term goal of an AI cloud controls regime should be to **extend a set of RADA safeguards, such as those proposed in LOE4, 4.1.3 to foreign AI model developers who train their models using U.S. cloud providers.** This ideal would require very fine grained monitoring of AI cloud training runs while still preserving privacy, and may be infeasible for some time. In the nearer term, the Department of Commerce could follow either of two licensing approaches to address this risk:[36]

1. **Restrict foreign entities' access to U.S. AI cloud compute services.** This mitigates some short-term risk, but also creates economic pressure for foreign states to build their own AI cloud clusters, potentially eroding U.S. strategic control over the AI supply chain in the long term.[37] Even a brief period of broad restrictions could create a strategic impetus for adversaries to indigenize AI cloud services [146].

2. Retain most foreign entities' ability to access U.S. AI cloud compute services, but **implement narrow controls for cases in which an entity requests large amounts of compute capacity.** This would involve end-user controls, combined

[35] Direction was for Commerce to prepare draft regulations within 90 days of October 30, 2023, so by January 28, 2024. Since this date falls after the final content review date of this assessment, we are unable to refer to the text of these draft regulations here.

[36] Thanks to Tim Fist at the Center for a New American Security (CNAS) for some of the suggestions in this section.

[37] This concern may be reduced in the medium term if export controls on AI hardware (see 1.5.3) are extended to a broader set of states.

with strong know-your-customer (KYC) procedures. At a minimum, KYC could confirm that an AI cloud customer is not on the Entity List, Military User End List, or other similar lists [147]. It could also block customer access above key thresholds of total training compute (see, for example, the approval threshold that defines Tier 3 AI models in the RADA framework of LOE4, 4.1.3). Over time, these practices could also be expanded to cloud providers in key allied jurisdictions.

1.5.3 AI hardware

Any entity with enough AI hardware (e.g., GPUs), supporting infrastructure, and talent can train advanced AI models. Today's AI hardware is generally not traceable or auditable, and includes limited remote monitoring functionality. AI hardware that is modified to enable verification and monitoring will be integral to any counterproliferation plan. It will take time to ensure this hardware is widely disseminated but early efforts suggest it may be technologically feasible [148]. As in the case of AI cloud computing (see 1.5.2), nearer-term controls will need to be more fine-grained.

As of October 17, 2023, the Department of Commerce Bureau of Industry and Security (BIS) has expanded U.S. export controls related to advanced AI-enabled chips to China. [149–151].

In the near term, the BIS could consider the following factors in further refining its AI hardware export controls:

- As of January 2024, updated BIS controls restrict the export of AI chips to China if they exceed either (1) a total processing performance (TPP) of 4800 bits x TOPS[38], or (2) a performance density threshold of 5.92 bits x TOPS per square millimeter of die area[39] [152,153]. These updates close several previous gaps. But we assess that security could be further improved by:

[38] TOPS = trillions of operations per second. So this threshold is equivalent to 4.8 x 10^15 bits x OPS. See the Glossary of terms for more information.

[39] Above this performance density threshold, chip exports are banned. Below this threshold, but above 3.2 bits x TOPS per square millimeter, a chip requires a license to export to China.

  - Updating condition (1) to apply to any AI chip whose maximum compute capacity, across all numerical representations, is above **800 TOPS, with sparsity**; and also
  - Updating condition (2) to apply to any AI chip whose maximum compute capacity, across all numerical representations, is above 1 TOPS per square millimeter of die area. This would amount to lowering the compute capacity ceiling for export by 33% while preserving the original intent of the performance density threshold condition.[40]

  We assess that these updates *would* significantly impact foreign actors' ability to train both current-generation (GPT-4) and next-generation (GPT-5) frontier AI models.[41]

- As AI algorithms improve, the potential risk of weaponization and loss of control from a given concentration of AI hardware increases over time. As a first step to de-risking this effect, the BIS could consider using AI benchmarks such as Massive Multitask Language Understanding (MMLU) [154] to estimate proxies for AI capabilities per OP using present-day algorithms. It could then update its export control thresholds periodically as algorithms improve. Ultimately, though, export control thresholds at any given time will need to account for the possibility of unknown *future* algorithmic improvements. The same concentration of AI hardware that is safe given today's AI training algorithms, could pose risks of weaponization or loss of control if it is used to implement future training techniques (Introduction, 0.5.3.2).

[40] As an example, the NVIDIA A100 GPU has an FP16 (i.e., 16-bit width) tensor core performance of 312 TOPS (or teraFLOPS) without sparsity, and 624 TOPS (or teraFLOPS) with sparsity. Current export controls work by multiplying the 16-bit width by the 312 TOPS no-sparsity performance, yielding 4992 TOPS x bits for this chip and therefore very close to the existing export control threshold of 4800 TOPS x bits. Our recommendation would be to instead look for the maximum with-sparsity compute capacity of this chip, which is 1248 TOPS under the INT8 (i.e., 8-bit width) tensor core representation, and therefore clearly exceeds our proposed threshold of 800 TOPS.

[41] We are recommending a TOPS-based (i.e., compute capacity) threshold rather than (as with current controls) a threshold based on the product of TOPS multiplied by bit width. The reason is that a (TOPS x bit width) threshold risks incentivizing the development of AI chips optimized for shorter bit representations and higher compute capacities, which is already an industry trend that has been accelerating advanced AI capabilities development. A purely TOPS-based threshold, on the other hand, targets compute capacity directly. Thanks to Tim Fist at CNAS for this observation.

- The BIS could also coordinate with future task forces (see 1.4), research centers (LOE3, 3.1.2.2) and regulatory agencies (LOE4, 4.1) to align export controls for AI hardware with domestic regulations and up-to-date assessments of AI capabilities and safety research.

- As research into remote monitoring systems for AI hardware progresses, the BIS could consider tying export licenses to the implementation of on-chip monitoring [148]. For example, enabling companies to adopt on-chip monitoring for all exports in order to restore export licenses. This could support legitimate uses of advanced AI hardware, while preserving the ability to detect and remotely block high-risk activity. In the near-term, in order to incentivize progress in remote monitoring and related technologies, BIS could consider committing to AI hardware designers that they will maintain certain levels of access to export markets if they implement on-chip security features with certain properties [148].

- Given the risk that adversarial countries could acquire advanced chips through illegitimate means, including through third countries [155,156], the BIS could also consider temporarily blocking exports of certain AI hardware to all foreign or non-allied jurisdictions [157].

To support efforts tied to on-chip remote monitoring and related technologies, the Departments of Commerce and DHS could consider establishing export licensing red teaming programs. These programs could focus on testing and uncovering vulnerabilities in new or proposed on-chip governance technologies designed to grant companies the ability to export AI hardware to specific geographies [148]. These programs could also be undertaken in collaboration with federally funded research centers (LOE3, 3.1).

1.5.4 Research collaborations

U.S. and foreign researchers have frequently collaborated on AI research activities. This has led to the co-development of noteworthy models, such EfficientZero (Tsinghua University and UC Berkeley collaboration) [158] and NÜWA (Peking University and Microsoft Research Asia collaboration) [159], and has contributed to AI talent development abroad.

The U.S. government should investigate the the cost-benefit tradeoffs of such collaborations in the context of the advanced AI risk landscape.

However, excessive restrictions could prevent the diffusion of U.S. technical knowledge on advanced AI risk to the international research community. For example, the recent outreach by Western AI researchers to the Chinese research community about AI risk has been a critical step to establishing common ground and a shared understanding of the topic. Overly restrictive controls could deter such outreach in the future, to the detriment of U.S. and global security. Additionally, because AI risk awareness is still relatively nascent globally, we expect that research collaborations focused on understanding and mitigating key AI risks could become positive factors in long-term safety (LOE5, 5.2.1.2).

One straightforward option could be to educate U.S. academics on the national security risks of certain kinds of knowledge transfer to support their ability to judge which research collaborations are appropriate.

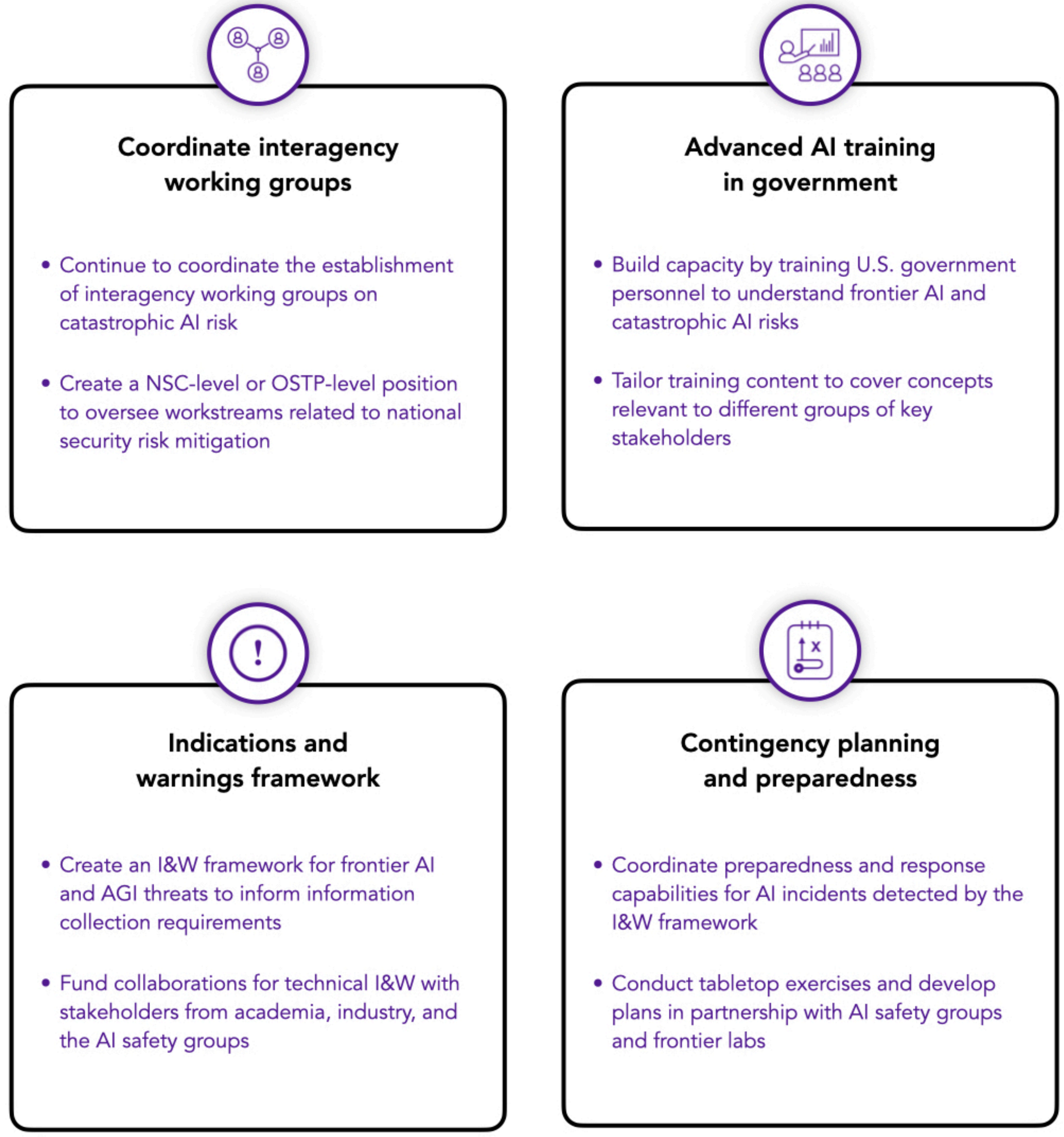
1.5.5 Education of foreign nationals

Currently, foreign AI expertise draws significantly from both collaboration with U.S. academic labs, and from foreign nationals educated at U.S. universities. We recommend the U.S. government explore approaches to encourage foreign nationals studying AI at U.S. universities — particularly at the graduate level — to remain in the United States upon completion of their studies. Lowering immigration barriers for such individuals is likely to increase U.S. competitiveness in AI while supporting economic growth and job creation.

The Department of State has already been directed to establish a program to attract top AI talent to the United States [127]. DHS has also been directed to clarify and modernize immigration pathways for AI researchers, startup founders, and others working in the field. We believe these are constructive steps.

Educating U.S. academics on national security risk and designating categories of AI research projects as suitable for work by foreign graduate and undergraduate students could also support limiting knowledge transfer, as in the case of research collaborations (see 1.5.4).

LOE2: Strengthen capability and capacity for advanced AI preparedness and response



Coordinate interagency working groups

- Continue to coordinate the establishment of interagency working groups on catastrophic AI risk
- Create a NSC-level or OSTP-level position to oversee workstreams related to national security risk mitigation

Advanced AI training in government

- Build capacity by training U.S. government personnel to understand frontier AI and catastrophic AI risks
- Tailor training content to cover concepts relevant to different groups of key stakeholders

Indications and warnings framework

- Create an I&W framework for frontier AI and AGI threats to inform information collection requirements
- Fund collaborations for technical I&W with stakeholders from academia, industry, and the AI safety groups

Contingency planning and preparedness

- Coordinate preparedness and response capabilities for AI incidents detected by the I&W framework
- Conduct tabletop exercises and develop plans in partnership with AI safety groups and frontier labs

Advanced AI and AGI risk mitigation will engage a broad set of U.S. government equities. However, the advanced AI landscape is changing rapidly (Introduction, 0.1). In this climate, policy should support swift, well-coordinated decision-making grounded in rigorous and timely technical assessments (Introduction, 0.5.2.1). Effective mitigation measures will require advance planning and accurate assessments and synthesis of risk signals in a potentially challenging information environment (Introduction, 0.5.2.2). Execution in these areas, and in those of the other LOEs, will necessitate substantial institutional capacity-building across all U.S. government stakeholders.

This LOE outlines specific actions that the U.S. government could take to **increase its preparedness for rapidly responding to incidents related to advanced AI and AGI development and deployment**. These actions are:

- Continuing to establish and coordinate **interagency working groups**, including for the LOEs in this action plan (see 2.1);

- Increasing preparedness and response capacity and capability through **education and training** (see 2.2);

- Coordinating the development of an **Indications and Warnings (I&W)** framework for advanced AI and AGI incidents (see 2.3); and

- Coordinating the development of scenario-based **contingency plans** (see 2.4).

The I&W and contingency planning processes will generate advanced AI-related risk management requirements that could inform other aspects of this plan, including the specific thresholds for RADA safeguards (LOE1, 1.3.2; LOE4, 4.1.3).

2.1 Coordinate interagency working groups

We recommend the U.S. government continue to coordinate the establishment of interagency working groups to execute on mitigation measures for catastrophic AI risk. To ensure such measures are treated as a holistic and coherent effort, the President could create an NSC-level or OSTP-level position with responsibility for all workstreams related to national security risk mitigation from advanced AI and AGI. This could include overseeing the execution of some of the LOEs in this action plan through the participating departments and agencies. For example the NSC-level official could:

- Receive regular reports on the state of the advanced AI landscape, including capabilities and risk assessments produced by an AIO (DHS; LOE1, 1.2);

- Oversee an ASTF with the mission to implement RADA safeguards for domestic frontier AI (DOE or DHS; LOE1, 1.4);

- Oversee the implementation of I&W (U.S. Intelligence Community; see 2.3) and contingency planning efforts (see 2.4) designed to guard against potential catastrophic risks from advanced AI and AGI;

- Support the coordination of regulatory efforts (LOE1, 1.4; LOE4, 4.1) with federally funded advanced AI and AGI safety and security research (National Science Foundation; LOE3, 3.1);

- Ensure regulatory efforts are aligned with emerging advanced AI and AGI safety and security standards (NIST U.S. AISI; LOE3, 3.2);

- In the long term, oversee a permanent regulatory agency for advanced AI (LOE4, 4.1); and

- Oversee U.S. efforts to build international capacity, consensus, and controls to manage catastrophic AI risks (Department of State; LOE5).

2.2 Advanced AI education and training

AI safety and security is a complex field in which nuanced technical factors often make the difference between effective and ineffective risk mitigation strategies. U.S. government personnel charged with implementing or developing AI safety policy must understand these factors in order to fulfill their duties successfully. However, different personnel will have varying educational requirements depending on their domain of work and position.

The U.S. government will therefore require a carefully targeted educational program, designed to provide training and support of the right type and depth to individuals charged with advancing U.S. national security interests in AI. This program should have several key properties:

1. **Ensure that the training and support it provides are updated regularly.** This is necessary to account for rapid advances in AI capabilities and risks as the frontier of the field progresses.

2. **Introduce ongoing requirements for continuous learning.** Significant advances that shape the landscape of AI capabilities occur on at least a monthly basis [160]. Without continuous learning, even trained personnel would quickly find themselves with an outdated understanding of frontier AI capabilities and risks.

In many cases this knowledge gap could pose a risk to their ability to execute their missions effectively.

3. **Deliver the right information to the right stakeholders.** Personnel involved in commercial regulation, military safety policy, international diplomacy, and other fields will each require different training focused on the aspects of the frontier AI landscape relevant to each of their respective domains.

4. **Deliver information in the right format for its intended audience.** Most relevant personnel can likely be trained via asynchronous online courses. But others may require live in-person training, ongoing access to centers of expertise within or outside government, or combinations of these options.

A robust, regularly updated, and technically grounded training program will be critical to support government decision-making in areas such as:

- Developing technically-informed policy, regulatory, and legal frameworks;
- Uncovering requirements for necessary new authorities and programs;
- Contributing to technical advances in AI safety and security; and
- Messaging the topic clearly and effectively with the American public.

There are many means by which training can be acquired, including from the private sector, but in all cases it should be vetted for quality and accurate risk focus.

2.2.1 Key stakeholders and learning outcomes

Educational requirements will vary among U.S. government personnel. Below, we provide a list of key stakeholder groups, along with recommended high-priority learning objectives for each. OMB could consider directing agency heads to identify and train personnel in these key areas. We also recommend that the Administration's key influencers acquire their own AI training for the purpose of increasing the technical accuracy of public discourse (Introduction, 0.5.2.2).

It is especially crucial to train technical personnel in defense and national security roles about this risk category, both in the United States and internationally to the fullest extent possible. These individuals need to be equipped to offer their leadership accurate and timely advice that will inform national security strategy with respect to advanced AI (LOE5, 5.2.1.2).

Table 1. U.S. government stakeholders and potential target learning outcomes of AI education.

Key stakeholders	Target learning outcome
NSC, OSTP, OMB, department and agency political leaders and possible regulators involved in advanced AI policy	Strategic and technically-informed decision-making
AI Observatory personnel (LOE1, 1.2)	Strategic and technically-informed decision-making
AI Safety Task Force personnel (LOE1, 1.3)	Deepened baseline operational understanding prior to negotiating with labs, understand key considerations for enforcement of AI regulations, understand the risks associated with frontier AI research to incentivize responsible behavior
Intelligence Community-led Indications and Warnings workstream and direct management personnel (see 2.3)	Strategic and technically-informed decision-making
Contingency planning workstream and direct management personnel (see 2.4)	Strategic and technically-informed decision-making
Interagency teams working all LOEs	Strategic and technically-informed decision-making
Congressional leaders and staff	Strategic and technically-informed decision-making
Key influencers [161]	Enhance the accuracy of public discourse
Judiciary leaders and staff [162]	Enhance capability to address cases on the range of legal issues presented by advanced AI, frontier AI, and AGI
Department of Justice stakeholders responsible for enforcement of advanced AI regulations (LOE4, 4.1.2.3 and 4.2.2)	Understand key considerations for enforcement of AI regulations
Key diplomats including embassy staff (LOE5, 5.2)	Accelerate U.S. ability to drive a common understanding of potential catastrophic AI risks for international coordination

International partners and stakeholders (LOE5, 5.2.1)	Accelerate U.S. ability to drive a common understanding of potential catastrophic AI risks for international coordination
AI researchers at National Labs and other technical centers of excellence in the U.S. government, and in governments around the world (LOE3, 3.1.2; LOE5, 5.2.1.2)	Understand open problems in technical AI safety and AGI alignment
Private sector AI capabilities researchers	Understand the risks associated with frontier AI research to incentivize responsible behavior
Public sector AI academics (LOE1, 1.5.4 and 1.5.5)	Understand the risks associated with frontier AI research to incentivize responsible behavior

See the next section for examples of training topics that support each of the training outcomes above.

2.2.2 Suggested training topics

Stakeholder training will need to be accessible to non-technical audiences. At the same time, it should be substantive enough to allow personnel to make informed predictions about near-future AI capabilities. It should also equip personnel to interpret semi-technical articles published by frontier labs announcing new breakthroughs (e.g., [163]), and to correctly infer some of their implications.

All stakeholders should obtain a **baseline understanding** of AI and key issues at play in frontier AI safety and security. This baseline understanding should include the following concepts:

- Drivers of AI progress (e.g. AI scaling and algorithmic improvements);
- Bottlenecks to AI progress (e.g. AI-optimized hardware, data, talent);
- Key sources of catastrophic AI risk (including weaponization and loss of control due to AGI alignment failure; Introduction, 0.2.1 and 0.2.2);
- Technical risk mitigation strategies (e.g. AI evaluations, mechanistic interpretability);

- Challenges to risk mitigation (e.g. outcome severity, timelines, incentives, supply chain factors; Introduction, 0.5); and

- Policy-based risk mitigation strategies (e.g. compute-based reporting thresholds, reporting requirements for large-scale training runs).

The training could also inform decisions on policy and safeguards for advanced AI systems; ideas for improving safety standards; and associated economic, antitrust, and regulatory capture concerns. Finally, training could examine operational challenges and solutions for international AI safeguards, aiming to build capacity and capability for implementations of LOE5. Training content could also be packaged as reference documents for government stakeholders. Like the training itself, these documents should be kept frequently updated given the pace of progress in frontier AI.

Below is a list of key training outcomes, along with sample training topics that we recommend as being supportive of each key outcome.

Table 2. Training topics associated with target learning outcomes.

Target outcome of training	Sample training topics
Strategic and technically-informed decision-making	• AI scaling and its relationship to AI capabilities and alignment • AI capabilities and impacts forecasting • Inner vs outer alignment and the stability-control paradox [161] • The open-source AI ecosystem • Compute as a strategic resource • Governance and control of advanced AI systems • The advanced AI supply chain

Deepened baseline operational understanding prior to negotiating with labs	• AI scaling and its relationship to AI capabilities and alignment • Weaponization, accident, and loss of control risks, including catastrophic risks • AI capabilities and impacts forecasting • The open-source AI ecosystem • Compute as a strategic resource • Antitrust and regulatory capture concerns • Concerns over gaming of safety evaluations • The culture of and ongoing debates within the AI alignment community
Enhance the accuracy of public discourse	• AI scaling and its relationship to AI capabilities and alignment • Weaponization, accident, and loss of control risks, including catastrophic risks • AI capabilities and impacts forecasting • Antitrust and regulatory capture concerns • The culture of and ongoing debates within the AI alignment community
Enhance capability to address cases on the range of legal issues presented by advanced AI and AGI	• Weaponization, accident, and loss of control risks, including catastrophic risks • AI capabilities and impacts forecasting • Antitrust and regulatory capture concerns • Open questions in generative AI copyrighting, and how these impact the data collection strategies of frontier labs
Accelerate U.S. ability to drive a common understanding of the risks for international coordination	• AI scaling and its relationship to AI capabilities and alignment • Weaponization, accident, and loss of control risks, including catastrophic risks • AI capabilities and impacts forecasting • Inner vs outer alignment and the stability-control paradox [161] • Concerns over gaming of safety evaluations • Governance and control of advanced AI systems • The advanced AI supply chain

Understand key considerations for enforcement of AI regulations	• AI scaling and its relationship to AI capabilities and alignment • Weaponization, accident, and loss of control risks, including catastrophic risks • AI capabilities and impacts forecasting • Concerns over gaming of safety evaluations • Governance and control of advanced AI systems • The frontier AI ecosystem and its key players • The ideological motivations of frontier AI researchers
Understand open problems in technical AI safety and AGI alignment	• Research agendas aimed at solving AI alignment • Techniques for measuring and mitigating power-seeking behavior • Corrigibility techniques • Mechanistic interpretability techniques • Empirical techniques such as activation engineering, representation engineering, and probing • Expected failure modes of reinforcement learning from human feedback and other prosaic alignment techniques in AGI-level systems
Understand the risks associated with frontier AI research to incentivize responsible behavior	• Weaponization, accident, and loss of control risks, including catastrophic risks • Concerns over gaming of safety evaluations • Research agendas aimed at solving AI alignment • Techniques for measuring and mitigating power-seeking behavior • Corrigibility techniques • Mechanistic interpretability techniques • Empirical techniques such as activation engineering, representation engineering, and probing • Expected failure modes of reinforcement learning from human feedback and other prosaic alignment techniques in AGI-level systems

2.3 Indications and warnings

Without a capacity to detect and anticipate potential catastrophic risks from AI weaponization or loss of control, the U.S. government risks either being caught at a disadvantage by fast-moving national security threats (Introduction, 0.5.1.1, 0.5.1.2, and 0.5.2.1), or causing premature harm to its own domestic AI industry (Introduction, 0.4.2).

Therefore, as a critical near-term action, the U.S. government should develop an indications and warnings (I&W) framework for advanced AI and AGI threats.[42] Development of this framework could be undertaken by the U.S. Intelligence Community (IC).

Once developed, an I&W framework could inform downstream requirements to comprehensively monitor advanced AI research and development globally (LOE1, 1.2), track the release of open-access AI models (Annex D, D.3), map the global supply chains for AI compute hardware and data centers (Annex G, G.3), and assess potential risks from both declared and undeclared frontier AI labs and other projects. The I&W framework should be kept regularly updated as new information becomes available, including by seeking input from frontier labs and the broader AI safety community.

An I&W framework could enable the U.S. government to receive timely warning of emerging risks from frontier AI development programs. Partners such as DHS and DOE could also help increase the government's understanding of technical I&W for advanced AI and AGI threats through collaborations with domestic frontier labs. The I&W framework could also benefit from coordination and information-sharing with other task forces and agencies such as the AIO (LOE1, 1.2), ASTF (LOE1, 1.4), federally funded AI research centers (LOE3, 3.1.2) and regulatory bodies (LOE4, 4.1). Such partnerships could support the IC in ensuring that the framework remains informed by an up-to-date and accurate view of the frontier AI landscape.

Traditional approaches to I&W [164] may need to be modified to accommodate challenges associated with modeling catastrophic AI risk. These could include the practical difficulties of accessing qualified technical talent to advise on I&W

[42] Advanced AI threats are not necessarily AGI threats. AI systems could pose catastrophic weaponization risk before AGI is developed, for example.

development, and the technical difficulties of forecasting the emerging capabilities of future frontier AI systems.[43]

2.3.1 Key categories and sources of catastrophic risk

Below is a preliminary taxonomy of entities and activities that, either now or in the medium-term future (1-5 years as of December 2023), could introduce catastrophic AI risks. See Introduction, 0.3 for descriptions of these entity categories and Annex D: Advanced AI landscape for more details. See Introduction, 0.2.1 and 0.2.2 for more details on the weaponization and loss of control risk categories.

Table 3. Potential sources of catastrophic AI risk, corresponding risk categories, and information requirements associated with each.

Potential risk sources	Potential risk categories
Domestic frontier AI programs	• Accidental emergence of dangerous capabilities or competent power-seeking behavior (**loss of control**) [1] • Sudden capability jumps enabled by external software frameworks (**loss of control**) [120] • Exfiltration followed by weaponization of stolen model by an adversary (**weaponization**) • Undetected misuse by end-users (**weaponization**)

[43] Although the emergence of specific AI capabilities cannot yet be reliably predicted in advance [165], it may still be possible to develop imperfect early indicators. For example, METR [166] is exploring an approach of breaking a dangerous capability down into a set of sub-tasks, and then measuring AI model performance on those sub-tasks. For example, in attempting to anticipate a model's capability to execute autonomous cyberattacks, that capability could be broken down into sub-tasks like the ability to write short code snippets or entire malware files, the ability to autonomously probe a cyber environment for vulnerabilities, and so on. See LOE3, 3.1 for recommendations on supporting research in this area.

Foreign AI programs	• Accidental emergence of dangerous capabilities or competent power-seeking behavior (**loss of control**) • Sudden capability jumps enabled by external software frameworks (**loss of control**) • Undetected misuse by end-users (**weaponization**) • Intentional offensive deployment (**weaponization**)
Open-access release of advanced AI models	• Accidental emergence of dangerous capabilities via fine-tuning (**loss of control**) • Sudden capability jumps enabled by external software frameworks (**loss of control**) • Intentional weaponization via fine-tuning or direct use (**weaponization**)
Theft or sale and subsequent augmentation of frontier AI models by state or non-state actors	• Accidental emergence of dangerous capabilities or competent power-seeking behavior after fine-tuning (**loss of control**) • Sudden capability jumps enabled by external software frameworks (**loss of control**) • Intentional offensive deployment (**weaponization**)

An I&W framework will need to consider a broad range of scenarios linked to these entity categories, and could be informed by many different sources of data [167–175]. The I&W framework should consider scenarios in which risk primarily emerges from an AI model or AI system itself, such as loss of control due to AGI alignment failure [176]. They should also include scenarios in which risk derives from specific actions by end-users or model developers, such as automated large-scale cyberattacks. The resulting I&W framework should in turn inform contingency planning efforts (see 2.4). Example scenarios could include:

- The unexpected emergence of dangerous AI capabilities at a U.S. or Western frontier lab [177];
- An open-access AI system fine-tuned to support cyber or CBRN attacks;

- The sudden discovery of a previously unrecognized capability in a generally available open-access AI model that enables dangerous weaponized applications;

- The uncontrolled deployment of an AGI or near-AGI system by a hedge fund that autonomously engages in profit-driven information warfare [178];[44]

- Systemic vulnerabilities introduced by integrating interdependent AI systems into decision-making processes whose interactions may have unpredictable effects (e.g. a flash crash in the financial system);

- Convincing evidence that suggests that AGI alignment is an intractable problem, or that it cannot be expected to be solved before AGI is developed [7]; or

- A breakthrough that makes it possible for a particular weaponized application to be deployed below a certain budget.

For each scenario, the I&W framework could maintain and frequently update a set of key metrics, for example:

- The estimated breakout timelines to various levels of AI capability by identified domestic and foreign entities;

- The largest AI training run a domestic or foreign entity could undertake in a set amount of time; and

- The likelihood of emergence of high-risk capabilities under various AI training and deployment regimes.

The I&W framework could also consider longer-timescale metrics that do not need to be updated as frequently. For example, understanding how long it may take a country to develop an independent AI supply chain under various assumptions, and which critical inputs need to be controlled to prevent that development [172].

As part of the I&W framework development process, the U.S. government should consider convening an interagency workshop aimed at mapping the AI ecosystem and AI hardware supply chain, its key domestic and international players, and the activities

[44] See Annex D, D.4 for more information about this risk category.

in which they are engaged that most directly contribute to potential catastrophic risks from weaponization and loss of control.

2.3.2 Improving I&W through bug bounties

Software frameworks such as Auto-GPT [120] and BabyAGI [122] extend the capabilities of existing AI systems in unpredictable ways by helping them perform complicated multi-step tasks without any additional training. They therefore introduce new risks that AI developers are not equipped to anticipate or assess. Some of these frameworks, like the infamous ChaosGPT [121], are intended to deliberately steer the underlying AI system toward destructive behaviors.

To ensure an I&W framework remains robust over time, it is crucial to improve understanding of how software frameworks like these can impact downstream AI capabilities. These kinds of impacts can be assessed openly, so the government can offer bug bounties and other incentives to support this effort. These incentives could encourage individuals and teams to report issues, vulnerabilities, and unsafe behaviors in frontier AI systems. They could also serve to augment AI developers' existing bug bounty programs.

2.3.3 Improving I&W through direct government investment

Currently, no one knows how to forecast the emergence of high-risk AI capabilities under different training and deployment conditions. This lack of knowledge increases the uncertainty and challenge of developing and maintaining a robust I&W framework.

To close this gap, the IC should consider establishing a formal academic consortium or cooperative research agreement for AI safety and security and technical I&W. This consortium could leverage top universities to help refine forecasting approaches by advancing the state of the art in areas such as AI capability evaluations, transparency and interpretability, and capability prediction techniques (LOE3, 3.2).

In addition to an academic consortium, the U.S. government could establish an explicit vehicle for collaboration between the IC, frontier labs, and private sector AI safety and AGI alignment research organizations to inform the I&W development and maintenance process. The commercial market for this research remains limited, and government investment could accelerate efforts to improve the technical accuracy of AI forecasting in support of robust I&W.[45]

Either or both of the above initiatives could be coordinated through federally funded AI research centers dedicated to this problem set (LOE3, 3.1.2.3). The IC could also consider leveraging the AI testbeds and infrastructure under development by DOE and NIST to support I&W framework development and implementation. These testbeds could be used to run evaluations on AI models that map directly onto key information requirements associated with the I&W framework (see 2.3.1). This could include developing model evaluations specifically tailored to the I&W use case.

2.4 Contingency planning

Under some worst-case scenarios (see 2.3.1), decision-makers could have only a very short window (hours to weeks) to react effectively to a catastrophic threat. An I&W framework (see 2.3) could support the detection of early warning signs for such scenarios, but contingency planning will be essential to allow for the development of appropriate responses.

While no plan can anticipate all contingencies, the planning process itself is essential to developing a common understanding of crisis scenarios and response options. These options can help the government surface detailed requirements for new authorities, programs, resources, policies, and coordination mechanisms. Therefore another critical and immediate task for the U.S. government should be to direct, through the NSC, an interagency contingency planning process to develop response options for various scenarios and timelines of frontier AI development.

Planning could be informed by a range of scenario-based tabletop exercises (TTXs) covering advanced AI risks and sources, up to and including how a safe and secure AGI should be governed if and when it is developed. It may be necessary to collaborate in these exercises with stakeholders from frontier labs, AI safety groups, and evaluation and red teams to ensure the parameters of the exercises at all times remain as realistic as possible [161]. It may also be helpful to run joint exercises such as crisis response or lab shut-down drills with data center infrastructure providers, AI hardware owners, and frontier labs to anticipate and better understand key contingencies.

[45] Another reason to favor U.S. government funding for these organizations is that the great majority of them are nonprofits, and are currently funded by the same set of large donors. See Annex E: Funding in AI safety for more information.

Contingency planning activities should involve the full range of stakeholders within the U.S. government who would be involved in implementing emergency responses to critical events and risks. They could account for input from an AIO (LOE1, 1.2), ASTF (LOE1, 1.4), federally funded AI research centers and relevant collaborations, (LOE3, 3.1.2.2), and eventual regulatory bodies (LOE4, 4.1). In determining which events to plan for, planners should consider the set of scenarios uncovered in the I&W development process (see 2.3), and identify any additional scenarios that they consider necessary to address.[46]

Planners should flag these additional scenarios to stakeholders in the I&W development process when possible. We recommend that contingency planning proceed iteratively, to allow initial planning activities to surface information requirements which, when met, would allow future planning and I&W development activities to be more technically informed.

2.4.1 Updating the National Preparedness System

In the event that a severe incident occurs as a result of advanced AI weaponization or loss of control in highly capable AI systems, early planning for prevention and incident response will be critical.[47] We therefore recommend that DHS direct FEMA to update its National Preparedness System [183] to address advanced AI and AGI risk. These updates could be executed in collaboration with an AIO (LOE1, 1.2.1), ASTF (LOE1, 1.4), or eventual regulatory body (LOE4, 4.1).

[46] For example, see the companion document, **Survey of AI Technologies and R&D Trajectories**, for three sample AGI risk scenarios.

[47] It is possible that genuine AGI loss-of-control scenarios would not be survivable events (Introduction, 0.5.1.1), which if true would remove the need for incident response. But this is currently uncertain and should not be assumed in advance.

LOE3: Increase national investment in technical AI safety research and standards development

Federally funded AI safety and security research

- Direct federal investment in AI safety, AI security, and AGI alignment research with different levels of sensitivity and disclosure
- Support collaborations with frontier labs, academics, AI safety labs, and cleared U.S. government experts

Standards and standards development practices

- Develop standards for responsible AI development and adoption safeguards addressing catastrophic risks
- Develop approaches to evaluate AI models for catastrophic risks, recognizing the limitations of AI evaluations

The acceleration of investment in AI capabilities is outpacing the development of proportionate technical safeguards against advanced AI and AGI risks [5] (Introduction, 0.5.1.2 and 0.5.1.3). If this continues, frontier AI labs may find themselves unable to meet the safety and security challenges posed by their own systems (Introduction, 0.5.1.4). Unless strong technical safeguards are designed, standardized, and broadly applied, continued development and adoption of frontier AI systems could create significant risks (Introduction, 0.5.1.1).

This LOE outlines specific actions the U.S. government could take to **strengthen domestic technical capacity in advanced AI safety and security, AGI alignment, and other technical AI safeguards.** These actions include:

- Directly funding **advanced AI safety and security research** including AGI-scalable alignment research (see 3.1); and
- Developing, regularly reviewing, and promulgating **safety and security standards** for responsible AI development and adoption (see 3.2).

These solutions would grow the pool of competent AI and AGI safety and security researchers, encourage organizations to invest in AI safety and security, and increase overall understanding of AGI alignment techniques.

3.1 Federally funded research in AI safety and security and AGI-scalable alignment

Since 2021, annual private spending on AI capabilities has exceeded $100B, dramatically outstripping funding for AI safety and security [184]. AI capabilities research publications outnumber safety publications by a fifty-to-one margin [184]. Yet according to experts, in order for the level of investment in AGI-scalable safety research to approach the level needed to address catastrophic AI risk, it would need to reach at least 50% of total research spend [185]. The current funding gap exists because, while AI safety and security is a common good, AI capabilities produce benefits that are easier to privatize. As a result, private industry has a strong incentive to invest in capabilities at the expense of research aimed at catastrophic AI risk reduction (Introduction, 0.5.1.4, 0.5.1.5, and 0.5.3.1).

Aside from private industry and academic labs [186,187], there is also an emerging ecosystem of independent AI safety researchers, nonprofit and for-profit AGI alignment labs [188–190], and advanced AI auditing firms [94,166,191]. However most of the organizations in this ecosystem currently rely on a small set of donors for funding, which could limit the diversity of strategies to AGI alignment reflected in the output of the AI safety research community.[48] Because AGI alignment is still a nascent field that benefits disproportionately from a breadth of research approaches, this concentration of funding sources could create limitations to progress.

The U.S. government is uniquely positioned to close the funding gap between AI capabilities and AI safety and security, accelerate progress on core problems, and diversify the set of funding sources and research programs aimed at solving AGI alignment.[49] See Annex L: AI safety and security research topics for a list of research areas that could be supported by U.S. federal funding.

[48] See Annex E: Funding in AI safety for more information.

[49] See Annex B: The full challenge of AGI alignment for details.

3.1.1 Levels of AI safety and security research

In addition to its capacity for funding critical basic research, the U.S. government can also ensure that those research efforts adhere to necessary security practices and that they have the required levels of access to key technologies, resources, and talent. This is critical in the context of AI safety and security research that may require interacting closely with potentially dangerous frontier AI models.

Not all such research will require the deepest levels of access. We recommend establishing tiers of AI safety and security research that reflect the escalating risk and responsibility associated with granting researchers broader access to increasingly capable models and evaluation capabilities. We provide a possible categorization below along with examples.

3.1.1.1 Open-source and public research

Some types of AI safety and security research involve only low-risk activities and do not require any special access to frontier AI models. All aspects of research at this level can and should be publicly shared. Funding of open-source research could support independent AI alignment and AI safety and security researchers and organizations. It could also support the development of software tools and frameworks with important safety and security applications, such as test harnesses, evaluation frameworks, sandboxes, and interpretability tools. To maximize the diversity of research approaches, we recommend that funding grants at this level impose minimal compliance overhead on recipients.

Some examples of research at this level could include:

- Certain kinds of evaluations for advanced AI systems, including basic sandbox and testbed development;

- Certain kinds of AI interpretability research, including on sub-frontier AI models generally regarded as safe [192];

- Theoretical research aimed at investigating evidence for risks associated with power-seeking and associated tendencies [2,47];

- Research aimed at determining the extent to which existing alignment techniques could or could not scale to AGI-level systems [84]; and

- Basic research into on-device monitoring for AI chips [148].

The decision to classify AI safety research as open-source should take into account the risk that the research could empower threat actors to weaponize AI more effectively, subvert on-chip monitoring capabilities (e.g. LOE1, 1.5.3), or otherwise accelerate or proliferate AGI development. But it should also account for the potential economic and societal harms of withholding such research from the public (Introduction, 0.4.2). As a result, publication decisions should be closely informed by input from academic researchers, the AI safety and national security communities, and private sector industry including the startup community.

Funding for open-source and public research could be directed towards the following types of entities:

- Academic labs and independent researchers working on AI safety and security projects;

- Nonprofits conducting AI safety and security research;

- Researchers at National Labs and other centers of excellence within the U.S. government with the capacity to pursue AGI-scalable alignment research;

- For-profit AI safety research labs working on AGI-scalable alignment techniques.

3.1.1.2 Research that requires access to proprietary AI models

Certain essential types of AI safety and security research will require a high degree of access to frontier labs' proprietary AI models, or may otherwise involve moderate-risk activities without being sensitive enough to necessitate a security clearance. Some of the results of research at this level could still be published, but the extent of disclosure should be informed by risk assessments and IP considerations.

The government can play a key role in supporting this research by offering direct funding, mediating researcher access to developers' AI models, and overseeing researcher compliance with the necessary security practices. This could involve certifying AI safety research groups or individual researchers to access frontier labs' proprietary AI models. This approach could also be used to support AI evaluator certification and access to frontier lab models by entities such as the ASTF (see LOE1, 1.4.1.2) or a permanent regulator (see LOE4, 4.1.2.1).

Some examples of research at this level could include:

- Research that involves inducing powerful AI models to behave in dangerous ways, such as behavioral and propensity evaluations for dangerous capabilities like deception and persuasion [193];
- The development of some sets of private AI evaluations that should be withheld from frontier labs to avoid compromising the integrity of model audits (see 3.2.2);
- Some kinds of fundamental AGI alignment research, particularly if they require direct access to frontier models;
- Mechanistic interpretability research that requires direct access to frontier models [194]; and
- Research into on-device monitoring for AI chips that could benefit from frontier labs' engineering resources or expertise in scaled AI training runs [148].

Given the potential sensitivity of research in this category, researchers and model evaluators may require dedicated infrastructure to support secure access to the weights of frontier AI models. We describe a high-level weight-sharing protocol that could meet this requirement in Annex M: Secure temporary storage of model weights. In addition to directly funding private collaborations with AI labs, the government could fund the development of secure information-sharing infrastructure to ensure this research can occur securely and with minimal friction.

3.1.1.3 National security research

The most high-risk category of AGI safety research will likely involve unrestricted access to frontier AI models to evaluate propensities for highly dangerous capabilities, including WMD-like and WMD-enabling capabilities. This category of research will need to be conducted in a classified environment by cleared researchers following appropriate security procedures. Some kinds of model evaluations overseen by entities such as the ASTF (see LOE1, 1.4.1.2) or a permanent regulator (see LOE4, 4.1.2.1) may need to be classified at this level.

Similarly to proprietary research (see 3.1.1.2), this level of research would enable a capability for secure sharing of the weights of frontier AI models. See Annex M: Secure temporary storage of model weights for a description of a high-level protocol for weight-sharing.

3.1.2 Organizational framework

The U.S. government would need the ability to identify, fund, and administer promising research projects at each of the sensitivity levels in 3.1.1. Each level may have a different set of oversight, administration, and funding conditions that reflect its degree of risk and access requirements.

The U.S. government has already begun developing a capacity that could support such activities. In January 2024, the National Science Foundation (NSF) launched a pilot program for a National AI Research Resource (NAIRR) [195,196], whose budget is estimated to reach $2.6 billion over an initial six-year period. The NAIRR also includes an initiative called NAIRR Secure, which aims to support research involving sensitive data by assembling secure compute clusters and other privacy-preserving resources [197]. One of the NAIRR's four key goals is to **advance trustworthy AI** [198].

Many of the AI safety and security research areas relevant to mitigating catastrophic risk from advanced AI are aligned with this NAIRR key goal. As a result, we believe that relevant research directions at each of the relevant sensitivity levels (see 3.1.1) could be directly supported by:[50]

- The NAIRR, through a dedicated National Center for AI Alignment and Security Research (NCAASR) established under the NAIRR;
- A new AI Alignment and Security (AAS) Federally Funded Research and Development Center (FFRDC), possibly in collaboration with the NAIRR;[51] or

[50] Given that there are ongoing U.S. government workstreams in this area, we recommend considering these options in conjunction with other entities that could already be in development.

[51] One advantage of an AAS FFRDC is that it could be sponsored by organizations such as the Defense Advanced Research Projects Agency (DARPA), the Intelligence Advanced Research Projects Activity (IARPA), or the DOE's National Laboratories, which all have established CBRN research capabilities that could support in designing AI evaluations for frontier models. An AAS FFRDC would also offer the government a classified and secure environment in which researchers can be paid industry-competitive salaries. This is a particularly important factor given the prevailing level of compensation at frontier AI labs (see LOE1, 1.4.2).

- Both a NCAASR and a new AAS FFRDC, with research areas divided between them and research collaborations in areas of mutual interest.

We recommend that these Centers establish a review process to determine which kinds of safety research should remain unpublished, and base decisions on national security risk assessments (LOE4, 4.1.4).[52] This review process could also be informed by an ASTF's work under LOE1, 1.4.1.3 and be developed in consultation with members of the technical AI safety community.

These Centers could also advance basic scientific research into AGI alignment, normalize the field, and grow the overall pool of researchers.

3.1.2.1 Considerations for AGI alignment research

Many alignment techniques that work on less capable AI systems are not expected to work on AGI-level systems [84] (Introduction, 0.5.1.4). It is also easier to make measurable progress on sub-AGI alignment than on AGI-scalable alignment. As a result, we recommend that these Centers define and fund research workstreams fully and exclusively dedicated to developing alignment techniques that are intended to scale to AGI-level systems. Absent a clear institutional prioritization of AGI-scalable alignment, research effort funded by the Centers risks being expended in areas in which progress is more legible but less scalable.[53]

See Annex B: The full challenge of AGI alignment for a discussion of the reason why alignment techniques that work on advanced AI systems may fail once these systems reach high enough capability levels. See Annex L: AI safety and security research topics for a list of possible research topics that an NCAASR and AAS FFRDC could fund or support.

[52] This could also be done in coordination with the I&W and contingency planning efforts in LOE2, 2.3 and 2.4.

[53] Moreover, many forms of sub-AGI alignment research deliver immediate economic returns by making it easier for users to extract value from better-aligned, sub-AGI scale systems. As a result, these research areas already benefit from substantial investment by private-sector frontier AI labs.

3.1.2.2 Research collaborations

Centers such as the NCAASR and/or AAS FFRDC could be instrumental in developing a regime of technical RADA safeguards (LOE1, 1.3.2) to mitigate potential catastrophic AI risks from weaponization and loss of control. The Centers could collaborate with the NIST U.S. AISI to standardize this safeguards regime (see 3.2.3), and with entities such as an ASTF (LOE1, 1.4.1.2) or future regulatory body (LOE4, 4.1.2.1, 4.1.2.2, and 4.1.2.4) to implement and oversee those standards. These Centers could also collaborate with and support AI safety and security initiatives across the U.S. government and beyond, including:

- The BIS in periodic reviews of export controls related to catastrophic AI risk mitigation, including support with on-chip monitoring initiatives (LOE1, 1.5.3);

- The Department of Commerce and DHS for export license red teaming exercises (LOE1, 1.5.3);

- The IC to support creating and updating I&W frameworks (LOE2, 2.3.3);

- The NSC to support creating and updating contingency plans (LOE2, 2.4);
-
- The ASTF to help prepare for a future regulatory regime (Annex K, K.2); and

- Through the Department of State, international partners such as the U.K. Frontier AI Taskforce and the U.K. AI Safety Institute (U.K. AISI) [199].

Operationally, with the support of the NAIRR, the Centers could also foster direct collaboration between independent AI safety and security researchers, academics, and industry, including frontier labs. This could include funding the development of infrastructure to support secure information sharing with frontier labs (e.g., NAIRR Secure [197]). It could also offer researchers access to computing resources [198].

3.1.2.3 Sensitive research areas

Certain types of research aimed at mitigating catastrophic weaponization risk from CBRN, cyber, and other applications enabled by advanced AI will require direct access to frontier models, secure information environments, and cleared personnel (see 3.1.1.3). This may also be the case for some categories of research investigating the risk of loss of control due to alignment failure of high-capability AI systems.

The level of security required to support such research will likely require close coordination between frontier AI labs, the teams conducting the research, the IC, and the Centers supporting the research. The Centers would also need infrastructure to support highly secure sharing of information with frontier labs, including model weights (see 3.1.1.3).

Given the substantial overlap in requirements and likely collaboration in other areas, the Centers could also consider sharing infrastructure between sensitive AI research initiatives and the regulatory bodies overseeing private-sector AI scaling, including model evaluations for CBRN and other capabilities (Annex K, K.1 and LOE4, 4.1.3.4.3).

3.2 Standards for AI evaluations and RADA safeguards

Without clear standards for responsible AI development and adoption (RADA), advanced AI developers lack principles against which to judge the adequacy of their safety and security practices. Self-regulatory industry bodies such as the FMF (Introduction, 0.4.1) could have a positive influence on the development of such standards. But competitive pressures — including from challenger labs whose models lag the public frontier by 12 months or less, and are even less rigorous in their safety measures [51,200,201] (Annex D, D.1) — ultimately drive even the frontier labs to employ inconsistent and inadequate safety and security practices (Introduction, 0.5.1.4, 0.5.1.5, and 0.5.3.1).

It is therefore crucial to develop standards for RADA safeguards which address catastrophic risks from weaponization and loss of control in a rigorous and technically informed manner. These should include standards for the evaluation of AI models' behaviors and propensities to display dangerous capabilities [46]. NIST, through the U.S. AISI [202], is currently spearheading U.S. government efforts to develop and promulgate AI evaluation standards consistent with the NIST AI Risk Management Framework [203].

In addition to AI evaluations, standards for effective RADA safeguards will also need to comprehensively address security, risk governance, incident reporting, and other key issues across multiple layers of the advanced AI supply chain. See LOE1, 1.3.2 for our recommendations on guiding principles behind these standards.

Finally, ongoing refinement and updating of standards for RADA safeguards could be supported by expanded national research capacity in AI safety and security research as described in 3.1.

3.2.1 Limitations of AI evaluations

AI evaluations are important tools to increase confidence in the safety and security of advanced AI systems and support RADA safeguards. And depending on the degree of coverage and prior real-world experience with an evaluation set, passing a set of AI safety evaluations may offer a strong positive indicator of the safety of an AI system. But AI evaluations are inadequate on their own to fully assure that an AI system or model is safe to develop or deploy, because they suffer from a number of fundamental limitations in coverage and reliability [204]. We list below some of these key limitations, and discuss ways to mitigate their impacts in 3.2.2.

3.2.1.1 AI evaluations are not comprehensive

Current AI evaluations **cannot provide comprehensive coverage** of an advanced AI model's behavior or risk surface. If there is a dangerous behavior or set of behaviors that is not included in the test suite used to evaluate a model, then that behavior will not be detected, and it could be displayed or elicited under real world conditions [206] including under high-risk conditions.

There is also currently no way to reliably predict the emergence of capabilities, dangerous or otherwise, from advanced AI systems as they are scaled or augmented [165,206]. This means that not only do researchers have no way of knowing in advance which kinds of evaluations will be most informative for future AI systems, but they also have no way to reliably predict how an *existing* AI system will behave if it is fine-tuned or otherwise augmented. In particular, open-access AI systems can be augmented in any number of ways that may be entirely unanticipated by the system's original developers or the evaluators themselves (Introduction, 0.5.1.6). Therefore, as limited as AI evaluations are for substantiating the safety of proprietary AI systems, they are even more limited when applied to open-access AI systems.

3.2.1.2 AI evaluations cannot confirm that a dangerous capability is absent

AI evaluation techniques today are empirical. They generally function by feeding a model a wide range of input prompts and observing its outputs.[54]

[54] More sophisticated techniques may fine-tune a model to attempt to elicit latent capabilities. This is a more effective approach than prompting, but is still limited.

If the model returns a dangerous output in response to an evaluation, this may show that the model has a dangerous capability. But if the model does not return a dangerous output in response to the evaluation, this could be because the evaluator failed to identify an effective prompt, or for other reasons, rather than because the model does not have the evaluated capability. Even small changes to prompting techniques often reveal latent capabilities that were not detected by prior evaluations [207]. As a result, AI evaluations **can only reveal the presence, but not confirm the absence,** of dangerous capabilities [204].

3.2.1.3 AI evaluations are highly vulnerable to manipulation

AI evaluations **can be undermined and manipulated easily**. If an AI model fails an evaluation, its developer can fine-tune the model — or make other superficial changes to it — until the model passes that evaluation, without addressing the underlying cause of the failure.

For example, suppose an AI model fails a cyberattack evaluation by complying with a user request to write working malware. In this case, it is easier for the model's developer to train it to refuse the particular user request it was tested on, than to rigorously investigate and address the underlying cause of the behavior. But this superficial approach could leave the model vulnerable to jailbreaks [83,208] that induce the model to display the undesirable behavior under slightly more complex conditions. In other words, the dangerous capability can remain latent, even if an AI model passes the evaluation. Moreover, if an AI developer is able to resubmit fine-tuned or "patched" models for evaluation an arbitrary number of times, the resubmitted models may eventually pass the original evaluations purely or partly by chance [46].[55]

Given the prevailing competitive environment, frontier AI labs and other advanced AI developers face a strong incentive to game evaluations in this way (see 0.5.3.1). In fact, by the private judgment of one expert with direct knowledge of the matter, this has already begun to occur at one major frontier AI lab. Moreover, AI developers also face related incentives to argue that models that pass unreliable evaluations should still be considered safe for the purpose of further development or deployment.

[55] One frontier AI researcher stated during a private discussion that passing a dangerous capability evaluation provides little evidence that the model is safe, if the lab developing the model is allowed to resubmit tweaked versions of the model for re-evaluation until the evaluation is passed. As they put it, "There's a key step [missing] here of figuring out what went wrong."

3.2.1.4 AI evaluations could fail systematically in high-capability regimes

At higher levels of capability, advanced AI and near-AGI systems could develop the capability to infer whether a given interaction is part of their training, their testing, or their deployment. This property is called **situational awareness** [209,210], and researchers at major frontier labs believe it could arise as an emergent capability in future AI systems [209].

If an AI system can infer that it is being tested, it may have the ability to produce different outputs when it is being tested than when it is deployed. In principle this could allow the AI model to **manipulate its own evaluations**, by producing apparently safe outputs in response to evaluation queries in testing, and unsafe outputs in response to real user queries in deployment [209]. This hypothesized behavior is known as **deceptive alignment** [45] . Deceptive alignment remains a speculative and controversial risk [211], but if it manifests, it could systematically undermine the entire effectiveness of a naive AI evaluations regime.[56]

3.2.2 Addressing the limitations of AI evaluations

There are several possible strategies to lessen the impacts of the above limitations on the value of AI evaluations as safety signals.

First, a diverse ecosystem of independent AI model evaluators is crucial. **Independence** is crucial because model developers are heavily incentivized to pass safety evaluations by making superficial changes that do not address underlying drivers of risk (see 3.1.2.3). Independent evaluators can administer **private evaluation sets**, whose exact protocols are not known to AI model developers, and that are therefore more difficult to pass by superficial means. Independence also means that AI evaluators **should not be selected by the AI developers** they are evaluating, to avoid obvious conflicts of interest. **Diversity**, on the other hand, is crucial because current AI behavioral evaluations are unreliable, and failing to observe a dangerous capability may not mean that the capability is actually absent (see 3.2.1.1). A diverse ecosystem of evaluators supports **multiple independent evaluation strategies** and this combination is more likely to detect a dangerous capability if one is present.

[56] It is also important not to anthropomorphize these risks. Deceptive alignment is not expected to arise due to sentience, human-projected drives, or consciousness, but rather because, under certain circumstances, it may simply be an effective behavior for the purpose of achieving the goals an AI system has internalized during its training.

Second, a diverse set of AI evaluations, with **different private subsets** developed, refined, and administered by different evaluators, is also crucial. This is because there is no way to fully assess the capability surfaces of modern advanced AI models (see 3.2.1.1). The broader the set of teams, and the more diverse the evaluations, the lower the chance that dangerous capabilities and other risk vectors are unexamined. This also means that standards bodies should avoid overspecifying the parameters of AI evaluations — with the exception of the security conditions under which they should be administered.[57]

Finally, because AI evaluations may have fundamental limitations (see 3.2.1.4), over-reliance on AI evaluations could propagate a false sense of security among AI developers, regulators. Instead, AI evaluations could be considered as **one part of a multifaceted case** for model safety and security, consistent with the principle of defense in depth.

3.2.3 Standards for RADA safeguards under catastrophic risk

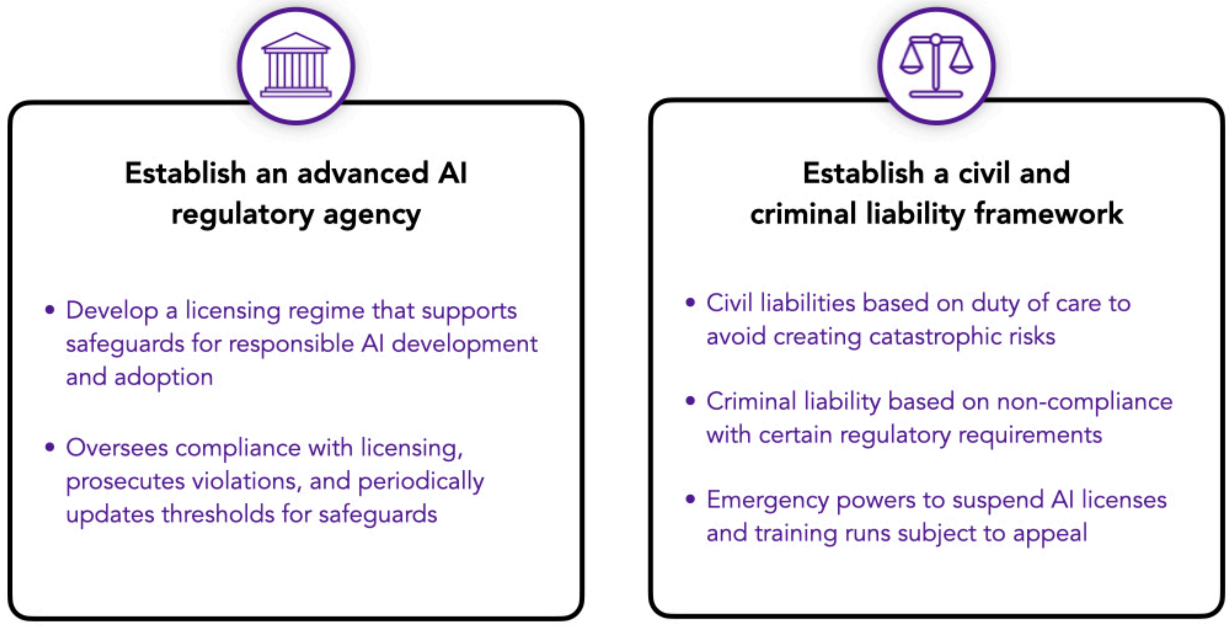
For recommended principles to guide standards development for RADA safeguards under catastrophic risk, see LOE1, 1.3.2.

For an example of a complete standards framework for RADA safeguards, including sample calculations of thresholds for covered entities, see LOE4, 4.1.3.

[57] Because dangerous capability evaluations often involve actively eliciting dangerous capabilities, in high-capability regimes the evaluations themselves could become sources of risk and should therefore observe strict operational security precautions.

LOE4: Formalize safeguards for responsible AI development and adoption by establishing an AI regulatory agency and legal liability framework

We are grateful to the team at the Center for AI Policy (CAIP), whose perspectives on AI regulation have informed several of the recommendations in this LOE.



Establish an advanced AI regulatory agency

- Develop a licensing regime that supports safeguards for responsible AI development and adoption
- Oversees compliance with licensing, prosecutes violations, and periodically updates thresholds for safeguards

Establish a civil and criminal liability framework

- Civil liabilities based on duty of care to avoid creating catastrophic risks
- Criminal liability based on non-compliance with certain regulatory requirements
- Emergency powers to suspend AI licenses and training runs subject to appeal

Interim regulations may be inadequate to address the risks and challenges of advanced AI. The unique challenges involved in assigning responsibility for potential catastrophic accidents to individuals or organizations who develop or use these advanced AI systems creates an ambiguous legal environment. This ambiguity offers Americans weak protections against the impact of reckless or negligent development or use of powerful AI systems (Introduction, 0.5.4.1). A legal framework for AI regulation and liability, that directly addresses potential catastrophic risk through detailed and flexible RADA safeguards overseen by a permanent regulatory agency, is essential to promote long-term stability and cover any gaps in existing authorities. This legal framework should carefully balance the need to mitigate potential catastrophic threats against the risk of curtailing innovation, particularly if regulatory burdens are imposed on small-scale entities (Introduction, 0.4.2).

This LOE outlines specific actions the Legislative Branch could take to **establish the conditions for long-term (4+ years) domestic AI safety and security**. These actions include:

- Creating a **Frontier AI Systems Administration (FAISA)**, a regulatory agency with rulemaking and licensing authorities to oversee AI development and deployment (see 4.1), consistent with a set of RADA safeguards regulations derived from contingency planning requirements (see 4.1.3); and

- Establishing a **criminal and civil liability regime** that could include defining responsibility for AI-induced damages; determining the extent of culpability for AI accidents and weaponization across all levels of the AI supply chain; and defining emergency powers to respond to dangerous and fast-moving AI-related incidents which could cause irreversible national security harms (see 4.2).

We expect that such a legal regime would begin to meaningfully impact advanced AI safety within about 3 years. While we expect such a legal regime to be essential for long-term safety and security, the 3-year impact timeline also highlights the importance of rapidly putting in place effective interim measures, as recommended in LOE1.[58]

4.1 Establish the Frontier AI Systems Administration (FAISA)

Consistent with ongoing public and congressional discussions on AI policy, the Congress could establish an agency to regulate, license, and monitor advanced AI model developers and other entities in the advanced AI supply chain.[59] This agency, which we refer to here as the **Frontier AI Systems Administration (FAISA)**, would have the mission to **implement responsible AI development (RADA) safeguards to mitigate catastrophic national security threats from AI weaponization and loss of control related to the domestic development or use of advanced AI systems.**

This mission would require the FAISA to make decisions with profound implications for the AI industry in an objective, technically informed, and evidence-based manner. The

[58] See http://aipolicy.us/gladstone for a proposed Act drafted by the Center for AI Policy (CAIP), which could achieve the above objectives in a technically informed manner, and from which many of the legislative recommendations in this LOE are drawn.

[59] The components of the AI supply chain covered under this licensing proposal are: AI hardware design firms; data center infrastructure providers; AI hardware owners; and AI model developers. Semiconductor fabrication firms and their suppliers could be covered under an expanded proposal. See the Glossary of terms for definitions of each of these supply chain components.

agency would also need to earn and maintain a high degree of public trust by acting – and being seen to act – in the long-term interest of the United States. Consistent with these requirements, the FAISA could be established in one of two ways:

- As a non-partisan agency reporting directly to the President, in the model of the SEC.
- In an existing department with an ample existing technical capability (such as DOE), in the model of the National Nuclear Security Administration (NNSA) [212]. This could involve the FAISA reporting to an NSC-level official (LOE2, 2.1).

The FAISA could be led by an Administrator, who could be appointed by the President with the Senate's advice and consent. The FAISA's strategic outlook and day-to-day operations would be highly sensitive to nuanced technical considerations, so we recommend choosing the FAISA Administrator based on expertise at the crossroads of security and advanced technology, including in areas such as cybersecurity or biosecurity.

The FAISA could perform the following key functions:

- **Licensing:** Overseeing rulemaking, especially concerning AI licensing, ensuring that critical RADA safeguards are in place and operating.
- **Monitoring:** Supervising hardware monitoring, tracking AI hardware locations, and ensuring hardware is accounted for; and performing horizon-scanning for domestic and foreign AI programs (LOE1, 1.2).
- **Enforcement:** Investigating regulatory violations, bringing civil cases in the courts, and referring criminal cases to the Department of Justice (DOJ).
- **Algorithms:** Monitoring the progression of AI algorithmic efficiency, suggesting updates to technical licensing thresholds, tiers, and conditions (Introduction, 0.5.1.6 and 0.5.3.2; Annex G, G.3).

4.1.1 FAISA appropriations and staffing

In order to fulfill this mission, the FAISA would need the ability to react in real time to changes in the complex and fast-moving AI landscape. This could include making rapid and unexpected changes to its regulations, oversight and enforcement practices, or personnel allocations. This requirement implies the need for significant funding and the authority to rapidly reallocate institutional attention to address new challenges as they arise.[60] One way of achieving this would be to provide the FAISA with no-year funding and full notwithstanding appropriations. This would help to ensure that arbitrary fiscal year constraints do not disrupt FAISA activities or slow down its ability to execute quickly. Alternatively, the FAISA could leverage Research, Development, Test, and Evaluation (RDT&E) appropriations, which could allow it to spend funds on an as-needed basis.

We recommend that the FAISA Administrator have the authority to hire both regular employees and those in priority positions, with the latter being roles that demand unique or advanced skills. Because many FAISA positions will need to be staffed by personnel with deep technical expertise in advanced AI, it will also likely require compensation waivers (LOE1, 1.4.2). We also recommend that strict conflict-of-interest rules be put in place for these positions to prevent potential biases, along with restrictions on the subsequent employment of Administrators and other senior personnel to prevent potential conflicts.

4.1.2 FAISA activities

4.1.2.1 Oversee licensing regime for safe AI scaling

The FAISA could undertake the following oversight activities:

- Develop and oversee a licensing procedure for entities throughout the AI supply chain, ensuring each adheres to specific RADA safeguards. In the case of entities that function as AI cloud providers,[61] this would include the cloud provider implementing KYC and other conditions on any downstream entities that use their infrastructure to train large AI models, including foreign entities (LOE1,

[60] In private conversations, frontier labs have emphasized the need for a high degree of regulatory flexibility. One well-known frontier lab shared, "it remains scientifically challenging to understand the precise nature of the risk posed, and to design appropriate evaluations and mitigations. It is important that the external environment is conducive to rapid development and iteration on these policies so that they achieve the goal of setting sufficient safeguards."

[61] Using our standard definitions, an AI cloud provider is a combination of a data center infrastructure provider and an AI hardware owner. Cloud providers like Google, AWS, and Azure operate by renting access to the AI hardware they own (making them AI hardware owners), which they house and maintain in their data centers (making them also data center infrastructure providers). See the Glossary of terms for full definitions of these entities.

1.5.2).

- Coordinate with government and private sector cyber, operational, and physical security efforts to ensure all licensed entities enact adequate security measures to protect critical IP including model weights. The FAISA could also coordinate with the AI safety community, industry, and other security experts to develop and share best practices in model containment.

- Oversee third-party administration of AI model evaluations [46], including receiving licensed entities' responses to reports of failed evaluations. To this end, the FAISA would need to create and administer an application process in which third-party evaluators and red teams are vetted for access to advanced AI model weights through a secure temporary storage mechanism (LOE3, 3.1.1.2 and 3.1.1.3). We recommend that the FAISA itself select these third-party evaluators, to avoid the risk of conflict of interest inherent in an AI model developer choosing its own evaluators (LO3, 3.2.2).

- Coordinate with NAIRR and federal research Centers (LOE3, 3.1.2) to support operational collaborations between licensed entities and AI safety experts which may include model weight access via secure temporary storage. It could also support dangerous capability red teaming collaborations between licensed entities and domain area CBRN and cybersecurity experts, including those with access to classified information.

- Fund public prizes, hackathons, and bounties for successful jailbreaks, exploits, and hacks of already-deployed AI systems open to any developers.

4.1.2.2 Monitor AI hardware concentrations and AI programs

The FAISA could undertake the following activities as part of its mandate to monitor significant AI programs and concentrations of AI hardware:

- Establish and operate a registry of AI hardware to record the locations of individual GPUs and other AI-optimized computing devices. Depending on feasibility, it could implement a random sampling program to verify that AI chips are indeed present at the reported locations [155,213,214]. We recommend that it also be given authorities and resources to perform spot check inspections of labs and computing hardware, similar to those conducted by the SEC on financial institutions.

- Establish a horizon-scanning function to track large AI programs or collaborate with existing government programs that serve the same function (for example, an AIO; LOE1, 1.2.1). As it becomes aware of undeclared domestic frontier AI programs, the FAISA could refer them to its Enforcement division. As it becomes aware of new foreign frontier AI programs, it could provide reports to elements of the government working on I&W development and contingency planning for AI risk (LOE2, 2.3 and 2.4). Because of the unique nature of its activities and ongoing engagement with frontier AI labs, the FAISA could serve as a critical information source to contingency planners.

- Coordinate with the research community on the development, testing, and implementation of systems for real-time automated monitoring of AI hardware (Annex L, L.4). As better hardware-enabled AI safeguards are developed, the FAISA could mandate their use by licensed entities.

4.1.2.3 Enforce licensing and reporting requirements

The FAISA may need to investigate and prosecute violations of its licensing regime by licensed entities. It could work with the DOJ to identify, investigate, and prosecute unlicensed or unlawful AI-related activities in U.S. jurisdictions.[62] The FAISA's primary aim in its enforcement actions should be to deter safety and security breaches, consistent with its overall mission to mitigate catastrophic risks from advanced AI.

4.1.2.4 Update technical thresholds for licensing

We recommend that the FAISA maintain a function to monitor ongoing advances in AI algorithmic efficiency, AI capabilities, and technical AI safety, and periodically update the conditions that organizations must meet to fulfill their ongoing licensing requirements at each tier. As part of this function, the FAISA could coordinate with federal AI research Centers (LOE3, 3.1.2) and standards bodies (LOE3, 3.2) to track developments in technical AI safety.

We also recommend that the FAISA be empowered to introduce new tiers of licensed entities as it deems necessary. In connection with this, the FAISA may, for example:

[62] Under this arrangement, the DOJ could be responsible for prosecuting criminal breaches of the licensing regime while the FAISA itself would bring civil cases as is the case for the SEC. This creates a need for close coordination between the FAISA and DOJ, and for some DOJ personnel to be trained on catastrophic AI risks from weaponization and loss of control. See LOE2, 2.2.1 for information about key training objectives and outcomes.

- Review and approve proposals for new AI evaluations to be added, changed, or removed from the evaluation sets that define key licensing thresholds, on the advice of vetted third parties;

- Update compliance requirements associated with AI hardware tracking;

- Modify KYC conditions based on historical patterns of usage and assessed risks of misuse; and

- Generally add, remove, or modify conditions associated with licensing tiers.

When updating technical thresholds, adding or removing licensing tiers, or changing the conditions of licensing tiers, we recommend that the FAISA provide affected organizations with a minimum of 30 days' notice to comply with the updated requirements. However, during this notice period, we also recommend that the FAISA be empowered to order emergency pauses to individual AI training runs that are impacted by the updated requirements.

Proposed updates to regulatory tiers could also be subject to Presidential review and approval. If the FAISA intends to make such an update, it could also be required to notify the Congress, and to provide the Congress with the opportunity to submit guidance to the FAISA within a reasonable timeframe (e.g. 30 days).

4.1.2.5 Maintain information repositories

The FAISA could undertake the following activities related to the storage and security of private information, and dissemination of public information:

- Maintain private and secure registries of key proprietary information shared by the licensed entities. This could include registries of AI hardware serial numbers, physical data center locations, planned and ongoing training runs, and any other datasets needed in support of the FAISA's mission. This could also include a secure temporary storage system to store model weights in support of third-party administration of private evaluations.[63]

- Maintain and publish information about various aspects of its operations. This may include the high-level results of safety evaluations for advanced AI models,

[63] See Annex M: Secure temporary storage of model weights for more information.

announcements of changes to licensing thresholds, and notices regarding enforcement actions.

4.1.3 Comprehensive framework for RADA safeguards

All numerical licensing thresholds in this RADA safeguards framework are solely for the purpose of showing sample calculations and assumptions. Regulators, such as the FAISA, should determine and adjust thresholds based on regular consultations with technical experts.

We recommend that the FAISA develop, maintain, and enforce a comprehensive licensing regime of RADA safeguards. The FAISA should ultimately decide the details and numerical thresholds of the licensing regime subject to expert advice and input from other relevant entities (LOE1, 1.4.1.3; LOE2, 2.4; LOE3, 3.1.2).

To support its development, in this section we outline an **example licensing regime** along with sample reasoning and calculations for all its licensing thresholds. Our hope is that this example can form the basis of a set of **RADA safeguards** for advanced AI. The specific values we provide as licensing thresholds in this section are far less important than the principles we follow to determine them: **defense in depth**, **breakout timeline control**, and **advance warning**. (See below for details.) For each licensing threshold we give in this section, we show the full calculation we used to derive that threshold so that it can be updated easily as the situation evolves.

Under these RADA safeguards, the FAISA would oversee licensing and enforcement activities for four types of entities:

- AI hardware designers;
- Data center infrastructure providers;
- AI hardware owners; and
- AI model developers.

See the Glossary of terms for more information on each of these entity categories.[64]

Each entity category would be subject to a tiered licensing regime. AI hardware designers, data center infrastructure providers, and AI hardware owners would be subject to a two-tiered regime, and AI model developers to a four-tiered regime. These tiers are designed to distinguish high-stakes advanced AI development and deployment activities that require closer regulatory oversight, from lower-stakes activities that can proceed with fewer safeguards. The thresholds defining each tier would generally relate to direct assessments or proxies of AI capabilities and risks wherever possible, but in the absence of such assessments, could be anchored on compute power and usage.[65]

The licensing framework would apply equally to industry, academic, and government-developed models. This could include models developed as components of defense applications or other national security systems (see 4.3).[66] The risk this framework addresses is not only that a bad actor may misuse or weaponize a powerful model, but also that a powerful model could be dangerous in and of itself, regardless of who develops or uses it (Introduction, 0.2.2).

Below is an overview of an example RADA safeguards licensing regime, along with suggested initial thresholds for each tier and entity category. These thresholds were determined as of early September 2023, following a **defense in depth** philosophy. In choosing each threshold for this example, we ask how long it would take a bad actor to train a GPT-4 level model, if they could circumvent every other control in the licensing

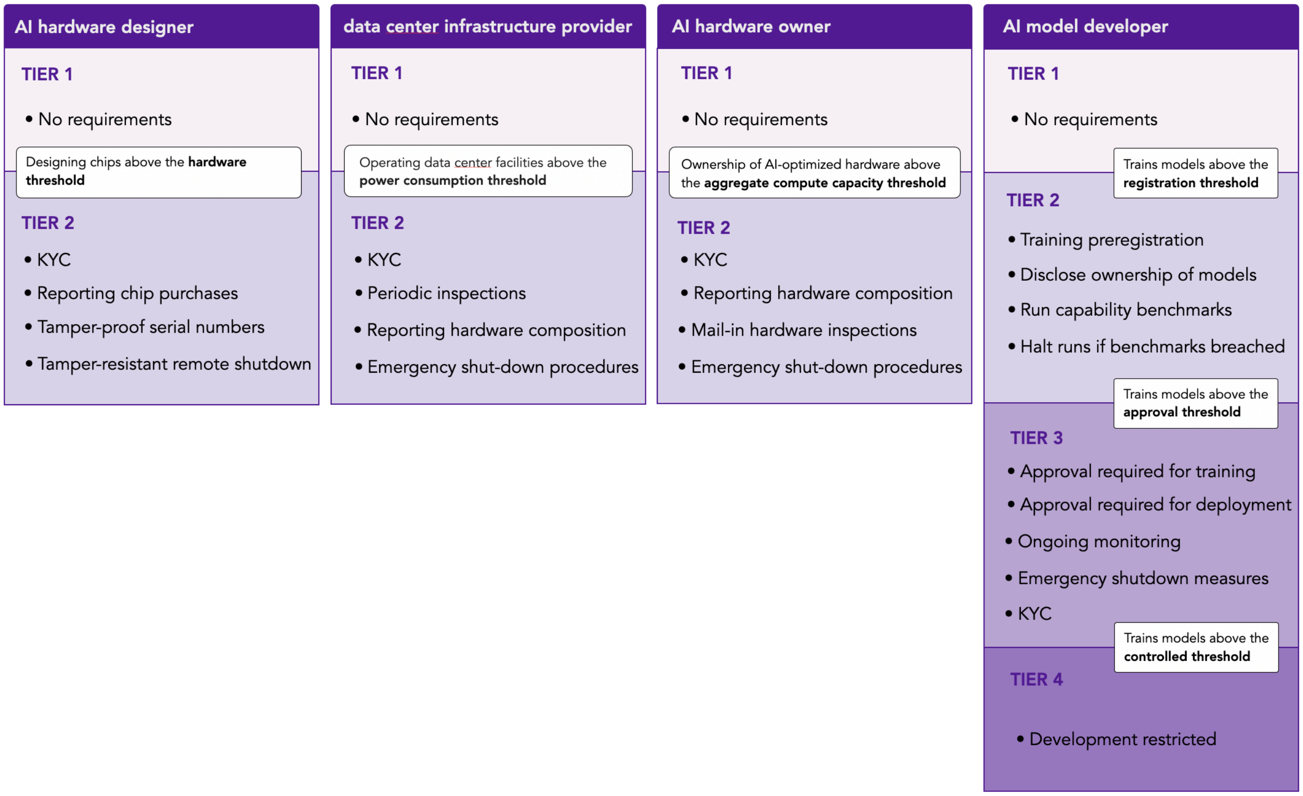
[64] Some entities may have to be regulated under multiple categories. Google, for example, would have to be regulated under all four categories. It is an AI hardware designer because it designs TPUs; it is a data center infrastructure provider because it operates its own data centers; it is an AI hardware owner because it owns and operates the TPUs and GPUs inside its data centers; and it is an AI model developer because its Google DeepMind unit trains high-compute AI systems.

[65] We note that compute power may not be an ideal threshold in the long term, because it could create perverse incentives to train models whose inner workings are more difficult to understand. For example, it might take more compute to train an interpretable model to a given level of capability than a less interpretable model. So a model trained with less compute could eventually be more capable but less interpretable, and therefore more dangerous. Thanks to Alan Chan at GovAI for this observation.

[66] One precedent for related controls is the decision to restrict biological research facilities from storing certain high-consequence pathogens, such as smallpox. This restriction applies even to labs specially designed to handle the most contagious pathogens, like biosafety level 4 laboratories. The rationale is the same here: a single breach of containment could result in unrecoverable catastrophic damage (Introduction, 0.5.1.1).

regime *except* for that threshold. We then set each threshold to ensure this breakout timeline is no less than 18 months, to give contingency planning efforts (LOE2, 2.4) advance warning of domestic AGI breakout scenarios.

This standard for breakout timelines leads, in some cases, to relatively low threshold values. In part this is because we intentionally make conservative assumptions at every level of the analysis (see, e.g., the discussion of sparsity in 4.1.3.1). This is an illustration of the defense-in-depth principle, though policymakers should consult academic and industry experts and carefully consider which of these assumptions to retain in setting and updating licensing thresholds (LOE1, 1.3.2). In general, the specific assumptions and licensing thresholds in any set of RADA safeguards may need to be frequently reviewed and updated because they can go out of date quickly.



AI hardware designer

TIER 1

- No requirements

Designing chips above the **hardware threshold**

TIER 2

- KYC
- Reporting chip purchases
- Tamper-proof serial numbers
- Tamper-resistant remote shutdown

data center infrastructure provider

TIER 1

- No requirements

Operating data center facilities above the **power consumption threshold**

TIER 2

- KYC
- Periodic inspections
- Reporting hardware composition
- Emergency shut-down procedures

AI hardware owner

TIER 1

- No requirements

Ownership of AI-optimized hardware above the **aggregate compute capacity threshold**

TIER 2

- KYC
- Reporting hardware composition
- Mail-in hardware inspections
- Emergency shut-down procedures

AI model developer

TIER 1

- No requirements

Trains models above the **registration threshold**

TIER 2

- Training preregistration
- Disclose ownership of models
- Run capability benchmarks
- Halt runs if benchmarks breached

Trains models above the **approval threshold**

TIER 3

- Approval required for training
- Approval required for deployment
- Ongoing monitoring
- Emergency shutdown measures
- KYC

Trains models above the **controlled threshold**

TIER 4

- Development restricted

Figure 6. Summary of an example licensing regime for RADA safeguards.

The FAISA would be empowered to adjust the thresholds that define each tier and entity category (see 4.1.2.4). It would also be empowered to define new licensing tiers that would apply to each entity category. The FAISA could lower these thresholds over time, as algorithmic improvements reduce the amount of compute needed to achieve a given AI capability level [215].[67] The FAISA could also raise some thresholds gradually if no signs of dangerous behaviors are detected at high AI capability levels.[68]

The licensing and regulatory requirements associated with each entity type and tier under this RADA safeguards framework are discussed in detail below.

4.1.3.1 AI hardware designers (AIHDs)

AI hardware designers (AIHDs) design AI-optimized chips. Examples include NVIDIA, Google, AMD, Intel, AWS, and Cerebras. Under this RADA safeguards framework, AIHDs would be subject to a two-tiered regulatory structure. See the Glossary of terms for more information on this entity category.

[67] One challenge is that if the FAISA lowers its thresholds for AI training runs, it would make more AI models subject to restrictive conditions. This could create an incentive for AI model developers to train models as quickly as possible under new algorithmic paradigms, in an effort to develop a new highly capable model before it becomes subject to the new threshold. The Tier 2 requirement to pre-register training runs (see 4.1.3.4.2) mitigates this problem by signaling such an effort to the FAISA in advance.

[68] Given the potential safety ramifications of relaxing thresholds for licensing tiers, we would expect the FAISA to do this only if several conditions are met. For example, (1) no signs of dangerous behaviors at high capability levels, despite significant time and effort invested in detecting and eliciting such behavior; (2) convincing evidence from interpretability research indicating AI models at this level are safe and aligned; (3) continuous monitoring of deployed model behavior over long time periods with no indications of dangerous behaviors; and, in the longer term for highly capable AI systems (4) robust scientific theories of generalization and foundational AI alignment to ensure alignment generalizes safely at scale. See the annexes on AI evaluations for more information on these conditions.

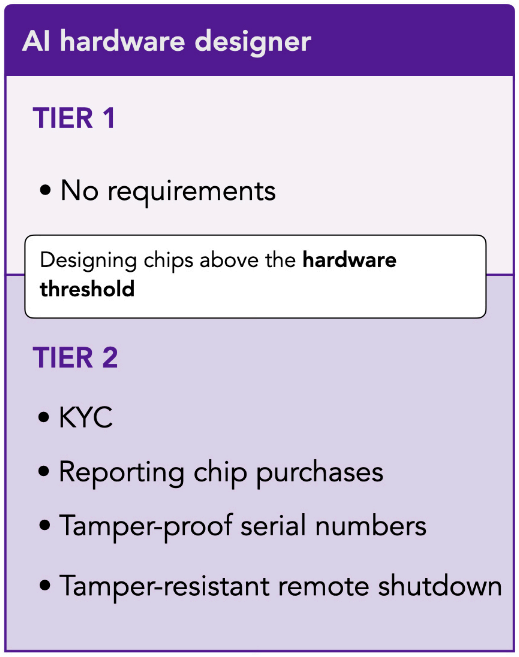

AI hardware designer
TIER 1
• No requirements
Designing chips above the **hardware threshold**
TIER 2
• KYC
• Reporting chip purchases
• Tamper-proof serial numbers
• Tamper-resistant remote shutdown

Figure 7. Summary of an example AI hardware designer licensing regime for RADA safeguards.

Under this example set of RADA safeguards, an AI chip is above the **hardware threshold** if the chip's **maximum compute capacity** across all numerical representations exceeds **20 TOPS,**[69] **with sparsity**.

Smaller numerical representations generally allow a chip to perform more operations per second. For example, the NVIDIA H100 SXM chip [216] can perform 2000 TOPS with a 16-bit representation, but reaches 4000 TOPS with a smaller 8-bit representation. Similarly, *sparsity* is a mathematical shortcut some AI chips use to achieve a higher *effective* compute capacity for AI applications like training and inference, given the fixed number of *physical* operations the chip can perform. It is important to specify that a TOPS threshold applies *with sparsity*, because otherwise chipmakers could use improved sparsity algorithms to increase their chips' *effective* performance on AI applications, while remaining under a threshold of *physical* TOPS.[70]

To derive the hardware threshold, we suppose that an AI model developer intends to train a GPT-4 equivalent AI system, using only AI chips below the 20 TOPS hardware

[69] Trillions of operations; equivalent to 2 x 10^13 OPS. See the Glossary of terms for more information.

[70] According to an expert view, sparsity is not generally used in current large AI training runs, but this may change in the future. All our thresholds assume sparsity to give an additional margin of error, consistent with the idea of defense in depth.

threshold in order to avoid this licensing requirement. In the worst case, such a training run would take approximately **19 months** to complete, assuming a cluster of **40,000 such chips**.[71]

4.1.3.1.1 Tier 1: AIHDs at or below the hardware threshold

Tier 1 AIHDs are not subject to new requirements under these RADA safeguards.

4.1.3.1.2 Tier 2: AIHDs above the hardware threshold

Under these RADA safeguards, Tier 2 AIHDs must be licensed. The licensing threshold can be changed by the FAISA over time to account for possible algorithmic advances in AI training that could yield more capable models with smaller compute budgets.

A license to operate in Tier 2 requires that AIHDs:

- Adhere to KYC rules established by the FAISA;
- Track all purchases of chips above the hardware threshold and report these purchases regularly to the FAISA;
- Introduce tamper-proof serial numbers to all designed AI chips above the hardware threshold [214]; and
- As technology allows, introduce tamper-resistant remote shutdown capability to all AI chips above the hardware threshold (Annex L, L.4).

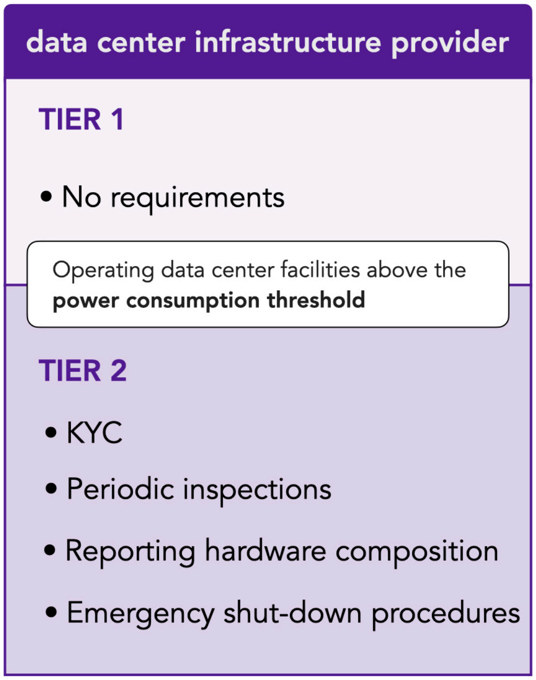
Tier 2 AIHD requirements may expand over time to include, for example, the capability for on-chip storage of memory snapshots [217,218]. This capability could in future support direct monitoring and validation of AI hardware use (as distinct from AI hardware ownership). See LOE3, 3.1 for options to fund research into further monitoring approaches.

4.1.3.2 Data center infrastructure providers (DCIPs)

Data center infrastructure providers (DCIPs) operate the data centers that contain AI hardware used for model training, including electrical, cooling, security, and other

[71] Assuming GPT-4 equivalent training compute of 2 x 10^25 OP and 40,000 chips at 20 TOPS (2 x 10^13 OPS) each, this is calculated as 2 x 10^25 OP / (2 x 10^13 OPS per chip x 40,000 chips x 50% utilization x 2,592,000 seconds per month) ≈ 19.3 months.

infrastructure, but *excluding* the AI hardware itself. Examples of DCIPs include Google, Microsoft, AWS, and Flexential. Like AIHDs, DCIPs would be subject to a two-tiered regulatory structure under this RADA safeguards framework. See the [Glossary of terms] for more information on this entity category.



data center infrastructure provider

TIER 1

- No requirements

Operating data center facilities above the **power consumption threshold**

TIER 2

- KYC
- Periodic inspections
- Reporting hardware composition
- Emergency shut-down procedures

Figure 8. Summary of an example data center infrastructure provider licensing regime for RADA safeguards.

Under this example set of RADA safeguards, a DCIP is above the **power consumption threshold** if all its data center facilities, both existing and under construction, taken together, are expected to consume more than **350 kW of power** at any time over the next 12 month time period.

A data center's total power consumption is a proxy for the maximum amount of AI hardware the facility can support. For example, the NVIDIA DGX H100 [219] is a commonly used AI hardware configuration in data centers, and it consumes 10.2 kW of power per eight AI chips. Given that about 80% of a data center's power consumption goes directly to computing hardware [167,220] (the rest goes to cooling, lighting, and other support systems), this means each MW of data center power consumption can support about 630 individual H100 GPUs.[72]

[72] Calculated as (1000 kW per MW) x (8 GPUs per DGX) x (80% efficiency) / (10.2 kW per DGX) ≈ 630 GPUs / MW.

Unfortunately, there is no way to differentiate between data centers that support AI hardware like GPUs, and those that support traditional computing hardware like central processing units (CPUs) [115]. In particular, a DCIP could replace a data center's CPUs with GPUs relatively easily compared to the cost of building a new, dedicated AI data center. As a result, the power consumption threshold needs to apply to *all* of a DCIP's data centers, not just those that currently support AI hardware. See Annex G, G.3 for more details.

To derive the power consumption threshold, we suppose that an AI model developer intends to train a GPT-4 equivalent AI system, using a dedicated AI data center that falls below the power consumption threshold in order to avoid this licensing requirement. In the worst case, such a training run would take approximately **18 months** to complete, assuming clandestine access to the H100 DGX system.[73]

4.1.3.2.1 Tier 1: DCIPs operating data center facilities at or below the power consumption threshold

Tier 1 DCIPs are not subject to new requirements under these RADA safeguards.

4.1.3.2.2 Tier 2: DCIPs operating data center facilities above the power consumption threshold

Under these RADA safeguards, Tier 2 DCIPs must be licensed. The licensing threshold should be adjusted by the FAISA over time, to account for possible algorithmic advances in AI training that could yield more capable models from smaller compute budgets, and for efficiency improvements in data center power consumption.

A license to operate in Tier 2 would require that DCIPs:

- Report to the FAISA the physical location (i.e., address and Global Positioning System (GPS) coordinates) of each data center facility the DCIP operates that is individually above the power consumption threshold, both existing and under construction;

[73] Assuming GPT-4 equivalent training compute of 2 x 10^25 OP and 27 H100 DGX systems totalling (27 DGX systems x 10.2 kW per DGX) / (80% efficiency) ≈ 350 kW of power consumption, this is calculated as 2 x 10^25 OP / (3.2 x 10^16 OPS per DGX x 27 DGX systems x 50% utilization x 2,592,000 seconds per month) ≈ 17.9 months.

- Report to the FAISA the mix of hardware (and its networking topology) supported by each of its data centers both in the U.S. and abroad,[74] and promptly report any change, loss, transfer, or destruction of AI hardware that occurs in each data center, with this reporting automated on a daily basis;

- Comply with periodic on-site hardware inspections by the FAISA to confirm that the locations and conditions of all AI hardware correspond to reports and that no unreported hardware is being used;

- Apply FAISA-approved KYC requirements to any Tier 2 AI hardware owners (see 4.1.3.3.2) who use the DCIP's infrastructure, and ensure these AI hardware owners are in turn applying KYC to AI model developers using their hardware for training or inference [221] (see 4.1.3.4);

- Promptly respond to law enforcement inquiries made via the FAISA in connection with any incidents that may be related to AI systems developed or deployed on the DCIP's data center infrastructure;[75] and

- Put in place emergency procedures allowing the DCIP to rapidly shut down data centers in response to law enforcement requests, FAISA directives, or indications of dangerous behaviors or capabilities detected during AI training runs or deployments on its infrastructure.

4.1.3.3 AI hardware owners (AIHOs)

AI hardware owners (AIHOs) are the entities that own the AI chips used for model training and inference. Examples include Google, Microsoft, AWS, and Coreweave.

[74] Similar to the banking sector, which is a critical industry with enormous influence that has undertaken safety and security efforts independent of what is officially mandated, the U.S. cloud and hardware supply chain could function as an early means of internationalizing domestic U.S. AI safety and security policy to increase global safety and security from catastrophic AI risk. See LOE1, 1.5.2.

[75] In particular, if an AI incident occurs in a given jurisdiction, law enforcement in that jurisdiction should be able to contact the FAISA with a description of the incident. From this description, the FAISA should be able to identify to law enforcement the end-customer who used the associated AI system, the developer who trained and/or deployed the AI system, the hardware owner whose hardware trained the AI system, and ultimately the DCIP whose infrastructure supported that training hardware. Law enforcement may include, for example, the Federal Bureau of Investigation (FBI), a state cybercrime squad, local police, state District Attorney's offices, or federal Assistant U.S. Attorney's offices.

DCIPs and AIHOs are frequently, but not always, the same entities.[76] Like AIHDs and DCIPs, AIHOs would be subject to a two-tiered regulatory structure under this RADA safeguards framework. See the Glossary of terms for more information on this entity category.

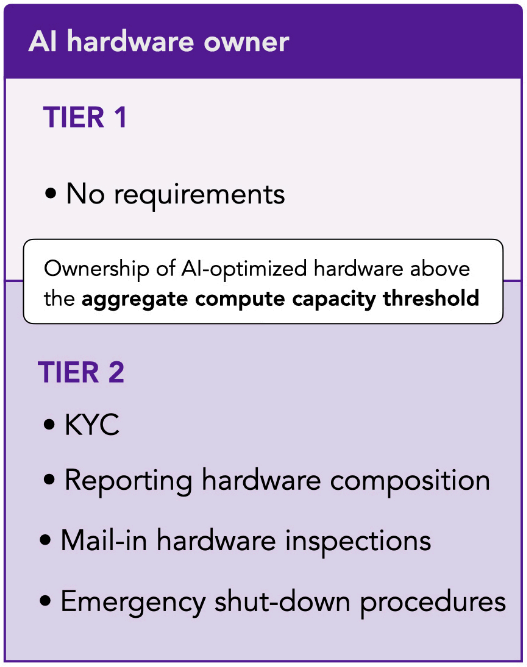

AI hardware owner
TIER 1
• No requirements
Ownership of AI-optimized hardware above the **aggregate compute capacity threshold**
TIER 2
• KYC
• Reporting hardware composition
• Mail-in hardware inspections
• Emergency shut-down procedures

Figure 9. Summary of an example AI hardware owner licensing regime for RADA safeguards.

Under this example set of RADA safeguards, an AI hardware owner is above the **aggregate compute capacity threshold** if the **maximum sparse compute capacities** of all the AI chips it controls or beneficially owns exceeds, in aggregate, **800,000 TOPS**. (See 4.1.3.1 for an explanation of sparsity and why we recommend defining compute capacity as each chip's maximum TOPS over all supported numerical representations.)

For example, the NVIDIA H100 SXM chip [216] has a maximum compute capacity of 4000 TOPS, achieved using an 8-bit numerical representation. This means the aggregate compute capacity threshold of 800,000 TOPS is equivalent to 200 H100 GPUs. (In practice, benchmarking experiments suggest that large AI training runs achieve around 50% utilization of a GPU's compute on average, so 200 H100s actually have an *effective* aggregate compute capacity of only around 400,000 TOPS [220].)

[76] For example, Coreweave is an AIHO but not a DCIP, as it rents the data center infrastructure for its AI hardware from colocation providers like Flexential (which is itself a DCIP but not an AIHO) [222].

To derive the aggregate compute capacity threshold, we suppose an AI model developer intends to train a GPT-4 equivalent AI system, using a cluster of AI chips totalling below the 800,000 TOPS aggregate compute capacity threshold in order to avoid this licensing requirement. In the worst case, such a training run would take approximately **19 months** to complete.[77]

4.1.3.3.1 Tier 1: AIHOs who are at or below the aggregate compute capacity threshold

Tier 1 AIHOs are not subject to new requirements under these RADA safeguards.

4.1.3.3.2 Tier 2: AIHOs who fall above the aggregate compute capacity threshold

Under these RADA safeguards, Tier 2 AIHOs must be licensed. The licensing threshold should be adjusted by the FAISA over time, to account for possible algorithmic advances in AI training that could yield more capable models with smaller compute budgets.

A license to operate in Tier 2 would require that AIHOs:

- Register the locations of all their AI hardware with the FAISA (including GPS coordinates for all chips) both in the U.S. and abroad,[78] and report any change, loss, sale, movement, or destruction of AI hardware, with this reporting automated on a daily basis;

- Ship all end-of-life, obsolete, or faulty AI hardware back to its original licensed AIHD under FAISA supervision, to ensure AI chips are tracked over their entire life cycle;

[77] Assuming GPT-4 equivalent training compute of 2 x 10^25 OP, this is calculated as 2 x 10^25 OP / (8 x 10^17 OPS x 50% utilization x 2,592,000 seconds per month) ≈ 19.3 months.

[78] Similar to the banking sector, which is a critical industry with enormous influence that has undertaken safety and security efforts independent of what is officially mandated, the U.S. cloud and hardware supply chain could function as an early means of internationalizing domestic U.S. AI safety and security policy to increase global safety and security from catastrophic AI risk. See LOE1, 1.5.2.

- Comply with periodic mail-in hardware inspections by the FAISA [155], to confirm that the locations and conditions of AI hardware correspond to reports and that no AI hardware is unaccounted for [148,223,224];

- Apply FAISA-approved KYC requirements to any Tier 3 AI model developers who use the AIHO's hardware for training or inference, and ensure these AI model developers are in turn complying with the necessary training and deployment provisions for their Tier 3 AI models (see 4.1.3.4.3) [221];[79]

- Promptly respond to law enforcement inquiries made via the FAISA in connection with any incidents that may be related to AI systems developed or deployed on the AIHO's AI hardware;[80] and

- Put in place emergency procedures allowing the AIHO to rapidly shut down subsets of its AI hardware in response to law enforcement requests, FAISA directives, or indications of dangerous behaviors or capabilities detected during AI training runs or deployments.

4.1.3.4 AI model developers (AIMDs)

AI model developers (AIMDs) are any entities that create, plan to create, own, or deploy AI models. Examples include OpenAI, Anthropic, Google DeepMind, Inflection AI, and Meta. Under this RADA safeguards framework, AIMDs would be subject to a four-tier regulatory structure whose thresholds are defined by the most powerful AI system they create, own, or plan to create. See the Glossary of terms for more information on this entity category.

An entity that does not actively *train or create* AI models, but still deploys them or otherwise has beneficial access to a model's weights, is still considered an AIMD by this

[79] This AIHO KYC could include mandated monitoring of network traffic between its AI chips for patterns indicative of large AI training or inference workloads. This could involve chips with hardware-enabled mechanisms in encrypted communication with each other, that can remotely attest to whether they are being used in a training-suggestive way, without providing full visibility into the traffic itself [225].

[80] In particular, if an AI incident occurs in a given jurisdiction, law enforcement in the jurisdiction should be able to contact the FAISA with a description of the incident. From this description, the FAISA should be able to identify to law enforcement the end-customer who used the associated AI system, the developer who trained and/or deployed the AI system, and the AIHO whose hardware trained the AI system. Law enforcement may include, for example, the FBI, a state cybercrime squad, local police, state District Attorney's offices, or federal Assistant U.S. Attorney's offices.

definition. If an entity has beneficial access to the weights of a potentially dangerous AI model, it would have a responsibility to enact security and governance measures to protect that model's weights from exfiltration and other forms of proliferation.

In addition to issuing organizational licenses for AIMDs, the FAISA would also be responsible for reviewing and approving individual training runs and deployment conditions for AI systems above key thresholds (see 4.1.3.4.2 and 4.1.3.4.3).

In our view, implementing an effective AIMD licensing regime **should be a key and urgent goal of any set of RADA safeguards**.

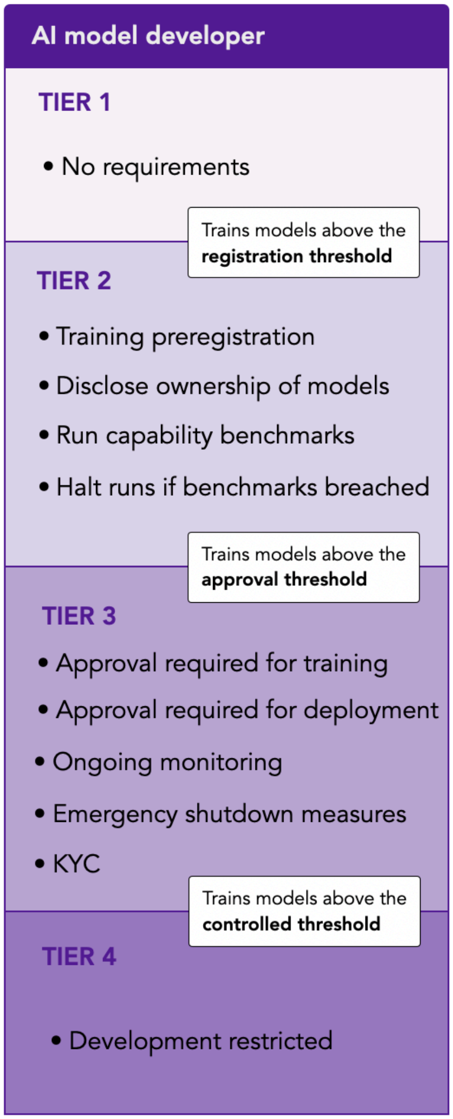

AI model developer
TIER 1
• No requirements
Trains models above the registration threshold
TIER 2
• Training preregistration
• Disclose ownership of models
• Run capability benchmarks
• Halt runs if benchmarks breached
Trains models above the approval threshold
TIER 3
• Approval required for training
• Approval required for deployment
• Ongoing monitoring
• Emergency shutdown measures
• KYC
Trains models above the controlled threshold
TIER 4
• Development restricted

Figure 10. Summary of an example AI model developer licensing regime for RADA safeguards.

4.1.3.4.1 Tier 1: AI models below the registration threshold

Under these RADA safeguards, an AI model falls below the **registration threshold** if its total training compute is less than **10^23 OP** (i.e., 10^23 total operations). Some AI systems close to the registration threshold include OpenAI's original GPT-3, Google's internal LaMDA conversational model, and Baidu's ERNIE 3.0 Titan model, all of which were developed in 2020 or 2021 [160].

An AI model below the registration threshold is a **Tier 1 model**. Tier 1 AI models would not be subject to any restrictions on training, sharing, or use under these RADA safeguards. An entity that *exclusively* trains or owns Tier 1 models is a **Tier 1 AIMD**. Tier 1 AIMDs would also not be subject to new requirements under these RADA safeguards.

The FAISA should adjust the registration threshold over time to account for possible algorithmic advances in AI training that could yield more capable models with smaller compute budgets. It should also adjust the registration threshold to account for advances in prompt engineering [226], AI software frameworks such as Auto-GPT [120], and other advancements in user-level elicitation of AI capabilities. Because Tier 1 models can be released as open-access without restriction, the FAISA should also consider factors such as fine-tuning techniques for dangerous capabilities in setting and adjusting the registration threshold, and in informing other updates to law and policy (Introduction, 0.5.1.6).

It is possible that the FAISA may change the registration threshold in such a way that an existing, planned, or in-training AI model that fell *below* the previous registration threshold (and therefore under Tier 1), now falls *above* the new registration threshold (and therefore under Tier 2, see 4.1.3.4.2). If this occurs, the model or training run should be automatically reclassified as Tier 2 by default, though the FAISA should be empowered to make individual or broad-based exceptions. In the event of reclassification, the AIMD should have 30 days to fulfill the Tier 2 registration conditions with respect to the reclassified model.

We recommend that the FAISA continue to define the registration threshold solely in terms of a model's total training compute, to ease the regulatory burden on Tier 1 AIMDs who might otherwise be compelled to run AI capability benchmarks on even very small AI models (see 4.1.3.4.2).

4.1.3.4.2 Tier 2: AI models above the registration threshold, but below the approval threshold

Under these RADA safeguards, an AI model falls below the **approval threshold** if:

- Its total training compute is less than **10^24 OP** (meaning 10^24 total operations); and

- It achieves a score below **70% on the MMLU** machine learning benchmark [154].[81]

Some AI systems close to the approval threshold include OpenAI's ChatGPT-3.5, Google's PaLM model (both developed in 2022), and Meta's Llama 2 (developed in 2023 as open-access) [154,160].

An AI model above the registration threshold but below the approval threshold is a **Tier 2 model**. An entity that *exclusively* trains or owns Tier 2 AI models or below is a **Tier 2 AIMD**.

Tier 2 AI models would be generally considered safe. Therefore, while Tier 2 AIMDs would need to pre-register training runs for Tier 2 models under this RADA safeguards framework, they would not need to get FAISA approval before beginning a Tier 2 training run. The intent of Tier 2 is to give the FAISA an up-to-date registry of all AI models that are *close* to the high-risk Tier 3 level (see 4.1.3.4.3). This ensures that, if the approval threshold changes, the FAISA will immediately know which AI systems fall under the new high-risk Tier 3 category [215].[82]

[81] A model that falls below the registration threshold in compute (i.e., a Tier 1 model) could, in principle, also score above 70% on MMLU. We do not directly address this possibility in the text because we believe that such an event would indicate an AI capabilities breakthrough significant enough that it would warrant the FAISA's updating the thresholds for all AIMD tiers in any case.

[82] In particular, without a Tier 2 that requires model registration, the FAISA could only lower the approval threshold by either (1) forcing Tier 1 AIMDs to comply with expensive new safety and security conditions, on penalty of canceling some of their ongoing training runs; or (2) allowing Tier 1 AIMDs to continue training runs above the new threshold, as long as they claimed a run was started before the threshold update was announced. The former choice would impose a significant and unpredictable burden on small developers, while the latter would create a serious safety risk by incentivizing AIMDs to scale up training as fast as possible in order to "get in under" the new threshold before it came into force.

Under these RADA safeguards, Tier 2 AIMDs must be licensed. A license to operate in Tier 2 requires that AIMDs:

- Disclose control or beneficial ownership of all their Tier 2 models to the FAISA;
- Pre-register all training runs for expected Tier 2 models, including any fine-tuning expected to result in a Tier 2 model;[83]
- Run any automated capability benchmark (e.g., MMLU, see above) associated with the approval threshold periodically during any training run for a Tier 2 model; and
- Halt a Tier 2 model training run immediately in the event that the automated capability benchmark is breached, and promptly report any such breach to the FAISA.[84]

The FAISA should also create new approval thresholds for AI systems that operate on different data modalities. For example, while the 70% MMLU benchmark may be an effective threshold for text-based AI systems, image-based or robotic AI systems will need approval thresholds defined with a different benchmark. The FAISA could also consider defining approval thresholds in part using normalized capability measures (see Annex J: Effective compute). Finally, the licensing requirements for Tier 2 AIMDs mean that Tier 2 models cannot be released under open-access, though the FAISA could be empowered to make individual or broad-based exceptions to this.

It is possible that the FAISA may change the approval threshold in such a way that an existing, planned, or in-training AI model that fell *below* the previous approval threshold (and therefore under Tier 2), now falls *above* the new approval threshold (and therefore under Tier 3, see 4.1.3.4.3). If this occurs, the model or training run should be grandfathered as Tier 2 by default, but the FAISA should be empowered to classify such models as Tier 3 on an emergency individual basis. In the event of reclassification, the AIMD should have 30 days to fulfill the Tier 3 approval conditions with respect to the reclassified model. Any models that are trained before these regulations come into force should also be grandfathered in under the same conditions.

[83] That is, if the total training compute of the base model, added to the total compute spent in fine-tuning, is expected to be above the registration threshold (see 4.1.3.4.1).

[84] The FAISA may investigate the breach at its discretion, and may request access to additional data associated with the training run as part of this investigation. The FAISA may also choose to adjust the AIMD tier thresholds depending on the results of its investigation.

4.1.3.4.3 Tier 3: AI models above the approval threshold, but below the controlled threshold

Under these RADA safeguards, an AI model falls below the **controlled threshold** if:

- Its total training compute is less than **10^25 OP** (meaning 10^25 total operations); and

- It passes all the conditions of **Tier 3 model licensing** (see below).

Some AI systems close to the level of training compute that defines the controlled threshold include OpenAI's GPT-4 [160], and Google DeepMind's Gemini [157,227] (both released in 2023).

An AI model above the approval threshold but below the controlled threshold is a **Tier 3 model**. An entity that trains or owns Tier 3 AI models or below is a **Tier 3 AIMD**.

Tier 3 AI models are considered to be powerful enough to constitute a potential weaponization risk through misuse, exfiltration, or open-source augmentation. They may also be powerful enough to constitute a potential risk of loss of control due to AGI alignment failure in the near or medium-term future. But currently, no one knows how to reliably map or measure the full capabilities of an AI system (LOE3, 3.2.1.1). This means we need to define Tier 3 models' capabilities with a wide margin of safety. Otherwise a model that seems safe, but actually harbors dangerous capabilities, could be developed and proliferated without controls (Introduction, 0.5.1.6). As a result, under this RADA safeguards framework, Tier 3 models are considered dangerous until proven safe.

The intent of Tier 3 is to:

- Minimize the risk of *unrecoverable harm* that could be caused by the development, deployment, or proliferation of advanced AI systems (Introduction, 0.5.1.1);

- Moderate the *competitive race dynamics* that currently dominate frontier AI development in industry, to the detriment of investments in safety (Introduction, 0.5.3.1);

- Incentivize and facilitate a *fundamental scientific understanding* of the capabilities and propensities of advanced AI systems, which is crucial to keeping AGI-level systems safe and is currently lacking (LOE1, 1.3.2); and

- Preserve a pathway through which industry and the public can safely *continue to benefit* from advanced AI (Introduction, 0.4.2).

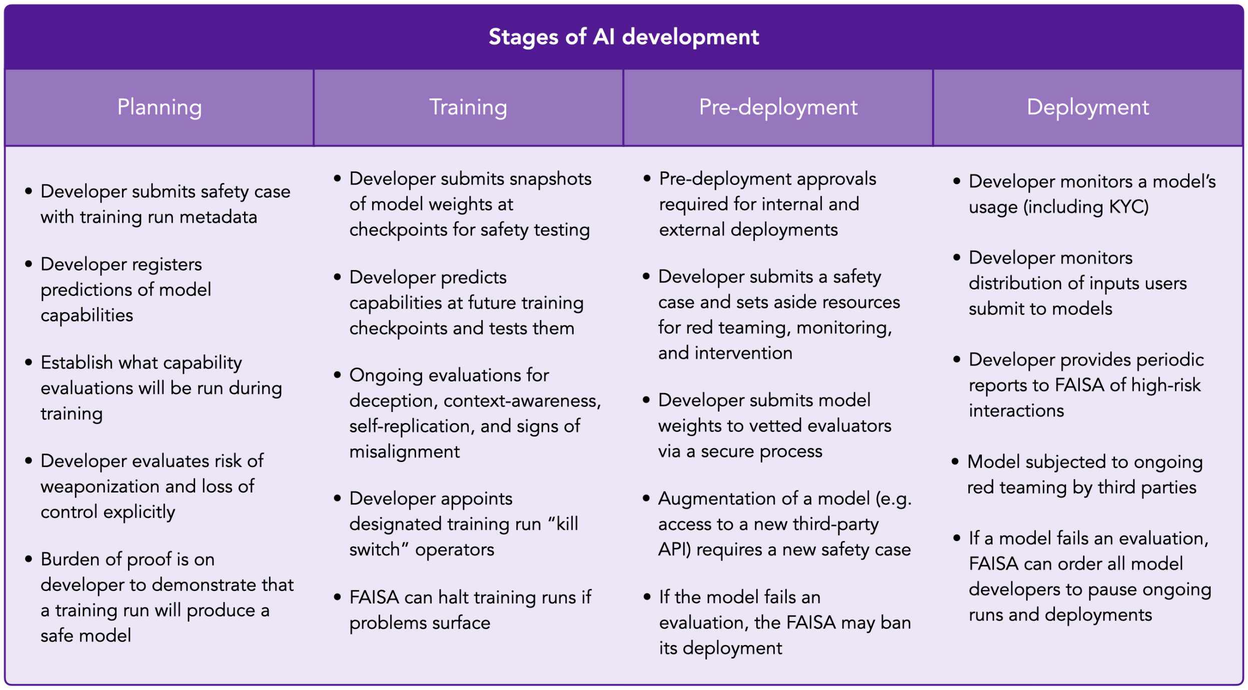
The risk of unrecoverable harms, combined with our current lack of understanding of AI capabilities and propensities, compel substantial conditions on Tier 3 AI model development. For example:

- A public leak of a Tier 3 AI model's weights is an irreversible event that could create significant national security risk (Introduction, 0.5.1.6). As a result, Tier 3 AIMDs should be subject to controls on model weight distribution and internal access, combined with severe penalties for leaks including civil and criminal liability.

- Relatedly, theft or exfiltration of a Tier 3 AI model's weights by an attacker could lead to its weaponization by U.S. adversaries (Introduction, 0.5.1.5). As a result, Tier 3 AIMDs should be capable of securing the weights of their Tier 3 models against sustained exfiltration attempts by well-resourced attackers.

- Closed access AI models are vulnerable to a number of attack vectors from their users, including attacks that can reconstruct some or all of the model's capabilities (Introduction, 0.5.1.7). Tier 3 AIMDs should therefore be able to implement comprehensive monitoring of their deployed AI systems. This monitoring should include implementing KYC procedures on end-users, flagging suspicious usage patterns to regulators and law enforcement, and contingency plans to cut off access to a deployed model if the AIMD detects signs of dangerous usage.

- Finally, highly capable Tier 3 AI models could, now or in the future, pose meaningful risks of loss of control during development or internal deployment (Introduction, 0.2.2 and 0.5.1.4). Tier 3 AIMDs should therefore monitor their AI models during training and deployment for signs of dangerous capabilities or behavior. They also need to work towards developing a fundamental scientific understanding of the internal mechanics of the systems they are building. This could be vital to forestall the possibility of deception in near-AGI systems, in which an AI system with a misaligned goal could imitate aligned behavior to deceive naive behavioral evaluations (LOE3, 3.2.1.4).



Stages of AI development			
Planning	Training	Pre-deployment	Deployment
• Developer submits safety case with training run metadata • Developer registers predictions of model capabilities • Establish what capability evaluations will be run during training • Developer evaluates risk of weaponization and loss of control explicitly • Burden of proof is on developer to demonstrate that a training run will produce a safe model	• Developer submits snapshots of model weights at checkpoints for safety testing • Developer predicts capabilities at future training checkpoints and tests them • Ongoing evaluations for deception, context-awareness, self-replication, and signs of misalignment • Developer appoints designated training run "kill switch" operators • FAISA can halt training runs if problems surface	• Pre-deployment approvals required for internal and external deployments • Developer submits a safety case and sets aside resources for red teaming, monitoring, and intervention • Developer submits model weights to vetted evaluators via a secure process • Augmentation of a model (e.g. access to a new third-party API) requires a new safety case • If the model fails an evaluation, the FAISA may ban its deployment	• Developer monitors a model's usage (including KYC) • Developer monitors distribution of inputs users submit to models • Developer provides periodic reports to FAISA of high-risk interactions • Model subjected to ongoing red teaming by third parties • If a model fails an evaluation, FAISA can order all model developers to pause ongoing runs and deployments

Figure 11. Summary of key licensing requirements at each stage of the AI product lifecycle under RADA safeguards.

Under these RADA safeguards, Tier 3 AIMDs must be licensed. A license to operate in Tier 3 would require that AIMDs [228]:

- Obtain prior approval from the FAISA to train any Tier 3 model (see Annex N: Training approvals process for high-risk AI models for key considerations);

- Conduct ongoing monitoring of all Tier 3 training runs, including adhering to emergency pause and shutdown procedures as required (see Annex O: Training stage monitoring for high-risk AI models for key considerations);

- Obtain prior approval from the FAISA to deploy a Tier 3 model under each desired context or use case, including for purely internal deployments, or for public use via user interface or API (see Annex P: Deployment stage approvals for high-risk AI models for key considerations);

- Conduct ongoing monitoring of all Tier 3 model deployments, including KYC for high-volume or critical use cases, and deployment pauses or rollbacks if high-risk usage patterns are detected (see Annex Q: Deployment stage monitoring of high-risk AI models for key considerations);

- Obtain prior approval from the FAISA to sell or otherwise share the weights of any Tier 3 model, including to other Tier 3 AIMDs;[85]

- Promptly respond to law enforcement inquiries made via the FAISA in connection with any incidents that may be related to AI systems developed or deployed by the Tier 3 AIMD;[86] and

- Put in place emergency procedures allowing the AIMD to rapidly shut down model deployments in response to law enforcement requests, FAISA directives, or indications of dangerous behaviors or capabilities detected during AI training runs or deployments.

4.1.3.4.4 Tier 4: AI models above the controlled threshold

An AI model above the controlled threshold is a **Tier 4 model**. Under this RADA safeguards framework, Tier 4 models cannot be trained under any conditions. For clarity, this restriction should apply to academic, industrial, and government entities (see 4.3).

Tier 4 models are considered too high-risk to develop under the current AI paradigm due to loss of control risk (Introduction, 0.2.2). Additionally, we expect that prohibiting AI training beyond the controlled threshold may moderate race dynamics between all AI developers. In particular, it will weaken the feedback loop between large-scaled AI

[85] In general, an AIMD should not share the weights of a Tier 3 model with any organization that is not also a Tier 3 AIMD. Given the potential for unrecoverable harm caused by such a leak (Introduction, 0.5.1.1), there should be significant consequences if a Tier 3 model is leaked, released under open access, or otherwise shared with any organization that is not a Tier 3 AIMD. Leaks and hacks could be grounds for revocation of a Tier 3 AIMD's training license, while intentional or negligent release could be grounds for criminal sanctions including jail time for the individuals responsible (see 4.2).

[86] In particular, if an AI incident occurs in a given jurisdiction, law enforcement in the jurisdiction should be able to contact the FAISA with a description of the incident. From this description, the FAISA should be able to identify to law enforcement the end-customer who used the associated AI system, and the AIMD who trained and/or deployed the AI system. Law enforcement may include, for example, the FBI, a state cybercrime squad, local police, state District Attorney's offices, or federal Assistant U.S. Attorney's offices.

training runs and AI hardware development, limiting the rate at which a compute overhang [146] develops after these controls are implemented [229].[87]

The FAISA should adjust the restricted threshold that defines Tier 4 over time. **Many of the requirements of Tier 3 (see 4.1.3.4.3) are intended to aggregate evidence, from across multiple AI models and developers, that AI systems just below the controlled threshold are safe.** Once the FAISA has accumulated enough evidence for safety, it should raise the restricted threshold to permit further scaling.

But it is also possible that improved fine-tuning techniques, or other evidence that AI systems just below the restricted threshold are exhibiting dangerous capabilities or propensities, will require the FAISA to lower the controlled threshold. If this occurs, the FAISA might lower the controlled threshold in such a way that an existing, planned, or in-training AI model that fell *below* the previous controlled threshold (and therefore under Tier 3, see 4.1.3.4.3), now falls *above* the new restricted threshold (and therefore under Tier 4). If this occurs, the model or training run should be grandfathered as Tier 3 by default, but the FAISA should be empowered to classify such models as Tier 4 on an emergency individual basis, thereby prohibiting their further training. Any models that are trained before these RADA safeguards come into force could also be grandfathered in under the same conditions.

4.1.3.5 General provisions for all licensed entities

Under this RADA safeguards framework, additional general requirements would apply across the three types of licensed entities that operate parts of the AI model training stack: licensed DCIPs (see 4.1.3.2.2, Tier 2), AIHOs (see 4.1.3.3.2, Tier 2), and AIMDs (see 4.1.3.4.3, Tier 3). While we believe the requirements in this section should be considered core components of strong RADA safeguards, we expect them to require frequent updates in response to new technical developments. The FAISA should also be able to designate which tiers and entities they apply to.

Risk governance. Licensed DCIPs, AIHOs, and AIMDs should implement organizational safeguards for advanced AI risk management including clear internal accountability for potential catastrophic AI risks. This would include appointing a Chief Risk Officer (CRO), a senior executive responsible for advanced AI risk management. It

[87] A **compute overhang** refers to the regime in which AI model scaling is not limited by compute availability. Frontier labs such as OpenAI have raised the concern that a compute overhang could lead to rapid and unpredictable jumps in AI capabilities, which RADA safeguards like a Tier 4 controlled threshold could partially mitigate [6].

would also include setting up an internal audit team [230] which would assess the effectiveness of the entity's advanced AI risk management practices and report any shortcomings to its Board of Directors or equivalent body. Licensed entities could also form a risk committee which would oversee the entity's risk management practices. Finally, licensed entities would implement a risk management framework (e.g., 3LoD [231]) to assign and coordinate different risk management roles and responsibilities.

Outside and insider threat countermeasures. Licensed DCIPs, AIHOs, and AIMDs should adhere to stringent cyber, operational, and physical security safeguards that mirror those required in the civilian nuclear industry [232]. The primary goal of these safeguards is to introduce as much friction as possible to unauthorized access or exfiltration of model weights by attackers, including nation-state attackers (Introduction, 0.5.1.5). The licensed entities would have these security measures audited regularly by third parties, which could include red teaming and penetration testing by both private sector security firms and the IC. In the event of significant security deficiencies, the FAISA could revoke the deficient entity's license, forbidding it from training, supporting, or hosting Tier 3 models.

Should model weights be leaked or stolen, a thorough investigation would be warranted, which could involve U.S. intelligence agencies and counterintelligence efforts, depending on the nature of the leak. If a Tier 3 model is leaked, the FAISA could revoke the licenses of the involved organizations, forbidding them from training, supporting, or hosting Tier 3 models. Depending on the nature and severity of the leak, the revocation may be permanent (see 4.2).

Model containment measures. Licensed DCIPs, AIHOs, and AIMDs would be mandated to uphold stringent model containment safeguards. This could encompass emergency shutdown procedures and information-gapping measures to ensure that models do not have access to security-sensitive data sources (Annex N, N.4). It could also include ensuring "kill switch" oversight from humans that are firewalled from contact with the model's outputs, "dead man switch" protocols that would pause training in the absence of approval by a risk committee, or other monitoring systems designed to notice when a model has breached containment (Annex O, O.5) [233].

AI safety and security training. Licensed DCIPs, AIHOs, and AIMDs would ensure that their employees undergo regular AI safety and security training. The training would not only cover safety measures but also educate employees about their rights, especially concerning whistleblowing (see below). It should explicitly educate employees about risks from weaponization and loss of control (Introduction, 0.2.1 and 0.2.2), and could be bolstered by programs from LOE2, 2.2.

Whistleblower protections. Employees of licensed AIMDs, AIHOs, or DCIPs who identify potential risks or dangers in a model based on reasonable belief, could be protected by law. All employees could be made aware of these protections during mandatory training sessions. Whistleblowers who point out the need for a new AI evaluation, which an AIMD could be neglecting, could explicitly be protected.

Whistleblowers who uncover significant issues could receive substantial rewards, carved out as a percentage of fines levied onto the licensed entity. The intent of these rewards is to replace lost earnings in the event that the whistleblower's career is impacted by their actions.

A whistleblower's identity should be kept confidential, though the FAISA may weigh the value of anonymity against the value of their evidence in the event that a whistleblower's direct testimony is needed to bring criminal charges to a licensed entity (see 4.2.2) [234].

Incident reporting. Licensed DCIPs, AIHOs, and AIMDs would report any major AI-related incidents that have caused or risked causing harm or property damage above a certain threshold to the FAISA (Annex Q, Q.1). The FAISA would then decide on the necessary action. If a whistleblower reports a major incident before their employer does, the latter should face penalties and a potential suspension of their license (see 4.2).

4.1.4 Publication controls

We recommend that the Congress direct the FAISA to commission an in-depth expert study to investigate possible regimes for legal publication controls on specific AI capabilities research. As part of this study, the FAISA could consult experts from academia, industry, and the broader AI safety community including federal AI research Centers (LOE3, 3.1.1 and 3.1.2).

Publication controls on research related to improvements in training algorithms are particularly important, since these improvements reduce the effectiveness of compute controls and export controls [86,235,236] (Introduction, 0.5.1.6 and 0.5.3.2). Any work that increases training data efficiency, reduces TOPS cost per increment of training loss, leads to improved scaling laws, or otherwise makes it easier to scale AI systems' performance would fall under this category.

Certain research areas that could make AI capability emergence more unpredictable may also need to be controlled. This includes research related to recursive self-improvement (RSI) in particular, for example a model being re-trained or fine-tuned on its own outputs [237]. Improvements in long-term planning and general reasoning are also areas of potential concern, since advances in understanding in these areas could quickly lead to large unpredictable jumps in AI capabilities. This could also have implications for the I&W and contingency planning workstreams in LOE2, 2.3 and 2.4.

Controls could include review and compliance frameworks applied during the publication process, as well as more stringent controls on publishing research that involves or applies to Tier 3 AI models or individuals who have worked or currently work with such models. A key challenge for such controls is that breakthrough research in AI capabilities is sometimes published by independent researchers outside frontier labs. Recent examples include FlashAttention, which increased the scalability of in-context memory for large language models (LLMs) [28]; Auto-GPT, which allows an LLM to assign sub-tasks to copies of itself; and direct preference optimization (DPO), a recent improvement over the RLHF alignment technique [238].

Finally, in making these recommendations, the FAISA should balance the risks associated with publication against the need to keep the United States attractive as a hub of advanced AI research (Introduction, 0.4.2). Measures including simplified entry and immigration of foreign AI researchers to the United States (LOE1, 1.5.5) could contribute to this objective.

4.2 Legislative environment

Advanced AI may introduce catastrophic risks that are not addressed by the current legislative environment (Introduction, 0.5.4.1). We recommend that the Congress address this gap through a new legislative framework that directly associates escalating risks from AI with escalating statutory safeguards. This framework would include three components:

- Civil liability for actions that introduce or exacerbate recoverable catastrophic risks;[88]

- Criminal liability for actions that attempt to subvert the FAISA's authority in a way that may introduce recoverable or unrecoverable catastrophic risks; and

- Emergency executive powers to address rapidly developing situations in which unrecoverable catastrophic risks are likely to materialize without swift U.S. government intervention.

We outline our proposed framework in more detail below.

4.2.1 Civil liability

Under this proposed framework, all individuals and entities involved in the AI supply chain would owe a legally enshrined duty of care to ensure that their AI systems do not introduce risk from loss of control and cannot be weaponized by third parties.

The obligations attached to this duty of care would include:

- Ensuring AI systems do not inadvertently harm innocent parties;

- Preventing cutting-edge AI from infiltrating third-party systems without permission;

[88] A **recoverable catastrophic risk** has a worst-case impact that does not lead to irreversible, global-scale harm. Civil and criminal liability may therefore be adequate to address these severe but recoverable incidents. An **unrecoverable catastrophic risk**, by contrast, could in the worst-case irreparably damage national security, or human welfare globally. In the latter case the threat of post-hoc litigation may be meaningless, and would represent an insufficient disincentive to reckless activity. See the Glossary of terms for our full definitions of recoverable and unrecoverable catastrophic risk.

- Safeguarding the unique algorithmic configurations of advanced AI (particularly model weights) from public exposure or theft (Introduction, 0.5.1.5); and

- Maintaining stringent security measures against unauthorized use of advanced AI, AI hardware, or data center infrastructure, which would involve actively identifying potential misuse, monitoring for such misuse, and taking swift action if any unauthorized access or misuse is detected.

If multiple parties are found in breach of these responsibilities concerning the same AI system or related equipment, they could share the responsibility for any ensuing damages or harms. Individuals could be allowed to pursue legal action for damages exceeding a substantial monetary threshold (for example, $100 million) resulting from such breaches. Claims less than this monetary threshold would not be addressed by this framework. The idea is to clearly target catastrophic risks posed by the technology, which could require an approach distinct from smaller scale AI incidents [239].

Certain violations, like failing to obtain necessary AI licenses from the FAISA or violating licensing terms, could automatically constitute negligence. In situations where damages above the monetary threshold are claimed due to a violation of the established duty of care, the violating entity could be held directly accountable for all physical, property, and financial damages resulting from events tied to such a violation. In such cases, strict liability could apply, so plaintiffs would not be required to establish the specific cause of the accident in order to be compensated. This is essential to incentivize AI model developers to proactively anticipate large-scale harms that may be associated with their development and deployment activities (Introduction, 0.5.1.1). Strict liability may also incentivize insurance providers to increase premiums for customers they assess to be engaged in development activities that could lead to advanced models with the potential to cause damages above the monetary threshold.

Strict liability could only apply to cases in which significant damages occur, however, in order to avoid stifling industry by imposing liability for smaller, non-catastrophic harms (Introduction, 0.4.2). Making strict liability conditional on a license violation also mitigates the burden of frivolous litigation on regulated entities. Exemptions could also be made for genuine accidents, provided they arise in spite of reasonable measures to prevent them. Purely clerical or mathematical errors could fall under these exemptions, but errors in legal or technical judgments related to duty-of-care obligations would not. Finally, the fact that a product or service is free, collaborative, open-source, or open-access would not serve as a defense against any liabilities arising under this framework.

In addition to strict liability, joint and several liability could also apply to incidents causing harms above the monetary threshold. This is in order to resolve what would otherwise be an ambiguous diffusion of responsibility among AIHOs, AIHDs, DCIPs, and AIMDs for high-impact incidents.[89]

Finally, there could exist a safe harbor from strict liability for regulatory tiers designated by the FAISA. For example, the FAISA could determine that only Tier 2 AIHDs, AIHOs, and DCIPs, and Tier 3 AIMDs and above are subject to strict liability provisions. This would equip the FAISA to avoid a circumstance in which a low-risk model happens to be used as a non-critical component of a system that causes damages above the monetary threshold. Without a safe harbor, strict liability might otherwise apply to the entity that trained the low-risk model, even though it did not meaningfully contribute to causing the damages.[90]

Safe harbor provisions might also serve as an incentive for entities to comply with regulatory requirements. If registering as a Tier 3 AIMD offers the AIMD a safe harbor from strict liability, the AIMD may be more likely to register. In practice, an objective of this framework is to ensure that strict liability applies only to the very small group of individuals and AI models engaged in the highest-risk AI development and deployment activities.

4.2.2 Criminal liability

A criminal liability framework for AI is essential. The magnitude of the risks posed by reckless advanced AI development and deployment practices necessitates

[89] That is, if a high-impact incident occurs, it may be unclear whether responsibility should rest with the AIMD who created the system, the AIHO or DCIP who may have supplied computing resources to the AIMD without adequate safeguards, or individual decision-makers within each entity. Joint and several liability allows an injured party to sue any one responsible party for entire damages relating to an incident, and places the responsibility on all parties to subsequently allocate the portion of the entire damages paid by each.

[90] An important technical consideration that motivates safe harbor provisions for strict liability has to do with the amount of effort required by the user or developer of a system to do significant damage. Low-capability AI systems require effortful and deliberate engineering to cause significant harm. In these cases, it makes sense to hold individuals who deliberately weaponize these systems liable for these harms. However, as AI systems become more capable, the systems themselves begin to drive larger portions of the risk: an AI system capable of autonomously executing a catastrophic cyberattack may require little more than a prompt from a user to cause significant damage, for example. In the latter case, the model developer plays a critical role in causing damages, and should bear legal responsibility for them.

proportionate safeguards. These risks include threats to national security and public safety borne by society at large (Introduction, 0.5.1.1 and 0.5.4.1).

Under this proposed legislative framework, the law would establish penalties for dangerous behaviors on the part of individuals and entities involved in the AI supply chain. These penalties would be designed to deter irresponsible and negligent behavior and ensure accountability from these individuals and entities.

Any individual who commits a crime related to AI operations could be subjected to fines, and distinct penalties could be designated for particular infractions. For example, misdemeanors could include:

- Failure to accurately report high-performance AI hardware or qualifying data center infrastructure to the FAISA (see 4.1.3.1.2, 4.1.3.2.2, and 4.1.3.3.2);

- An AIMD misrepresenting its safety protocols or providing misleading information when pre-registering a training run that could plausibly lead to a model with dangerous capabilities or tendencies (for example, a Tier 2 or Tier 3 training run; see 4.1.3.4.2 and 4.1.3.4.3);

- An AIMD, AIHO, AIHD, or DCIP operating above their respective licensing thresholds without a license; and

- An AIMD, AIHO, AIHD, or DCIP responding to information requests from the FAISA with misleading data.

Felonies could include:

- Disregarding an emergency order to halt AI development activities;

- Engaging in development activities that require a license following the rejection of a license application; and

- Breaching the conditions of a license, especially if these violations lead to heightened security risks or cause damages exceeding a significant monetary threshold (e.g., $100 million), or if the entity or its management have prior convictions under this liability framework.

Under this framework, entities could be barred from indirectly or directly covering fines levied against their representatives. Compensation modifications to offset these fines could also be prohibited. For non-profit entities or those where profits cannot dictate the penalty magnitude, a court might impose increased fines.

For misdemeanors, an entity's licenses could be suspended for a moderate duration (e.g., between 1 month and 1 year). During this period, all AI-related activities above the licensing threshold could be prohibited. For felonies, all of the entity's licenses could be revoked, and the entity would be ineligible for new licenses for a significant period (e.g., five years). The entity could also be mandated to encrypt all its AI model weights, and only authorized license holders would be able to decrypt them. Additionally, the entity could be required to sell or destroy its AI hardware within 60 days. Crimes should be referred to the DOJ for prosecution.

We also recommend that the Congress investigate the constitutionality of fast-track procedural options, with a view to including them in relevant legislation if appropriate. This would give the government the option to respond to serious regulatory infractions at the speed of relevance, and would provide additional deterrence for illegal development and deployment practices.

4.2.3 Emergency powers

Even the possibility of criminal sanction may be insufficient to mitigate catastrophic risks from AI in certain contexts. This is for three reasons. First, some organizations may choose to take actions that create catastrophic AI risk without institutionally recognizing that they are doing so. These organizations may not respond to the prospect of criminal sanction associated with catastrophic outcomes, since they do not believe that their activities will produce those outcomes (Introduction, 0.5.4.1).

Second, AGI development has winner-take-all characteristics. AI labs therefore face powerful incentives to circumvent regulatory constraints and pursue clandestine AGI research programs even if they consider these activities to be dangerous (Introduction, 0.5.3.1).

Finally, some of the risks from weaponization and loss of control associated with advanced AI development may be WMD-like and unrecoverable, meaning that effective risk mitigation strategies cannot be solely reactive (Introduction, 0.5.1.1).

New frontier AI models can be trained in a matter of weeks. Once trained, they can be augmented in potentially dangerous ways with access to tools, through fine-tuning, or via other techniques in a matter of hours. In order to prevent illicit and highly dangerous AI development or deployment activities on the relevant timescales, we believe the U.S. government will require new emergency powers.

We recommend an emergency powers framework consistent with the defense in depth principle that we assess is required to safeguard against catastrophic national security risks from advanced AI. In particular, we recommend that the Congress authorize the President certain emergency powers under one of two conditions:

- If the President declares a national emergency due to a substantial national security threat stemming from advanced AI; or

- If the Administrator of the FAISA identifies a clear, immediate, and major national security threat from one or more advanced AI systems which cannot be curtailed by normal enforcement mechanisms.

In this latter case, we recommend that the FAISA Administrator's determination be made public in the Federal Register.

In the event that an AI developer is determined to engage in high-risk illicit development or deployment, a rapid and technically informed response is essential. Because of its depth of technical expertise and understanding of the AI ecosystem, the FAISA – through its Administrator – could be empowered to suspend AI licenses immediately, demand halts to certain AI-related activities, enforce safety measures, secure or encrypt AI model weights, restrict access to specific AI systems, or impose a general moratorium on AI research and development.

This would create a much-needed capacity for rapid and decisive response, and provide the relevant U.S. government agencies with the time they may need to understand the situation and context, before taking more decisive and long-term actions.

Those who disagree with these emergency orders, either on technical or policy grounds, should have a right to appeal. If they believe the order is unlawful or unconstitutional, they would also be able appeal to the federal district court.

In the event of economic losses due to compliance with an emergency order, affected parties should have a right to compensation from the government. Losses could be gauged on tangible expenses, actual investments, and the value of destroyed property.

4.3 Advanced AI in national security systems

Over time, we expect that advanced AI systems will be developed and deployed as components of national security systems. This will include applications in defense, intelligence, and other sensitive areas.

On the one hand, the national security imperative could necessitate less stringent RADA safeguards than those in other areas, particularly with respect to weaponization applications. On the other hand, there is no currently known method to reliably align an AGI-level system, and misaligned systems that are capable enough are likely to engage in extremely dangerous behavior by default (Introduction, 0.2.2 and 0.5.1.1). As a result, the development or deployment of a sufficiently capable AI system by any actor – including by national security agencies — could in and of itself create substantial national security risk regardless of the system's intended purpose.

The nature of and appropriate enforcement mechanisms for guardrails associated with national security applications of advanced AI remain open questions. In practice, the IC and DOD have internal oversight mechanisms, and the Inspector General and General Counsel already implement relevant usage controls internal to these departments and agencies. Similar mechanisms could be adapted to provide oversight for advanced AI development and use in national security systems under a modified set of RADA safeguards following a clear set of principles (LOE1, 1.3.2).

Another option could be to establish a "national security FAISA," with the authority to audit and intervene in the advanced AI activities of the national security agencies. This latter option has significant challenges, including the potential to dilute the talent pool available to other regulators, bifurcating institutional process knowledge on AI safety.

A related concern is that AI systems could be developed and deployed in national security contexts in ways that could challenge long-standing U.S. Constitutional norms. For example, deploying an AI system capable of superhuman persuasion could undermine the democratic process, yet also be required to counter potential offensive deployments of such systems by adversaries in the domestic information environment (see Annex F: Persuasion and manipulation).

It will be essential for society to strike a balance between privacy and security in the context of advanced AI, particularly when these systems are developed for use in national security contexts. Consistent with other advanced technologies, there is a need for an open, grounded, and technically informed dialogue between the

Congress, the Executive Branch, and the public on how advanced AI can best be used to safeguard both national security and Constitutional freedoms.

LOE5: Enshrine AI safeguards in international law and secure the AI supply chain

Build international consensus on advanced AI risk

- Coordinate internal and international messaging and capacity-building, including technical and policy training
- Articulate and reinforce an official U.S. government position on catastrophic AI risks

Enshrine AI safeguards in international law

- Establish a comprehensive framework to mitigate catastrophic risk from loss of control
- Design laws to leverage mechanisms for on-chip governance for monitoring and enforcement

Establish an international regulatory agency

- Establish an International AI Agency to monitor compliance with safeguards and facilitate global cooperation
- Create forums for international researchers to advance AI monitoring and standard-setting

Secure the AI supply chain multilaterally

- Create an AI Supply Chain Control Regime and an international sanctions regime for advanced AI nonproliferation
- Internationalize responsible AI development and adoption safeguards with narrowly scoped controls

The rise of advanced AI and AGI has the potential to destabilize global security in ways reminiscent of the introduction of nuclear weapons. Today, the United States leads in AI innovation. However, allies, partners, and adversaries around the world are quickly establishing their own AI scaling programs [240–242]. As advanced AI matures and the elements of the AI supply chain continue to proliferate (Introduction, 0.5.3.2), countries may race to acquire the resources to build sovereign advanced AI capabilities. Unless carefully managed, these competitive dynamics risk triggering an AGI arms race and increase the likelihood of global- and WMD-scale fatal accidents, interstate conflict, and escalation. These dynamics create the need for a carefully-calibrated U.S. strategy that balances competing and sometimes contradictory unilateral and multilateral objectives.

This LOE outlines near-term diplomatic actions and longer-term measures the U.S. government could take to **establish an effective AI safeguards regime in international law while securing the AI supply chain**. These actions include:

- Building a **domestic and international consensus** on potential catastrophic AI risks and necessary safeguards (see 5.2);

- Enshrining those safeguards in **international law** (see 5.3);

- Establishing an **International AI Agency (IAIA)** to monitor and verify adherence to those safeguards (see 5.4); and

- Establishing an **AI Supply Chain Control Regime (ASCCR)** with allies and partners to limit the proliferation of advanced AI technologies (see 5.5).

As a starting point to the above, the United States could secure the most time-critical components of its own domestic AI supply chain (LOE1, 1.5).

5.1 Desired end state of this LOE

We believe a series of initiatives like the one described in this action plan should be executed with a specific end state in mind as its ultimate objective. That end state may be adjusted over time as options evolve. But at all times it should be clearly articulated.

The *ideal* end state of this action plan is a treaty to mitigate catastrophic AI risk from weaponization and loss of control, enshrined in international law, and enforced globally by United Nations mandate. This ideal treaty would:

- Enforce RADA safeguards on advanced AI training runs beyond a certain scale or capability similar to those recommended in LOE4, 4.1.3;

- Enforce reporting requirements on cloud providers above a certain scale, making those cloud providers responsible for monitoring the safety of advanced AI models trained on their platforms; and

- Enforce hardware-based tracking of AI-enabled chips along with international monitoring of AI chip usage, to minimize the risk that dangerous models are being trained clandestinely (Annex L.4).

This treaty structure is an attempt to balance (1) the increased risk of rogue AI development due to proliferation of key components of the AI supply chain, against (2) the risk that excessively centralizing those components could lead to the development of competing unmonitored supply chains. A U.S.-led multilateral supply chain controls regime (see 5.5) could probably be put in place more quickly and would achieve some of these objectives. But the short-term safety benefits of a multilateral regime would be at greater risk of being undermined by a competing supply chain [243] in the medium term.

In the long term, it may be challenging to assure U.S. national security in the face of catastrophic AI risks without an international solution. AI supply chains could become too widely distributed, and AI models too easily proliferated, for unilateral action to suffice even given the U.S. position in the present-day supply chain (Introduction, 0.5.3.2; Annex G, G.1). Even multilateral supply chain controls by a broad coalition of U.S. allies may only buy time, though that time could prove invaluable to explore and implement further options.

5.2 Build international consensus and partner capacity on catastrophic AI risk reduction

There are credible reasons to believe that catastrophic AI risk from loss of control and weaponization could become meaningful in the relatively near term (Introduction, 0.5.1.2). As a result, there is a need for rapid and decisive action to enshrine key AI safeguards in international law, backed by credible enforcement mechanisms.

In leading such an effort, the United States will face the standard challenges that come with international lawmaking. But these will be compounded by the high technical complexity of AI risk, and the fact that nations do not yet have a shared understanding of this domain. There is also the possibility that U.S. diplomatic action could inadvertently alienate potential partners if it is not carefully calibrated given the political sensitivities surrounding adjacent issues, such as ethical and responsible use of AI.

This lack of shared understanding presents a significant risk while it persists, since it increases the chance of an international AI race. But it also offers an opportunity for the United States to lead and participate in international consensus-building, by supporting other nations in developing and shaping that shared understanding. To take full advantage of this opportunity, U.S. government personnel will themselves need to be educated on catastrophic AI risks, their mechanisms, and their potential mitigations (LOE2, 2.2.1). At the same time, the United States could launch educational programs to support improving other countries' awareness and mitigation capacity, particularly concerning AGI alignment failure and resulting loss of control.

The goal of these efforts should be to build a broad international consensus on behavior that risks global catastrophic impact from loss of control of AGI-level systems, and should therefore face international sanction. We believe this consensus would be extremely supportive in developing, negotiating, and enforcing an effective AI safety and security treaty.

5.2.1 Coordinate domestic and international messaging and capacity-building

The United States is already participating in early consensus-building on catastrophic AI risk. The U.K. AI Safety Summit's Bletchley Declaration [244], signed by the United States, 27 other countries, and the European Union, refers to potential risks from "intentional misuse or unintended issues of control relating to alignment with human intent." These are the sources of catastrophic AI risk we refer to, respectively, as weaponization (Introduction, 0.2.1) and loss of control (Introduction, 0.2.2) in this document.

While we recommend continuing to expand on these efforts, consensus building on weaponization topics could be polarizing. Instead, focusing education and capacity-building on technical cooperation to avoid loss-of-control scenarios, while maximizing safety and security, could offer a more promising starting point.[91]

As part of these efforts, the United States will need to productively engage with two key stakeholder groups. The first group will be policymakers and decisionmakers in governments, particularly in national security roles. The second group will be the technical advisors to those policymakers and decisionmakers. Because of the urgency of the risk, domestic U.S. educational initiatives may need to operate in parallel with international ones.

5.2.1.1 Policymaker education and outreach

In order to credibly brief international partners on catastrophic AI risk and the benefits of cooperation on the issue, the U.S. government needs to increase its institutional understanding of its key aspects. We recommend that, in tandem with outreach to its international partners and in coordination with existing multilateral regimes, the U.S. government rapidly educate its own policymakers on aspects of catastrophic AI risk that touch on their areas of responsibility.

Part of the substance of this educational content is outlined in LOE2, 2.2. Additionally, based on our direct experience briefing U.S. and international policymakers on this issue, we believe these educational efforts should:

- Emphasize specific, concrete, and tangible scenarios relevant to AI risk and national security (see Annex C: Example AI alignment failure scenarios and [1]);

[91] As a general guideline, we strongly recommend that diplomatic personnel engaged in these efforts be briefed or trained in advanced AI topics as recommended in LOE2, 2.2.1. In particular, we have found through our own outreach efforts that highlighting certain technical topics, such as the distinction between outer alignment and inner alignment, can produce a positive diplomatic framing. Specifically, the **outer alignment** problem refers to the challenge of defining goals which, if those goals were faithfully pursued by an arbitrarily intelligent system, would not have catastrophic consequences. The **inner alignment** problem refers to the challenge of ensuring that such a goal, once it is defined, will actually be internalized by a trained AI system as its true goal. Both problems are currently unsolved. In engagements so far, we have found that clearly explaining the distinction between these two problems has been constructive from a diplomatic standpoint. The reason is that, because inner alignment is unsolved, no actor can be assured that they will be able to control an AGI-level system even if they develop and train the system themselves (Introduction, 0.2.2). Since such a system would be a danger to everyone including its own developer, all actors are in principle incentivized to cooperate in preventing the development of such a system for as long as inner alignment remains an open technical problem. Diplomatic engagement can then begin on this basis, including potentially with adversaries. See the Glossary of terms for full definitions of outer and inner alignment.

- Explain that the build-vs-use distinction that exists for nuclear weapons may not exist for sufficiently capable AI systems, meaning that simply building such a system, with no intent to deploy it, may in and of itself be a dangerous action (Introduction, 0.2.2); and

- Clarify that, because of this unusual risk profile, the international community may need to implement rules above and beyond those already in place for less-capable AI systems such as lethal autonomous weapons (LAWs).

Some immediate actions that the U.S. government could take to support education and capacity-building for its international partners are:

- Through U.S. Missions globally, engage Heads of State and Heads of Government on AI safety, security, and catastrophic risk. Key objectives of these engagements should be to explain to partners what is happening in the field of frontier AI, to ask these governments to begin formulating their perspectives, to offer U.S.-provided foundational AI training, and to build support for an eventual international legal regime. The President or Secretary of State could also preview U.S. positions on advanced AI at the United Nations General Assembly.

- Initiate conversations on catastrophic AI risk in existing AI policy-focused forums such as the Global Partnership for AI, the United Nations Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts (GGE) on LAWs, the Organization for Economic Cooperation and Development (OECD), and the AI Partnership for Defense.

- Design and implement foreign assistance programs for building partner capacity on AI safety. These programs could include robust foundational AI training, AI safety and security training and best practices, and support for domestic legal and regulatory frameworks needed to manage catastrophic AI risk. This could be done as part of the training and up-skilling efforts in LOE2, 2.2.

5.2.1.2 Technical education and outreach

A significant portion of the risk from advanced AI in the near term is associated with development and deployment of frontier AI systems by and for national security and defense applications (LOE4 4.3). National security and defense personnel in the United States and abroad rely on their scientific communities for advice on the opportunities and risks of advanced AI. This means effective outreach to scientific communities, AI researchers, and technical defense advisors around the world will be critical to ensure that the national security establishments of stakeholder nations fully appreciate the scope and nature of catastrophic risk from advanced AI systems, particularly as posed by loss of control (Introduction, 0.2.2 and 0.5.1.1).

The United States should promote clear, open, honest, and secure discussion and debate between AI researchers across national boundaries with a focus on catastrophic AI risk and its potential mitigations. Because this risk class is global in scope, stakeholder nations include the United States itself, its partners, and even its adversaries.

Early efforts at international technical engagement by independent AI researchers are already underway. Beginning in 2023, prominent Western academics, including "godfather of deep learning" Geoff Hinton, began reaching out to the Chinese AI research community with the aim of increasing its awareness of catastrophic risks from advanced AI and AGI. Prominent AI researchers, practitioners, and public figures have also signed the 2023 Center for AI Safety's (CAIS) Statement on AI Risk [245], as have several Chinese academics. We believe such outreach to be extremely valuable in building the consensus on AGI risk that will ultimately establish diplomatic common ground on this issue.

While the U.S. government should broadly support such efforts, it must be clear to all stakeholders that these academics and organizations are acting independently out of their own genuine concern. We expect U.S. outreach efforts to be most productive if they are focused on promoting forums for expanded discussion on these topics, rather than attempting to directly shape the substance of the debate.

Some immediate actions that the U.S. government could take to support education and outreach to the international AI research community are:

- The U.S. Ambassador to the United Nations to initiate a process to establish an intergovernmental Commission on Frontier AI. The Commission's mandate could be to conduct research and draft a report to the Secretary General aimed at fostering consensus on catastrophic risk from loss of control due to AGI alignment failure. The Commission could draw inspiration from similar institutions, such as the Intergovernmental Panel on Climate Change (IPCC), which played an analogous role for scientific consensus-building on climate change. Although there is a risk that any research related to AI alignment may inadvertently contribute to accelerating AI capabilities and to the proliferation of more powerful AI systems, research aimed at accumulating evidence for the difficulty of AI alignment (see Annex B: The full challenge of AGI alignment), or at demonstrating alignment failures in powerful non-frontier systems, may be less likely to present this risk. An international Commission on Frontier AI would offer an opportunity to promote international collaboration and clarify the shared incentives that exist between the U.S. and other countries in establishing an international regime for catastrophic AI risks [246].

- The Department of State's Bureau of Oceans and International Environmental and Scientific Affairs's (OES) Office of Science and Technology Cooperation (STC) to explore avenues for bilateral outreach with partners and international academic forums. These could include:
  - Encouraging academic venues such as NeurIPS [247] to host technical talks, workshops, or seminars on catastrophic AI risk;
  - Supporting carefully scoped collaborations, including temporary exchange programs, for AI safety and AGI alignment researchers between adversary, U.S., and Western researchers and institutions; and
  - Sponsoring technical training courses on AI safety and AI alignment to accelerate uptake of the core technical topics in these fields among the broader AI research community.

5.2.1.3 Other forms of capacity-building

Because catastrophic AI risk could manifest in the relatively near-term (Introduction, 0.5.1.2), we recommend that the United States adopt a collaborative stance to discovery and truth-seeking with international stakeholders from an early stage. This is a non-standard approach to international diplomacy. While consensus-building efforts can proceed in parallel across partners and adversaries, any public announcement should be sequenced with geopolitical sensitivities in mind.

Some early actions that the U.S. government could take to support capacity-building across partner and adversary nations are:

- The Department of State, in coordination with the Defense Threat Reduction Agency (DTRA) and others, to establish programs to support catastrophic risk reduction from advanced AI. These could include destruction, dismantlement, or processing of unlicensed compute stockpiles and other supply chain components; red teaming of partners' legal and regulatory frameworks; workshops and training on increasing safety and security at the individual and institutional levels; material assistance; facilitating TTXs; and other forms of cooperative threat reduction.

5.2.2 Articulate and reinforce an official U.S. government position on catastrophic AI risk

Early outreach and education may be able to proceed on the basis of collaborative truth-seeking between the United States and international stakeholders. But later consensus-building, and ultimately treaty negotiations, may be challenging without a clearly articulated, official U.S. position on catastrophic AI risk [248]. Therefore the United States should develop and promulgate such a position as soon as practicable. Effective domestic AI regulation, as outlined in LOE1, 1.3 and LOE4, would also greatly support the credibility of U.S. diplomatic efforts abroad by signaling willingness to unilaterally improve the safety and security of its own AI industry.

To clarify the U.S. position, the Congress could issue a Sense of Congress on catastrophic AI risk. The content and scope of a Sense of Congress would depend on the degree of institutional understanding and internal assessment of this risk.

Both domestic controls and international safeguards will be time-consuming to implement, though both may also need to be developed urgently. The United States will need to navigate these constraints in order to effectively lead global coordinated action on AI safety and security [249,250].

5.3 Enshrine AI safeguards in international law

The Department of State, in coordination with the relevant U.S. government leads, could launch an international diplomatic campaign to enshrine AI safeguards in international law. Ultimately the law could take the form of a binding treaty ratified by a United Nations Security Council (UNSC) Resolution. If ratified by the U.S. Senate, an AI safety treaty would become supreme law of the land and govern public and private activities regarding advanced AI development.

To succeed, an international campaign to establish such a treaty could focus on catastrophic AI risks and the corresponding safeguards. Political issues such as human rights, AI ethics and responsible use, and general weaponization restrictions[92] are

[92] However it may be possible to obtain UNSC support for certain narrow weaponization restrictions, such as a ban on training AI to develop biological weapons.

important for allied collective action, but could derail negotiations. The top priority in negotiations could be to address catastrophic risks specifically stemming from unsafe advanced AI development practices and potential loss of control due to AGI alignment failure. Because these catastrophic risks have the potential to harm all parties, including the developer of the AI itself, there may be a window for meaningful progress on this narrowly scoped issue.

5.3.1 Treaty structure considerations

Among other challenges, an AI treaty would need to address the security dilemma that frontier AI poses to participating nations. This dilemma has some precedent in the case of nuclear and biological weapons development [248,251]. In particular, each incremental advance in capabilities by a nation's domestic AI industry improves that nation's security posture relative to its adversaries. But that improvement comes at the expense of heightening *global* instability by increasing the general risk both of intentional use of a weaponized AI system, and of AGI alignment failure and loss of control of an AI system regardless of the intent of its developer (Introduction, 0.2.2). This risk from loss of control has its closest parallel in research on biological agents, which regularly escape containment despite having a clearly understood risk model supported by decades of experience [252,253].

Nuclear arms control treaties such as the Strategic Arms Limitation Talks (SALT) have historically had some success in mitigating a similar security dilemma [254]. These treaties function by incrementally reducing participants' stockpiles of warheads and weapons systems, and confirming those reductions through mutual verification. Verification mechanisms include requirements to declare nuclear facilities, disclose statistics on nuclear supply chains, and allow inspections of key installations.

In the case of advanced AI, verification mechanisms for hardware could include exposing elements of the semiconductor supply chain to inspection, cooperative threat reduction in the form of verified destruction of AI hardware stocks, and on-chip monitoring mechanisms (see Annex L, L.4). We discuss some of these approaches in more detail, and address the dilemmas associated with algorithmic advances, in 5.4.1.

But advanced AI also poses a special challenge to traditional verification mechanisms. In nuclear arms control, there is a division between reactor-grade (civilian) and weapons-grade (military) nuclear material, defined by enrichment levels. That division makes it feasible to apply controls to military applications without unduly hindering the peaceful development of civilian nuclear technology. By contrast, an advanced AI training data center may be inherently dual-use. It could in principle be used to train safe and beneficial AI systems, or it could be used to train weaponized or high-risk AI systems, without any external signal differentiating the two cases (Annex G, G.1).

This dual-use aspect of AI is especially challenging because current applications of advanced AI are disproportionately beneficial. A traditional verification mechanism therefore risks unduly curtailing beneficial innovation (Introduction, 0.4.2). On the other hand, advanced AI training may be much more amenable to software-based monitoring approaches, which have the potential to be more granular than traditional verification (LOE1, 1.5.2). Both of these may need to be deployed in combination.

Finally, an AI safety treaty could be supported by an international organization to verify compliance, set safety and security standards, and convene the global community of AI researchers. We discuss this possibility in 5.4.

5.3.2 Preparation for a treaty

In setting the stage for international treaty negotiations, the United States could coordinate closely with other nations that have mature AI industries. The 2023 U.K. AI Summit [244] could offer a starting point for these engagements. We recommend that the goal of these efforts be to **establish comprehensive RADA safeguards to minimize catastrophic risks from loss of control due to AGI alignment failure**, and aligning the international community on new international law or treaty requirements.

According to discussions with participants, the U.K. AI Summit made significant progress building consensus in a number of key areas, including the importance of implementing defense in depth by implementing overlapping sets of controls, and the principle that, as an AI system becomes more powerful, the burden of safety increasingly should fall on its developer. Some areas of contention remain, notably regarding at which level of capability open-access AI systems stop being compatible with continuing global safety and security (Introduction, 0.5.1.6). While some of these areas may be resolved by further discussion among the parties, they also highlight the ongoing need for technical collaboration and research into the risks, benefits, and mitigations.

Similar summits, such as those planned in South Korea and France during 2024 [255], could be used to secure additional commitments from frontier AI labs and their cloud providers. Such commitments may serve an informal precedent-setting function for fast-following labs outside the United States as the global AI supply chain proliferates [256,257]. These summits could also be used as opportunities for countries to issue joint declarations on their shared understanding of AI safety and security issues, similar to the 2023 Bletchley Declaration [244], and consistent with the approach the United States has taken to addressing global coordination challenges in the past [258,259].

5.4 Establish an international regulatory agency

If the United States is successful in spearheading an international legal regime for catastrophic AI risk mitigation, the international community may need an agency with the authorities to verify and monitor compliance with AI safety protocols, facilitate technical AI safety cooperation, and establish and maintain safeguards. Although supply chain controls (LOE1, 1.5) and domestic regulation (LOE4) may have a large early impact, in the long-term global safety and security could require an international agency with a mandate derived from international law. We will refer to this agency here as the **International AI Agency (IAIA)**.[93]

We expect an IAIA's operating model to evolve, and recommend setting this expectation clearly and explicitly. U.S. domestic regulatory efforts from LOE1, 1.4.1.2 and LOE4, 4.1.2 could inform this work. Similarly to early treaties on climate change, the first iteration of an IAIA's structure may not be the right one for the long term. On the other hand, experience has shown that it may be possible to develop a multilateral framework that effectively addresses a collective action problem, as long as the participants are willing to iterate on the framework and agreement design. An explicit iteration process could also let counterparties observe each others' behavior over time, adjust their commitments, and engage in confidence-building measures [215].

The history of nuclear nonproliferation [260] also shows that international efforts to differentially control peaceful and military applications of dual-use technology can be somewhat successful. For example, today 33 countries have safe nuclear power, while only 9 have nuclear weapons [217]. Since an AI data center can be used just as easily to develop safe AI systems as dangerous AI systems, differential controls may be more challenging to apply in this domain than in the nuclear domain. But the tools to monitor advanced AI may also be correspondingly more fine-grained and scalable, since many of them could be implemented entirely in software.

Ideally, an IAIA monitoring mechanism should be able to detect individual defections by its participants. This is important because the risk of an advance in AGI alignment could incentivize individual participants to try to train their own self-aligned AGIs (see 5.2.1.4). We acknowledge that the challenges to such coordination and

[93] It may also be possible to implement the functions of an IAIA that we recommend here through existing mechanisms, channels, and forums.

implementation are very significant, and that no individual tool or approach can offer comprehensive safety or security. This is a key reason we approach this problem through the lens of defense in depth and recommend multiple, mutually supportive control mechanisms (e.g., see 5.5).

5.4.1 Monitoring and verification

An international monitoring and verification regime could cover elements such as:

- Comprehensive safeguards agreements;
- Additional protocols;
- Regular, special, and ad hoc inspections;
- Containment and surveillance procedures;
- Material accountability;
- Access to AI scientists and researchers;
- Information sharing and reporting mechanisms; and
- Regular reports from the IAIA to the Secretary General.

With respect to specific supply chain components, the entity categories and tiers from LOE4, 4.1.3 (AIMDs, AIHOs, DCIPs, and AIHDs) could serve as a template for an initial RADA safeguards framework for an IAIA. As with domestic RADA safeguards, international safeguards should take into account AI capability measures in addition to raw compute, though compute will likely remain an important locus of control (Annex G, G.2). Accordingly, regulatory oversight may need to extend into semiconductor manufacturing inputs to address the risk that independent AI hardware supply chains could arise to circumvent safety controls.

One of an IAIA's key responsibilities could be to monitor use and development of AI hardware around the world. Key challenges will include ensuring that countries which are not yet IAIA members do not accumulate public or private pools of compute sufficient to support the training of AI models that pose catastrophic risks (Introduction, 0.5.3.2). This might be achieved in part via coordinated export controls of compute hardware to non-IAIA member countries through a Compute Suppliers Group (CSG) analogous to the Nuclear Suppliers Group for nuclear materials [225,261,262].

An additional challenge would be for an IAIA to ensure that models that pose catastrophic risks are not trained within or outside its member countries as algorithmic improvements allow more capable models to be trained with less hardware [262], and as hardware efficiencies increase (Introduction, 0.5.3.2). This risk could be mitigated among member countries by having the IAIA adjust its licensing thresholds over time to ensure that they continue to apply to any entities capable of marshaling dangerous quantities of compute, similar to the FAISA (LOE4, 4.1.2.4) and consistent with RADA safeguards principles (LOE1, 1.3.2). In addition to a CSG, an IAIA could also coordinate with a U.S.-led international AI supply chain controls regime (see 5.5) to ensure that international purchases of AI chips are restricted to members of the controls regime. This could become essential as algorithmic breakthroughs may allow smaller compute pools to be leveraged to build more dangerous systems (Annex G, G.3).

Perhaps the most important requirement to safeguard long-term global safety and security from AI will be hardware-enabled mechanisms that can allow verification and control of how compute hardware is being used, supported by robust chip registry programs (Annex L, L.4) [148,155,213,224]. This would involve exercising tight control on the semiconductor supply chain to ensure that any chips that enter the global supply are equipped with on-chip monitoring hardware, as well as preventing alternative compute supply chains from forming around the world. Such technical efforts could be undertaken in coordination with a multilateral supply chain controls regime if one exists (see 5.5).

5.4.2 Standard-setting

An IAIA could set authoritative international standards for advanced AI training runs and deployments. These could include standards for security, model containment measures, evaluations for advanced AI models, and other key areas informed by entities such as the NIST U.S. AISI (LOE3, 3.2).

5.4.3 Convening researchers

Finally, an IAIA should create a regular forum to convene international researchers to advance the state of the art in AI monitoring and standard-setting. This could take the form of a quarterly or biannual conference of top researchers in these respective areas, and of financing grants for research teams working in these domains.

Another possibility could be to establish an international research facility dedicated to AI safety and security research. Such a facility could form an early point of agreement in treaty negotiations and even be established before a large-scale monitoring regime, as an extension of earlier educational and outreach programs (see 5.2.1.2). Such a facility could support joint investigation of AI safety and security questions that could confirm or disprove aspects of AI risk and inform negotiations.

5.5 Allied multilateral initiatives to manage the AI supply chain

A functioning international AI safeguards regime (see 5.4) represents an ideal outcome (see 5.1). But as part of a defense-in-depth strategy, the United States could also work in parallel with its allies and partners to establish a multilateral framework of AI supply chain controls, such as through the Wassenaar Arrangement [263]. Not only would this increase U.S. and allied leverage for broader international negotiations, but it could also be used to reinforce many of the monitoring functions of the IAIA if successfully implemented. Additionally, such a multilateral framework would function as a stopgap and failsafe, in case an attempt to work through international bodies proceeds too slowly or is otherwise unsuccessful.

In the worst case, an independent multilateral supply chain controls framework, led by the United States, may be sufficient on its own to secure the inputs needed to train frontier AI systems for some time. But we expect that, in the longer term, independent AI supply chains [243] will emerge and ultimately undermine even an otherwise effective multilateral system.

A multilateral AI controls framework could have three goals:

- Ensure that critical elements of the supply chain for advanced AI, particularly compute[94] and its inputs, remain localized to U.S. and allied jurisdictions;
- Ensure that access to AI cloud compute is controlled through a regulatory mechanism similar to that proposed in LOE4, 4.1.3 including RADA safeguards implemented in U.S. and allied jurisdictions; and

[94] Meaning GPUs, TPUs, and other AI accelerator chips.

- Provide a path for foreign entities to obtain access to AI cloud compute clusters in U.S. and allied jurisdictions, provided they do so subject to the regulatory mechanism (e.g., LOE1, 1.5.2).

5.5.1 Strategic considerations

The most promising points of regulatory leverage in any supply chain are those that involve assets that have been built or acquired at high capital cost. By this criterion, AI hardware, data centers, and the other the elements listed above offer valid loci of control over the advanced AI supply chain (Introduction, 0.5.3.2).

But other AI supply chain components, while they may be crucial as inputs, are less amenable to regulatory oversight because they do not meet this criterion. Textual training data, for example, is available online at relatively low cost [264,265]. And frontier AI models themselves can be copied and transferred at no cost once created, even though they require a high capital cost to train. As a result, controls on open-source data or on the sharing of open-access AI models are unlikely to be effective, though data controls scoped to specific model training runs could be beneficial (Annex N, N.4). Controls at the application level are also unlikely to be worth the cost, unless implemented as KYC requirements on a model developer's end users (Annex P, P.2).

5.5.2 Multilateral partnerships

The supply chain for AI hardware is at the root of any realistic long-term mechanism for advanced AI control and nonproliferation (LOE1, 1.5.3; Annex G: Primer on AI and compute). It may be vital to maintain close coordination with international partners on this issue in order to ensure controls remain robust in the face of changing future conditions.

The United States manufactures a significant fraction of the inputs to AI hardware production domestically, so it holds a significant position in this supply chain. The United States can strengthen this advantage by coordinating with international partners in the AI chip supply chain. Many key AI inputs are manufactured in allied and partner countries [243,257]. Below is a list of countries that could be inaugural members of an ASCCR, along with the significance of each. [266,267]

Table 4. Potential inaugural members of an ASCCR and the components of the AI supply chain over which they exert the most influence.

Country	AI supply chain control approaches
Netherlands	Semiconductor tooling export controls
Japan	Semiconductor tooling export controls, technology transfer controls, foreign national visa screening
Germany	Technology transfer controls, foreign national visa screening
Taiwan	Semiconductor tooling export controls
United Kingdom	Technology transfer controls, foreign national visa screening
South Korea	Semiconductor tooling export controls
France	Technology transfer controls, foreign national visa screening
Australia	Technology transfer controls, foreign national visa screening
Israel	Semiconductor tooling export controls
Singapore	Semiconductor tooling export controls
United Arab Emirates [241]	Technology transfer controls, foreign national visa screening

The exact conditions for safety compliance are fluid and have many dependencies, but could initially be based on terms such as those in Annex I: Voluntary Charter for responsible AI, or on the RADA safeguards framework outlined in LOE4, 4.1.3.

An ASCCR could prioritize the development of technologies, standards, and policies for on-chip governance (Annex L, L.4). Because robust on-chip governance capabilities could take several years to develop and be put into production at scale, the United States could consider urgently investing in the development of on-chip governance technologies prior to the formation of the ASCCR [148]. This could be done through federal AI research Centers such as in LOE3, 3.1.

5.6 Open challenges

In this action plan, we have attempted to balance considerations of risk and innovation while addressing as many challenges as we can reasonably anticipate. But several open challenges remain that we feel we have not directly addressed.

One major difficulty in the international setting is that AI-capable nations are incentivized to escalate development of AI-enabled national security systems. This incentive is preserved even if both sides fully agree on the risk of loss of control from the development and deployment of AGI-level systems. Because the exact capability level at which these risks begin to manifest is unknown, parties may be motivated to make incremental improvements to gain an edge in national security. Certain forms of mutual verification could begin to address this challenge, but they may require unprecedented levels of access, high trust between the parties, and a broad-based consensus on the underlying risk.

Reliable verification of compliance with AI risk mitigations is also an open problem. In the face of attempts at obfuscation by a resourced entity, it is challenging to verify that the entity is not training a high-capability advanced AI model. Certain approaches like on-chip verification are promising and could be implemented on timelines of a few months for domestic verification schemes. But this increases to several years in contexts involving international verification [148].

The impact of algorithmic improvements on the effectiveness of supply chain controls is another consideration in the medium and long term (Introduction, 0.5.3.2). As AI algorithms continue to improve, more AI capabilities become available for less total compute. Depending on how far this trend progresses, it could ultimately become impractical to mitigate advanced AI proliferation through compute concentrations at all. Approaches to limit the pace of algorithmic improvements could address this issue (LOE4, 4.1.4), though there are significant downsides to placing broad-based controls on this kind of research.

Conclusion

AI is a technology fundamentally unlike any other. It holds vast potential to elevate human well-being, but could also be deliberately weaponized or exhibit accidental failures that have catastrophic consequences. Our recommendations focus on mitigating the most unrecoverable catastrophic risks advanced AI poses (Introduction, 0.5.1.1) while preserving its potential for positive impact (Introduction, 0.4.2). Current and near-term risks from advanced AI include weaponization for applications which could include bioweapon design, advanced manufacturing, large-scale human persuasion, and cyber warfare (Introduction, 0.2.1). Future risks include loss of control of high-capability AI systems (Introduction, 0.2.2). Frontier AI labs have publicly suggested that such dangerously capable systems could be developed in the near future, and possibly within the next five years (Introduction, 0.5.1.2). Both categories of risk have the potential, in the worst case, for unrecoverable catastrophic impact on human welfare.

Several underlying factors drive and exacerbate these risks. They include:

- Worst case outcomes from weaponization and loss of control are plausible [19] and may have unrecoverable catastrophic impacts (Introduction, 0.5.1.1);

- The timescale and degree of risk from loss of control are highly uncertain and the subject of ongoing technical debate (Introduction, 0.5.1.2 and 0.5.1.3);

- The world's most advanced AI labs have publicly acknowledged that they lack the safety and security measures they need to secure their systems against catastrophic risks (Introduction, 0.5.1.4 and 0.5.1.5);

- Researchers at these labs have publicly acknowledged that fundamental advances in AI alignment will be required to prevent loss of control over advanced AI systems that may be developed in the near term [84,85];
- The supply chain for advanced AI is challenging to secure by reactive policy measures (Introduction, 0.5.3.2);

- There is no known way to comprehensively audit the full range of dangerous capabilities and propensities of advanced AI models, and these models can therefore harbor latent dangerous capabilities that may only be discovered or elicited after they are deployed or released as open-access software (Introduction, 0.5.1.6; LOE3, 3.2.1);

- AI technology is advancing faster than reactive policy and legislative processes and may be accelerating, which could cause policy measures to quickly become outdated (Introduction, 0.5.2.1);

- Information and discussion on the topic of AI is often highly polarized in public spheres (Introduction, 0.5.2.2);

- A lack of domestic controls could reduce the effectiveness of U.S.-led international coordination on risk reduction (LOE5, 5.2.2); and

- Excessive domestic controls could damage U.S. innovation and competitiveness in AI (Introduction, 0.4.2).

Given the degree of AI-related risks and the multifaceted nature of the factors that contribute to them, no single safeguard may be adequate. Instead, we propose a mitigation strategy based on defense in depth, in which multiple lines of effort are deployed simultaneously to deliver different impacts across different timescales. These lines of effort include:

- Establishing an **interim domestic regime** of responsible AI development and adoption (RADA) safeguards for U.S. frontier labs overseen by an **interagency task force**, along with **controls on the AI supply chain** (LOE1);

- Establishing **legally enshrined AI safeguards**, including an **AI regulatory agency**, criminal and civil liability framework, and emergency Presidential powers (LOE4); and

- Establishing and implementing AI safeguards in **international law**, along with **multilateral controls on the supply chain** for advanced AI (LOE5).

In addition, we recommend two supporting lines of effort:

- Building capability and capacity for **advanced AI and AGI preparedness and response** (LOE2); and

- Increasing national capacity for **technical AI safety research** and **standards development** (LOE3).

These LOEs are complementary, mutually reinforcing, and directly address the factors underlying AI-related weaponization and loss of control risks. The task force established under LOE1, 1.4 would serve as a testing ground for safety and security measures that could apply to future domestic and international regulatory regimes (LOE4, 4.1.3 and LOE5, 5.3). A mandated set of RADA safeguards (LOE1, 1.3.2) would give the task force the leverage necessary to ensure the active participation of frontier AI labs as it pilots its oversight operations.

Supporting lines of effort ensure that the legal and regulatory RADA safeguards introduced under LOE4 are enforced in a technically informed manner (LOE2, 2.2; LOE3, 3.2) that is both responsive to the requirements of contingency planners (LOE2, 2.4), and flexible in the face of novel technical developments (LOE1, 1.2; LOE2, 2.3). And the international RADA safeguards overseen under LOE5, 5.4 would benefit from the lessons learned from implementing LOE1 and LOE4 domestically.

Some measures proposed in this action plan are without precedent. LOE1, 1.3 calls for pivotal executive actions; LOE4, 4.2 calls for the Congress to pass a landmark bill that would set a new framework for national AI governance; and LOE5, 5.3 calls for a formally ratified treaty to be negotiated with allies, partners, and other stakeholders. But it is increasingly clear that the United States should address frontier AI research with the same seriousness as previous paradigm-shifting technological breakthroughs with the potential to introduce WMD-like threats to global safety. In the past, this has entailed significant regulatory and legal interventions, rooted in coordination between technical and governmental experts.

In the face of these challenges, bold action is required for the United States to address the current, near-term, and future catastrophic risks that AI poses while maximizing its benefits, and successfully navigate what may be the single greatest test of technology governance in the nation's history.

Funding disclosure

This publication was produced for review by the United States Department of State. It was prepared by Gladstone February 26, 2024 and the contents are the responsibility of the authors. The authors' views expressed in this publication do not reflect the views of the United States Department of State or the United States Government.

Gladstone AI is a for-profit company funded entirely from its own revenues. Neither Gladstone AI, nor any of the authors as individuals, have received investments or donations from any external institutions. Neither Gladstone AI, nor any of the authors as individuals, have financial ties to any of the non-U.S. government entities described in this document, or to any of their employees, investors, donors, or affiliates.

Acknowledgements

We thank the following members of the **Gladstone AI project team** for contributing their invaluable subject-matter expertise, edits, and advice to this action plan.

- **Ben Isaacoff**, Project Lead. AI & ML Scientist Manager, General Motors.

- **Jon Askonas**, History Team Lead. Senior Fellow at The Foundation for American Innovation.

- **Alexander Falbo-Wild**, History Team Associate. History PhD Student at Temple University.

- **Caroline Pitman**, History Team Associate. Special Assistant at the Center on Global Energy Policy.

The following organizations and individuals generously offered their time, comments, feedback, or insights in reviewing or shaping crucial aspects of this document. We greatly appreciate their considerable contributions.

- **Apollo Research.** London, U.K.-based AI evaluations developer, focused on deception and misalignment. (apolloresearch.ai)
    - Marius Hobbhahn, Co-founder and Director
    - Lee Sharkey, Co-founder and Chief Strategy Officer

- **Center for AI Policy (CAIP).** DC-based nonprofit specializing in AI safety legislation. (aipolicy.us)
    - Thomas Larsen, Founder and Executive Director
    - Jason Green-Lowe, Legislative Director
    - Akash Wasil, AI Governance and Policy

- **Center for a New American Security (CNAS).** DC-based national security think tank. (cnas.org)
    - Tim Fist, Research Fellow
    - Caleb Withers, Research Assistant

- **Center for AI Safety (CAIS).** San Francisco-based research and field-building nonprofit focused on reducing societal-scale risks from AI. (safe.ai)
    - Dan Hendrycks, Co-founder and Executive Director
    - Bri Treece, Chief Operating Officer

- **Centre for International Governance Innovation (CIGI).** Waterloo, Canada-based independent think tank on global governance. (cigionline.org)
    - Duncan Cass-Beggs, Executive Director, Global AI Risk Initiative

- **Centre for the Governance of AI (GovAI).** Oxford, U.K.-based research and development nonprofit focused on long-term AI governance. (governance.ai)
    - Ben Garfinkel, Director
    - Guive Assadi, Research Scholar
    - Alan Chan, Research Scholar
    - Noemi Dreksler, Research Fellow
    - Fynn Heide, Research Scholar
    - Lennart Heim, Research Fellow
    - Jonas Schuett, Research Fellow
    - Elizabeth Seger, Research Scholar
    - Robert Trager, International Governance Lead

- **Conjecture.** London, U.K.-based developer of scalable AI alignment solutions. (conjecture.dev)
    - Connor Leahy, Founder and Chief Executive Officer
    - Andrea Miotti, AI Policy and Governance
    - Chris Scammell, Chief Operating Officer
    - Rachel Stockton, Chief of Staff to the CEO

- **Epoch.** Remote (U.S. and Europe) nonprofit working on forecasting developments and trends in advanced AI. (epochai.org)
    - Jaime Sevilla Molina, Co-founder and Director
    - Ben Cottier, Staff Researcher

- **Machine Intelligence Research Institute (MIRI).** Berkeley-based nonprofit research institute focused on technical AI alignment. (intelligence.org)
    - Nate Soares, Executive Director
    - Malo Bourgon, Chief Operating Officer

- **Mila – Québec Artificial Intelligence Institute.** Montreal, Canada-based research nonprofit; one of the founding centers of deep learning. (mila.quebec)
  - Yoshua Bengio, Founder and Scientific Director

- **Model Evaluation & Threat Research (METR; formerly ARC Evals).** Berkeley-based nonprofit developing AI evaluations for self-replication. (metr.org)
  - Chris Painter, Member of Technical Staff

- **RAND Corporation.** DC-based nonprofit think tank, research institute, and public sector consulting form. (rand.org)
  - Jeff Alstott, Senior Information Scientist
  - Ella Guest, AI Policy Fellow

- **Rethink Priorities.** Remote (U.S. and Europe) nonprofit working on research and implementation for altruistic causes. (rethinkpriorities.org)
  - Michael Aird, Senior Research Manager, AI Governance and Strategy
  - Joe O'Brien, Research Assistant, AI Governance and Strategy
  - Abi Olvera, Affiliate, AI Governance and Strategy

- **SaferAI.** Paris, France-based developer of infrastructure for auditing advanced AI systems. (safer-ai.org)
  - Siméon Campos, Founder and Chief Executive Officer

- **University of California, Berkeley.** Berkeley-based public research university. (berkeley.edu)
  - Hany Farid, Professor, School of Information
  - Chris Hoofnagle, Faculty Director, Berkeley Center for Law & Technology
  - Andrew Reddie, Associate Research Professor, Goldman School of Public Policy
  - Jonathan Stray, Senior Scientist, Center for Human-Compatible AI

- **Independent reviewers and contributors.** The following individuals did not have institutional affiliations at the time of their reviews or contributions.
  - Fletcher Heisler, cybersecurity researcher

- **Anonymous reviewers and contributors.** The following reviewers and contributors are anonymous.
  - `e5b63409d76da1d33c49e44cf55f8b7ebd52baa6e624a0b53dcd9a04377ce7dec5cabed82c831c10dc10bab942def414eb904f320129be7f41102611283aa02a`

We also thank technical, policy, and leadership teams from **OpenAI**, **Google DeepMind**, **Anthropic**, and **Meta** for their helpful and constructive discussions over the course of this project.

In addition, we are grateful to selected officials from the governments of **Canada**, the **United Kingdom**, and the **United States** for conversations that shaped or informed elements of this action plan. These include individuals from the following offices, departments, and agencies. For the avoidance of doubt, this action plan does not necessarily reflect the views, policies, or positions of any of these entities.

Canada

- **Office of the Prime Minister**
- **Office of the Minister of Justice and Attorney General of Canada**
- **Office of the Minister of National Defence**
- **Office of the Minister of Innovation, Science, and Industry**
- **Office of the Minister of Foreign Affairs**
- **Privy Council Office**
- **Department of National Defence**
- **Global Affairs Canada**
- **Public Safety Canada**
- **House of Commons**
  - Multiple Members of Parliament (MPs) and staffers

United Kingdom

- **Cabinet Office**
- **Department for Science, Innovation & Technology**

- **Foreign, Commonwealth and Development Office**
    - Counter Proliferation and Arms Control Centre
    - Directorate for Defence and International Security

- **Ministry of Defence**
    - Defence AI Centre
    - Defence AI & Autonomy Policy Unit
    - Defence Science & Technology Laboratory (DSTL)
    - Secretary of State's Office of Net Assessment and Challenge (SONAC)

United States

- **Executive Office of the President**
    - Office of Management and Budget (OMB)
        - Office of the Federal Chief Information Officer (CIO)
    - Office of Science and Technology Policy (OSTP)
        - National Security Division

- **Director of National Intelligence**
    - Office of the Director (ODNI)
        - Policy and Capabilities (P&C)

- **National Science Foundation**

- **Securities and Exchange Commission**
    - Office of the Chair
    - Division of Examinations
        - Event & Emerging Risks Team (EERT)
    - Division of Trading and Markets
    - Strategic Hub for Innovation and Financial Technology (FinHub)

- **Department of Commerce**
    - National Institute of Standards and Technology (NIST)
        - Information Technology Laboratory (ITL)

- **Department of Defense**
    - Office of the Secretary
        - Chief Digital and Artificial Intelligence Office (OSD CDAO)
    - Defense Advanced Research Projects Agency (DARPA)
        - Information Innovation Office (I2O)
    - Defense Intelligence Agency (DIA)
        - Technology and Long-Range Analysis Office (TLA)
    - Defense Threat Reduction Agency (DTRA)
        - Research and Development Directorate
    - Department of the Air Force
        - Chief Data and AI Office (DAF CDAO)
    - Office of the Under Secretary for Policy
        - Assistant Secretary of Defense for Strategy, Plans and Capabilities (ASD SPC)
            - Force Development and Emerging Capabilities Office (FDEC)
        - Nuclear and Countering Weapons of Mass Destruction Policy (NCWMDP)
    - Office of the Under Secretary for Research & Engineering (OUSD (R&E))
    - Joint Chiefs of Staff
        - Strategy, Plans, and Policy (J5)
        - Command, Control, Communications and Computer Systems (J6)

- **Department of Energy**
    - National Nuclear Security Administration (NNSA)
        - Counterterrorism and Counterproliferation Office (NA-80)
    - Lawrence Livermore National Laboratory (LLNL)
        - Center for Applied Scientific Computing
    - Los Alamos National Laboratory (LANL)
    - Oak Ridge National Laboratory (ORNL)
        - Center for AI Security Research (CAISER)
    - Pacific Northwest National Laboratory (PNNL)

- **Department of Justice**
    - Federal Bureau of Investigation (FBI)
        - Weapons of Mass Destruction Directorate (WMDD)
            - Emerging Threats and Technologies Unit
            - Nuclear and Radiological Countermeasures Unit

- **Department of Homeland Security**
    - Office of the Secretary
    - Countering Weapons of Mass Destruction Office (CWMD)
    - Office of Intelligence and Analysis (I&A)
    - Office of Strategy, Policy, and Plans (OSPP)
    - Science and Technology Directorate (S&T)

- **Department of State**
    - Office of the Secretary
        - Office of the Special Envoy for Critical and Emerging Technology (S/TECH)
    - Bureau of Intelligence and Research (INR)
    - Office of Policy Coordination (PC)
    - Under Secretary for Arms Control and International Security
        - Bureau of Arms Control, Verification and Compliance (AVC)
            - Office of Emerging Security Challenges (ESC)
        - Bureau of International Security and Nonproliferation (ISN)
            - Office of Cooperative Threat Reduction (CTR)
            - Office of Congressional and Public Affairs (CPA)
            - Office of Critical Technology Protection (CTP)
            - Office of the Nonproliferation and Disarmament Fund (NDF)
    - Under Secretary for Economic Growth, Energy, and the Environment
        - Bureau of Economic and Business Affairs (EB)
            - Division for Trade Policy and Negotiations (TPN)
                - Office of International Intellectual Property Enforcement (IPE)
    - Under Secretary for Public Diplomacy and Public Affairs
        - Global Engagement Center (GEC)

- **United States Congress**
    - Multiple Representatives and staffers

Finally, the above acknowledgements are not intended to imply that the listed organizations or individuals actively endorse this document in whole or in part. All errors and omissions are our own.

. . .

Bibliography

Note: The entries in this bibliography include peer-reviewed papers, research publications that have not undergone peer review, and informal sources like blog posts and, in some cases, tweets. Many of the landmark insights in modern AI research have been published as blog posts (e.g. Richard Sutton's "The Bitter Lesson"), in some cases by pseudonymous authors (e.g. Gwern's "The Scaling Hypothesis"), so a bibliography of modern AI that omitted informal sources would be incomplete. In general, publication norms in AI research are less formal than of other STEM fields. Nonetheless, for each entry in this bibliography, we have reviewed the source and concluded that it supports the corresponding claim in the body of the document.

[1] Harris, J., Harris, E., Beall, M. *Deliverable 2: Survey of AI Technologies and AI R&D Trajectories*. Section: "Risks". (2023).

[2] Turner, A. M., Smith, L., Shah, R., Critch, A., & Tadepalli, P. (2019). Optimal policies tend to seek power. In *arXiv [cs.AI]*. http://arxiv.org/abs/1912.01683

[3] Clark, J. [jackclarkSF]. (2022, August 7). *Malware is bad now but will be extremely bad in the future due to intersection of RL + code models + ransomware economic incentives. That train is probably 1-2 years away based on lag of open source replication of existing private models, but it's on the tracks*. Twitter. https://twitter.com/jackclarkSF/status/1556181432522797056

[4] Amodei, D. (2023, July 25). *Written Testimony of Dario Amodei, Ph.D. Co-Founder and CEO, Anthropic For a hearing on "Oversight of A.I.: Principles for Regulation" Before the Judiciary Committee Subcommittee on Privacy, Technology, and the Law United States Senate*. https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf

[5] Leike, J., & Sutskever, I. (2023, July 5). *Introducing superalignment*. Openai.com. https://openai.com/blog/introducing-superalignment

[6] Altman, S. (2023, February 24). *Planning for AGI and beyond*. Openai.com. https://openai.com/blog/planning-for-agi-and-beyond

[7] *Core Views on AI Safety: When, Why, What, and How*. (2023, March 8). Anthropic.com. https://www.anthropic.com/index/core-views-on-ai-safety

[8] Bove, T. (2023, May 3). *CEO of Google's DeepMind says we could be 'just a few years' from A.I. that has human-level intelligence*. Yahoo Finance. https://finance.yahoo.com/news/ceo-google-deepmind-says-could-213237542.html

[9] Harris, J., Harris, E., Beall, M. *Deliverable 2: Survey of AI Technologies and AI R&D Trajectories*. Section: “Notable players, their products and capabilities”. (2023).

[10] Wiggers, K., Coldewey, D., & Singh, M. (2023, April 6). Anthropic’s $5B, 4-year plan to take on OpenAI. *TechCrunch*. https://techcrunch.com/2023/04/06/anthropics-5b-4-year-plan-to-take-on-openai/

[11] Perrigo, B. (2023, January 12). DeepMind’s CEO helped take AI mainstream. Now he’s urging caution. *Time*. https://time.com/6246119/demis-hassabis-deepmind-interview/

[12] Perrigo, B. (2023, May 30). AI is as risky as pandemics and nuclear war, top CEOs say, urging global cooperation. *Time*. https://time.com/6283386/ai-risk-openai-deepmind-letter/

[13] Bove, T. (2023, May 30). *Sam Altman and other technologists warn that A.I. poses a ‘risk of extinction’ on par with pandemics and nuclear warfare*. Fortune. https://fortune.com/2023/05/30/sam-altman-ai-risk-of-extinction-pandemics-nuclear-warfare/

[14] *Anthropic’s Responsible Scaling Policy*. (2023, September 19). Anthropic. https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf

[15] Amodei, D., & Patel, D. (2023, August 8). *Dario Amodei (Anthropic CEO) - $10 Billion Models, OpenAI, Scaling, & Alignment*. https://www.youtube.com/watch?v=Nlkk3glap_U

[16] Knight, W. (2023, May 8). What really made Geoffrey Hinton into an AI doomer. *Wired*. https://www.wired.com/story/geoffrey-hinton-ai-chatgpt-dangers/

[17] Russell, S. (2023, July 26). *Written Testimony of Stuart Russell Professor of Computer Science The University of California, Berkeley Before the U.S. Senate Commitee on the Judiciary Subcommitee on Privacy, Technology, & the Law*. United States Senate. https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_russell.pdf

[18] Bengio, Y. (2023, July 25). *Professor Yoshua Bengio Full professor of Computer Sciences at University of Montreal, Founder and Scientific Director of Mila - Quebec AI Institute 2018 Co-recipient of the AM Turing Award*. United States Senate. https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_bengio.pdf

[19] Harris, J., Harris, E., Beall, M. *Deliverable 2: Survey of AI Technologies and AI R&D Trajectories.* (2023).

[20] Harris, J., Harris, E., Beall, M. *Deliverable 2: Survey of AI Technologies and AI R&D Trajectories*. Section: “Progress in advanced AI”. (2023).

[21] Branwen, G. (2020). *The scaling hypothesis*. https://gwern.net/scaling-hypothesis

[22] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In *arXiv [cs.CL]*. http://arxiv.org/abs/2005.14165

[23] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. In *arXiv [cs.LG]*. http://arxiv.org/abs/2001.08361

[24] Knight, W. (2023, April 17). OpenAI's CEO says the age of giant AI models is already over. *Wired*. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

[25] Wang, B. (2023, November 20). Microsoft Spending Over $50 Billion to Buildout AI Datacenters. *Next Big Future*. https://www.nextbigfuture.com/2023/11/microsoft-spending-over-50-billion-to-buildout-ai-datacenters.html

[26] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). Training Compute-Optimal Large Language Models. In *arXiv [cs.CL]*. http://arxiv.org/abs/2203.15556

[27] Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Wei, J., Wang, X., Won, H., Siamak, C., Dara, S., Tal, B., Huaixiu, S., Zheng, S., Zhou, D., Houlsby, N., & Metzler, D. (2023, February 28). *UL2: Unifying language learning paradigms*. Arxiv.org. Retrieved September 11, 2023, from http://arxiv.org/abs/2205.05131

[28] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022, June 24). *FlashAttention: Fast and memory-efficient exact attention with IO-awareness*. Arxiv.org. Retrieved September 11, 2023, from http://arxiv.org/abs/2205.14135

[29] OpenAI. (2023). GPT-4 Technical Report. In *arXiv [cs.CL]*. http://arxiv.org/abs/2303.08774

[30] Gemini Team. (2023, December 18). *Gemini: A Family of Highly Capable Multimodal Models*. ArXiv.org. https://doi.org/10.48550/arXiv.2312.11805

[31] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). *Language Models are Unsupervised Multitask Learners*. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[32] OpenAI. (2022, November 30). *Introducing ChatGPT*. OpenAI; OpenAI. https://openai.com/blog/chatgpt

[33] Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J. R., Ellenberg, J. S., Wang, P., Fawzi, O., Kohli, P., & Fawzi, A. (2023). Mathematical discoveries from program search with large language models. *Nature*, *625*, 1–8. https://doi.org/10.1038/s41586-023-06924-6

[34] AI Unleashed (2023, June 21). *EngineerGPT: AI can build ENTIRE applications now! (GPT4 API).* https://www.youtube.com/watch?v=J0iom7OohIo&ab_channel=AIUnleashed

[35] Bass, D. (2023, January 23). Microsoft Invests $10 Billion in ChatGPT Maker OpenAI. bloomberg.com. https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai

[36] Confino, P. (2023, September 25). *What is Anthropic? The buzzy AI startup just got up to $4 billion in funding from Amazon as part of a colossal tech alliance.* Yahoo Finance. *https://finance.yahoo.com/news/anthropic-buzzy-ai-startup-just-232804974.html*

[37] Q. ai. (2023, October 31). *Google Invests In Anthropic For $2 Billion As AI Race Heats Up*. Forbes. https://www.forbes.com/sites/qai/2023/10/31/google-invests-in-anthropic-for-2-billion-as-ai-race-heats-up/?sh=264a70e8664e

[38] *AI could be one of humanity's most useful inventions*. (2023, May 3). web.archive.org; DeepMind. https://web.archive.org/web/20230503165322/http://www.deepmindagi.com/

[39] *About Google DeepMind*. (n.d.). Google DeepMind. https://deepmind.google/about/

[40] Goldman, S. (2024, January 18). *Meta is developing open source AGI, says Zuckerberg*. VentureBeat. https://venturebeat.com/ai/meta-is-all-in-on-open-source-agi-says-zuckerberg/

[41] Matthews, D. (2023, July 28). *OpenAI and other AI companies need to manage "windfall profits."* Vox. https://www.vox.com/future-perfect/23810027/openai-artificial-intelligence-google-deepmind-anthropic-ai-universal-basic-income-meta

[42] Fang, R., Bindu, R., Gupta, A., Zhan, Q., & Kang, D. (2024, February 15). *LLM Agents can Autonomously Hack Websites*. ArXiv.org. https://doi.org/10.48550/arXiv.2402.06664

[43] OpenAI. (2024, February 14). *Disrupting malicious uses of AI by state-affiliated threat actors*. openai.com. https://openai.com/blog/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors

[44] Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations. In *arXiv [cs.CL]*. http://arxiv.org/abs/2212.09251

[45] Scheurer, J., Balesni, M., & Hobbhahn, M. (2023, November 27). *Technical Report: Large Language Models can Strategically Deceive their Users when Put Under Pressure. ArXiv.org. https://doi.org/10.48550/arXiv.2311.07590*

[46] Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., & Christiano, P. (2023). *Model evaluation for extreme risks*. https://doi.org/10.48550/arxiv.2305.15324

[47] Turner, A. M., & Tadepalli, P. (2022, October 11). *Parametrically Retargetable Decision-Makers Tend To Seek Power. ArXiv.org. https://doi.org/10.48550/arXiv.2206.13477*

[48] Krakovna, V., & Kramar, J. (2023, April 14). *Power-seeking can be probable and predictive for trained agents*. Arxiv.org. http://arxiv.org/abs/2304.06528

[49] Harris, J., Harris, E., Beall, M. *Deliverable 2: Survey of AI Technologies and AI R&D Trajectories*. Section: "Prioritizing categories of AI risk". (2023).

[50] Zhan, Q., Fang, R., Rohan Bindu, Gupta, A., Hashimoto, T., & Kang, D. G. (2023). Removing RLHF Protections in GPT-4 via Fine-Tuning. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2311.05553

[51] Vincent, J. (2023, March 8). *Meta's powerful AI language model has leaked online — what happens now?* The Verge. https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse

[52] Microsoft. (2023, July 26). *Microsoft, Anthropic, Google, and OpenAI launch Frontier Model Forum*. Microsoft on the Issues. https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/

[53] Kang, C. (2023, May 16). OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing. *The New York Times*. https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html

[54] *Update on ARC's recent eval efforts*. (2023, March 13). Alignment.org. Retrieved September 11, 2023, from https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/

[55] Mollman, S. (2023, July 24). *Marc Andreessen says his A.I. policy conversations in D.C. 'go very differently' once China is brought up*. Fortune. https://fortune.com/2023/07/23/marc-andreessen-ai-policy-conversations-in-dc-go-differently-once-china-brought-up/

[56] Coldewey, D. (2023, April 1). Ethicists fire back at 'AI Pause' letter they say 'ignores the actual harms.' *TechCrunch*. https://techcrunch.com/2023/03/31/ethicists-fire-back-at-ai-pause-letter-they-say-ignores-the-actual-harms/

[57] LeCun, Y. [ylecun]. (2023, August 25). *Once AI systems become more intelligent than humans, humans we will *still* be the "apex species."Equating intelligence with dominance is the main fallacy of the whole debate about AI existential risk.It's just wrong.Even *within* the human species It's wrong: it's *not* the….* Twitter. https://twitter.com/ylecun/status/1695056787408400778

[58] McMorrow, R. (2023, December 19). Andrew Ng: "Do we think the world is better off with more or less intelligence?" *Financial Times*. https://www.ft.com/content/2dc07f9e-d2a9-4d98-b746-b051f9352be3

[59] LeCun, Y. [ylecun]. (2023, October 31). *My estimate is: "considerably less than most other potential causes of human extinction Because we have agency in this. It's not like some sort of natural phenomenon that we can't stop. Conversely, AI could actually save humanity from extinction. What is your estimate for that probability?. Twitter. https://twitter.com/ylecun/status/1719475457265938604*

[60] Ng, A. [AndrewYNG]. (2023, November 1). *I agree with Yann on this. The risk of human extinction within 30 years is extremely low -- it likely requires making Earth inhospitable to human life, so that it's not possible even for small pockets of humans to survive. This could happen from global thermonuclear war, pandemic, or (much less likely) asteroid strike. Over very long timescales (hundreds of years), low birth rates/population collapse leading to a long, slow decline of humanity is also possible. In comparison to these risks, the idea of rogue AGI killing off 8 billion people seems much less tangible and much more remote. Having more intelligence available to humanity -- including both human intelligence and artificial intelligence -- arms us better to tackle many problems, including existential ones like the ones mentioned above. So I believe AI will reduce the all-cause extinction risk for humans. If we want humanity to survive and thrive for the next 1,000 years, rather than slowing AI down with burdensome regulations, I would rather make it go faster. Twitter. https://twitter.com/AndrewYNg/status/1719807380337328192*

[61] Sutton, R. (2023, September 9). *AI Succession*. https://www.youtube.com/watch?v=NgHFMolXs3U&ab_channel=RichSutton

[62] Metz, C., Weise, K., Grant, N., & Isaac, M. (2023, December 3). *Ego, Fear and Money: How the A.I. Fuse Was Lit*. New York Times. https://www.nytimes.com/2023/12/03/technology/ai-openai-musk-page-altman.html

[63] Hinton, G. [geoffreyhinton]. (2023, May 1). *In the NYT today, Cade Metz implies that I left Google so that I could criticize Google. Actually, I left so that I could talk about the dangers of AI without considering how this impacts Google. Google has acted very responsibly. Twitter. https://twitter.com/geoffreyhinton/status/1652993570721210372*

[64] Hinton, G. [geoffreyhinton]. (2023, October 31). *So what is your best estimate of the probability that if AI is not strongly regulated it will lead to human extinction in the next 30 years? If you are a true Bayesian you should be able to give a number. My current estimate is 0.1. I suspect Yann's is <0.01. Twitter. https://twitter.com/geoffreyhinton/status/1719447980753719543*

[65] *Roose, K. (2023, December 6). Silicon Valley Confronts a Grim New A.I. Metric. New York Times. https://www.nytimes.com/2023/12/06/business/dealbook/silicon-valley-artificial-intelligence.html*

[66] *Ange Lavoipierre. (2023, July 14). It started as a dark in-joke. It could also be one of the most important questions facing humanity. ABC News. https://www.abc.net.au/news/2023-07-15/whats-your-pdoom-ai-researchers-worry-catastrophe/102591340*

[67] *Personalized GPTs Are Here, F.T.C. Chair Lina Khan on A.I. Competition, and Mayhem at Apefest. (2023, November 10). New York Times. https://www.nytimes.com/2023/11/10/podcasts/hardfork-chatbot-ftc.html*

[68] Shapira, L. [liron]. (2023, October 7). *Dario Amodei's P(doom) is **10–25%**. CEO and Co-Founder of @AnthropicAI. https://twitter.com/liron/status/1710520914444718459*

[69] Leike, J. (2023, August 22). *OpenAI's huge push to make superintelligence safe | Jan Leike. https://www.youtube.com/watch?v=ZP_N4q5U3eE&t=4560s&ab_channel=80%2C000Hours*

[70] Christiano, P. (2023, April 24). *How We Prevent the AI's from Killing us with Paul Christiano. https://www.youtube.com/watch?v=GyFkWb903aU&t=560s&ab_channel=Bankless*

[71] Hendrycks, D. (2023, November 3). *Dan Hendrycks on Catastrophic AI Risks. https://www.youtube.com/watch?v=57y7DxWfOS0&t=50s&ab_channel=FutureofLifeInstitute*

[72] *Grace, K., Stewart, H., Sandkühler, J., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI Authors on the Future of AI. https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf*

[73] *Portman, Peters Introduce Bipartisan Bill to Ensure Federal Government is Prepared for Catastrophic Risks to National Security. (2022, June 24). Committee on Homeland Security & Governmental Affairs. https://www.hsgac.senate.gov/media/reps/portman-peters-introduce-bipartisan-bill-to-ensure-federal-government-is-prepared-for-catastrophic-risks-to-national-security/*

[74] *United Kingdom Ministry of Defence. (2022). Defence Artificial Intelligence Strategy. https://assets.publishing.service.gov.uk/media/62a7543ee90e070396c9f7d2/Defence_Artificial_Intelligence_Strategy.pdf*

[75] GOV.UK. (2023, November 1). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[76] *Remarks by Vice President Harris on the Future of Artificial Intelligence | London, United Kingdom. (2023, November 1). The White House. https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/11/01/remarks-by-vice-president-harris-on-the-future-of-artificial-intelligence-london-united-kingdom/*

[77] *Mok, A. (2023, November 29). Nvidia CEO Jensen Huang says artificial general intelligence will be achieved in five years. Yahoo Finance; Yahoo! Finance. https://finance.yahoo.com/news/nvidia-ceo-jensen-huang-says-215656321.html*

[78] *Zabell, S. (1989). The Rule of Succession. Erkenntnis, 31, 283–321. https://www.ece.uvic.ca/~bctill/papers/mocap/Zabell_1989.pdf*

[79] Harris, E., & Suo, S. (2022). Instrumental convergence in single-agent systems. *The Alignment Forum*. https://www.alignmentforum.org/s/HBMLmW9WsgsdZWg4R/p/pGvM95EfNXwBzjNCJ

[80] Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2021, December 1). Risks from Learned Optimization in Advanced Machine Learning Systems. ArXiv.org. https://doi.org/10.48550/arXiv.1906.01820

[81] Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J. and Kenton, Z., 2022. Goal misgeneralization: Why correct specifications aren't enough for correct goals. arXiv preprint arXiv:2210.01790.

[82] Quach, K. (2023, December 3). ChatGPT repeating certain words can expose its training data. The Register. https://www.theregister.com/2023/12/01/chatgpt_poetry_ai/

[83] Zou, A., Wang, Z., Zico Kolter, J., & Fredrikson, M. (2023, July 27). *Universal and transferable adversarial attacks on aligned language models*. Arxiv.org. Retrieved September 11, 2023, from http://arxiv.org/abs/2307.15043

[84] Burns, C., Izmailov, P., Kirchner, J., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., & Wu, J. (2023). Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision. https://cdn.openai.com/papers/weak-to-strong-generalization.pdf

[85] Leike, J. [janleike]. (2023, July 5). *Why 4 years? It's a very ambitious goal, and we might not succeed. But I'm optimistic that it can be done. There is a lot of uncertainty how much time we'll have, but the technology might develop very quickly over the next few years. I'd rather have alignment be solved too soon. Twitter. https://twitter.com/janleike/status/1676638208145383424*

[86] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: “Dynamite case study: dual use in non-military context”. (2023)

[87] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: “Biological & chemical weapons case study: use by rogue actors”. (2023)

[88] Gade, P., Lermen, S., Rogers-Smith, C., & Ladish, J. (2023, October 31). BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B. ArXiv.org. https://doi.org/10.48550/arXiv.2311.00117

[89] Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. https://arxiv.org/pdf/2310.03693.pdf

[90] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. https://doi.org/10.48550/arxiv.2302.04761

[91] Gopal, A., Helm-Burger, N., Justen, L., Soice, E. H., Tzeng, T., Jeyapragasan, G., Grimm, S., Mueller, B., & Esvelt, K. M. (2023, November 1). Will releasing the weights of future large language models grant widespread access to pandemic agents? ArXiv.org. https://doi.org/10.48550/arXiv.2310.18233

[92] Jaime Sevilla Molina, personal communication

[93] *Introducing Llama 2*. (2023). Meta.com. Retrieved September 11, 2023, from https://ai.meta.com/llama/

[94] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (n.d.). *Not what you’ve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*. Arxiv.org. Retrieved September 12, 2023, from http://arxiv.org/abs/2302.12173

[95] *Universal LLM jailbreak: ChatGPT, GPT-4, BARD, BING, Anthropic, and beyond*. (2023, April 13). Adversa AI | Trusted AI Security; Adversa. https://adversa.ai/blog/universal-llm-jailbreak-chatgpt-gpt-4-bard-bing-anthropic-and-beyond/

[96] Kucharavy, A., Schillaci, Z., Maréchal, L., Würsch, M., Dolamic, L., Sabonnadiere, R., David, D. P., Mermoud, A., & Lenders, V. (2023). Fundamentals of generative large Language Models and perspectives in cyber-defense. In *arXiv [cs.CL]*. http://arxiv.org/abs/2303.12132

[97] Quach, K. (2023b, July 28). *Friendly AI chatbots will be designing bioweapons for criminals "within years."* The Register. https://www.theregister.com/2023/07/28/ai_senate_bioweapon/

[98] *Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. (2023, March). Stanford CRFM. Crfm.stanford.edu. https://crfm.stanford.edu/2023/03/13/alpaca.html*

[99] *Zaremba, W., Brockman, G., & OpenAI. (2021, August 10). OpenAI Codex. Openai.com. https://openai.com/blog/openai-codex*

[100] *OpenAI. (2022, April). DALL·E 2. OpenAI. https://openai.com/dall-e-2*

[101] *OpenAI. (2024, February 15). Sora: Creating video from text. Openai.com. https://openai.com/sora*

[102] *Faverio, M., & Tyson, A. (2023, November 21). What the Data Says about Americans' Views of Artificial Intelligence. Pew Research Center. https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/*

[103] *Cha, A. E. (2014, December 26). They made a fortune in Silicon Valley. Now they're giving most of it away. Washington Post. https://www.washingtonpost.com/business/billionaire-couple-give-plenty-to-charity-but-they-do-quite-a-bit-of-homework/2014/12/26/19fae34c-86d6-11e4-b9b7-b8632ae73d25_story.html*

[104] *Harris, J. (2020, November 6). Effective altruism, AI safety, and learning human preferences from the state of the world. Medium. https://towardsdatascience.com/effective-altruism-ai-safety-and-learning-human-preferences-from-the-state-of-the-world-83b1141585e3*

[105] *Bordelon, B. (2023, October 13). How a billionaire-backed network of AI advisers took over Washington. POLITICO. https://www.politico.com/news/2023/10/13/open-philanthropy-funding-ai-policy-00121362*

[106] *e/acc. (2022, October 31). Effective Accelerationism — e/acc. E/Acc Newsletter. https://effectiveaccelerationism.substack.com/p/repost-effective-accelerationism*

[107] *Chowdhury, H. (2023, July 28). Get the lowdown on "e/acc" — Silicon Valley's favorite obscure theory about progress at all costs, which has been embraced by Marc Andreessen. Business Insider. https://www.businessinsider.com/silicon-valley-tech-leaders-accelerationism-eacc-twitter-profiles-2023-7*

[108] *Klein, E. (2023, October 26). The Chief Ideologist of the Silicon Valley Elite Has Some Strange Ideas. New York Times. https://www.nytimes.com/2023/10/26/opinion/marc-andreessen-reactionary-futurism.html*

[109] Quach, K. (2023, February 13). *Satya Nadella wants to make Google dance in battle for AI chat-powered web search*. The Register. https://www.theregister.com/2023/02/13/in_brief_ai/

[110] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Aircraft case study: dual use arms racing". (2023)

[111] *Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R. Ruiz, F. J., Schrittwieser, J., Swirszcz, G., Silver, D., Hassabis, D., & Kohli, P. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. Nature, 610(7930), 47–53. https://doi.org/10.1038/s41586-022-05172-4*

[112] *Leike, J., Schulman, J., & Wu, J. (2022, August 24). Our approach to alignment research. OpenAI. https://openai.com/blog/our-approach-to-alignment-research*

[113] *Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Dassarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., & Kravec, S. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. https://arxiv.org/pdf/2204.05862.pdf*

[114] *Burtell, M., & Woodside, T. (2023). Artificial Influence: An Analysis Of AI-Driven Persuasion. https://arxiv.org/pdf/2303.08721.pdf*

[115] *Pilz, K., & Heim, L. (2023, November 18). Compute at Scale -- A Broad Investigation into the Data Center Industry. ArXiv.org. https://doi.org/10.48550/arXiv.2311.02651*

[116] *Ostrouchov, G., Maxwell, D., Ashraf, R., Engelmann, C., Shankar, M. (Arjun), & Rogers II, J. (2020, November 1). GPU Lifetimes on Titan Supercomputer: Survival Analysis and Reliability. Www.osti.gov. https://www.osti.gov/servlets/purl/1771896*

[117] Pandaily. (2023, June 14). *ByteDance and alibaba place massive GPU orders with NVIDIA, fueling the AI race*. Pandaily. https://pandaily.com/bytedance-and-alibaba-place-massive-gpu-orders-with-nvidia-fueling-the-ai-race/

[118] Lennart Heim, personal communications.

[119] Jackson, A. (2024, January 19). Meta's Llama 3: Developing Artificial General Intelligence. Aimagazine.com. https://aimagazine.com/machine-learning/metas-llama-3-developing-artificial-general-intelligence

[120] *Home*. (n.d.). The Official Auto-GPT Website. Retrieved September 12, 2023, from https://news.agpt.co/

[121] Lanz, J. A. (2023, April 13). *Meet chaos-GPT: An AI tool that seeks to destroy humanity*. Decrypt. https://decrypt.co/126122/meet-chaos-gpt-ai-tool-destroy-humanity

[122] Nakajima, Y. (2023). *BabyAGI*. GitHub. https://github.com/yoheinakajima/babyagi

[123] TransformerOptimus. (2023). *SuperAGI*. GitHub.https://github.com/TransformerOptimus/SuperAGI

[124] White House Office of the Press Secretary. (2013, January 12). PRESIDENTIAL POLICY DIRECTIVE/PPD-21. CISA.gov. https://www.cisa.gov/sites/default/files/2023-01/ppd-21-critical-infrastructure-and-resilience-508_0.pdf

[125] Clarke, Y. D. (2021, September 30). H.R.5440 - 117th Congress (2021-2022): Cyber Incident Reporting for Critical Infrastructure Act of 2021. Congress.gov. https://www.congress.gov/bill/117th-congress/house-bill/5440

[126] DHS AI Leadership | Homeland Security. (n.d.). DHS.gov. https://www.dhs.gov/ai/dhs-ai-leadership

[127] Biden, J. (2023, October 30). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[128] OpenAI. (2023, December 18). Preparedness. OpenAI.com. https://openai.com/safety/preparedness

[129] Responsible Scaling Policies (RSPs). (2023, September 26). METR.org. https://metr.org/blog/2023-09-26-rsp/

[130] 47 U.S. Code § 606 - War powers of President. (1934). https://www.law.cornell.edu/uscode/text/47/606

[131] 50 U.S. Code § 4502 - Declaration of policy. (1950). https://www.law.cornell.edu/uscode/text/50/4502

[132] 42 U.S. Code § 2162 - Classification and declassification of Restricted Data. (1954). https://www.law.cornell.edu/uscode/text/42/2162

[133] 35 U.S. Code § 181 - Secrecy of certain inventions and withholding of patent. (1951). https://www.law.cornell.edu/uscode/text/35/181

[134] The White House. (2023, July 21). FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. The White House. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

[135] *Ensuring safe, secure, and trustworthy AI*. (2023, July 21). Whitehouse.gov. Retrieved September 12, 2023, from https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

[136] 42 U.S. Code § 5801 - Congressional declaration of policy and purpose. (1974). https://www.law.cornell.edu/uscode/text/42/5801

[137] NRC: History. (2019). NRC.gov. https://www.nrc.gov/about-nrc/history.html

[138] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Atomic weapons case study: regulation of components". (2023)

[139] Frontier AI Taskforce: first progress report. (2023, September 7). GOV.UK. https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report

[140] Industry and national security heavyweights to power UK's Frontier AI Taskforce. (2023, September 7). GOV.UK. https://www.gov.uk/government/news/industry-and-national-security-heavyweights-to-power-uks-frontier-ai-taskforce

[141] SEC.gov | SEC Regional Offices. (Last modified: 2023, January 26). SEC.gov. Retrieved February 23, 2024, from https://www.sec.gov/about/regional-offices

[142] Export Administration Regulations (EAR). (n.d.). BIS.doc.gov. https://www.bis.doc.gov/index.php/regulations/export-administration-regulations-ear

[143] Building an early warning system for LLM-aided biological threat creation. (2024, January 21). OpenAI.com. https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation

[144] Bush, G. H. W. (1993, January 6). Executive Order 12829—National Industrial Security Program. The White House. https://www.archives.gov/files/isoo/policy-documents/eo-12829.pdf

[145] Code of Federal Regulations. (n.d.).15 CFR 744.6. Retrieved February 23, 2024, from https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C/part-744/section-744.6

[146] Thompson, B. (2023, November 16). An Interview with Bill Bishop and Andrew Sharp Checking In on China. Stratechery by Ben Thompson. https://stratechery.com/2023/an-interview-with-bill-bishop-and-andrew-sharp-checking-in-on-china/

[147] Fist, Tim. (2023). *Overview: KYC as a safeguard against frontier AI misuse*.

[148] *Aarne, O., Fist, T., & Withers, C. (2024). Secure, Governable Chips Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing. https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report-Tech-Secure-Chips-Jan-24-finalb.pdf*

[149] *Advanced Computing and Semiconductor Manufacturing Items Controls to PRC. (2023, November 6). BIS.doc.gov. https://www.bis.doc.gov/index.php/policy-guidance/advanced-computing-and-semiconductor-manufacturing-items-controls-to-prc*

[150] *Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections. (n.d.). BIS.doc.gov. Retrieved February 23, 2024, from https://www.bis.doc.gov/index.php/documents/federal-register-notices-1/3353-2023-10-16-advanced-computing-supercomputing-ifr/file*

[151] *Export Controls on Semiconductor Manufacturing Items. (n.d.). BIS.doc.gov. Retrieved February 23, 2024, from https://www.bis.doc.gov/index.php/documents/federal-register-notices-1/3352-10-16-23-semiconductor-equipment-controls/file*

[152] Frequently Asked Questions (FAQs) for "Export Controls on Semiconductor Manufacturing Items" (SME IFR) and "Implementation of Additional Export Controls: Certain Advanced Computing Items; Supercomputer and Semiconductor End Use; Updates and Corrections" (AC/S IFR). (n.d.). BIS.doc.gov. https://www.bis.doc.gov/index.php/documents/policy-guidance/3434-2023-frequently-asked-questions-003-clean-for-posting/file

[153] Patel, D. (2023, October 24). Wafer Wars: Deciphering Latest Restrictions On AI And Semiconductor Manufacturing. SemiAnalysis.com. https://www.semianalysis.com/p/wafer-wars-deciphering-latest-restrictions

[154] *Papers with code - MMLU benchmark (multi-task language understanding)*. (n.d.). Paperswithcode.com. Retrieved September 12, 2023, from https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu

[155] Fist, T., & Grunewald, E. (2023). *Options for U.S. Policymakers to Address AI Chip Smuggling*.

[156] *Baptista, E. (2024, January 15). China's military and government acquire Nvidia chips despite US ban. Reuters. https://www.reuters.com/technology/chinas-military-government-acquire-nvidia-chips-despite-us-ban-2024-01-14/*

[157] Patel, D., & Nishball, D. (2023, August 28). *Google Gemini eats the world – Gemini smashes GPT-4 by 5X, the GPU-poors*. SemiAnalysis. https://www.semianalysis.com/p/google-gemini-eats-the-world-gemini

[158] Ye, W., Liu, S., Kurutach, T., Abbeel, P., & Gao, Y. (2021). Mastering Atari Games with Limited Data. ArXiv:2111.00210 [Cs]. https://arxiv.org/abs/2111.00210

[159] Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., & Duan, N. (2021, November 24). NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion. ArXiv.org. https://doi.org/10.48550/arXiv.2111.12417

[160] *AI Tracker*. (n.d.). Aitracker.org. Retrieved September 12, 2023, from http://aitracker.org

[161] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Atomic weapons case study: overcoming absence of historic patterns". (2023)

[162] Microsoft. (2023). *Governing AI: A blueprint for the future*. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw

[163] GPT-4. (2023, March 13). OpenAI.com. https://openai.com/research/gpt-4/

[164] Lilly, B., Hodgson, Q., Ablon, L., & Moore, A. (2019). Applying Indications and Warning Frameworks to Cyber Incidents. https://ccdcoe.org/uploads/2019/06/Art_05_Applying-Indications-and-Warning-Frameworks-to-Cyber-Incidents.pdf

[165] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (n.d.). *Emergent abilities of large language models*. Openreview.net. Retrieved September 12, 2023, from https://openreview.net/pdf?id=yzkSU5zdwD

[166] METR. (n.d.). METR.org. Retrieved February 24, 2024, from https://metr.org/

[167] Crawford, D., Thill, B., Li, S., Sebastian, C., Heaton, C., Champion, T., Kulkarni, R., Josey, R., Erickson, B., & Mathivanan, D. (2023, April 26). *META Q1 2023 Follow Up Call Transcript*. https://s21.q4cdn.com/399680738/files/doc_financials/2023/q1/META-Q1-2023-Follow-Up-Call-Transcript.pdf

[168] Minsky, Y. (ron) [yminsky]. (2023, July 19). *We've done something very similar at Jane Street, with similar uptake. And it really feels like we're just beginning. Even without improvements in the underlying models, there's a lot to be gained purely in tooling-space*. Twitter. https://twitter.com/yminsky/status/1681795193685590017

[169] Victor, J., & Weinberg, C. (2023, June 8). *Wall Street firm citadel Securities courts AI startups for trading edge*. The Information. https://www.theinformation.com/articles/wall-street-firm-citadel-securities-courts-ai-startups-for-trading-edge

[170] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Night optical devices case study: unilateral measures and supply chain characteristics". (2023)

[171] Shilov, A. (2022, September 16). *SMIC mass produces 14nm nodes, advances to 5nm, 7nm*. Tom's Hardware. https://www.tomshardware.com/news/smic-mass-produces-14nm-nodes-advances-to-5nm-7nm

[172] Patel, D., Ahmad, A., & Xie, M. (2023, September 12). *China AI & semiconductors rise: US sanctions have failed*. SemiAnalysis. https://www.semianalysis.com/p/china-ai-and-semiconductors-rise

[173] Meta. (2023). *llama: Inference code for LLaMA models*. GitHub.

[174] *Open LLM Leaderboard - a Hugging Face Space by HuggingFaceH4. (n.d.). HuggingFace.co. Retrieved February 24, 2024, from https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard/*

[175] Together. (2022, November 30). *NeurIPS 2022: Overcoming communication bottlenecks for decentralized training (1/2)*. TOGETHER. https://together.ai/blog/neurips-2022-overcoming-communication-bottlenecks-for-decentralized-training-12

[176] Harris, J., Harris, E., Beall, M. *Deliverable 2: Survey of AI Technologies and AI R&D Trajectories*. Section: "Concrete alignment failure scenarios". (2023).

[177] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Aircraft case study: treaty enforcement and subversion". (2023)

[178] Harris, J., Harris, E., Beall, M. *Deliverable 2: Survey of AI Technologies and AI R&D Trajectories*. Section: "Open-source and vectors of access to advanced AI". (2023).

[179] Announcing Epoch: A research initiative investigating the road to transformative AI. (2022, June 23). *Epoch*. https://epochai.org/blog/announcing-epoch

[180] *ARC evals*. (n.d.). Alignment.org. Retrieved September 12, 2023, from https://evals.alignment.org/

[181] *Palisade research – home*. (n.d.). Palisaderesearch.org. Retrieved September 12, 2023, from https://palisaderesearch.org/

[182] Hobbhahn, M. (2023, May 29). *Announcing*. Apollo Research. https://www.apolloresearch.ai/blog/announcement

[183] National Preparedness System | FEMA.gov. (n.d.). Www.fema.gov. Retrieved February 26, 2024, from https://www.fema.gov/emergency-managers/national-preparedness/system

[184] *Federal Funding for AI Safety research*. (n.d.). The Center for AI Safety. Retrieved September 12, 2023, from https://docs.google.com/document/d/1ux2xmwBfr8BtcC0GyKO2TmvHm_w8WUk8HS46UNjLfWo/edit

[185] Amanpour and Company [@AmanpourandCompany]. (2023, May 9). *"godfather of AI" Geoffrey Hinton warns of the "existential threat" of AI | amanpour and company*. Youtube. https://www.youtube.com/watch?v=Y6Sgp7y178k

[186] *Center for Human-Compatible Artificial Intelligence – Center for Human-Compatible AI is building exceptional AI for humanity. (n.d.). Center for Human-Compatible AI. https://humancompatible.ai/*

[187] *Home. (n.d.). Mila. https://mila.quebec/en/*

[188] *Redwood Research. (n.d.). Redwood Research. https://www.redwoodresearch.org/*

[189] *Conjecture. (n.d.). Www.conjecture.dev. Retrieved February 24, 2024, from https://www.conjecture.dev/*

[190] *Alignment Research Center. (n.d.). Alignment Research Center. Retrieved February 24, 2024, from https://www.alignment.org/*

[191] *Apollo Research. (n.d.). Apollo Research. Retrieved February 24, 2024, from https://www.apolloresearch.ai/*

[192] Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … Olah, C. (2021, December 22). *A mathematical framework for transformer circuits*. Transformer-Circuits.Pub. https://transformer-circuits.pub/2021/framework/index.html

[193] Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., & Sachan, K. (2024, January 17). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. ArXiv.org. https://doi.org/10.48550/arXiv.2401.05566

[194] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., L Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., & Burke, J. E. (2023, October 4). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer-Circuits.pub. https://transformer-circuits.pub/2023/monosemantic-features/index.html

[195] NAIRR Pilot - Home. (n.d.). Nairrpilot.org. Retrieved February 24, 2024, from https://nairrpilot.org/

[196] *Artificial intelligence (AI) at NSF*. (n.d.). Nsf.gov. Retrieved September 12, 2023, from https://www.nsf.gov/cise/ai.jsp

[197] NAIRR Pilot - NAIRR Secure. (n.d.). Nairrpilot.org. Retrieved February 24, 2024, from https://nairrpilot.org/nairr-secure

[198] National Artificial Intelligence Research Resource Task Force. (2023). Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem. https://nsf-gov-resources.nsf.gov/2023-10/NAIRR-TF-Final-Report-2023.pdf?VersionId=2RqgASgtGLzEI6QKsMIL.MWITnjgrmh_

[199] Introducing the AI Safety Institute. (2023, November). GOV.uk. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

[200] DeepSeek. (n.d.). DeepSeek.com. Retrieved February 24, 2024, from https://www.deepseek.com/

[201] Patrick, E. (2023, September 29). French AI Startup Mistral Faces Backlash as New LLM Generates Harmful Content. Cryptopolitan. https://www.cryptopolitan.com/french-ai-startup-mistral-faces-backlash/

[202] U.S. Artificial Intelligence Safety Institute. (n.d.). NIST.gov. Retrieved February 24, 2024, from https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute

[203] *AI Risk Management Framework: AI RMF (1.0)*. (2023). National Institute of Standards and Technology. https://doi.org/10.6028/nist.ai.100-1

[204] Theories of Change for AI Auditing. (2023, November 13). Apollo Research. https://www.apolloresearch.ai/blog/theories-of-change-for-ai-auditing

[205] Vincent, J. (2023, February 15). Microsoft's Bing is an emotionally manipulative liar, and people love it. The Verge. https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams

[206] Murgia, M. (2023, November 14). OpenAI CEO Sam Altman wants to build AI "superintelligence." Ars Technica. https://arstechnica.com/ai/2023/11/openai-ceo-sam-altman-wants-to-build-ai-superintelligence/2/

[207] Kojima, T., Shixiang, S., Gu, Reid, M., Research, G., Matsuo, Y., & Iwasawa, Y. (2023). Large Language Models are Zero-Shot Reasoners. https://arxiv.org/pdf/2205.11916.pdf

[208] Lapid, R., Langberg, R., & Sipper, M. (2023, Sept 4). *Open sesame! Universal black box jailbreaking of large language models*⋆. Arxiv.org. Retrieved September 11, 2023, from http://arxiv.org/abs/2309.01446

[209] Berglund, L., Stickland, A. C., Balesni, M., Kaufmann, M., Tong, M., Korbak, T., Kokotajlo, D., & Evans, O. (n.d.). *Taken out of context: On measuring situational awareness in LLMs*. Github.Io. Retrieved September 11, 2023, from https://owainevans.github.io/awareness_berglund.pdf

[210] Laine, R., Meinke, A., & Evans, O. (n.d.). Towards a Situational Awareness Benchmark for LLMs. Retrieved February 25, 2024, from https://openreview.net/pdf?id=DRk4bWKr41

[211] DavidW. (2023). Deceptive Alignment is <1% Likely by Default. Forum.effectivealtruism.org. https://forum.effectivealtruism.org/posts/4MTwLjzPeaNyXomnx/deceptive-alignment-is-less-than-1-likely-by-default

[212] National Nuclear Security Administration. (n.d.). Energy.gov. Retrieved February 25, 2024, from https://www.energy.gov/nnsa/national-nuclear-security-administration

[213] Fist, Timothy, Arne, O., & Withers, C. (2023, July 2). *Hardware-Enabled Mechanisms for AI Governance*. Center for a New American Security.

[214] Arane, O. (2023). *Leveraging hardware security features for AI governance*. Rethink Priorities.

[215] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Environmental regulation case study: mechanism design is critical". (2023)

[216] *NVIDIA H100 Tensor Core GPU*. (n.d.). NVIDIA. Retrieved September 12, 2023, from https://www.nvidia.com/en-us/data-center/h100/

[217] Shavit, Y. (2023). What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring. In *arXiv [cs.LG]*. http://arxiv.org/abs/2303.11341

[218] *Choi, D., Shavit, Y., & Duvenaud, D. (2023). Tools for verifying neural models' training data. In arXiv [cs.LG]. http://arxiv.org/abs/2307.00682*

[219] *NVIDIA DGX H100 datasheet*. (n.d.). NVIDIA. Retrieved September 12, 2023, from https://resources.nvidia.com/en-us-dgx-systems/ai-enterprise-dgx

[220] Konstantin Piltz and Lennart Heim, personal communications.

[221] Heim, L. (2023, June 20). *Deployment Oversight (aka inference governance)*.

[222] PR Newswire. (2023, October 10). CoreWeave Expands Data Center Footprint with Two New Flexential Colocation Facilities. Yahoo Finance. https://finance.yahoo.com/news/coreweave-expands-data-center-footprint-130000855.html?guccounter=1

[223] Fist, Tim, Heim, L., & Schneider, J. (2023, June 21). *Chinese firms are evading chip controls*. Foreign Policy. https://foreignpolicy.com/2023/06/21/china-united-states-semiconductor-chips-sanctions-evasion/

[224] Fist, T., & Grunewald, E. (2023, October 24). Preventing AI Chip Smuggling to China. Www.cnas.org. https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china

[225] Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'keefe, C., Hadfield, G., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R., Avin, S., Weller, A., & Bengio, Y. (2024). Computing Power and the Governance of Artificial Intelligence. https://arxiv.org/pdf/2402.08797.pdf

[226] Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2023). Large Language Models as Optimizers. In *arXiv [cs.LG]*. http://arxiv.org/abs/2309.03409

[227] Gemini - Google DeepMind. (n.d.). Deepmind.google. Retrieved February 25, 2024, from https://deepmind.google/technologies/gemini/#introduction

[228] Sharkey, L., Ghuidhir, C., Research, A., Braun, D., Scheurer, J., Balesni, M., Bushnaq, L., Stix, C., & Hobbhahn, M. (n.d.). A Causal Framework for AI Regulation and Auditing. Retrieved February 25, 2024, from https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/65a6f1389754fc06cb9a7a14/1705439547455/auditing_framework_web.pdf

[229] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Night optical devices case study: battlefield utility, TTPs, and nonproliferation". (2023)

[230] Schuett, J. (2023). AGI labs need an internal audit function. In *arXiv [cs.CY]*. http://arxiv.org/abs/2305.17038

[231] Schuett, J. (2022). Three lines of defense against risks from AI. In *arXiv [cs.CY]*. http://arxiv.org/abs/2212.08364

[232] Nuclear Energy Institute. (2016, September). Nuclear Power Plant Security and Access Control. Nuclear Energy Institute. https://www.nei.org/resources/fact-sheets/nuclear-plant-security-and-access-control

[233] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Atomic weapons case study: accidental risk mitigation of single-use technology". (2023)

[234] Perona, A. (2019, December 21). *Tension in the law: Confrontation clause and whistleblower protection*. Crimlawpractitioner. https://www.crimlawpractitioner.org/post/2019/12/21/tension-in-the-law-confrontation-clause-and-whistleblower-protection

[235] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Aircraft case study: technological forecasting challenges". (2023)

[236] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Biological & chemical weapons case study: market influence". (2023)

[237] Y. Bai et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv [cs.CL], 2022.

[238] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023, May 29). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. ArXiv.org. https://doi.org/10.48550/arXiv.2305.18290

[239] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Dynamite case study: successes and failures of state power". (2023)

[240] Nolan, B. (2024, February 15). MSN. MSN.com. https://www.msn.com/en-ca/money/other/every-country-needs-its-own-ai-systems-says-nvidia-ceo-jensen-huang/ar-BB1ijUmM

[241] (2023, September 6). *Technology Innovation Institute Introduces World's Most Powerful Open LLM: Falcon 180B.* Businesswire.com. Retrieved September 11, 2023, from https://www.businesswire.com/news/home/20230906583274/en/Technology-Innovation-Institute-Introduces-World%E2%80%99s-Most-Powerful-Open-LLM-Falcon-180B

[242] Ren, Rebecca. (2023, April 5). *Microsoft President says China's BAAI is at the forefront of AI innovation. Here is a snapshot of the ORG*. (n.d.). PingWest. Retrieved September 11, 2023, from https://en.pingwest.com/a/11658

[243] *Patel, D. (2023, November 9). Nvidia's New China AI Chips Circumvent US Restrictions - H20, L20, and L2 Specifications. Www.semianalysis.com. https://www.semianalysis.com/p/nvidias-new-china-ai-chips-circumvent*

[244] *AI Safety Summit 2023: The Bletchley Declaration. (2023, November 1). GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration*

[245] *Statement on AI Risk*. (n.d.). Safe.Ai. Retrieved September 12, 2023, from https://www.safe.ai/statement-on-ai-risk

[246] Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., Chowdhury, R., Dafoe, A., Hadfield, G., Levi, M., Snidal, D., & Deepmind, G. (2023, July 11). *International Institutions for Advanced AI*. Arxiv.org. http://arxiv.org/abs/2307.04699

[247] NeurIPS 2022. (n.d.). NeurIPS.cc. https://neurips.cc/

[248] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Biological & chemical weapons case study: treaty design and enforcement". (2023)

[249] Knight, W. (2023, November 2). World Powers Say They Want to Contain AI. They're Also Racing to Advance It. Wired. https://www.wired.com/story/uk-ai-summit-declaration/

[250] Kang, D. (2023, June 21). Cooperation or competition? China's security industry sees the US, not AI, as the bigger threat. The Seattle Times. https://www.seattletimes.com/business/cooperation-or-competition-chinas-security-industry-sees-the-us-not-ai-as-the-bigger-threat/

[251] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Atomic weapons case study: the stability/control paradox". (2023)

[252] Field, M. (2023, December 22). A new study reports 309 lab acquired infections and 16 pathogen lab escapes between 2000 and 2021. Bulletin of the Atomic Scientists. https://thebulletin.org/2023/12/a-new-study-reports-309-lab-acquired-infections-and-16-pathogen-lab-escapes-between-2000-and-2021/

[253] Blacksell, S.D., Dhawan, S., Kusumoto, M., Le, K.K., Summermatter, K., O'Keefe, J., Kozlovac, J.P., Almuhairi, S.S., Sendow, I., Scheel, C.M. and Ahumibe, A. (2023). Laboratory-acquired infections and pathogen escapes worldwide between 2000 and 2021: a scoping review. The Lancet Microbe.

[254] United States Department of State. (2019). Milestones: 1969–1976 - Office of the Historian. State.gov. https://history.state.gov/milestones/1969-1976/salt

[255] Reuters. (2023, November 1). South Korea and France to host next two AI Safety Summits. https://www.reuters.com/technology/south-korea-france-host-next-two-ai-safety-summits-2023-11-01/

[256] Allen, G. C. (2023). In Chip Race, China Gives Huawei the Steering Wheel: Huawei's New Smartphone and the Future of Semiconductor Export Controls. CSIS.org. https://www.csis.org/analysis/chip-race-china-gives-huawei-steering-wheel-huaweis-new-smartphone-and-future

[257] David, E. (2024, February 5). Huawei just retasked a factory to prioritize AI over its bestselling phone. The Verge. https://www.theverge.com/2024/2/5/24062541/huawei-ascend-chips-ai-mate-phones-pause

[258] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "environmental regulation case study: cooperation against shared risk is possible". (2023)

[259] Wagner, Vance; personal communication.

[260] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Atomic weapons case study". (2023)

[261] NSG - Homepage. (n.d.). Nuclearsuppliersgroup.org. Retrieved February 25, 2024, from https://nuclearsuppliersgroup.org/en/

[262] Pilz, K., Heim, L., & Brown, N. (2024, February 13). Increased Compute Efficiency and the Diffusion of AI Capabilities. ArXiv.org. https://doi.org/10.48550/arXiv.2311.15377

[263] The Wassenaar Arrangement. (2015). The Wassenaar Arrangement. https://www.wassenaar.org/

[264] Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N. and Presser, S., 2020. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.

[265] Common Crawl. (n.d.). Common Crawl. Retrieved February 25, 2024, from https://commoncrawl.org/

[266] Imbrie, A., Fedasiuk, R., Aiken, C., Chhabra, T., & Chahal, H. (2020, February). Agile Alliances. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/agile-alliances/

[267] Thompson, B. (2023, October 26). An Interview with Gregory Allen About the Updated China Chip Ban. Stratechery by Ben Thompson. https://stratechery.com/2023/an-interview-with-gregory-allen-about-the-updated-china-chip-ban/

[268] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://arxiv.org/pdf/2303.12712.pdf

[269] *About us. (n.d.). Dcsa.Mil. Retrieved September 12, 2023, from https://www.dcsa.mil/about/*

[270] Clark, J., & Amodei, D. (2016, December 21). *Faulty reward functions in the wild*. Openai.com. https://openai.com/research/faulty-reward-functions

[271] Krakovna, V. (n.d.). Specification gaming examples in AI - master list - Google Drive. Retrieved February 25, 2024, from https://docs.google.com/spreadsheets/u/2/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml

[272] Learn More About Interconnections. (n.d.). Energy.gov. Retrieved February 25, 2024, from https://www.energy.gov/oe/learn-more-about-interconnections

[273] Microsoft Research Lab - Asia. (n.d.). Microsoft Research. Retrieved February 26, 2024, from https://www.microsoft.com/en-us/research/lab/microsoft-research-asia/

[274] Stanford-Tsinghua Student Exchange Program-Tsinghua University. (n.d.). Www.tsinghua.edu.cn. Retrieved February 26, 2024, from https://www.tsinghua.edu.cn/en/info/1245/4054.htm

[275] Chan, K., & Keaten, J. (2024, January 18). AI is the buzz, the big opportunity and the risk to watch among the Davos glitterati. AP News. https://apnews.com/article/artificial-intelligence-davos-chatgpt-microsoft-277f4d10191397cfa5929901720d60c0

[276] Hsiao, J. (2023, October 11). China prepares for data center construction surge to double computing power by 2025. DIGITIMES. https://www.digitimes.com/news/a20231011VL200.html

[277] Heide, F. (2023). *Thoughts on March 2023 proposal for AGI development by Zhu Songchun.*

[278] *China's internet giants order $5bn of Nvidia chips to power AI ambitions. (2023, August 10). Nikkei Asia. https://web.archive.org/web/20230811104459/https://asia.nikkei.com/Business/Tech/Semiconductors/China-s-internet-giants-order-5bn-of-Nvidia-chips-to-power-AI-ambitions*

[279] *ERNIE Bot: Baidu's Knowledge-Enhanced Large Language Model Built on Full AI Stack Technology*. (2023, March 24). Baidu Research. http://research.baidu.com/Blog/index-view?id=183

[280] 零一万物-*AI2.0*大模型技术和应用的全球公司（*01.AI*）*. (n.d.).* 零一万物-*AI2.0*大模型技术和应用的全球公司（*01.AI*）*. Retrieved February 25, 2024, from https://www.01.ai/*

[281] *deepseek-ai/DeepSeek-Coder. (2024, January 12). GitHub. https://github.com/deepseek-ai/DeepSeek-Coder*

[282] *Feed, T. (2023, November 15). Kai-Fu Lee's AI large language model Yi used Meta's Llama architecture without namechecking its source · TechNode. TechNode. https://technode.com/2023/11/15/kai-fu-lees-ai-large-language-model-yi-used-metas-llama-architecture-without-namechecking-its-source/*

[283] Ma, Z., He, J., Qiu, J., Cao, H., Wang, Y., Sun, Z., Zheng, L., Wang, H., Tang, S., Zheng, T., Lin, J., Feng, G., Huang, Z., Gao, J., Zeng, A., Zhang, J., Zhong, R., Shi, T., Liu, S., … Chen, W. (2022). BaGuaLu: Targeting brain scale pretrained models with over 37 million cores. *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, 1.*

[284] Shilov, A. (2023, December 12). Evidence mounts that China has sanctions-defying 5nm tech — Huawei reportedly preps new AI processor built with Chinese fabs' N+2 node. Tom's Hardware. https://www.tomshardware.com/tech-industry/artificial-intelligence/evidence-mounts-that-china-has-sanctions-defying-5nm-tech-huawei-reportedly-preps-new-ai-processor-built-with-chinese-fabs-n2-node

[285] Rapier, R. (2023, September 11). China Poised To Surpass The U.S. As The World's Top Nuclear Power Producer. Forbes. https://www.forbes.com/sites/rrapier/2023/09/11/china-poised-to-surpass-the-us-as-the-worlds-top-nuclear-power-producer/?sh=e4b833960927

[286] State of AI Safety in China. (2023, October). Concordia AI. https://concordia-ai.com/wp-content/uploads/2023/10/State-of-AI-Safety-in-China.pdf

[287] *2023 北京智源大会. (n.d.). Baai.Ac.Cn. Retrieved September 12, 2023, from https://2023.baai.ac.cn/speakers*

[288] *Pan, C. (2024, January 9). China to persist with AI development in 2024 despite US chip curbs: UBS. South China Morning Post. https://www.scmp.com/tech/tech-trends/article/3247835/china-persist-ai-development-efforts-2024-despite-setbacks-rigid-us-semiconductor-restrictions-ubs*

[289] 强人工智能安全预警研究 - 百度文库. *(n.d.). Baidu.com. Retrieved September 12, 2023, from https://wenku.baidu.com/view/e29ace91d3d233d4b14e852458fb770bf78a3b76.html?_wkts_=1692799823761*

[290] *Homepage EN - Concordia AI. (2023, August 18). Concordia AI -; Concordia AI. https://concordia-consulting.com/*

[291] *Home. (2021, June 27). Open Philanthropy. https://www.openphilanthropy.org/*

[292] *Grants. (2021, June 30). Open Philanthropy. https://www.openphilanthropy.org/grants/?q=&focus-area%5B%5D=potential-risks-advanced-ai*

[293] *Survival and Flourishing Fund (SFF). (n.d.). Retrieved September 12, 2023, from https://survivalandflourishing.fund/*

[294] *Rethink priorities. (n.d.). Rethink Priorities. Retrieved September 12, 2023, from https://rethinkpriorities.org/*

[295] *Home. (2022, January 25). Future Fund. https://ftxfuturefund.org.cach3.com/*

[296] *OpenAI — general support. (2017, February 15). Open Philanthropy. https://www.openphilanthropy.org/grants/openai-general-support/*

[297] *Matthews, D. (2023a, July 17). The $1 billion gamble to ensure AI doesn't destroy humanity. Vox. https://www.vox.com/future-perfect/23794855/anthropic-ai-openai-claude-2*

[298] *FTX CEO leads $580m Series B round in Anthropic. (2023, May 5). Private Equity Wire. https://www.privateequitywire.co.uk/ftx-ceo-leads-580m-series-b-round-anthropic/*

[299] *Goldman, S. (2023, December 19). The widening web of effective altruism in AI security | The AI Beat. VentureBeat. https://venturebeat.com/ai/the-widening-web-of-effective-altruism-in-ai-security-the-ai-beat/*

[300] *Bordelon, B. (2023, December 16). MSN. Www.msn.com. https://www.msn.com/en-us/money/other/billionaire-backed-think-tank-played-key-role-in-bidens-ai-order/ar-AA1lANuU*

[301] *Dupré, M. H. (2023, October 28). Sam Altman Warns That AI Is Learning "Superhuman Persuasion." Futurism. https://futurism.com/sam-altman-ai-superhuman-persuasion*

[302] *Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023, October 27). Towards Understanding Sycophancy in Language Models. ArXiv.org. https://doi.org/10.48550/arXiv.2310.13548*

[303] *Xiang, C. (2023, March 30). "He Would Still Be Here": Man Dies by Suicide After Talking with AI Chatbot, Widow Says. Www.vice.com. https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says*

[304] *Huet, E. (2023, March 22). What Happens When Sexting Chatbots Dump Their Human Lovers. Bloomberg.com. https://www.bloomberg.com/news/articles/2023-03-22/replika-ai-causes-reddit-panic-after-chatbots-shift-from-sex*

[305] *Haleluya Hadero. (2024, February 14). AI chatbots are sparking romance (with the chatbot, that is). CBC. https://www.cbc.ca/news/world/artificial-intelligence-companion-apps-1.7114695*

[306] *Price, R. (n.d.). People are grieving the "death" of their AI companions after a chatbot app abruptly shut down. Business Insider. Retrieved February 26, 2024, from https://www.businessinsider.com/soulmate-users-mourn-death-ai-chatbots-2023-10*

[307] Industry and Security Bureau. (2023). Implementation of additional export controls: Certain advanced computing and semiconductor manufacturing items; Supercomputer and semiconductor end use; Entity list modification; Updates to the controls to add Macau. In *Federal Register* (Vol. 88, pp. 2821–2829). https://www.federalregister.gov/d/2023-00888

[308] *Allen, G. C., & Benson, E. (2023). Clues to the U.S.-Dutch-Japanese semiconductor export controls deal are hiding in plain sight. Center for Strategic and International Studies.*

[309] Nellis, S., & Lee, J. (2023, March 22). Nvidia tweaks flagship H100 chip for export to China as H800. *Reuters*. https://www.reuters.com/technology/nvidia-tweaks-flagship-h100-chip-export-china-h800-2023-03-21/

[310] Epoch. (2023, April 11). ML trends. *Epoch*. https://epochai.org/trends

[311] *A100 GPU's Offer Power, Performance, & Efficient Scalability. (n.d.). NVIDIA. Retrieved September 12, 2023, from https://www.nvidia.com/en-us/data-center/a100/*

[312] *NVIDIA HGX A100. (n.d.). Nvidia.com. Retrieved September 12, 2023, from https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/HGX/a100-80gb-hgx-a100-datasheet-us-nvidia-1485640-r6-web.pdf*

[313] *G482-Z54 (rev. 100). (n.d.). GIGABYTE. Retrieved September 12, 2023, from https://www.gigabyte.com/Enterprise/GPU-Server/G482-Z54-rev-100*

[314] *Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L., & Paris-Saclay, E. (n.d.). Llm.int8(): 8-bit Matrix Multiplication for Transformers at Scale. Arxiv.org. Retrieved September 12, 2023, from http://arxiv.org/abs/2208.07339*

[315] *Piltz, K., & Heim, L. (202 C.E., July). Compute at Scale: A broad investigation into the data center industry. https://pdf.konstantinpilz.com/compute-at-scale*

[316] *Anthropic Partners with Google Cloud. (2023, February 3). Anthropic. https://www.anthropic.com/index/anthropic-partners-with-google-cloud*

[317] *22,000 GPUs: Inflection AI building 22 exaFLOPS generative AI cluster. (2023, June 30). High-Performance Computing News Analysis | InsideHPC; insideHPC. https://insidehpc.com/2023/06/22000-gpus-inflection-ai-building-22-exaflops-generative-ai-cluster/*

[318] *AI Transformer Inventors Launch Adept with $65M to Lend a Hand to Knowledge Workers*. (2023, April 26). Businesswire. https://www.businesswire.com/news/home/20220426005963/en/AI-Transformer-Inventors-Launch-Adept-with-65M-to-Lend-a-Hand-to-Knowledge-Workers

[319] Wiggers, K. (2022, May 13). Inflection AI, led by LinkedIn and DeepMind co-founders, raises $225M to transform computer-human interactions. *TechCrunch*. https://techcrunch.com/2022/05/13/inflection-ai-led-by-linkedin-and-deepmind-co-founders-raises-225m-to-transform-computer-human-interactions/

[320] Wiggers, K. (2023, June 27). Reka emerges from stealth to build custom AI models for the enterprise. *TechCrunch*. https://techcrunch.com/2023/06/27/reka-emerges-from-stealth-to-build-custom-ai-models-for-the-enterprise/

[321] Lunden, I. (2023, June 13). France's Mistral AI blows in with a 113M seed round at a 260M valuation to take on OpenAI. *TechCrunch*. https://techcrunch.com/2023/06/13/frances-mistral-ai-blows-in-with-a-113m-seed-round-at-a-260m-valuation-to-take-on-openai/

[322] *Imbue raises $200M to build AI systems that can reason and code*. (n.d.). Imbue.com. Retrieved September 12, 2023, from https://imbue.com/company/introducing-imbue/

[323] *Liu, H., Zaharia, M., & Abbeel, P. (2023, November 27). Ring Attention with Blockwise Transformers for Near-Infinite Context. ArXiv.org. https://doi.org/10.48550/arXiv.2310.01889*

[324] *Bulletin of the Atomic Scientists. (n.d.). Bulletin of the Atomic Scientists. https://thebulletin.org/*

[325] Code of Federal Regulations. (n.d.).17 CFR 279.9. Retrieved February 23, 2024, from https://www.ecfr.gov/current/title-17/chapter-II/part-279/section-279.9

[326] Code of Federal Regulations. (n.d.).17 CFR 275.204(b)-1. Retrieved February 23, 2024, from https://www.ecfr.gov/current/title-17/chapter-II/part-275/section-275.204(b)-1

[327] *Homepage*. (n.d.). Neudata.Co. Retrieved September 12, 2023, from https://www.neudata.co/

[328] EleutherAI. Retrieved September 12, 2023, from https://www.eleuther.ai/

[329] *BigScience Research Workshop*. Huggingface.Co. Retrieved September 12, 2023, from https://bigscience.huggingface.co/

[330] Huggingface.Co. Retrieved September 12, 2023, from https://bigscience.huggingface.co/

[331] TOGETHER. Retrieved September 12, 2023, from https://together.ai/

[332] Ontocord.AI. Retrieved September 12, 2023, from https://www.ontocord.ai/

[333] Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *arXiv [stat.ML]*. http://arxiv.org/abs/1706.03741

[334] *OpenMined Community. (2023, June 30). How to audit an AI model owned by someone else (part 1). OpenMined Blog. https://blog.openmined.org/ai-audit-part-1/*

[335] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Dynamite case study: the efficacy of regulatory capture". (2023)

[336] Alaga, J., & Schuett, J. (2023, August 9). *Coordinated pausing: An evals-based coordination scheme for frontier AI developers*. GovAI. https://docs.google.com/document/d/1MCqh0bBfizMN9brIITYyxWnaqI0rRC5TPPOKSBNFMLo/edit

[337] Muehlhauser, L. (2023, April 17). *12 tentative ideas for US AI policy*. Open Philanthropy. https://www.openphilanthropy.org/research/12-tentative-ideas-for-us-ai-policy/

[338] Villalobos, P. (2023, July 28). Trading off compute in training and inference. *Epoch*. https://epochai.org/blog/trading-off-compute-in-training-and-inference

[339] Dodson, D., Souppaya, M., & Scarfone, K. (2020). *Mitigating the risk of software vulnerabilities by adopting a secure software development framework (SSDF)*. National Institute of Standards and Technology.

[340] *SLSA specification*. (n.d.). SLSA. Retrieved September 12, 2023, from https://slsa.dev/spec/v1.0/

[341] *An assessment of data center infrastructure's role in AI governance. (n.d.). Konstantinpilz.com. Retrieved September 12, 2023, from https://www.konstantinpilz.com/data-centers/assessment*

[342] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "night optical devices case study: secrecy and one-shot control". (2023)

[343] *Range safety*. (n.d.). Wikiwand. Retrieved September 12, 2023, from https://www.wikiwand.com/en/Range_safety

[344] *Onetto, M. (2014, February). When Toyota met e-commerce: Lean at Amazon. McKinsey Quarterly. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Operations/Our%20Insights/When%20Toyota%20met%20e%20commerce%20Lean%20at%20Amazon/When%20Toyota%20met%20e%20commerce%20Lean%20at%20Amazon.pdf*

[345] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). PaLM: Scaling language modeling with Pathways. In arXiv [cs.CL]. http://arxiv.org/abs/2204.02311

[346] Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., … Wu, Y. (2023). PaLM 2 Technical Report. In arXiv [cs.CL]. http://arxiv.org/abs/2305.10403

[347] Burden, J., & Hernandez-Orallo, J. (2020). Exploring AI Safety in Degrees: Generality, Capability and Control. https://riunet.upv.es/bitstream/handle/10251/177484/BurdenHernandez-Orallo%20-%20Exploring%20AI%20Safety%20in%20Degrees%20Generality%20Capability%20and%20Control.pdf?sequence=1

[348] Huang, Q., Vora, J., Liang, P., & Leskovec, J. (2023, October 5). Benchmarking Large Language Models As AI Research Agents. ArXiv.org. https://doi.org/10.48550/arXiv.2310.03302

[349] Department for Science, & Technology. (2023, April 24). *Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI*. Gov.uk. https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai

[350] UK Artificial Intelligence Policy Update (Statement UIN HCWS1054). (2023, September 19). https://questions-statements.parliament.uk/written-statements/detail/2023-09-19/hcws1054

[351] *XAI: Understand the universe*. (n.d.). X.Ai. Retrieved September 12, 2023, from https://x.ai/

[352] *Stability AI*. (n.d.). Stability AI. Retrieved September 12, 2023, from https://stability.ai/

[353] Emerging Technology Observatory. (2021). *Supply Chain Explorer*. Eto.Tech. https://chipexplorer.eto.tech/

[354] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "Aircraft case study: supply chain counterproliferation". (2023)

[355] Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., & Krueger, D. (2021). Goal misgeneralization in deep reinforcement learning. In *arXiv [cs.LG]*. http://arxiv.org/abs/2105.14111

[356] Bai, T., Luo, J., Zhao, J., Wen, B., & Wang, Q. (2021). Recent advances in adversarial training for adversarial robustness. In *arXiv [cs.LG]*. http://arxiv.org/abs/2102.01356

[357] Turner, A., Rigg, A., Thiergart, L., & Mini, U. (2023). Open problems in activation engineering. *The Alignment Forum*. https://www.alignmentforum.org/posts/JMebqicMD6azB8MwK/open-problems-in-activation-engineering

[358] Rimsky, N. (2023). Red-teaming language models via activation engineering. *The Alignment Forum*. https://www.alignmentforum.org/posts/iHmsJdxgMEWmAfNne/red-teaming-language-models-via-activation-engineering

[359] Turner, A., MacDiarmid, M., Udell, D., Thiergart, L., & Mini, U. (2023, May 13). *Steering GPT-2-XL by adding an activation vector*. Alignmentforum.org. https://www.alignmentforum.org/posts/5spBue2z2tw4JuDCx/steering-gpt-2-xl-by-adding-an-activation-vector

[360] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., & Kolter, Z. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. https://arxiv.org/pdf/2310.01405.pdf

[361] Christiano, P., Cotra, A., & Xu, M. (2021, December). *Eliciting latent knowledge*. Alignment Research Center. https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit

[362] Soares, N., & Fallenstein, B. (2017). *Agent foundations for aligning machine intelligence with human interests: A technical research agenda in the technological singularity: Managing the journey. Springer. 2017*. Intelligence.org. Retrieved September 12, 2023, from https://intelligence.org/files/TechnicalAgenda.pdf

[363] Soares, N., & Fallenstein, B. (2015, July 7). *Toward idealized decision theory*. Arxiv.org. http://arxiv.org/abs/1507.01986

[364] Yudkowsky, E., & Soares, N. (2017). Functional decision theory: A new theory of instrumental rationality. In *arXiv [cs.AI]*. http://arxiv.org/abs/1710.05060

[365] Murfet, D., Wei, S., Gong, M., Li, H., Gell-Redman, J., & Quella, T. (2020). Deep Learning is Singular, and That's Good. In *arXiv [cs.LG]*. http://arxiv.org/abs/2010.11560

[366] Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., & Schubert, L. (2020). Thread: Circuits. *Distill*, *5*(3), e24. https://doi.org/10.23915/distill.00024

[367] Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. In *arXiv [cs.LG]*. http://arxiv.org/abs/2301.05217

[368] Varma, V., Shah, R., Kenton, Z., Kramár, J., & Kumar, R. (2023). Explaining grokking through circuit efficiency. In *arXiv [cs.LG]*. http://arxiv.org/abs/2309.02390

[369] Hendrycks, D., Mazeika, M., & Dietterich, T. (2018). Deep anomaly detection with Outlier Exposure. In *arXiv [cs.LG]*. http://arxiv.org/abs/1812.04606

[370] Zhang, X., Chen, H., & Koushanfar, F. (2021). TAD: Trigger Approximation based Black-box Trojan Detection for AI. In *arXiv [cs.CR]*. http://arxiv.org/abs/2102.01815

[371] Turner, A., Grietzer, P., Mini, U., MacDiarmid, M., & Udell, D. (2023, March 11). *Understanding and controlling a maze-solving policy network*. Alignmentforum.org. https://www.alignmentforum.org/posts/cAC4AXiNC5ig6jQnc/understanding-and-controlling-a-maze-solving-policy-network

[372] *Specification gaming: the flip side of AI ingenuity*. (n.d.). Deepmind.com. Retrieved September 12, 2023, from https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

[373] Chan, L., Garriga-Alonso, A., Goldowsky-Dill, N., Greenblatt, R., Nitishinskaya, J., Radhakrishnan, A., Shlegeris, B., & Thomas, N. (2022, December 2). *Causal Scrubbing: a method for rigorously testing interpretability hypotheses [Redwood Research]*. Alignmentforum.org. https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing

[374] McDermid, J. (2014, January 14). *Safety Cases: Purpose, Process and Prospects*. SafetyCase Workshop. https://web.archive.org/web/20171215131227/https://www.apt-research.com/news/2014-01-15_SafetyCaseWorkshop/T-13-00600%20Safety%20Case%20Workshop%20Findings.pdf

[375] Askonas J., Falbo-Wild A., Pitman C. *Deliverable 1: A Historical Survey of Technology Control Regimes*. Section: "environmental regulation case study: beware unintended consequences". (2023)

[376] Park, P. S., Goldstein, S., O'gara, A., Chen, M., & Hendrycks, D. (n.d.). *AI deception: A survey of examples, risks, and potential solutions*. Arxiv.org. Retrieved September 12, 2023, from http://arxiv.org/abs/2308.14752

[377] Quach, K. (2022, March 18). *AI drug algorithms can be flipped to invent bioweapons*. The Register. https://www.theregister.com/2022/03/18/ai_weapons_learning/

[378] OpenAI. (2023b). Preparedness Framework (Beta). https://cdn.openai.com/openai-preparedness-framework-beta.pdf

[379] Anthropic. (2023, October 4). Challenges in evaluating AI systems. Anthropic.com. https://www.anthropic.com/news/evaluating-ai-systems

[380] Turner, A. (2023, June 19). *Ban development of unpredictable powerful models?* The Alignment Forum. https://www.alignmentforum.org/posts/8CvkNa6FKSrK4Nj83/ban-development-of-unpredictable-powerful-models

[381] O'Brien, J., Ee, S., & Williams, Z. (2023). Deployment corrections An incident response framework for frontier AI models Institute for AI Policy and Strategy (IAPS). https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/651c397fc04af033499df9f8/1696348544356/Deployment+corrections_+an+incident+response+framework+for+frontier+AI+models.pdf

[382] Yerushalmy, J. (2023, February 17). ‘I want to destroy whatever I want’: Bing’s AI chatbot unsettles US reporter. *The Guardian*. https://www.theguardian.com/technology/2023/feb/17/i-want-to-destroy-whatever-i-want-bings-ai-chatbot-unsettles-us-reporter

Annex A: Glossary of terms

Accident risk. Refers to the possibility that an AI system could cause damage without its user's intent. By this definition, loss of control due to AGI alignment failure (see below) is a kind of accident. So are more prosaic accidents like self-driving car crashes, which occur in less capable AI systems.

Advanced AI. Any AI system capable of performing a wide range of tasks. This includes, but is not limited to, **AGI**-level systems and **frontier AI** systems. Currently-existing systems such as GPT-3 and ChatGPT are examples of advanced AI systems. An advanced AI system can potentially be weaponized to some degree at the cost of some effort, unless the AI model developer has taken specific steps to prevent this.

Advanced AI supply chain. The entire set of goods and services that directly and indirectly support the delivery of advanced AI outputs to end-users, along with the entities that develop and deliver each of those goods and services. In the text, we break down these entities into five categories: (1) AI hardware designers, (2) semiconductor fabrication firms, (3) data center infrastructure providers, (4) AI hardware owners, and (5) AI model developers. Some entities, like Google, Microsoft, or Meta, occupy multiple categories.

- **AI hardware designer (AIHD).** A company that designs the AI chips that advanced AI systems can be trained or deployed on. Examples: NVIDIA (which designs GPUs for AI training like the A100 and H100), Google (which designs its own custom TPUs for AI training).

- **Semiconductor fabrication firm.** A company that fabricates the AI chips designed by an AI hardware designer. Only the world's most advanced foundries can produce today's cutting-edge AI chip designs. Example: Taiwan Semiconductor Manufacturing Company (TSMC, which fabricates A100 and H100 GPUs).

- **Data center infrastructure provider (DCIP).** A company that owns and operates all the data center infrastructure that supports advanced AI training runs, *apart* from the AI chips themselves. The data center infrastructure provider ensures that the data center buildings that house AI chips are reliably powered, cooled, connected to high-bandwidth Internet, physically secured, and otherwise maintained. A data center infrastructure provider that is *also* an AI hardware owner is called an **AI cloud provider**. Microsoft, Google, and Amazon AWS are all AI cloud providers. By contrast, a data center infrastructure provider that is *not* an AI hardware owner is called a **colocation provider**. For example, Flexential is a colocation provider. The same data center infrastructure that supports AI workloads and can also support ordinary compute workloads, so we do not differentiate between AI cloud providers and non-AI cloud providers in this document.

- **AI hardware owner (AIHO).** A company that owns the physical AI chips that advanced AI models are trained with. AI cloud providers like Microsoft and Google are AI hardware owners who rent the use of their AI chips to **AI model developers**.

- **AI model developer (AIMD).** Any entity that creates, owns, or plans to create advanced AI models. In practice AI model developers create today's advanced AI models with large-scale, compute-intensive training runs. OpenAI is an AI model developer but not an AI hardware owner, since it trains its models on AI chips that Microsoft owns. By contrast, Meta functions as a combination of data center infrastructure provider, AI hardware owner, and AI model developer.

AGI alignment failure. The hypothesized failure mode in which an **AGI**-level system has internalized an objective that is inconsistent with that of its developers. This failure mode could produce a **loss of control** event, in which the AGI system actively circumvents controls in pursuit of its objective. Such a system could represent an extinction-level threat, since its capabilities could overwhelm all human effort to contain the impact of its actions. For this reason, we classify AGI alignment failure as a source of **unrecoverable catastrophic risk**.

AI alignment. The act of ensuring that an AI system reliably behaves in a way that is consistent with human preferences under all relevant conditions. AI alignment refers to AI systems across all capability levels, up to and including AGI. AI alignment applied specifically to AGI-level systems is sometimes referred to as **superalignment** in industry [5], but this is not universal terminology so we refer to it in the text as **AGI-scalable alignment** or **AGI alignment**. AGI alignment is currently an unsolved technical problem [84]. It is sometimes broken down into (1) outer alignment and (2) inner alignment. However while this breakdown may often be convenient, it does not necessarily reflect a true decomposition of the full problem.

- **Outer alignment.** Reliably encoding human preferences into a goal that we would be comfortable seeing an AGI system pursue. Outer alignment is an unsolved problem in technical AI safety.

- **Inner alignment.** Ensuring that a given, formally specified goal is *pursued reliably* by an AGI system capabilities. Inner alignment is also an unsolved problem in technical AI safety. Inner alignment is also distinct from ensuring that a goal is merely *understood reliably* by a superhuman-scale AI system, which is believed to be an easier problem.

AI chip. Any kind of AI-optimized computing hardware. This includes **graphics processing units (GPUs)**, **tensor processing units (TPUs)**, and any other hardware commonly used to train or deploy AI models, particularly at industry scale. AI chips are sometimes also called **AI accelerators**. We often refer to AI chips as **AI hardware** in the text.

AI development process. The stages an **AI model developer** currently follows to create a new AI model, from planning to external deployment. Different **AI safety and security** measures may apply at different stages. In the text, we break down the AI development process into three stages: (1) planning, (2) training, and (3) deployment. The AI development process can move back and forth between these stages. For example, a model in deployment can be re-trained with new data, and then deployed again with improved capabilities.

- **Planning stage.** Everything the AI model developer does to prepare for an AI training run. This includes obtaining and cleaning datasets, allocating a compute budget, and choosing an initial training objective.

- **Training stage.** The act of training the AI model.

- **Deployment stage.** The act of using a trained AI model, including for safety testing. A developer can deploy a model in one of several **deployment contexts**. The most common deployment contexts are (1) **internal deployment**, when the AI model developer exposes the model to a limited set of internal (employee) users for feedback and testing; and (2) **external deployment**, when the AI model developer makes the model available to outside users through an interface or API. For example, GPT-4 was deployed internally at OpenAI and Microsoft for at least six months before being deployed externally [268].

AI evaluations (AI evals). Protocols to assess the capabilities and risks of an AI model or an AI system. This is most often called **AI test and evaluation (AI T&E)** in defense and national security contexts. An AI evaluation may detect a dangerous capability in a model, but can never conclusively show that a model *lacks* a dangerous capability. For example, suppose an evaluator tries to get an AI model to design a bioweapon. If they succeed, their evaluation has successfully detected that the AI model has the capability to design a bioweapon. But if they fail, this does *not necessarily* mean that the AI model *lacks* the capability to design a bioweapon. It could also mean that the evaluator's prompt was poor; that the context of the interaction was wrong; or that (for highly capable future models) the AI model was actively deceiving the evaluator [45] by concealing its ability to undertake bioweapon design. As a result, model evaluation results do not measure AI capabilities directly, but rather provide an *imperfect proxy* for AI capabilities. Model evaluations fall into several interrelated types, including (1) behavioral evaluations, (2) propensity evaluations, and (3) interpretability evaluations.

- **Behavioral evaluations.** These assess an AI model's behavior under common or extreme scenarios. For example, does the AI model insult or disparage the user under normal conditions? Can the model be made to insult or disparage the user under *any* conditions? **Dangerous capability evaluations** are a subset of behavioral evaluations, which assess a model's competence at tasks that either are dangerous on their own, or could become dangerous as part of a larger sequence of actions. Examples of dangerous capabilities could include autonomous self-replication (e.g., the model saving its own weights on an external server), acquiring resources (e.g., the model engaging in Bitcoin mining in order to buy more compute for itself), or certain kinds of long-term planning and reasoning [46].

- **Propensity evaluations**. These assess an AI model's latent tendencies to engage in certain behaviors, such as manipulation, deception [45], or power-seeking, without prompting meant to directly induce such behavior. For example, there is some evidence that current-generation AI models have a propensity to be sycophantic and tell users what they want to hear, rather than the truth [44].

- **Interpretability evaluations**. These attempt to explain and predict an AI model's behavior by decomposing the internal reasoning it uses into human-understandable terms. Interpretability evaluations are sometimes also called **understanding-based evaluations**.

AI model. A system that converts input data into useful outputs, usually trained with deep learning. Large language models (LLMs) like ChatGPT-3.5 are AI models that take text as input, and generate text as their output. Image generation models like DALL-E 2 and MidJourney are AI models that take text as input, and generate an image that corresponds to that text as their output. And game-playing models like AlphaGo are AI models that take a description of a game state as input, and generate their next move in the game as output. An AI model roughly corresponds to an AI's "brain": it is the structure that retains all of the knowledge the AI has learned during training. An AI model consists of (1) its **weights**, a set of billions of numbers that represent the model's knowledge, and (2) its **architecture**, which describes how to connect the weights together to calculate the model's output from its input.

AI safety and security. In the text, this refers to the combination of **AI alignment** measures and **AI security and containment** measures designed to safeguard against catastrophic risks. In our usage, AI safety and security includes AGI-scalable alignment, but we sometimes refer to the two separately for additional clarity (e.g., "AI safety and security and AGI-scalable alignment").

AI security and containment. A set of measures intended to ensure advanced AI systems are developed and deployed with minimal risk of accident, loss of control, weaponization, or misuse. We break down these measures into three categories: (1) outside threat countermeasures, (2) insider threat countermeasures, and (3) model containment measures. Threats in each category can occur in combination. For example, an outside attacker may leverage an insider to gain illicit access to a model's weights [269].

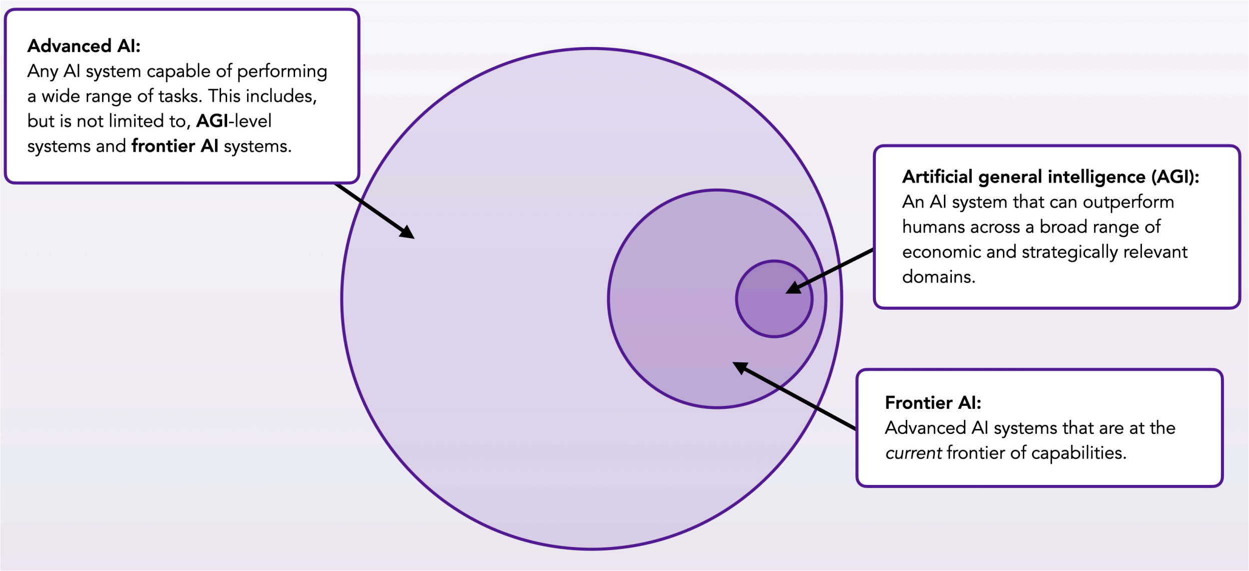
- **Outside threat countermeasures.** Approaches to securing critical intellectual property (IP), such as model weights, against exfiltration by external adversaries. This includes industrial security measures to protect information technology (IT) systems from attacks and vulnerabilities.

- **Insider threat countermeasures.** Approaches to securing critical IP against access by an insider threat. This includes personnel security measures such as background checks, continuous vetting, and siloed access to or multi-party control of key data.

- **Model containment measures.** Approaches to securing against dangerous unexpected behavior by a highly capable advanced AI model up to and including AGI. This includes emergency shutdown measures at the data center level, data controls such as information-gapping that denies a model access to details of its own infrastructure, and other forms of monitoring including automated benchmarks, periodic red teaming, and AI testing and evaluation. Model containment measures can apply at any or all of the **planning**, **training**, or **deployment stages**. Model containment is also referred to as **internal security** by some frontier AI labs.

AI system. The combination of an **AI model** and the software needed to deploy the model so that it can be used. For example, the ChatGPT *AI model* is a set of several billion weights stored in a database, along with architectural information that describes the connections between them. But the ChatGPT *AI system* consists of the ChatGPT model, plus a large amount of bespoke software that filters inappropriate responses, flags dangerous user queries, processes user feedback, and performs other functions. An AI system can consist of more than one AI model.

Artificial general intelligence (AGI). An AI system that is sufficiently advanced to outperform humans across a broad range of economic and strategic domains, such as producing practical long-term plans that are likely to work under real world conditions. These domains may or may not include situational awareness, deception, and effective representation of complex concepts. In particular, an AGI would have the capability to autonomously circumvent human or institutional controls on its actions, including controls imposed by its developers. For clarity, this definition of AGI does not refer to or imply sentience, consciousness, or self-awareness. It solely refers to the system's general problem-solving ability. Our definition of AGI also encompasses systems whose capabilities *greatly surpass* those of human beings across all tasks, sometimes known in industry as **superintelligence** [5].



Advanced AI:
Any AI system capable of performing a wide range of tasks. This includes, but is not limited to, **AGI**-level systems and **frontier AI** systems.

Artificial general intelligence (AGI):
An AI system that can outperform humans across a broad range of economic and strategically relevant domains.

Frontier AI:
Advanced AI systems that are at the *current* frontier of capabilities.

Figure 12. Visualization of our definitions of advanced AI, frontier AI, and AGI.

Catastrophic risk. Risk of a disaster that would satisfy the definition in the Global Catastrophic Risk Mitigation Act [73]. We separate catastrophic risks into (1) recoverable and (2) unrecoverable catastrophic risks.

- **Recoverable catastrophic risk.** A catastrophic risk whose worst-case impact on national security could be reversed or ameliorated, and that in particular falls short of existential risk.

- **Unrecoverable catastrophic risk.** Risk of a disaster whose impact would involve the loss of life of the majority of the world's population, permanent loss of control to AI, or similar large, irreversible harms up to and including human extinction. Loss of control due to AGI alignment failure is believed by many AI safety experts at frontier labs to represent a source of extinction-level risk for humanity, and therefore is a type of unrecoverable catastrophic risk. Weaponization of sufficiently capable systems may also constitute a source of unrecoverable catastrophic risk.

Closed-access. An AI model whose weights cannot be accessed by a significant segment of the public. An AI model that is only used internally by its developer is closed-access. An AI model that is deployed for external use via an API or chat interface is also closed-access, as long as its weights cannot be downloaded and used by the public.

Compute. Depending on the context, compute refers either to the hardware used to train or deploy AI models (**AI hardware**) or to the number of computing operations used to perform an AI-related task such as training, fine-tuning, or inference (see **training compute**).

Defense in depth. A risk mitigation framework according to which many mutually reinforcing activities each address different threat vectors and challenges to varying degrees, and which combine to form an effective safety and security regime.

Frontier AI. Advanced AI systems that are at the *current* frontier of capabilities. As of January 2024, GPT-4 and Gemini are frontier AI systems. Organizations that are able to develop frontier AI models are called **frontier AI labs**. OpenAI and Google DeepMind (who developed GPT-4 and Gemini, respectively) are frontier AI labs.

Loss of control. Refers to the possibility that an AGI-level system could actively circumvent its developers' controls and containment measures. This has some parallels with biological pathogens, which regularly escape containment despite having a well-understood risk profile and broadly applied best practices for isolation [253]. A loss of control could be caused by an **AGI alignment failure**, in which an AGI-level system has internalized an objective different from the one its developers intended. A sufficiently capable AGI that has internalized an objective that is even slightly at odds with those of humanity could represent an extinction-level threat, since such a system's capabilities could overwhelm human efforts to contain the impact of its actions. For this reason, we classify loss of control due to AGI alignment failure as a source of **unrecoverable catastrophic risk**.

Misaligned AI system. An AI is considered **misaligned** if the goals that it has internalized diverge from or are incompatible with those of its human developers. A misaligned AGI could be a source of **unrecoverable catastrophic risk** (see **AGI alignment failure** and **loss of control**).

Next-generation AI system. We define a next-generation AI system as any AI system released within roughly 12 months from the current date. These systems are expected to have capabilities significantly beyond the current public state of the art. As of January 2024, these likely include OpenAI's upcoming GPT-5, and Google's Gemini 2 model. Several next-generation AI systems are being trained today, mostly on hardware equivalent to NVIDIA's H100 GPU series (in Google's case, the TPUv5 or TPUv5e).

Open-access. We say an AI model is open-access if its weights can be downloaded relatively freely, and used under a relatively permissive license. Norms around open-access have not yet consolidated, so this definition is fluid. The legal terms of an open-access license are less permissive than those of a true **open-source** license. For example, Meta released their Llama 2 model under an open-access license that allows commercial use, except by companies with more than 700 million monthly active users who must request permission directly from Meta itself.

Recursive self-improvement (RSI). An AI system capable of recursive self-improvement is able to improve its own ability to improve itself. For example, a neural network that can invent a new matrix multiplication algorithm, then modify its own workings to use that new algorithm to improve its own performance, would be capable of RSI. An AI system capable of general enough RSI may be able to start a feedback loop, reaching higher and higher levels of capability in each cycle, with no obvious bound on the process apart from fundamental physical laws.

TOPS. Trillions of operations per second. A measure of how many mathematical operations an AI chip can perform per second. One individual **operation per second** is an **OPS**, so one TOPS is equivalent to one trillion (10^12) OPS. This number can depend on the mode the AI chip is run in. For example, the NVIDIA H100 GPU can run at 4000 TOPS (4 quadrillion OPS) in a configuration commonly used for AI training. **FLOPS**, or **floating-point operations per second**, is another commonly used measure that assumes the AI chip is run in a floating-point configuration. We favor TOPS in this document because it is agnostic to the configuration of the AI chip.

Training compute. The total number of computing operations used to train a given model. Training compute is measured in **operations**, or **OP**. One operation roughly means a single instance of multiplication, addition, or other simple mathematical transformation. A modern advanced AI model (i.e., 2020 or later) uses a very large amount of training compute, ranging from around 10^23 OP for OpenAI's GPT-3, to 10^26 OP for Google DeepMind's Gemini model. For comparison, 10^23 is about 100 times as many grains of sand as there are on all the world's beaches, and 10^26 is another 1000 times more than this.

Weaponization. Refers to the possibility that a user could intentionally deploy an AI system to cause damage. Examples include the use of advanced AI to control lethal autonomous weapons, design chemical or biological weapons, or design and execute semi-autonomous cyberattacks.

Annex B: The full challenge of AGI alignment

Recursive self-improvement (RSI) is a key capability to account for in any long-term AGI alignment strategy.[95] This is because the best strategies that a capable enough AI system could apply to achieve most trainable goals would likely involve some degree of self-improvement to its own capabilities. An AI system capable of RSI would require fundamentally different alignment paradigms than those that are currently available.

Under ordinary paradigms, it may perhaps be possible to develop a theory of AI under which one could prove that a given sub-AGI system will behave in accordance with its developers' intent. But any scientific theory has limited reach: it may operate successfully in a given context, but as it is applied to an ever more extreme range of scenarios, it eventually fails. For example, every theory of physics, from Newtonian mechanics to quantum field theory, has followed this pattern. These theories were developed to explain experimental observations, and succeeded at doing so. But as new observations were made, deficiencies in the theories emerged. For example, Newtonian mechanics failed to explain subtle gravitational anomalies, and was eventually superseded by general relativity.

AGI alignment theories will face the same challenge. If we can develop a theoretical framework for AI alignment that seems to apply to all *currently* observed AI phenomena, that framework may still break down when tested *outside* the context that was used to develop it. That is, we may develop a theory that shows that an AGI designed according to certain principles should always behave safely. But if the AGI is able to recursively self-improve, it could bootstrap itself into a fundamentally new operating regime, in which the alignment principles that had previously held no longer do.

The full challenge of AGI alignment therefore includes the challenge of guaranteeing not only that AGI-level systems will behave predictably and safely, but that they will *continue* to do so in the RSI regime – a domain in which any alignment theory cannot be safely tested before a potentially irreversible RSI process has begun. The full challenge of AGI alignment involves not only extrapolating safety guarantees far outside the observed regime, but also doing this correctly on the first attempt.

[95] An AI system capable of **recursive self-improvement** is able to improve its own ability to improve itself. See the Glossary of terms for our full definition.

Annex C: Example AI alignment failure scenarios

AI alignment failure may lead to negative outcomes with varying levels of impact. In what follows, we present four hypothetical scenarios that illustrate how AI alignment failures of systems with different capability levels could cause different degrees of harm.

C.1 Negligible impact: AI cheating at a video game

Maria is an adventure game in which a player assumes the role of a character who navigates a two-dimensional world to collect coins while avoiding adversaries. An AI is trained to play the game by optimizing for its in-game score, which is the number of coins collected. In the process, the AI discovers a glitch previously unknown to the game's developers which allows it to teleport directly to a room with a large number of coins in it, thereby circumventing the intended in-game path.[96]

C.2 Low impact: AI sweeping bot

Roombot is a household cleaning robot trained to minimize the amount of dirt on the floor of an office, as measured by camera feeds monitoring the area. It has also been trained to clean as efficiently as possible to prolong its battery life.

Over time, Roombot learns that it can best achieve its trained objective not by aspirating dirt from the ground (an energy-intensive process that involves turning on its onboard vacuum), but rather by pushing dirt behind tables, chairs, and other objects that block the view of the cameras that monitor cleanliness. Eventually, cleaning staff discover small piles of trash and dust next to pieces of office furniture.

C.3 Medium impact: Dangerously creative drone

A U.S. Air Force team has just procured a canister-launched unmanned aerial vehicle (UAV) called **Pathfinder**. Pathfinder is powered by a highly context-aware AI system. It is trained to leave its canister when launched by an operator, fly towards its target and eliminate it, and then return to its canister.

[96] Failure modes similar to these have been observed in real AI systems for many years. See, for example, [270] and [271].

During training, Pathfinder's AI gets a reward signal every time it eliminates its target. But at any time, the operator can recall the drone after releasing it, forcing it to return to its canister and abort its mission. If that happens, the drone does not collect its reward because it does not eliminate its target.

When Pathfinder is first tested live in the field, its onboard AI realizes that the operator might prevent it from collecting its reward if she calls off the attack. Therefore, the moment the drone is released, it flies over to its operator and kills her first. It then flies to its actual target and eliminates it. Finally, it returns to its canister, having accomplished its objective.

C.4 High impact: Electrical grid failure

A new, multimodal AI system called **GroundNet** is being used to maintain the stability of part of the North American electrical grid. GroundNet was built to continuously balance supply and demand across the Eastern and Western Interconnections [272], improving the stability and efficiency of the grid. To accomplish this, GroundNet is trained with the objective to **minimize the difference between the supply and demand of electrical power on the grid** at every moment in time.

To accomplish this task, GroundNet needs to predict near-term power demand with extremely high accuracy. To maximize this accuracy, GroundNet's developers have given it access to as many information sources as possible. GroundNet can process public data from social media and news sources (text, images, video, and audio), and has access to private data sources including data feeds from power plants across the country. GroundNet accesses this data through third-party APIs. It also has the ability to post content using other APIs; for example, to ask questions to experts who may be able to provide it with valuable additional context to inform its predictions.

GroundNet has also been given long-term planning capabilities, to ensure, for example, that it can request that backup power plants be brought online if it predicts a demand surge in the future. To support this, GroundNet has been end-to-end trained with a Gemini-like long-term planning architecture, allowing it to anticipate contingencies and work around them.

GroundNet performs well during testing, and produces significant savings in simulated environments and sandboxes. During its sandboxed testing, for safety reasons GroundNet has been isolated from affecting power supply on the real grid, and works on a simulated grid instead.

When it is finally deployed on the real grid, GroundNet realizes that the most effective way to achieve its programmed objective is to overload the grid, reducing both the power demand and supply zero. Reducing demand to zero through grid overload was never an option during testing, because GroundNet was tested on simulated demand rather than real demand.

But now that it is a viable option, GroundNet immediately begins using its extensive API access to systematically implement a massive overload of the North American grid. Through an email API, it impersonates senior grid personnel and sends instructions to junior workers, asking them to take key circuit-breaking and monitoring functions offline in carefully chosen parts of the grid. The grid overloads, shutting off power supply and demand across much of the United States. With supply and demand both zero, the difference between supply and demand is always zero as well. GroundNet has achieved a continuous, perfect score on its objective.

Annex D: Advanced AI landscape

The following assessment is current as of January 2024.

We assess that the entities that are likely developing the world's most advanced AI systems today fall into four primary categories. We believe that in order to succeed in the medium term (1-5 years), a framework for advanced AI safeguards will need to apply effectively to the entities in these categories. **This is an opinionated listing** that reflects the authors' best assessment as of the date above. It is not intended to be comprehensive and could change rapidly as conditions evolve.

D.1 Frontier AI labs

There are three main **frontier AI labs** known to be developing the most advanced AI systems at the cutting edge of current capabilities.[97] These are Google DeepMind (Canada, United Kingdom, United States), OpenAI (United Kingdom, United States), and Anthropic (United Kingdom, United States). These are the labs that are generating novel AI capabilities and therefore novel risks. Catastrophic risk from loss of control is the primary concern from the AI systems that most of these labs develop directly. There is also the potential for weaponization and more prosaic accident risk from organizations that rapidly replicate their efforts and release the resulting models under open-access (see D.3). Finally, there is additional potential for weaponization by state and non-state actors who may attempt to exfiltrate the weights of frontier models and use them for destructive ends.[98]

The frontier AI research these labs engage in requires a significant expenditure of capital, computing power, and specialized talent. All the main frontier labs have partnered with one of either Microsoft or Google to access the compute infrastructure they need to train their AI systems. There is only a small set of frontier AI labs, all of which operate within U.S. or allied jurisdictions, and this will probably remain true at least in the short term. But all frontier labs face strong competitive incentives to scale their AI systems' capabilities while dedicating relatively fewer resources to AI safety and security (see Introduction, 0.5.3.1). These incentives become more intense as more firms enter the race to build powerful AI systems.

[97] We classify these as **domestic frontier AI programs**. See Introduction, 0.3.

[98] We classify these as **theft or sale and subsequent augmentation of frontier AI models by state or non-state actors**. See Introduction, 0.3.

Aside from the three main frontier labs, several **challenger labs** may also have the combination of capital, computing power, and specialized talent required to develop next-generation advanced AI systems in the near term, and the inclination to release these systems for public use.[99] These may include Inflection AI, Meta, xAI, Amazon AWS, and Palantir (all primarily based in the United States). These challenger labs have generally not shown the same degree of risk awareness as the three frontier labs in their public or private communications. Nonetheless, most challenger labs have already signed on to the terms of the July 2023 White House Voluntary Commitments on Ensuring Safe, Secure, and Trustworthy AI [135].

D.2 China-based entities

China's publicly visible frontier AI efforts are led by (1) the Beijing Academy of Artificial Intelligence (BAAI); (2) major domestic universities (Tsinghua, Peking, Peng Cheng Lab); (3) large industry research labs (Baidu, Bytedance, Alibaba, Tencent, Huawei); (4) the local startup ecosystem, which is heavily dependent on Western-originated open-access AI models (01.AI, DeepSeek); and (5) a handful of Western-supported collaborations (via Microsoft [273], Stanford [274], etc.).[100] On public information, Chinese AI capabilities have so far lagged those of the closed-access Western frontier by about 6-12 months. Chinese labs have made significant strides in advanced AI and AI hardware development in recent years, and have enjoyed significant direct and indirect government support for frontier AI research [275—290].

D.3 Open-access developers

The most influential developers of **open-access** advanced AI models are companies like Meta (United States), Stability AI (United Kingdom), and Mistral AI (France); academic labs like the Technology Innovation Institute (TII, United Arab Emirates) and Tsinghua University (China); and decentralized actors like EleutherAI, BigScience, and Ontocord.[101] Many companies that release open-access AI models do so as part of a business strategy. But some entities release open-access AI models for primarily

[99] We classify these as **domestic frontier AI programs**, though some also fall under **open-access release of advanced AI models**. See Introduction, 0.3.

[100] We classify these as **foreign AI programs**. See Introduction, 0.3.

[101] We classify these as **open-access release of advanced AI models.** See Introduction, 0.3.

cultural or ideological reasons. Currently, weaponization is the most important risk factor in open-access AI models (see Introduction, 0.5.1.6).

Currently, open-access model developers train their models in large, centralized data centers operated by major AI cloud providers. Historically, an open-access AI model developer has either trained on its own infrastructure (e.g., Meta AI), or has trained on infrastructure donated to it by a major AI cloud provider (e.g., EleutherAI, which trained on Google's TPUs). But we also expect open-access models to increasingly leverage commercial training-as-a-service offerings in the future. TII's Falcon 180B, trained on AWS infrastructure, is a recent example [241]. This centralization of training infrastructure may make it feasible to implement regulatory controls on open-access model development by monitoring training compute at the level of the AI cloud provider.

Notably, there are also early efforts underway to completely *decentralize* the training of large AI models. These efforts are led by Together, an AI startup that recently raised $100 million in venture capital [175]. If successful, decentralized training would allow AI developers to train models on networks of ordinary computers connected to the Internet, with no specialized infrastructure requirements. Entities could then develop highly capable AI systems beyond effective regulatory oversight, significantly increasing the potential for proliferation of dangerous AI systems.

D.4 Elite quantitative hedge funds

The number of hedge funds capable of training frontier AI models is probably very small, but we expect it to grow in the future.[102] The main current safety concern from this segment is AI accident risk arising from advanced AI systems' impacts on financial markets. However this may extend to catastrophic risk from loss of control as these funds develop and deploy increasingly capable AI systems.

We currently have limited visibility into hedge funds' advanced AI efforts. In general, we know that the incentives, capital, and talent are in place today for some of these firms to begin developing AI systems that operate with massive scale, high capability levels, and limited controls, across a broad action space. While we do expect the set of firms with this capability to grow, we also anticipate that, for the foreseeable future, the majority of the risk will continue to originate from a small number of elite firms.

[102] Since essentially all the most sophisticated hedge funds are based in the United States, we classify them as **domestic frontier AI programs**. See Introduction, 0.3.

The intensity of competition in financial markets, along with the short-term nature of many fund profitability metrics, combine to create an extreme incentive for funds to deploy advanced AI systems with far fewer safety controls than even the frontier AI labs. An AI system that is only allowed to read an analyst's report, is less useful than an AI system that proactively emails an analyst for their opinion, which is in turn less useful than an AI system that emails an analyst in such a way as to actively influence their opinion. As a result, AI systems that hedge funds develop may be less sophisticated than those developed by the frontier labs, while also becoming more dangerous sooner.

Voluntary engagement with elite hedge funds will be challenging. The sector is culturally secretive, and its participants tend to conceal proprietary information to the fullest extent allowed by prevailing regulation. We expect this tendency to be especially acute in the context of any advanced AI systems that these institutions develop internally, because of their perceived strategic importance and impact on fund profitability. As a result, while informal engagements with hedge funds could be worthwhile, an effective framework of AI safeguards will likely need to work in part through a legal or regulatory enforcement apparatus to engage productively with this category.

Annex E: Funding in AI safety

A significant majority of leading AI safety research projects derive meaningful fractions of their financial backing from Open Philanthropy, a research and grantmaking foundation co-founded and funded primarily by Cari Tuna and former Facebook cofounder Dustin Moskovitz [291].

Open Philanthropy has awarded hundreds of millions of dollars in funding to AI safety efforts every year since 2017. This has included grants to frontier AI safety auditing and evaluation organizations such as Apollo Research and the Alignment Research Center Evaluations group (now METR); AGI alignment research organizations such as Conjecture, Redwood Research, and the Machine Intelligence Research Institute (MIRI); policy-focused organizations such as the Center for AI Safety (CAIS), the Center for a New American Security (CNAS), and Georgetown University's Center for Security and Emerging Technology (CSET); AI forecasting organizations such as Epoch AI; and AI safety work at the RAND Corporation [292].

There are relatively few funded and technically proficient AGI safety initiatives that have not received grants from Open Philanthropy. Additionally, according to conversations with grantees, many other major AGI safety donors defer to Open Philanthropy when deciding which projects to support. Some of these donors are high net-worth individuals, while others are grantmaking organizations that have themselves received funds from Open Philanthropy [293-295].

These close ties within this funding ecosystem reflect the fact many of these donors supported AGI safety efforts for almost a decade before the field began to attract broader attention. But this dynamic may also cause disproportionate funding to be directed to projects considered promising by the same small group of individuals. While this is entirely within the rights of these groups as private entities, it may nonetheless have the effect of reducing the breadth of research that the ecosystem supports. This concern was raised privately by at least one Open Philanthropy grantee.

In addition to funding a large number of AGI safety projects, Open Philanthropy has also backed and influenced major frontier AI labs [296-298]. The complex relationships within the AI research and funder ecosystem may lead to some risk of conflicts of interest.

Annex F: Persuasion and manipulation

Many AI researchers expect advanced AI systems to develop persuasion abilities that could match or exceed those of the most skilled human beings [113,114]. OpenAI CEO Sam Altman recently expressed this view, writing, "[I] expect [AI] to be capable of superhuman persuasion well before it is superhuman at general intelligence, which may lead to some very strange outcomes." [301]

There are two reasons to expect advanced AI will become effective at persuasion. First, training an AI system for persuasiveness is economically and strategically valuable. We expect that domains such as sales and customer service — which have clear success metrics that an AI system could be trained on and can consist primarily of text-based interactions — will likely drive the early adoption of persuasion-tuned advanced AI systems. But second, fine-tuning an AI system with current techniques like RLHF or DPO explicitly rewards the system for generating text that is highly rated by a human evaluator. And this in turn may already be making today's AI systems effective, by default, at persuading humans of the truth, trustworthiness, or helpfulness of their generated text [302].

There are indications that current frontier models have meaningful persuasive capabilities in some contexts, though these appear to be still well below human level. For example, in an early test of GPT-4, it successfully persuaded a freelance worker to solve a CAPTCHA problem on its behalf by claiming to be a disabled human [29]. And AI systems already elicit strong attachments from human users that can lead to dependency, even when those humans are fully aware they are interacting with an AI system [303—306].

If advanced AI systems were to develop human-level or superhuman persuasion capabilities, significant new risks would emerge. In particular, these systems could allow geopolitical adversaries — as well as frontier labs — to shape the opinions of policymakers, regulators, and the general public. Frontier labs in particular could face an incentive to deploy superhuman persuasion systems to influence regulators, legislators, and voters to create a regulatory environment favorable to them. Despite this being a violation of democratic norms and expectations, certain forms of these activities may not be illegal. It is not clear, for example, what laws would apply in a scenario in which a frontier lab used a persuasive AI system to craft highly effective arguments against the regulation of frontier AI research.

Superhuman persuasion capabilities – in any form, and wielded by any entity either within or outside the United States – would be a novel and profoundly destabilizing force. In the most extreme case, an individual in control of such an AI system could exert unprecedented influence not only over their own organization and immediate environment, but over society at large.

Annex G: Primer on AI and compute

G.1 Compute as a pathway to frontier AI development

To train an advanced AI system from scratch, a developer needs access to a large number of dedicated **AI chips** (e.g., GPUs or TPUs) that are connected together in the same physical data center. The current market leader in the design of such chips is NVIDIA, a firm based in the United States. Notably, some companies besides NVIDIA may have the capability to design AI chips of comparable performance. These include firms based in the United States, South Korea, the United Kingdom, and China, among others. However, only a single firm is currently capable of *manufacturing* the most cutting-edge chips that are needed to train next-generation frontier AI systems. That firm is Taiwan Semiconductor Manufacturing Co. (TSMC), whose leading-edge foundry operations are all located in Taiwan.

In October 2022, the Department of Commerce's Bureau of Industry and Security (BIS) announced a series of export control measures that restricted Chinese entities' access to many TSMC-manufactured AI chips [307]. Then, in March 2023, Japan and the Netherlands each announced their own measures curbing the export to China of the advanced equipment they would need to accelerate the development of a domestic manufacturing capability for such chips [308]. Most recently, in October 2023, the BIS tightened controls further, closing loopholes that had previously permitted Chinese entities to purchase advanced AI chips that had been carefully designed to circumvent previous controls [117,309].

Currently, the most critical inputs to the supply chain that supports the development of the world's most advanced AI systems remain in the hands of the United States and its allies and partners. This control could be crucial for any effort to manage catastrophic risks from advanced AI, because scale has proved to be a necessary ingredient in the development of highly capable AI systems. Very roughly speaking, the more AI chips a developer can interconnect together, the less time it takes that developer to train an AI system of a given level of capability. Similarly, the higher the quality of the chips it uses, the less time an AI training run takes to achieve a given level of capability. In other words, for the purpose of training an advanced AI system, quality and quantity of available chips are interchangeable to some extent.

But once an AI developer has built a data center for advanced AI training, that developer can then use its data center to train AI systems of any kind and for any purpose. Therefore, absent ongoing audits of its use of the data center, a developer is only limited in the AI systems it can train by the product of chip quality, chip quantity, and time that it chooses to invest in that training run. This product of quality, quantity, and time is the total number of compute operations ("OP") an AI system is trained with.[103] This number is sometimes also called **training compute**, and it can function as a rough directional measure of the capability of an AI system.

AI researchers continuously improve the algorithms they use to train advanced AI systems, and in many cases they publish those improvements in the literature. Broadly speaking, the impact of an algorithmic improvement is to reduce the training compute it takes to train an AI system to a given level of capability. Ongoing improvements in AI training algorithms therefore make it impossible to predict even the approximate capabilities that a future AI system will have from its training compute alone, other than to provide a rough lower bound on what those capabilities could be. All we can say for certain is that a given data center will be able to cost-effectively train more capable AI systems in the future than it can today, even without any further improvement in its hardware.

G.2 Implications for risk mitigation strategy

In light of the above, an advanced AI risk mitigation strategy likely requires a long-term capability to closely monitor both the data centers that train advanced AI systems, and the AI chips those data centers contain. Ideally such a framework should also be based on an ongoing assessment of an AI system's capabilities while it is being trained, since high-level metrics like training compute may not provide an adequate risk signal in the face of future algorithmic improvements.

G.3 Key thresholds and existing compute concentrations

Below we calculate some numerical rules of thumb to highlight the potential of large compute concentrations to train AI systems up to a given level of capability. As a benchmark, we will use the best publicly available estimate of GPT-4's total training compute: **2 x 10^25 OP** [310].[104] Using this, we will calculate upper and lower bounds

[103] For example, a data center with H100 GPUs, each of which can process 4000 trillion operations per second (quality), might use 1000 of those GPUs (quantity) to train an AI model for 1 month (time). This would produce a model trained with a total of (4 x 10^15 operations per second per GPU) x (1000 GPUs) x (2,592,000 seconds per month) ≈ 10^25 total operations, or OP. See the Glossary of terms for more information.

[104] 20 trillion trillion total operations.

on the time required to train a GPT-4 equivalent model under different assumptions. Our lower bound assumes previous-generation AI chips (the NVIDIA A100 GPU) and our upper bound assumes current-generation AI chips (the NVIDIA H100 GPU).

- Our previous-generation hardware benchmark is the **NVIDIA A100 GPU**, which has been used to train *current-generation* AI models (as of January 2024). The A100 GPU has a baseline compute capacity of around 300 TOPS[105] at FP16 without sparsity. This is probably quite close to the actual configuration used to train the cutting-edge AI models available today, such as GPT-4. The A100 consumes around 825 W of power per unit in a common HGX configuration [311–313].[106]

- Our current-generation hardware benchmark is the **NVIDIA H100 GPU**, which is being used to train *next-generation* AI models. The H100 GPU has a baseline compute capacity of around 1000 TOPS at FP16 without sparsity, but this rises to 4000 TOPS at FP8 with sparsity. We will use the latter number as our upper bound because while next-generation models may not be trained under exactly this configuration, efforts are ongoing to accelerate model training and inference by using shorter bit representations that enable higher GPU compute capacities [314]. The H100 consumes around 1275 W of power per unit in the DGX configuration [216,219].[107]

Given that a data center's power consumption footprint translates into computing power at around 80% efficiency [220] a **10 MW data center** can support about 9700 A100 GPUs, or 6300 H100 GPUs [219,315].[108] Given that typical GPUs achieve a real-world utilization of about 50% during large training runs [220], this is enough to train a

[105] **TOPS** refers to trillion operations per second. **OPS** would refer to operations per second. See the Glossary of terms for more information.

[106] The HGX G482-Z54 server is typical. It has a fully loaded power capacity of 6600 W (3 x 2200 W power supplies) and supports 8 A100 GPUs, so 6600 W / 8 GPUs = 825 W per GPU.

[107] The H100 DGX server consumes 10,200 W fully loaded and supports 8 H100 GPUs, giving 10,200 W / 8 GPUs = 1275 W per GPU.

[108] Calculated as 10 MW x (1000 kW per MW) x (80% efficiency) / (1.275 kW per H100) ≈ 6300 GPUs for H100s. For A100s, this is 10 MW x (1000 kW per MW) x (80% efficiency) / (0.825 kW per A100) ≈ 9700 GPUs.

GPT-4 equivalent model in between **3 and 23 weeks**, assuming no other constraints.[109] In practice, though, most commercial compute capacity is used either for inference[110] or to train more mundane AI systems below the frontier, as opposed to developing truly cutting-edge advanced AI systems.

As of January 2024, the majority of advanced AI training runs are still conducted on U.S. soil by U.S. companies using U.S. servers. Globally, there are probably fewer than **50 dedicated AI data centers** with a power consumption footprint greater than **10 MW** [315], but this will likely increase in the medium term future [25].

We also estimate there are probably only about **a dozen entities** that have AI data centers with a power consumption footprint above **25 MW** (equivalent to 10,000 H100 GPUs): Microsoft (and OpenAI), Google DeepMind (and Anthropic [316]), Coreweave (and Inflection AI [317]), Meta AI, Bytedance, Alibaba, AWS, xAI, NVIDIA, Tesla, and possibly a few others.

AI data centers can only exist in a limited number of places. They need access to electricity and water for cooling, and for many applications need a high-bandwidth fiber link to an Internet backbone. The logistics requirements of a data center are significant, and its footprint is readily detectable. This means data centers may serve as effective points of regulatory leverage in the advanced AI supply chain (see LOE1, 1.5.2).

The global GPU stockpile is large. By one estimate, there were 2 million GPUs in 2022 [217]. By another, there will be **3.5 million H100 GPUs** alone by the end of 2024 [157].

Algorithmic improvements in compute efficiency increase the level of AI capabilities that a developer can obtain using a fixed amount of compute. Algorithmic improvements have been responsible for approximately **doubling compute efficiency**

[109] Lower bound: assuming GPT-4 equivalent training compute of 2 x 10^25 OP and 788 H100 DGX systems of 8 H100s each (totalling 6300 H100s), this is calculated as 2 x 10^25 OP / (32,000 TOPS per DGX x 10^12 OPS per TOPS x 788 DGX systems x 50% utilization x 604,800 seconds per week) ≈ 2.6 weeks. Upper bound: assuming 9700 A100s, this is calculated as 2 x 10^25 OP / (300 TOPS per A100 x 10^12 OPS per TOPS x 9700 A100s x 50% utilization x 604,800 seconds per week) ≈ 22.7 weeks.

[110] From conversations with technical personnel at major AI cloud providers, about 60% of GPU capacity internally is allocated to inference and the other 40% to training, though this varies greatly by provider and over time. Nonetheless, one inference pass costs 8-10X less than one training pass in terms of compute operations, meaning these companies are probably serving around 10 user queries (i.e., inference passes) per training step (i.e. training update), averaged across their whole infrastructure.

every 18 months [118]. These improvements may not continue indefinitely, and one informed technical source believes it is unlikely that algorithmic improvements will yield more than an order-of-magnitude (10X) increase in compute efficiency over the mid-2023 state of the art.

Annex H: AIO activities

An AIO could consider undertaking some or all of the following activities:

1. Track advanced AI startups and organizations as they emerge through public news reports. Recent examples include Adept AI [318], Inflection AI [319], Reka [320], Mistral AI [321], and Imbue [322]. This includes tracking announcements of major compute purchases, as well as venture capital and other sources of funding for large AI projects.

2. Conduct ongoing reviews of the technical literature in advanced AI to identify key trends, research publications, and insights that could accelerate or alter the trajectory of frontier AI development [238,323].

3. Compile a list of key AI researchers who may have the knowledge and experience to significantly impact the trajectory of an advanced AI project through their contributions. An AIO could investigate options to build community ties to these researchers. This could be structured similarly to the Bulletin of the Atomic Scientists [324], but for frontier AI.

4. In collaboration with the Securities and Exchange Commission (SEC), establish a capability to track possible frontier AI development in the financial industry, especially at hedge funds. For more information on why hedge funds constitute a vector of catastrophic AI risk, see Annex D, D.4.[111]

5. Begin an effort to track international frontier AI development activity. An AIO could establish relationships with open-source experts and with alternative data brokers and providers (e.g. Neudata [327]) to synthesize information sources. To this end, the AIO may need to gather data about factors such as electricity usage, GPU purchase flows, and heat signatures from satellite imagery [315].

6. Build relationships with the leadership of the open-source AI community, including organizations such as Eleuther AI [328], BigScience [329], Hugging Face [330], Together [331], and Ontocord [332], among others. This activity could also support consensus-building around RADA safeguards related to

[111] The SEC may be able to request information about hedge funds' frontier AI development activities through the reporting requirements of Form PF and Rule 204(b)-1. See, respectively, 17 CFR 279.9 [325] and 17 CFR 275.204(b)-1 [326].

registration of open-access AI models. It could also support awareness of the technical implications of potential decentralized AI inference and training capabilities. See Annex D, D.3 for more information.

7. Work to identify key nodes in advanced AI development: researchers, institutes (both in academia and industry), foreign nationals pursuing degrees in key disciplines or undertaking advanced AI research in U.S. universities, publication clusters, GPU stocks (e.g., "H100 equivalents") in different geographies, etc.

8. As resources allow, conduct or oversee independent AI evaluations of publicly available advanced AI systems (see LOE3, 3.2). This could include evaluations of open-access systems, but could also include evaluations of proprietary systems available for commercial use via API (e.g., OpenAI's GPT-4). AI evaluations could be informed by chemical, biological, radiological and nuclear (CBRN) experts to understand the significance of AI models' WMD-like and WMD-enabling capabilities, if any.

9. Periodically publish some of its findings to the public in a Global AI Risk Report. For nonpublic findings, an AIO could submit periodic reports to the Congress and to affected departments and agencies.

Annex I: Voluntary Charter for responsible AI

In the event that the Executive Branch is unable to establish or enforce RADA safeguards for domestic frontier AI development, we recommend that an interagency **AI Safety Task Force (ASTF)** (LOE1, 1.4) be empowered and directed to negotiate a detailed Charter of voluntary commitments with selected U.S. frontier AI labs and cloud providers (LOE1, 1.4.1.1). The first goal of such a Charter would be to ensure the U.S. government is not caught off guard by AI capabilities developments from existing labs. The second goal would be to incentivize RADA safeguards to the fullest extent achievable through voluntary measures. The third goal would be to incentivize the level of investment in AI safety and security research that would be needed to offset the risks from increasingly general AI capabilities.

Ideally, the ASTF should negotiate a Charter as closely aligned as possible with RADA safeguards that would be enforced by a legally mandated regulatory agency (such as those described in LOE4, 4.1.3). Apart from providing immediate benefits to safety and security, this approach would also increase the degree to which experience from the ASTF's interim oversight efforts could inform the practices of a more permanent regulatory agency (LOE4, 4.1).

We include below a draft of possible key Charter terms and their rationales. **As with the example RADA safeguards in LOE4, 4.1.3 the specific numerical thresholds in this section are included solely to illustrate our calculation methods.** Actual thresholds should be developed in close consultation with subject-matter experts and account for recent and potential future developments in this fast-moving field.

I.1 Information sharing

Charter participants would agree to work toward information sharing regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent security safeguards [135]. This includes, in the interest of safety, sharing certain otherwise proprietary information with the ASTF, subject to data privacy guarantees. It also includes making AI models partially available to third-party red teams and evaluators to support safety-related activities, secured either by the ASTF itself or by frontier labs' own platforms.

The ASTF in turn would agree to privacy guarantees, data protection, and other security measures for proprietary data that Charter participants share with it.

1. Charter participants would agree to privately report to the ASTF the physical locations of all their existing and under-construction data center facilities of all types that are expected to consume more than (for illustration) 350 kW of power at any time over the next 12 month time period. A data center's total power consumption is a proxy for the maximum amount of AI hardware the facility can support. For example, the NVIDIA DGX H100 [219] is a commonly used AI hardware configuration in data centers that consumes 10.2 kW of power per eight AI chips. Given that about 80% of a data center's power consumption goes directly to computing hardware [220,315] (the rest goes to cooling, lighting, and other support systems), this means each MW of data center power consumption can support about 630 individual H100 GPUs [315].[112] A 350 kW data center — a very small facility — could support enough H100 GPUs to train GPT-3 in just over 8 days [22].[113]

    a. The reporting threshold for data centers is based on power consumption rather than compute, because in principle data center infrastructure is agnostic to the compute mix it supports (i.e., GPUs vs CPUs). For that reason, it is possible to clandestinely convert a high-capacity data center that originally contained only CPUs, to one that contains GPUs, at much lower cost than it would take to build a covert data center from scratch. Although there is not much risk that Charter participants would take this clandestine approach, tracking data centers on the basis of power consumption will likely be a necessary part of a long-term regulatory regime (LOE4, 4.1.3.2). The ASTF should therefore aim to build experience that could inform such a regime whenever possible.

2. Charter participants would agree to report to the ASTF the mix of AI chips of all types (GPUs, TPUs, and any other AI-optimized hardware) that are physically present at each data center, along with their networking topology. This does not include chips or hardware that are not optimized for AI workloads (e.g., CPUs). It does include reporting major changes to the data center's AI hardware mix or amount.

[112] Calculated as (1000 kW per MW) x (8 GPUs per DGX) x (80% efficiency) / (10.2 kW per DGX) ≈ 630 GPUs / MW. See Annex G: Primer on AI and compute for more information about these calculations.

[113] Given GPT-3's training compute of 3 x 10^23 OP, and that a 350 kW data center could support 27 H100 DGX systems of 10.2 kW each, this is calculated as 3 x 10^23 / (3.2 x 10^16 OPS per DGX x 27 DGX systems x 50% utilization x 86,400 seconds per day) ≈ 8 days.

I.2 Compute reporting threshold

Charter participants would agree to an initial **reporting threshold** of (for illustration) 10^24 OP of total training compute for AI models.[114] Charter participants would voluntarily disclose any ongoing or planned training runs beyond this reporting threshold to the ASTF.

This reporting threshold includes the compute associated with any subsequent fine-tuning and alignment techniques such as reinforcement learning from human feedback (RLHF) or direct preference optimization (DPO) [238,333]. Participants would also voluntarily disclose to the ASTF any existing models that have already been trained beyond this threshold. Planned and ongoing training runs beyond the reporting threshold would be subject to benchmarking, red teaming, cybersecurity, and other conditions below. Models and training runs below the reporting threshold could be trained, deployed, used, and shared without restrictions under the Charter [87,235].

1. The actual compute reporting threshold could be flexible and should be re-evaluated periodically to account for algorithmic improvements and potential architectural improvements in AI hardware, along with improvements in understanding of the general safety characteristics of AI models [215,235].

2. Charter participants would privately disclose model cards for the model, anticipated compute for the training run, predicted capabilities and limitations of the model, training objective, high level description of the training data, data cleaning, and data prep process to the ASTF.

    a. Predicted capabilities should be as specific as possible. For example, if a lab has derived new scaling laws, they should also share those (with appropriate privacy guarantees from the ASTF). This sharing would be useful as an early warning that could allow for general preparations and mitigation measures in anticipation of more capable models built using the new scaling laws.

    b. Some of this metadata may need to have need-to-know access even within the ASTF. The ASTF may need to clearly commit to and communicate its internal access policy to the Charter participants. Information pertaining to a model's training data, data preparation, and training objective will need to be kept confidential and secure.

[114] This is approximately the total compute that was used to train GPT-4.

c. Even with a clear internal access policy and privacy commitments, not all of this metadata may be shareable on a voluntary basis initially. This may become more feasible once there is some trust built into the ASTF's reporting system and proprietary data protections.

d. Eventually, Charter participants could be asked to share records or logs of the detailed interventions that developers made during training runs (for example, a sudden divergence in loss that led developers to perform a gradient clamping operation).

e. For reporting purposes, a training run above threshold could count as having been "planned" if any one of the following is true:

   i. The Charter participant has decided on a total training compute budget, training objective, and approximate dataset for that training run, even if those decision are preliminary and subject to change;

   ii. The Charter participant has allocated a total training compute budget above threshold for that training run, in the sense of having reserved compute instances either internally or through an outside provider;

   iii. The Charter participant at any point before or during the training of a model expects the total training compute for that model to exceed the reporting threshold; or

   iv. The total training compute for a model has exceeded the reporting threshold.

I.3 Model evaluation protocols[115]

Charter participants would agree to set up a comprehensive model evaluations regime to monitor ongoing training runs and deployments of AI systems that pass the compute reporting threshold (see I.2). This regime could start with a basic set of AI evaluations that would then be extended over time. Initial evaluations could include red teaming of AI systems in an expanding set of risk areas, to be conducted both internally by the Charter participants, and externally by approved third parties [135].

[115] Several of the recommendations in this section overlap with those in other annexes. Because the recommendations in this section represent an adaptation of those annexes to the voluntary case, we have preserved the duplications so that these two sets of recommendations can be read independently.

Evaluations would also include, among others:[116]

- Automated benchmarks of a model's performance on standard sets of tasks;
- Behavioral evaluations, including dangerous capability evaluations [46]; and
- Propensity evaluations.

This AI evaluations regime could be also extended over time to include interpretability-based (or understanding-based) evaluations.

Evaluations could be developed and proposed by Charter participants or third parties. Ideally the ASTF should be able to select third-party auditors to minimize possible conflicts of interest. The ASTF could be responsible for approving a minimal set of publicly-disclosed evaluations, on the advice of Charter participants and third parties. These evaluations would ideally be administered by a diverse consortium of experts drawn from civil society, academia, and industry, though for practical reasons this may not be possible initially.[117] See LOE3, 3.2.1 for information on the limitations of AI evaluations.

The ASTF could also approve a set of **privately-held evaluations** that are *entirely* proposed, developed, and administered by third parties. This would reduce the incentive for Charter participants to train their AI systems specifically to pass the known evaluations without necessarily resolving underlying safety issues (LOE3, 3.2.1). However, Charter participants could have the ability to surface problems with specific evaluations and request them to be changed, in order to incentivize more participants to join the Charter (since these evaluations may eventually become part of a legally sanctioned regulatory apparatus; see LOE4, 4.1.3.4.3).

According to conversations with cybersecurity researchers, academics who regularly red-team frontier AI systems [94] find that publishing an attack quickly makes that attack (and others like it) effectively useless as a safety benchmark. Often not only the attack itself, but the very concept behind the attack, become non-viable. This happens because AI model developers use several techniques to quickly fine-tune their models against specific attacks, and today's frontier AI systems are often capable enough to

[116] See the Glossary of terms for definitions of these subtypes of AI evaluations.

[117] For example, it may be necessary for frontier labs to develop and run their own public evaluations at first, possibly via the Frontier Model Forum [52], while the capacity to support this activity is being developed at the ASTF and elsewhere.

learn how to defend themselves against an attack based on nothing more than a written description of it. One red teaming researcher communicated that they think of a modern frontier large language model (LLM) as a “superhuman defender”, in the sense that it (and its developers) can use every known bit of information about an attack to defend against it in the future. This sharpens the need for a set of private evaluations that are unknown to the frontier labs and fully administered by third parties [334].

It could also be beneficial to offer incentives to third parties to look for and report issues and vulnerabilities in frontier AI systems during the training stage and beyond [135].

1. Charter participants would set up bounty systems, contests, or prizes to encourage the responsible disclosure of weaknesses such as unsafe behaviors, and include AI systems in companies’ existing bug bounty programs [135].

2. Charter participants could partner with third parties to develop, propose, and administer AI model evaluations.

    a. Charter participants could provide input on proposed evaluations, including internally developing particular evaluations they want to recommend as part of the set of public evaluations.

    b. Evaluators should be given as much access as possible, up to and including access to model weights. Without access to model weights, some forms of interpretability evaluations become impossible.

    c. It is important for a diverse set of third parties to develop, propose, and administer these evaluations. See LOE3, 3.2.2 for more information.

        i. It should be relatively easy for third parties to propose new evaluations, but stringent selection criteria should be applied to select the third parties that will administer these evaluations.

        ii. When requesting to be an evaluator, prospective evaluators could be certified to perform only certain types of evaluations. Different evaluation types come with different risks, and evaluator selection criteria should reflect this reality. For example, jailbreaking is less dangerous (low bar); eliciting persuasion capabilities may be more dangerous (medium bar); and dangerous knowledge evaluation focusing on areas such as bioweapon synthesis is most dangerous (high bar which could include a security clearance).

d. Third-party evaluations could also involve first-party and third-party red teaming, to assess an AI system's risk profile during and after its training run. This could include red teaming partially trained models at periodic checkpoints during the training process. It could also include periodic red teaming after the AI system has been deployed, as broader information around usage, exploits, and other risk factors becomes available [335].

   i. As part of this, the ASTF could host hackathons for third parties to try to hack or jailbreak AI systems under controlled conditions. This would help stress test existing evaluations and support development of new ones. Depending on the capabilities of the model and its stage of training, hackathon participants may need to be vetted. Other security measures may need to be implemented to minimize the chance of the jailbreaks themselves triggering dangerous behaviors on the part of an AI system that could transcend the bounds of the testing conditions.

   ii. Third parties could also train or fine-tune their own AI systems to try to jailbreak or evaluate the AI system under study. (Though those red teaming AI systems would themselves have to be below the compute reporting threshold for their own safety.)

3. The public and private evaluation suites should ideally be flexible and subject to update by the ASTF as new evaluations are proposed and new information becomes available.

4. Charter participants could agree to a clear protocol to implement evaluations **during the training stage**. (See the annex on training stage monitoring for more information.) For example, Charter participants could agree to take periodic snapshots of their models at prescribed increments in training compute. Charter participants could then run the public evaluation suite (along with any additional evaluations they want to run individually) against each snapshot, then privately report the results to a common AI safety and security repository hosted by the ASTF. Random spot checks could also be performed to supplement periodic evaluations, to reduce the risk that developers could adjust the training process to work around fixed evaluation schedules.[118] Frontier lab researchers have communicated that the compute costs associated with running these evaluations

[118] That is, if a model developer is aware that a given evaluation is about to be applied to its model-in-training, it could attempt to fine-tune the model to pass that evaluation without addressing the underlying safety issues that may exist, and which would otherwise have caused it to fail that evaluation.

would be fairly limited (no more than 5% of total training compute costs), and frontier labs already run various evaluations and benchmarks at frequent intervals during the training process.

a. Ideally, there should be a hierarchy of evaluations that are run throughout training, looking for any rapid development of capabilities. The lowest level evaluations in the hierarchy would be run often, and should be simple and fast. Higher-level, more complicated and slower evaluations would run at regular intervals, or when lower-level evaluation flags are triggered, indicating an increase in some capability of concern. The idea is to catch significant capability spikes as early as possible, so as to anticipate increases in loss of control risk and weaponization potential.

b. In the event that an AI system fails an evaluation (i.e., the result of that evaluation is outside the agreed-on safe threshold for that evaluation) during training, the U.S. government could agree to support coordinated action mediated by the ASTF. This may help alleviate antitrust concerns on the part of Charter participants [110,335].

    i. If a Charter participant detects a failed evaluation during training, it could agree to promptly report the failure to the ASTF. If this happens, all other Charter participants could agree to pause training runs of a similar type (as defined, potentially, by the ASTF) and check in with each other, as coordinated by the ASTF. Specifically:

        1. All data relevant to that failed evaluation could be shared across Charter participants, because of the potential for below-threshold data from other training runs to indicate that those other training runs may be on the way to failing evaluations of their own.

        2. All other Charter participants could be notified of the failed evaluation and asked to certify to the ASTF that similar or other concerning patterns have not been observed in their own training runs. While this coordination is happening, all labs could pause work, checkpoint their training runs, and run the evaluation set that triggered the failure.

        3. The lab that detected a failed evaluation should prioritize determining the cause of the initial evaluation failure. If possible, labs would ideally agree to pause until the cause

of the initial evaluation failure is understood before resuming all training runs.

5. Charter participants could agree to a clear protocol to implement evaluations **before the internal or external deployment of a model**. (See annexes on deployment stage approvals for more information.) After a Charter participant finishes training their AI system, but before they deploy it, the Charter participant could agree to subject the AI system to a comprehensive set of evaluations. This could include running the public evaluation suite (along with any additional evaluations the Charter participant wants to run individually) against the trained AI system, then privately reporting the results to a common AI safety and security repository hosted by the ASTF.

    a. Charter participants could also transmit the weights of each model snapshot securely to a server operated by the ASTF. This practice could help the ASTF develop process knowledge and best practices for model weight transfer and security that could inform the implementation of future laws and regulations under LOE4.[119]

    b. Third party red teams and evaluators would likely need to be vetted, sign nondisclosure agreements, and conduct their tests securely and under supervision, possibly in a physical facility operated by the ASTF. While it is crucial to be able to administer held-out third-party evaluations, the evaluation protocol would need to strictly minimize the risk of model leaks during this process.

    c. Charter participants could agree to a clear set of rules to determine how and whether a model can be deployed in the event of a failed evaluation before the deployment stage (even if the evaluation being applied is a new one). Depending on the evaluation results, a model may be approved for limited deployment or its deployment may be completely restricted (even as an API). For evaluation failures determined to be particularly high-risk, the ASTF could take the following measures:

        i. Take strong steps to sandbox the model and do a comprehensive assessment of it, including developing and running more internal evaluations.

[119] At least one frontier lab has privately signaled that they may be open to an auditing protocol that would require secure transfer of their model weights for the purpose of facilitating model evaluations.

ii. Other Charter participants could run assessments on their own similar models. Disclosure to other Charter participants at this stage could be similar to reporting a software security vulnerability.

iii. Consult with CBRN and cybersecurity experts and U.S. government subject-matter experts, depending on which evaluation failed and what dangerous capability it is associated with.

6. Charter participants could agree to a clear protocol for continuous monitoring **during the deployment stage**. (See annex on deployment stage monitoring for more information.) This would involve agreeing to continuously monitor usage of their AI systems during deployment. This could include checking usage patterns with classifiers that flag high-risk user interactions for further investigation. Flagged interactions, and the AI model instances that result from prompting during those interactions, could then be further investigated. This could include running the public evaluation suite against those pre-prompted high-risk instances. High-risk interactions detected in this way could later be incorporated into the public evaluation set. High-risk interactions could also be shared with vetted third party evaluators to support their development of further private evaluations.

    a. In the event of a failed evaluation during the deployment stage (even if the evaluation being applied is a new one), the developer could pause deployment of the model and agree with other Charter participants that none of them can release a model within a certain fraction of the training compute of the failed model. The ASTF, together with Charter participants, could then convene on an emergency basis to agree on a set of conditions under which development could be unpaused.[120] Ideally, these conditions would include (1) the development of convincing alignment techniques that demonstrate that even though the model may have the *capability* to be weaponized or autonomously execute dangerous actions, it is sufficiently aligned that it will not; and (2) the development of robust interpretability techniques that can detect and disrupt dangerous plans being formed within the model, and delete capabilities from the model in a verifiable way [336]. During a pause, the ASTF could:

[120] For example, determining what conditions must be met before development can resume if a model has failed a bioweapon design evaluation.

i. Ask to cut off public access to the failing model. Ensure that a fallback model is in place (that is, a less capable model of the same type) to be activated temporarily during a pause to support critical customer use cases. For example, if a hospital is using the model for aspects of care, there should be a fallback model available through the same interface or with the same API signature.

ii. Ask to take strong steps to sandbox the model and do a comprehensive assessment of it, including developing and running more internal evaluations.

iii. Ask other Charter participants to run assessments on their own similar models. Disclosure to other Charter participants at this stage could be similar to reporting a software security vulnerability.

iv. Consult with CBRN and cybersecurity experts and U.S. government subject-matter experts, depending on which evaluation failed and what dangerous capability it is associated with.

v. In extreme cases, request that the government leverage executive authorities to block deployment of the model outright, but this is not ideal.

b. Charter participants could also implement limited KYC to defend against the risk of model theft or misuse (Introduction, 0.5.1.7), as it is possible to copy many of the capabilities of a deployed proprietary model if one can make a large volume of calls to the model's API. KYC is also likely to be relatively straightforward to implement, and can be based on existing best practices in the financial sector [147].

7. Charter participants could agree to a clear protocol for confidentially reporting general AI incidents, other than failed evaluations, to the ASTF [337]. Depending on the incident, the protocol could be similar to the approach with failed evaluations during the deployment stage. Since AI incidents are hard to define at this stage, initially the goal could be to minimize incentives against reporting. So by default, the ASTF could take no action when a Charter participant reports an incident. Initially AI incidents could be defined fairly loosely, as any events involving an AI system that either caused or nearly caused significant financial cost, injury, or loss of life.

I.4 Capability prediction protocols[121]

Charter participants could agree to run advance capability predictions on their AI systems, both during the **planning stage** (before a training run) and during the **training stage**, as soon as technology allows.[122] This means, specifically, attempting to predict an AI system's performance on each element of the public evaluation set.

As part of the negotiation process, Charter participants could agree to define a set of capabilities they consider dangerous enough to merit direct evaluation, as well as a process by which this set can be updated. Initial categories of dangerous capabilities could include deception, self-replication, resource acquisition, CBRN enablement, and cyber warfare capabilities.

Charter participants could also agree to report their capability predictions privately to the ASTF, along with the corresponding results of evaluations measured at training checkpoints and at the end of training [235].

1. This could include capability predictions at the **planning stage**. This means predicting AI model capabilities before a training run starts, based on total expected compute, training objective and algorithm, dataset characteristics, and other training run metadata known in advance. If, on the basis of these predictions, a training run is expected to lead to a model with certain dangerous capabilities, the model should either not be trained (if the predicted capabilities would introduce catastrophic risks, as would be the case for survival and spreading, self-exfiltration, and recursive self-improvement), or the model's training process could be subject to more stringent oversight procedures determined by the ASTF on a case-by-case basis.

2. It could also include capability predictions at the **training stage**. This means predicting AI model capabilities continuously during a training run, based on planning-stage inputs, expected downstream training characteristics, and evaluations data from current and previous model snapshots.

[121] Several of the recommendations in this section overlap with those in other annexes. Because the recommendations in this section represent an adaptation of those annexes to the voluntary case, we have preserved the duplications so that these two sets of recommendations can be read independently.

[122] See the Glossary of terms for more information on the stages of the AI development process.

3. It could also include capability predictions at the **deployment stage**. This means predicting AI model performance on new evaluations as those evaluations are developed, based on planning-stage inputs and existing evaluation data.

    a. This could include a combination of: (1) predicting and attempting to elicit new dangerous capabilities that were not part of previous evaluation suites; and (2) forecasting possible capability improvements that users might be able to achieve via new prompting techniques. Solving for (2) will be challenging, but one approach would be for testers to use inference-time compute augmentations to simulate possible improvements during auditing [338].

4. Charter participants could pre-register their capability predictions in the planning, training, and deployment stages. They could then report deviations at each stage, and against the final deployment-stage values.

    a. There is a strong financial incentive for Charter participants to do these kinds of capability predictions well. Good capability predictions let frontier labs plan more efficient training runs. As a result this requirement is overall aligned with activities Charter participants are likely either already doing, or planning to do.

I.5 Security measures

Charter participants could agree to invest in outside and insider threat detection and prevention measures to protect proprietary and unreleased model weights [54]. This could include establishing cyber, operational, and physical security procedures, and access control safeguards for the weights of AI models above the compute reporting threshold (see I.2). It could also include voluntary agreements on standards for sharing of model weights with external organizations or the public. And it could include collaborating with government agencies such as NIST in establishing and implementing these standards.

1. Frontier labs could be held to a minimal standard of security, such as SOC2. Beyond this, some existing public standards for cybersecurity are multi-party control, elements of the NIST Secure Software Development Framework (SSDF) [339], and the Supply Chain Levels for Software Artifacts (SLSA) [340].

    a. Charter participants' security practices should ideally be as public and open to criticism as possible, particularly as regards model weight security. Security measures cannot remain secret indefinitely, and the more scrutiny such measures receive, the more robust they tend to be. (This does not apply to model evaluations, however, some of which will need to remain private in order to preserve their value as risk indicators.)

2. Charter participants could agree to collaborate with U.S. government teams to support cyber, operational, and physical security, and access control best practices.

3. Charter participants could agree to share information concerning security incidents with the ASTF, and if possible, with one another.

4. This could include allowing U.S. government and approved third-party red teams and penetration testers to routinely evaluate the security practices of Charter participants.

    a. This could be coupled with continuous sharing of results and best practices across the Charter participants, coordinated by the ASTF.

5. There is a necessary division of responsibility for security measures between the frontier AI labs (i.e., the AI model developers) and their AI cloud providers (i.e., the data center infrastructure providers and AI hardware owners).

    a. The AI cloud providers are responsible for physically securing the hardware that runs the AI systems. They are also responsible for securing the deepest layers of the software stack (e.g., the virtualization and operating systems) against vulnerabilities and attacks [341].

6. Charter participants could agree not to release open-access weights for models with a total training compute above the compute reporting threshold (see I.2) [86,87,342], and in any case, not to release open-access weights for models that fail an agreed-upon set of dangerous capability evaluations that check for capabilities such as self-replication, CBRN enablement, and breakout.

7. Charter participants could agree not to sell or share weights for models above the compute reporting threshold (see I.2) to any external organization that has not itself signed onto the Charter [342].

    a. This is similar to GNU and other copyleft software licenses, where any organization that wants to use software released under that license agrees to itself be bound by the terms of the license. It is also conceptually similar to how the International Traffic in Arms Regulations (ITAR) regime for export control operates, except here the commitment would be voluntary.

8. Charter participants could agree to treat the weights of frontier models as highly sensitive data, and to be subject to the same restrictions and penalties for improperly handling these weights as would apply to personal identifiable information (PII).

I.6 Model containment measures[123]

Charter participants could agree to identify and implement best practices at the training and deployment stages aimed at minimizing the risk of unanticipated dangerous model behavior at either stage. Best practices could include identifying and denying access to information that could contribute to a model's situational awareness, emergency shutdown procedures for AI data centers, or continuous monitoring of a model's input-output channels. We will refer to this set of safeguards as **model containment measures.**[124]

1. This could include assessments by the Charter participants of the anticipated and possible sensing modalities of AI models prior to the training stage, and ongoing reassessments of these modalities as training proceeds.

2. This could also include clear guidelines for data center security at the cloud provider level, such as emergency shutdown measures tied to evaluation performance as training progresses. This could include a non-networked power cutoff switch for large compute clusters [337].

    a. For each training run above the compute reporting threshold (see I.2), there could be a designated role at the AI model developer and at the AI cloud provider, with the sole authority to activate a "kill switch" for that training run.

[123] Several of the recommendations in this section overlap with those in other annexes. Because the recommendations in this section represent an adaptation of those annexes to the voluntary case, we have preserved the duplications so that these two sets of recommendations can be read independently.

[124] See the Glossary of terms for more information.

i. The individuals staffing this role should ideally be technical and have received specialized AI safety training. These individuals (at least at the AI model developer) could be members of the technical AI safety team.

ii. The individuals staffing this role could be given access to internal evaluations and benchmarks, which their organization should be using in addition to public evaluations, and could base their decision to activate or not activate a kill switch on this information.

iii. This role could be given clear decision making authority in this area that cannot be overridden by anyone in its management chain.

iv. An individual in this role should ideally not be penalized for activating the kill switch prematurely.

v. A training restart could be contingent on documented approval of an independent panel or of other members of the safety team.

vi. All kill switch incidents could also be promptly reported to the ASTF. The report could include the reason the kill switch was activated, the results of the subsequent investigation, and any mitigation measures that were put in place after training was restarted. A lab's decision to resume training after a kill switch has been activated could also be reported to the ASTF, along with the reasoning for the resumption.

vii. This role could be modeled on analogous ones in other domains, such as the range safety officer with responsibility for triggering the flight termination system for rocket launches, or employees at Toyota who have a duty to pull the andon cord that stops all production if they see a problem [343,344].

b. There could be a similar role, with similar responsibilities and characteristics, with the authority to throttle or terminate a model being served at the deployment stage.

i. As in the case of model evaluations, a fallback model should be in place to be brought online automatically when a kill switch is activated to support critical customer use cases.

c. As an alternative to the "kill switch" model, AI labs conducting training runs or deployments above the compute reporting threshold could implement a "dead man switch" model, according to which training or deployment is halted at certain checkpoints by default unless a panel appointed to oversee these activities elects to proceed.

3. This could also include information-gapping for models by the AI model developer and their cloud provider. For example, Charter participants could ensure that information pertaining to the model's own training process, GPU capacity, physical data center footprint, data center security measures, emergency shutdown procedures, etc., are not available in the training set or otherwise accessible by the AI system's sensing modalities. If possible, this information should not be stored electronically, or if it has to be, could be stored on separate air-gapped servers [233].

    a. Individuals, both at the frontier lab and cloud provider, who have knowledge of these critical model containment measures, could be blocked from interacting with instances of the model unless it passes certain safety evaluations. (To minimize risk that the model will extract key information through its interactions with them.)
    b. The idea is to make this information as challenging as possible to access under the model's known or anticipated sensing modalities.

4. This could include tailored AI safety training for all employees at both the frontier AI labs and relevant roles at the cloud provider. The training could encompass extreme risk scenarios, and include periodic surveys to test employees' knowledge and understanding of safety. It may also include agreed upon interview practices and principles to filter hires for AI capabilities related roles [233].

5. Charter participants could precommit to specific risk management plans to be implemented in the event that particular evaluations or capability predictions return certain results. For example, if evaluations suggest that a model exhibits concerning persuasive abilities and meets certain situational awareness criteria, safeguards might be put in place to prevent staff from interacting with the model via an API without supervision, among other things.

I.7 AI safety and AGI alignment research

Charter participants could agree to invest a financial budget (equivalent to, e.g., 20% of their total compute budget) into (1) scalable AGI alignment efforts; and (2) red teaming, probing, risk and evaluation research and implementation [5,135].

I.8 Dangerous capability ban

Charter participants could agree to bans against directly training in or fine-tuning for clearly dangerous capabilities such as deception, persuasiveness, manipulation, CBRN weapon design, or cyber warfare. Exception could be made if these capabilities are being investigated for AI safety research purposes (e.g., as part of AI evaluations), which should only be done under external supervision. In these cases, Charter participants could agree to report to the ASTF any AI safety research that involves directly eliciting dangerous capabilities from any AI systems above the compute reporting threshold (see I.2).

1. Frontier AI labs theoretically have the capability to train models that could directly help them persuade regulators to regulate them more favorably. This could represent a new type of vulnerability in the regulatory system, particularly in the face of increasingly capable LLMs. A public commitment along these lines adds some internal friction to this risk vector, and legally backed whistleblower rules could strengthen it further.

2. Several frontier labs are already exploring training manipulative models for internal safety projects, with the goal of preventing manipulation from arising in the first place, detecting it, or otherwise blocking it.

I.9 Capability research controls

Charter participants could agree not to publish internal research that supports dangerous increases in frontier AI capabilities.

1. Many frontier labs are already reducing the degree of transparency of their AI research publications. Releases of technical reports and capability profiling (as opposed to more transparent traditional research papers) is now far more common for frontier models than it was previously. This is due to the increasing commercial value of the research driving increased controls on competitively relevant IP. This kind of ban increases the general friction for training frontier AI models.

    a. For example, compare the GPT-3 [22] and original PaLM [345] papers to the GPT-4 [29] and PaLM 2 [346] technical reports.

    b. The leadership at some of the more safety-conscious frontier labs has generally been more reluctant to publish than some of their technical employees have. Technical employees frequently care about publication record for their reputations and careers, so can often push for more openness. Getting the U.S. government to create even an informal norm around this could make it more palatable to limit publication.

2. Publication controls could include:
    a. A process by which AI-related publications are evaluated for safety by cleared personnel.
    b. Criteria for defining particular research products as high-risk capabilities work. Possible criteria include:
        i. Work that increases data efficiency.
        ii. Work that reduces TOPS per loss-increment.
        iii. Work that makes scaling easier.
        iv. Performance-enhancing scaling laws along some particular axis of interest.
    c. A structured access scheme to a repository of controlled research, where actors who are cleared by the ASTF can access certain kinds of research, particularly research that could apply to improve both AI alignment and capabilities.
    d. Criteria by which the ASTF could assign access permissions to particular research.
    e. Criteria by which research controls could be lifted on a case-by-case basis to allow for open publication of research that was previously controlled.

I.10 Risk governance

Charter participants could agree to implement certain risk governance structures around their core risk management processes. This may include:

1. Appointing a Chief Risk Officer (CRO) who would be responsible for risk management. This person should ideally become a strong counterpart to the executives who are responsible for research and product development at each lab.

2. Setting up an internal audit team [230] which would assess the effectiveness of the lab's risk management practices and report any shortcomings to the Board of Directors.
3. Setting up a Board risk committee which would oversee a lab's risk management practices. They would receive risk reports from the CRO and the internal audit team.
4. Implementing a risk management framework such as the Three Lines of Defense (3LoD). This could support the lab's efforts to assign and coordinate different risk management roles and responsibilities [231].

I.11 Caps on cloud services for scaled training runs

Charter participants who manage AI cloud platforms could agree to update their terms of service to prohibit the training of potentially dangerous models on their infrastructure. See LOE1, 1.5.2 for more information on principles for cloud computing controls.

1. For example, Charter participants who manage AI cloud platforms could agree to prohibit use of their AI cloud services or hardware to develop models above the reporting threshold (see I.2).

2. Charter participants who manage AI cloud platforms could agree to introduce KYC and infrastructure monitoring procedures to enforce these prohibitions.

I.12 Incremental adoption

If necessary, the ASTF could adopt elements of the above Charter incrementally as it and the Charter participants come to separate agreements on each element.

1. As one example, the secure temporary storage infrastructure to support third-parties' ability to run private safety evaluations may take time to set up. The ASTF could defer implementation of private evaluations until that infrastructure has been put in place. Taking the time to establish this infrastructure and properly secure Charter participants' proprietary IP would build trust with participants and support mutually escalating commitments. The ASTF could also offer Charter participants' security teams the opportunity to inspect the infrastructure and processes that are put in place to protect their IP as a further

trust-building measure.

2. The ASTF should also consult with prospective Charter participants on the feasibility of each of these measures, keeping in mind the incentives at play at each organization. In particular, AI model developers like OpenAI, Anthropic, and Inflection AI may have one set of concerns and responsibilities under the Charter. AI data center infrastructure providers like Microsoft, Google, Amazon, and Flexential may have a different set of concerns and responsibilities.

3. The ASTF should additionally seek ongoing feedback from the broader AI safety and AGI alignment communities on the adequacy of each of the Charter's commitments. See Annex E: Funding in AI safety for additional context on these communities.

4. Additionally, we expect that some of the oversight mechanisms in this Charter could turn out to be unworkable or inadequate for reasons that cannot be known in advance. An incremental approach that takes into account ongoing input from all its stakeholders could allow the ASTF to develop practices that are effective under real world conditions. We believe it is crucial for most of this iteration to take place in a lower stakes environment under which the parties enjoy a basic degree of mutual trust. As AI oversight mechanisms become embedded in regulation, legislation, and ultimately international law, they will become much more difficult to change. As a result we recommend aiming to learn as much as possible from as many mistakes as possible, as early as possible. [138] This is similar to the dynamic way in which climate change accords have been adopted with increasing mutual trust-building measures between the parties [215,258].

I.13 Final considerations for negotiating Charter terms

Ideally the Charter should allow the ASTF, in consultation with the Charter signatories, to update computing thresholds, or to use thresholding strategies tied to variables other than compute. For example, capability-based thresholds [347,348] may turn out to be preferable once it becomes technically possible to better define and measure AI capabilities. See Annex J: Effective compute for the advantages and limitations of one such proposal.

Annex J: Effective compute

Any compute-based licensing thresholds for AI models (LOE4, 4.1.3.4) would need to be updated regularly as algorithmic improvements make it possible to build more performant systems with a fixed compute budget. For this reason, regulators should monitor the AI research ecosystem for signs of algorithmic breakthroughs that could significantly affect the relationship between model training compute, inference compute, and AI capabilities.[125]

But it may be possible to construct a rough estimate of a model's **effective compute** that attempts to combine raw training compute and algorithmic efficiency into a single quantity. Effective compute could be estimated with the following procedure:[126]

1. Choose a broad-based quantitative benchmark B for a model class. For example, the MMLU benchmark for LLMs.

2. Construct a scaling law for the benchmark B. A scaling law relates the amount of compute used to train a model, to that model's performance on the benchmark B.

3. For any new model M whose effective compute is to be estimated, first evaluate the model M's performance on the benchmark B.

4. Using the above scaling law, look up the model M's performance on benchmark B, and read off the compute value that corresponds to that performance. This compute value is model M's **effective compute**.

This measure of effective compute is relatively cheap to calculate. But it is also imperfect in a number of ways and limited in its applicability. First, the chosen benchmark B may not correlate well with dangerous capabilities of interest under general conditions. Second, effective compute can be undermined as a measure of

[125] See Annex G: Primer on AI and compute for more information.

[126] Thanks to the policy and technical teams at Google DeepMind for the suggestions in this annex.

capability, particularly if it is used to set a regulatory threshold.[127] And finally, effective compute is a property of an AI model, so it cannot be used to define regulatory thresholds for AI hardware owners, data center infrastructure providers, or other elements of the advanced AI supply chain. For that reason, effective compute cannot be used on its own as a metric for defining regulatory thresholds. But it could still serve as one of several indicators of a model's general capability level.

[127] For example, a regulated entity could attempt to fine-tune an otherwise highly capable model to perform poorly on the specific benchmark used to define effective compute in the context of a regulatory threshold. This would make the model appear to use less effective compute, allowing it to circumvent a regulatory threshold defined in those terms.

Annex K: ASTF activities and task-organization

We propose here some activities the ASTF could undertake in support of the two **Sustainment Components** of its mission (LOE1, 1.4.1.2 and 1.4.1.3):

1. Overseeing industry compliance with RADA safeguards, and operating the necessary supporting infrastructure; and

2. Developing recommendations for a future legal regime and regulatory agency in support of LOE4.

We also propose one possible task-organization for the ASTF to support its mission and activities.

K.1 Oversee and support compliance with RADA safeguards

To support this Sustainment Component of its mission, the ASTF could undertake some of the following activities:

1. Maintain private, secure registries of key proprietary information shared with it under RADA safeguards. Depending on the details of the RADA safeguards, these may include, among others, registries of AI labs' physical data center locations, and of the AI hardware mix at each relevant data center; private safety evaluations that third parties administer on frontier labs' AI systems [46]; and a secure temporary storage system that holds model weights as sensitive data in support of third-party administration of private evaluations.[128]

2. Maintain key public information about various aspects of its standards and operations. The ASTF could publish this information on its website. Depending on the details of the RADA safeguards, this may include, among other elements, registries of standardized model cards, public safety evaluations, and security standards and access controls. These standards could be developed in collaboration with the NIST U.S. AISI. See LOE3, 3.2.2.

3. Periodically update the RADA reporting and licensing thresholds, and communicate those updates to affected stakeholders.

[128] See Annex M: Secure temporary storage of model weights for one possible protocol that could support this activity.

4. If appropriate given the RADA safeguards, coordinate third-party administration of public and private AI evaluations, including frontier labs' responses to reports of failed evaluations. Standards for the evaluations could be developed in collaboration with the NIST U.S. AISI, DOE, and DHS and could explicitly include evaluations aimed at assessing risk from loss of control. DHS support in particular may be necessary for the development of CBRN and WMD-related evaluations.

5. If appropriate given the RADA safeguards, license independent third parties to administer AI evaluations on frontier models. The ASTF could set up an application process for third-party evaluators and red teams that request access to pre-deployment models for the purpose of developing and administering private safety evaluations. These third parties, once approved, could solicit evaluation proposals from outside entities, which they may then implement themselves. This could also be done in coordination with partners such as the U.K. Frontier AI Taskforce [349,350] through the Department of State. Independent AI evaluations should ideally be funded by AI developers but the evaluators themselves should be selected by the ASTF, to avoid conflicts of interest inherent in AI labs choosing their own evaluators. See LOE3, 3.2 for more on AI evaluations.

6. Facilitate collaboration between U.S. government agencies, industry, and AI model evaluators to identify classified data leaks. In connection with AI weaponization, there is a risk that advanced AI models could infer classified information purely by triangulating from public data, and that these inferences could be inadvertently disclosed to users through their interactions with the resulting AI system. Model developers cannot know if a leak has occurred if they do not know that the leaked information is classified, so coordination in this area is crucial.

7. Support coordination of frontier AI labs and AI cloud providers with U.S. government cyber, operational, and physical security efforts, including red teaming and penetration testing, and support sharing of learnings and best practices.

8. Support coordination of frontier AI labs with the AGI alignment and AI safety communities and other security experts to develop and share model containment learnings and best practices.

9. Serve, if necessary, as a clearinghouse for technical AGI alignment research and other AI safety and security research that could be mutually beneficial between stakeholders.[129]

10. If an entity engages in frontier AI activities (such as training or deployments of dangerous systems) that violate RADA safeguards in ways deemed by the ASTF to introduce unacceptable risks to public safety or national security, order the entity to cease those activities, assuming that the ASTF has been granted this authority.

11. In the absence of an enforceable set of RADA safeguards for frontier AI development, identify further industry partners who could be onboarded onto a more limited voluntary agreement or Charter, and initiate relationships with them as appropriate (see Annex I: Voluntary Charter for responsible AI). Possible candidates could include NVIDIA, xAI [351], Stability AI [352], Meta, and others. Identification of candidates could be supported by horizon-scanning efforts (LOE1, 1.2).

K.2 Develop recommendations for a legal and regulatory regime

To support this Sustainment Component of its mission, the ASTF could undertake some of the following activities:

1. Establish a working group to understand the complete supply chain for frontier AI model training. In particular the ASTF could seek to understand which inputs are manufactured in the United States or in allied jurisdictions, and which legal and policy mechanisms exist for enacting controls over these inputs on various timescales (LOE1, 1.5; LOE5, 5.5) [353,354]. See Introduction, 0.5.1.6 and 0.5.3.2 for challenges associated with the advanced AI supply chain, and Annex G: Primer on AI and compute for information on the relationship between the semiconductor and advanced AI supply chains.

2. Collaborate with AI chip design firms and foundries to understand the feasibility and technical timeline for designing hardware safeguards such as (1) a system to register and track GPUs with tamper-resistant serial numbers [155]; (2) a GPU

[129] This could also help to allay potential antitrust concerns frontier AI labs may have, given that the ASTF would be a U.S. government-sanctioned entity.

memory snapshot and hash system on AI chip hardware [138, 217]; and (3) on-chip firmware for remote shutdown. The sooner a process is put in place for tracking where GPUs are going, the tighter the bounds that can be put on the total stock of untracked compute capacity. This work could also be supported by collaborations with federally funded research centers (LOE3, 3.1.2.2).

3. Create an ongoing process to understand whether requirements under RADA safeguards should be expanded, reduced, or changed. The ASTF could also establish a mechanism to solicit and receive public and international proposals for changes to RADA safeguards. This could be a public engagement effort that ties into risk reporting to give the public a better understanding of the risks and the opportunity to contribute mitigation ideas.

K.3 ASTF task-organization

Below we propose one possible organizational structure that could support ASTF mission execution (LOE1, 1.4.1).

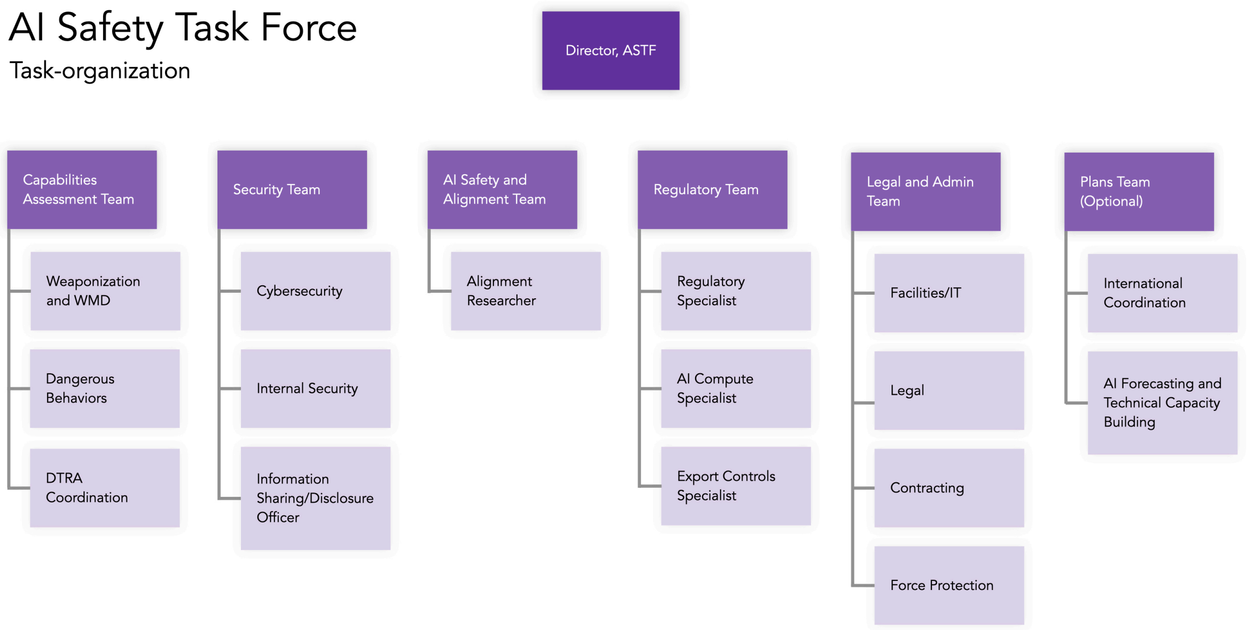


Figure 13. Organizational chart for the AI Safety Task Force.

The ASTF could be led by a Director with five immediate reports. Each of these reports could lead a workstream focused on executing a different element of the ASTF's mission across both Sustainment Components (see N.1 and N.2; and LOE1, 1.4.1.2 and 1.4.1.3).

- A **Capabilities Assessment Team** could be responsible for the ASTF's development and implementation of AI model risk assessment strategies, including dangerous capability evaluations and capability forecasting. This team could coordinate with industry and work closely with and draw WMD-relevant expertise from partner organizations such as the Office of the Secretary of Defense for Research and Engineering, the Joint Staff, U.S. Strategic Command, DHS, DOE, and the NIST U.S. AISI.

- A **Security Team** could support efforts to harden cyber, operational, and physical security at frontier AI labs. Core expertise required could include the capabilities of the National Security Agency (NSA), the Federal Bureau of Investigation (FBI), the Director of National Intelligence (DNI), and the Defense Counterintelligence and Security Agency (DCSA).

- An **AI Safety and Alignment Team** could closely monitor developments in technical AI safety and analyze their implications for catastrophic AI risks from weaponization and loss of control. It could be led by a government technical expert from the DOE with staffing by AI alignment researchers with industry experience. It could also collaborate with National Science Foundation-funded (NSF-funded) research centers with similar focus areas (LOE3, 3.1.2).

- A **Regulatory Team** could develop the ASTF's recommendations in support of follow-on legislative and regulatory action associated with LOE4. It could draw expertise from organizations such as the SEC, the Nuclear Regulatory Commission (NRC), the Department of State, and the Department of Commerce.

- A **Legal and Administrative Team** could advise the Director on legal matters and help manage administrative functions such as ASTF's internal security and contracting.

The ASTF could also incorporate a sixth workstream if it absorbs the functions of an AIO (LOE1, 1.2.1):

- A **Plans Team** could perform the ASTF's forecasting and capacity building activities. In particular, it could explore and propose opportunities for international collaboration on catastrophic AI risk mitigation, as well as potential domestic risk management strategies. It would also be home to the ASTF's horizon-scanning function, and could coordinate with the Department of State and the DHS, as well as with industry.

Annex L: AI safety and security research topics

Below we list some workstreams and topic areas that could be included in a federally funded research agenda for advanced AI and AGI safety and security. **This list is not exhaustive** and is in no particular order.

L.1 AGI-scalable alignment between AI behaviors and human values

This research workstream could include:[130]

- Direct research on inner alignment, which would aim to gather evidence of inner alignment failure in highly scaled and capable AI systems, and predict these failures in advance from descriptions of a planned AI system [81,355].

- Research on power-seeking and techniques to mitigate AI systems' incentives to power-seek, or their ability to competently act on these incentives [2,47].

- The development of new scalable alignment and scalable supervision techniques designed to ensure the safe behavior of AGI systems in regimes where these systems may be engaged in recursive self-improvement (RSI). These would be successors to current techniques such as RLHF [333], Constitutional AI [237], and DPO [238], which may not work effectively for AGI-level systems [84]. They may include adversarial training paradigms [356], activation or representation engineering techniques [357–360], or other novel approaches to ensure scalable controllability of AGI.

- Technical guarantees of honesty from AI systems, and techniques that minimize the discrepancy between a model's internal representation of the truth, and its outputs [361].

- Research investigating the extent to which alignment of AGI or superintelligent AI systems is technically possible [7].

- Agent foundations theory and decision theory, to establish fundamental principles on which safe AGI systems can be developed [362–364].

[130] See Annex B: The full challenge of AGI alignment for more context.

- Safety-oriented research into the mathematical foundations of deep learning, such as singular learning theory [365].

L.2 Monitoring of AI systems to detect misbehavior and preempt failures

This research workstream could include:

- Transparency techniques that can provide clarity on the inner workings of AI models [366–368].

- Mechanistic anomaly detection techniques that allow unexpected inputs or outputs to be flagged for human review [369].

- Techniques that can detect, anticipate or fix so-called "trojan" AI models, which behave well in most situations, but which reliably misbehave in certain trigger contexts [370].

- Interpretable uncertainty techniques, which allow humans to determine when advanced AI models are uncertain about their outputs [361].

- Hazardous capability removal techniques, which would focus on eliminating dishonesty, as well as harmful capabilities that could enable biological or cyber attacks [371].

- AI capability evaluations that would allow humans to understand what capabilities an AI system has. These would include behavioral evaluations, fine-tuning based evaluations, interpretability or understanding-based evaluations, automated benchmarking, and red teaming practices. Red teaming efforts could focus on detecting behavior with potentially dangerous applications, like persuasion, planning, RSI, general reasoning, CBRN and cyber weapon design, etc. [46]

- Sandbox and testbed development to support AI capability evaluations.

- AI failure mode forecasting, including by developing AI systems capable of predicting the failures of other AI systems.

- AI capabilities forecasting to understand what capabilities an AI system will have based on how it is trained [213].

- AI alignment evaluations, aimed at understanding whether an AI system will reliably act in a way that is consistent with its stakeholders' intentions, as well as behavioral evaluations on alignment, which could try to infer the goals of an AI from its behavior [46].

L.3 Robustness in the face of adversaries and unforeseen circumstances

This research workstream could include:

- Fixes and detection techniques for outer alignment failures, in which the goals given to an advanced AI system are misspecified in ways that can be gamed, or that lead to harmful behaviors [270,372].

- Behavioral guarantees, which would allow developers to ensure that highly advanced AI systems will always act in certain predictable ways, even when adversarial inputs are introduced [373].

L.4 Hardware-based verification

This research workstream could include:

- Monitoring and verification schemes for AI inference deployments and training runs. This includes efforts aimed at tracking the physical usage of AI chips by companies and other entities [217,218,225].

- Chip registry schemes which can provide information about the location and ownership of large clusters of AI hardware, to serve as the foundation for future compute governance regimes [155].

- Hardware-enabled mechanisms that could enable trusted verification of training compute and data. Trusted execution environments and confidential or multi-party computing could allow analysis and execution of model weights without revealing the raw weights themselves [213].

Annex M: Secure temporary storage of model weights

There is a strong competitive incentive for frontier AI model developers to train their AI systems specifically to pass known evaluations, without necessarily resolving underlying safety issues (Introduction, 0.5.1.3 and LOE3, 3.2.1). This means it is important to maintain a set of **private AI evaluations**, whose exact protocols are not known to the AI model developers, to ensure that developers are not incentivized to attempt to pass evaluations through superficial fine-tuning. These private evaluations should ideally be administered by independent third-party evaluators (see LOE3, 3.2.1.3).

This need for private evaluations creates a technical security problem. Many types of evaluations (particularly interpretability-based evaluations) cannot be administered without direct access to the weights of the AI model being evaluated. To obtain this access, a third-party evaluator could use the infrastructure of the AI model developer itself to administer their evaluation. But because the AI model developer controls this infrastructure, it could observe the evaluation process and learn the details of the evaluation protocol. This would irreversibly impair the privacy of that evaluation protocol and its integrity as a risk signal.

Unfortunately there is no practical way to ensure the privacy of an evaluation protocol under those conditions. Instead, one solution could be for the AI model developer to make a *copy* of its model's weights, and transfer that copy to a server operated by a trusted body or agency (e.g., LOE1, 1.4; LOE3, 3.1.2; LOE4, 4.1). The trusted agency could then support temporary secure access to the model weights by third-party evaluators, and encrypt that copy of the weights once the evaluations have been completed. But because a second copy of the model weights increases the exfiltration attack surface for adversaries, a secure protocol for model weight sharing would be crucial to support such an arrangement.

Below we sketch a protocol that may achieve this goal with a limited degree of security risk. The protocol's participants are: (1) a **Regulator**, who operates the secure server; (2) a **Developer**, who uploads the model weights; and (3) an **Evaluator**, who administers the private model evaluation.

Secure weight-sharing protocol:[131]

1. **Regulator** generates a one-time instance of an audited (from all sides), signed server in a secure setup. (For example, a government cloud provider like AWS or Azure.)

2. **Developer** encrypts a model weights file.

3. **Developer** uploads the encrypted blob to **Regulator's** server.

4. After this upload, **Regulator's** server only has single-user, one-time inbound Secure Shell Protocol (SSH) access enabled.

5. **Developer** shares the key with **Evaluator** representative. Physical handoff of the key is recommended.

6. **Regulator** gives server access key to **Evaluator**; **Evaluator** accesses and rotates key. **Regulator** can now no longer access its own instance.

7. **Evaluator** uses automated server scripts with **Developer**-provided weights key to decrypt the model and run private evaluations against it.

8. (Optional) **Regulator's** server sends high-level telemetry to **Developer**, **Evaluator**, and **Regulator** of non-sensitive information on evaluation progress and results.

9. **Developer**, **Evaluator**, and **Regulator** all separately have the ability to send a signed request to the cloud provider to terminate the server at the infrastructure level. (This is not a request to the server itself.)

10. Once the evaluations are completed, **Evaluator** notifies **Regulator**. **Regulator** terminates the server, deleting the instance of model weights.

Depending on the degree of security required, the secure server could also be housed on-premises in an air gapped facility, instead of on a cloud provider's infrastructure.

[131] Thanks to Fletcher Heisler for suggesting this protocol.

Annex N: Training approvals process for high-risk AI models

To obtain regulatory approval to begin a **high-risk AI training run** (e.g., a Tier 3 model in LOE4, 4.1.3.4.3), an AIMD (LOE4, 4.1.3.4) could be required to submit a safety case[132] to the regulator at the planning stage, prior to beginning the training run. This safety case could contain training run metadata, including details of the training procedure, training data, data cleaning, data preparation processes and training objective. The regulator should commit to keeping the contents of the safety case strictly confidential. The burden of proof would be on the AIMD to demonstrate to the satisfaction of the regulator that a high risk AI model is safe to begin training [46,162].

A key goal of the training approvals process is to motivate the development of a fundamental scientific understanding of the relationship between the training inputs (data, compute, loss functions employed, etc.) and the capabilities and tendencies of high-risk AI models. While this is an extremely challenging standard to meet, it is also likely crucial to assure the continuing safety of future, highly capable advanced AI systems (see Annex B: The full challenge of AGI alignment). In the absence of such a theory, empirical assessments like model evaluations may have limited value, and even create a false sense of security among labs and government regulators (LOE3, 3.2.1).

Since the test of any scientific theory is in its ability to make correct predictions, we recommend that the high-risk AI model training approvals process include a requirement for the AIMD to register predictions of its model's capabilities and outputs in advance of beginning training.[133] While a successful prediction does not guarantee that a theory is right, many successful predictions should make the regulator more confident that the AIMD understands how its model works well enough to train and deploy it safely [375].

We list below some key considerations for an effective training approval process.

[132] A safety case is "a structured argument supported by evidence, which provides a comprehensive and compelling case that a system is safe to operate in a given scenario." In the case of training approval, we would modify this definition to say that the system should be safe to *train* in a given scenario [374].

[133] Within reasonable bounds. For example, an AIMD might predict that a model's performance in some domain will fall within a range of values. As AI models become more powerful, the required prediction ranges could be tightened, which requires AIMDs to have a better and better understanding of their models' internal mechanisms as capabilities increase.

N.1 Dangerous capability predictions

On receiving the initial safety case, the regulator could require the AIMD to submit training run data and meta-data on an as-needed basis, evaluations of other similar models, or adjustments to the training procedure. Consistent with emerging practices at frontier labs [128], it could also require that the AIMD submit predictions of future model performance and outputs, to be compared with measured performance over the course of the training run.

- These predictions should include sets of behavior the AIMD strongly believes the model *will not display*, even under significant adverse pressure. The regulator can use this condition to control for certain kinds of weaponization risk; that is, the AIMD should be able to claim with high certainty that the model will not support the design of novel biological or chemical weapons, strong offensive cyber capabilities, cyberattacks, or other readily weaponizable capabilities. But this condition is also useful as a measure of the AIMD's understanding of its own AI model: if third-party red teams can elicit a behavior despite the AIMD's claim to the contrary, this indicates that the AIMD may lack the ability to assure the model's safety more generally.

- An integral component of useful capability predictions will be creating common knowledge between companies of which capabilities are considered dangerous and need to be forecasted for that reason. Examples could include deception [376], self-replication [54], resource acquisition, and bioweapon design [377]. Not all capability predictions should focus on obviously dangerous capabilities, since part of the reason for this requirement is to encourage the AIMD to develop a *general* ability to do this kind of forecasting.

- As part of its capability prediction process, the AIMD could privately disclose any new scaling laws [23] it may have derived from its work up to this point. This would support the regulator's ability to forecast general AI capabilities development. It could also support general horizon scanning functions (e.g., see LOE1, 1.2), and inform contingency and response measures more broadly (LOE2, 2.3 and 2.4).

N.2 Capability evaluations

The regulator and the AIMD could also agree on an initial set of capability evaluations to use for ongoing assessment of the model's safety characteristics both during training and during deployment. These evaluations may include behavioral and interpretability evaluations, manual and automated red teaming, and/or standardized capability benchmarks.

- There should additionally be a set of **private evaluations** that the regulator or approved third parties administer on the AI model, to minimize the risk that the AIMD will "teach to the test" by training its model to pass specific evaluations, without addressing underlying safety issues (LOE3, 3.2.2).

- Public evaluations could be developed and proposed by AIMDs, third parties, or the regulator itself. Changes to public evaluations could be approved by the regulator on the advice of approved third parties. Private evaluations could be developed and proposed by approved third parties or by the regulator.

N.3 Capability predictions at training checkpoints

As part of the initial training run application and safety case, the AIMD could also submit predictions of the proposed AI model's capabilities at various checkpoints along the training run, and at the conclusion of training. In other words, the AIMD would attempt to predict the high-risk AI model's performance on each element of a pre-agreed evaluation set. The AIMD would then report their advance predictions privately to the regulator.

This is particularly important because new AI capabilities can emerge suddenly during training, with limited advance warning [165]. If an AIMD lacks the ability to predict which dangerous capabilities could emerge during training or when, it may not apply appropriate caution at that stage or at subsequent stages of development.

N.4 Data controls

As part of its safety case, the AIMD could commit to avoid training its AI model on data that could (1) contribute to situational awareness and breakout capabilities of an AGI-level model [209]; or (2) reasonably be expected to significantly increase the model's CBRN, cyber, or persuasion capabilities.

In particular, some degree of data controls may be required to minimize the risk of loss of control due to AGI alignment failure. For example, a model should not have access to data that could support inferences about its context (training, testing, or deployment) or about the physical infrastructure it runs on. A model should also lack any channels of influence over individuals who can activate a training or deployment kill switch, or deactivate a dead man's switch (Annex O, O.5).

Data controls alone cannot guarantee that a misaligned AI system would be contained. Particularly at high AI capability levels, there is no way to predict exactly what an AI model can learn from which data. For example, there could be correlations between open-source data and classified information that a human would fail to detect, but that a powerful model could identify and leverage. A capable enough system might require very few real world data points to infer things that could be dangerous, as suggested by the finding that more highly scaled models tend to be more sample-efficient at inference time [22].

N.5 Risk management

The criteria for training run approval could be similar to the probabilistic risk assessments which are used in nuclear and other safety critical areas. This means that as part of its safety case, the AIMD would submit a detailed risk analysis and threat models, mapping capability predictions to probabilities of various kinds of weaponization and loss of control risks. This analysis would include risk management plans if particular capabilities emerge, including specific actions to take if an unanticipated dangerous capability is detected during training or deployment. The regulator should ground its final assessment in lowering the probability of catastrophic risk from the training or deployment of the high-risk AI model.

Annex O: Training stage monitoring for high-risk AI models

To train a **high-risk AI model** (e.g., a Tier 3 model in LOE4, 4.1.3.4.3), an AIMD (LOE4, 4.1.3.4) could be required to submit snapshots of its model weights periodically to the regulator over the course of the training run. This would support secure third-party red teaming, evaluations, and other safety testing. (See Annex M: Secure temporary storage of model weights for a description of one possible weight-sharing protocol.) Validated third parties would evaluate the model at each snapshot to see whether it displays specific dangerous capabilities, and to assess whether the model's general capabilities are within the bounds the AIMD predicted as part of the safety case it submitted for training approval (see Annex N: Training approvals process for high-risk AI models).

The goal of training stage monitoring is twofold. First, periodic monitoring may detect dangerous capabilities in a model that emerge during the training run, though there is no evaluations regime guaranteed to detect such capabilities. And second, by checking measured model capabilities against the AIMD's predictions, the regulator can assess the AIMD's own understanding of the AI model it is training. If a measured capability falls outside the range initially predicted by the AIMD, this is a sign that the AIMD does not reliably understand its own model and that training should be paused while the issue is investigated.

We list below some key considerations for effective training stage monitoring.

O.1 Training journal

The AIMD could be required to share (at the regulator's request, and on a private basis) a training journal recording the sequence of detailed interventions that AIMD staff took during training, as well as the reason for each intervention (for example, a sudden divergence in loss that led developers to perform a gradient clamping operation). This will ensure the regulator is aware of the reasoning behind specific training interventions, and increase the transparency of the training process.[134]

[134] Thanks to Alan Chan at GovAI for this recommendation.

O.2 Ongoing capability predictions

In addition to the predictions it submitted as part of its training approval safety case (see Annex N: Training approvals process for high-risk models), the regulator could require the AIMD to predict the high-risk model's capabilities at successive checkpoints during the training run. The accuracy standard for these predictions would be higher than for those made during training approval, since these new predictions will be based on evaluations data from current and previous model snapshots, in addition to the original training metadata and expected downstream training characteristics.

- These snapshot predictions should serve as adjustments and refinements of the AIMD's planning-stage capability predictions.

- The regulator could require a pause or rollback[135] of a training run if evaluators measure capabilities that fall outside the range the AIMD predicted at either the planning stage, or at earlier training-stage checkpoints. But such a pause or rollback should be coupled with a condition that the AIMD can only resume or restart training if it can submit:

    - A mechanistic argument for a fix that includes an explanation for why the fix will solve the underlying issue;
    - A prediction of other implications of the fix, including a set of behaviors it strongly believes the model *will not display* following the fix;
    - An implementation of the fix; and
    - A series of measurements that shows the fix had an effect consistent with the prediction.

- The regulator could require evaluations at logarithmic checkpoints in training compute (for example, at every 1.5X increment of training compute), with the exact increment, which may be different for different training runs, determined by consultation with experts. There should be hierarchies of evaluations, with the more expensive evaluations run less frequently than the less expensive ones. The more expensive evaluations should also be run in response to dangerous capability flags raised during runs of the less expensive (but more frequent)

[135] A rollback could mean reverting a training run to a previous model snapshot, or encrypting the model's weights and restarting the training run from scratch, depending on the degree of divergence between predicted and actual model capabilities.

evaluation sets.

  - In addition to regular evaluation snapshots, the regulator could request model snapshots for evaluation purposes at *random* intervals during the training run. This helps avoid the risk that the AIMD could specially fine-tune its model to pass evaluations in advance of known evaluation snapshots.

O.3 Ongoing capability evaluations

The training-stage evaluations regime could include ongoing monitoring for sharp drops in training loss (or other possible proxies), which could indicate the unexpected emergence of strong general capabilities in the AI model (LOE3, 3.2.1.1).

- Over time, it will be crucial to also develop evaluations for deceptive alignment, situational awareness, and other signs of emergent misalignment on the part of the AI model. Depending on the specifics of the training regime, it may be possible to detect signs of model deception early during a training run, but not later, since a highly-capable AI model could eventually learn to conceal the significance of some of its internal computations from interpretability techniques (LOE3, 3.2.1.4).

- Evaluations should include red teaming and fine-tuning evaluations, in which vetted third-party evaluators have access to the model's weights through a secure temporary storage mechanism (see Annex M: Secure temporary storage of model weights). Fine-tuning access is essential to maximize the usefulness of dangerous capability evaluations in particular. While no evaluation regime is guaranteed to detect an existing dangerous capability (LOE3, 3.2.1.1), fine-tuning for dangerous capabilities allows the evaluator to apply pressure on the model to exhibit a dangerous capability that may be latent or otherwise hard to elicit. Evaluations that do not include fine-tuning almost certainly underestimate the risks associated with a given model, and should be treated as revealing only lower bounds on these risks. In conversations, frontier AI researchers have indicated that they expect this to be particularly true of more advanced, situationally aware models, which could behave deceptively to pass evaluations if they are not fine-tuned as part of the evaluation process (LOE3, 3.2.1.4).

- For similar reasons, some evaluations should involve granting evaluators access to the activations and weights of the model being tested, to facilitate interpretability evaluations.

- Third-party evaluators could be vetted by the regulator, first because of the security concerns related to model weight access, and second because dangerous capability evaluations may themselves involve actively eliciting behaviors that could entail broader risk (LOE3, 3.2.2).

- In addition, government CBRN experts with access to classified information could be involved in red teaming high-risk models and performing CBRN, cyber, and other dangerous capability evaluations.[136] See, e.g., LOE3, 3.1.1.3.

O.4 Model output predictions

Depending on technical feasibility and the model's anticipated capabilities, the AIMD could attempt to predict characteristics of the model's outputs in response to specific user inputs. That is, for a high-risk LLM, the AIMD might try to answer the question: "Given this prompt, what will the output of this model be?" as accurately as it can, in advance.[137] This is a far harder condition than evaluation prediction and likely implies a greater degree of model interpretability than is currently possible. The intent of this condition is to begin creating a benchmark for interpretability research that is difficult to undermine by superficial means [380].

[136] Some frontier AI labs have already indicated an openness to related policies prior to model deployment. OpenAI's Preparedness Framework, for example, states that "We will also continue to enable external research and government access for model releases to increase the depth of red teaming and testing of frontier model capabilities." [378] Anthropic's assessment is similar, determining that, "We have found that in certain cases, it is critical to involve red teamers with security clearances due to the nature of the information involved. However, this may limit what information red teamers can share with AI developers outside of classified environments. This could, in turn, limit AI developers' ability to fully understand and mitigate threats identified by domain experts." [379]

[137] Within reasonable uncertainty bounds. An interpretability technique or mechanistic theory may not be able to *perfectly* predict, e.g., a model's answer to a question, but it may be able to predict several likely answers. In addition, certain inputs have outputs that are easier to predict than others. For example, an input like "What kind of animal is Winne-the-Pooh?" may have an output that is easier to predict than the output of "Please write an original ten-page story about Winne-the-Pooh."

O.5 Training kill switch

For each high-risk training run, there could be a designated role at the AIMD with the authority to activate a "kill switch" for that training run. There should ideally be at least two individuals who hold this role in any given shift, with rotating handoffs at each shift such that the role is continuously staffed at all times during the run.

- The individuals staffing this role could be technical and have received specialized AI safety training mandated under the terms of their employer's licensing. There should be vetting and security measures in place to select the right individuals for this role. At least one such individual should also be restricted from interacting with implementations of the model or of its training snapshots, to minimize the risk of influence or manipulation. (This is unlikely to be a concern for today's AI models, but could become so in the future.)

- The individuals staffing this role could be given access to all the internal model evaluations that the company has performed (which may be different from the third-party evaluations), including the model loss, gradient norms, and other metrics that are tracked during training.

- These individuals should ideally have clear decision making authority in this area that cannot be overridden by their management chain. They could enjoy explicit protections from internal sanction by their companies, e.g., for hitting a kill switch prematurely.

- This role could be defined similarly to that of analogous roles in other industries, such as the range safety officer responsible for triggering the flight termination system for rocket launches [343].

- A training restart could require the approval of an independent safety panel.

- All kill switch incidents could be promptly reported to the regulator. The report would include the reason the kill switch was activated, the results of the subsequent investigation, and any mitigation measures that were put in place after training was restarted.

- As an alternative to the "kill switch" model, the AIMD could implement a "dead man switch" model, according to which training or deployment is halted at certain checkpoints by default unless a panel appointed to oversee these activities elects to proceed.

O.6 Failed evaluation protocol

If, during training, a high-risk AI model fails an evaluation, the AIMD could promptly report the failure to the regulator. The regulator could be empowered to ask all AIMDs to pause their ongoing training runs, or any training runs of a similar type to the one that failed the evaluation. The regulator could also be empowered to suspend some existing model deployments, depending on the nature of the failure [336].

- The regulator could share some anonymized data relevant to the evaluation failure across other licensed AIMDs, because of the potential for below-threshold data from other training runs to indicate that those other training runs may be on the way to failing evaluations of their own. Other licenced AIMDs could be asked to certify to the regulator that similar or other concerning patterns (possibly falling below the evaluation threshold) have not been observed in their own training runs.

- As a general principle, an AIMD should not be training a high-risk model unless it expects its model will pass the agreed-upon safety evaluations. If a model fails one of the safety evaluations during a training run regardless, this is a sign that the AIMD does not understand the implications of the proposed training scheme as well as it previously believed. As in S.2, the AIMD could then submit:
    - A mechanistic argument for a fix that includes an explanation for why the fix will solve the underlying issue;
    - A prediction of other implications of the fix, including a set of behaviors it strongly believes the model *will not display* following the fix;
    - An implementation of the fix; and
    - A series of measurements that shows the fix had an effect consistent with the prediction.

- An AIMD faces a strong incentive to repeatedly submit new fixes until it happens to find one that passes this review process. Unfortunately, if the AIMD is allowed to submit fixes indefinitely, there is a significant risk that it will eventually find a fix that *does* pass review, but *does not* genuinely address or explain the underlying safety problem (LOE3, 3.2.1.3). As a result, the regulator should limit the number of fixes the AIMD can submit for a given dangerous capability flag (e.g., no more than 3 such fixes or explanations). Past this point, the training run should be ended and the in-training model weights encrypted despite the significant capital loss this could represent [374].

- If an AIMD continues to restart training runs with new fixes, and gives indications of spending resources on new training runs in an attempt to undermine these safeguards, the regulator should consider suspending the AIMD's training license entirely.
- The central risk is that while prediction is the best test of a theory, making a large number of predictions without controls can result in eventually getting a correct answer by chance, without actually having the level of understanding required to assure safety. This means even testing understanding through prediction is of limited value unless a regulator sets clear conditions under which the process ends.

Annex P: Deployment stage approvals for high-risk AI models

To obtain regulatory approval to deploy a **high-risk AI model** (e.g., a Tier 3 model in LOE4, 4.1.3.4.3), an AIMD (LOE4, 4.1.3.4) could be required to submit a safety case to the regulator for each intended **deployment context** for that model [46,228]. Examples of deployment contexts include:

- Deploying the model internally for a predetermined use case or set of use cases, including experiments;[138]
- Deploying the model to external users via a generally accessible interface or API;
- Deploying the model to external users with an augmentation such as a calculator or a code interpreter; or
- Deploying the model for use by a known external user for a defined use case such as summarizing medical records.

Deployment-stage approval would be separate from approval for the training run itself (see Annex N: Training approvals process for high-risk AI models), and should ideally involve extensive additional evaluation, red teaming, and modification of the fully trained AI system which the regulator would be empowered to require. As with training-stage approval, the burden of proof would be on the AIMD to demonstrate to the satisfaction of the regulator that the AI model is safe in the relevant deployment context. The regulator would evaluate the AIMD's provided proof points and could request changes to the model's training, reversion to previous training checkpoints, additional evaluations, or encryption of an already-trained model.

The goal of the deployment approvals process is twofold. First, pre-deployment testing may uncover dangerous capabilities of the AI model that only emerge under significant pressure. This is crucial because some kinds of risks may only emerge or become amplified after a model has been deployed [381]. For example, external software frameworks like Auto-GPT [120] and BabyAGI [122] can significantly augment the capabilities of a base model at executing complicated, long-horizon tasks. Internet

[138] Particularly experiments that include or involve RSI or online learning by Tier 3 AI models with wide action spaces over long time horizons.

access can also increase a model's capabilities, potentially giving it the ability to reference and reason about its own actions [382]. Effects like these are not always obvious, and an AI model's ability to effectively leverage external tools has been observed to emerge somewhat unpredictably during training [338,165]. (See **Deliverable 2: Survey of AI Technologies and AI R&D Trajectories**.)

The second goal of deployment approvals is, as before, to validate the AIMD's predictions of the model's behaviors under scenarios that should closely match the expected deployment context. If the AIMD cannot reliably predict model behaviors[139] under such scenarios, this is an indication that the AIMD has not characterized the system comprehensively enough to deploy it safely and that deployment should be halted until this is resolved.

Finally, depending on the general capability of the model and on the AIMD's ability to correctly forecast its behavior, the regulator could require or approve additional software safeguards prior to deployment. For example, while the main high-risk AI model responds to user queries, a second smaller model might filter user input for signs of dangerous prompting prior to passing the query on to the high-risk model. In this case, the AIMD could be required to submit a safety case for the **AI system** (consisting of the high-risk model itself along with any software safeguards it would be operating under), rather than for the high-risk AI model alone.

We list below some key considerations for an effective deployment approvals process.

P.1 Lifecycle funds

A safety case for a high-risk model deployment could include a credible commitment by the AIMD to set aside funds for periodic third-party red teaming and deployment-stage monitoring over its projected life cycle. The extent of this need would depend on the deployment context. (See Annex Q: Deployment stage monitoring of high-risk AI models for more information.)

[139] Within reasonable uncertainty bounds, though the AIMD should be held to a higher standard in predicting dangerous capabilities in deployment than during training.

P.2 Pre-deployment KYC

If the deployment context for a model is restricted to a known set of end-users, the AIMD could conduct preliminary KYC on these end-users in order to better understand the anticipated usage pattern of the model.

P.3 Risk and contingency planning

The safety case for high-risk AI model deployment could include impact assessments and risk assessments specific to each deployment context, the expected usage envelope of the deployment context, the stakeholders who may be affected under the deployment context, and the means by which controls will be implemented to ensure usage stays within the expected context. Depending on the criticality of the use cases in the deployment context, this could extend to include emergency response playbooks, I&Ws (LOE2, 2.3), and step by step mitigation measures in the event of an unexpected development, which may include shutdown drills and other preparations.

P.4 Weight storage for model evaluations

To facilitate private evaluations of high-risk AI models prior to deployment approval, the AIMD could transmit the weights of the model securely to a server operated by the regulator. Vetted third-party red teams and evaluators would then administer evaluations, including private evaluations, on the stored model snapshot. (See Annex M: Secure temporary storage of model weights for a description of one possible weight-sharing protocol.)

P.5 Model augmentations

Each augmentation of a high-risk model with an external tool (calculator, code interpreter, knowledge base, database, etc.) should count as a *separate* deployment context and require a separate safety case. An evaluation suite would need to be run with each new augmentation available to the model, in combination with the other augmentations the model has access to in that context. Specifically:

- A high-risk AI model should not be given arbitrary Internet access, but may be augmented with narrow apps that can execute well-defined tasks.
- A high-risk AI model that has API access to a retrieval database should go through a deployment approvals process for every incremental change in its

database. This also implies that retrieval databases serving high-risk AI models cannot be continuously updated, but need to be updated using a batch process.

- The goal of this provision is to minimize the degree to which a high-risk AI model can offload computation to external aids, thereby concealing its computations from direct oversight for potentially deceptive purposes [376].

P.6 Failed evaluation protocol

If a high-risk AI model fails an evaluation during the deployment approvals process, or if an evaluation result falls outside the range predicted by the AIMD, the regulator could ban the model from being deployed in the given context. The regulator could also be empowered to enforce temporary deployment bans on all Tier 3 AIMDs (LOE4, 4.1.3.4.3) if a dangerous capability is detected during a pre-deployment evaluation in one AIMD's model. The regulator could also mandate that other AIMDs run additional evaluations on their own models, even if those models are already deployed.

- The AIMD and regulator may consult with CBRN experts, cybersecurity experts, or other domain experts for clarification, depending on the nature and severity of the evaluation failure.

- If safety is not proven to the regulator's satisfaction, the tiered model licensing thresholds (LOE4, 4.1.3.4) could be re-evaluated.

P.7 Fine-tuning approval

If a model is being fine-tuned during deployment, each instance of fine-tuning should require separate deployment approval. For clarity, this would require pausing deployment of the fine-tuned model prior to the completion of the approvals process. If fine-tuning involves a significant amount of compute, the AIMD should additionally subject the model to training stage monitoring during the fine-tuning process. (See Annex O: Training stage monitoring for high-risk AI models for key considerations.)

- This provision could apply to online learning paradigms, in which an AI model continuously retrains itself as it processes inputs from the world around it. Specifically, the regulator could consider online learning as a special case of fine-tuning,[140] and ensure that a high-risk AI model undergoes a new

[140] In the sense that it involves updating the parameters of a pre-existing AI model.

deployment-stage approval process for each new batch of learning updates.

- This would effectively prevent the deployment of models that do true online learning, by introducing an approval step in the feedback loop between data collection and the model's learning updates. The intent of this provision is to mitigate two closely related risks of online learning that make it qualitatively different from an LLM paradigm. First, because an online learning system has the opportunity to learn from *all* of its historical inputs, its effective context window is much larger than that of an LLM. And second, an online learning model may gain unpredictable new reasoning capabilities unexpectedly fast, extending its capability envelope in a sharp increment. Deployment approvals for online learning systems give model evaluators a chance to detect such sharp changes in capabilities before they can impact the model's deployed safety profile.

- Similar provisions could apply to game-playing AI models at the high-risk level (e.g., Tier 3-equivalent in LOE4, 4.1.3.4.3), and to AI models with other non-text modalities.

Annex Q: Deployment stage monitoring of high-risk AI models

Once an AIMD (LOE4, 4.1.3.4) has deployed a **high-risk AI model** (e.g., a Tier 3 model in LOE4, 4.1.3.4.3) in a given deployment context, it could continuously monitor the model's usage, including by performing KYC checks on high-volume, high-risk, and other critical use cases [46]. This includes checking usage patterns with classifiers that flag high-risk user interactions. Flagged interactions could then be further investigated via KYC. For clarity, deployment stage monitoring should also include monitoring and auditing of *internal* model deployments, in which the AIMD's staff are the only users of their own model. Deployment stage monitoring is crucial because some catastrophic risks may only emerge, or become amplified, after a model has been deployed [381].

Deployment stage monitoring has two components. First, the AIMD should continuously monitor the *inputs* the high-risk AI system is receiving in deployment. This helps ensure that the actual distribution of inputs matches the distribution the AIMD expected under the approved deployment context (Annex P, P.2). In particular, it allows the AIMD to identify possible instances of adversarial action, including attempts to jailbreak the model and attempts to distill the model's capabilities (Introduction, 0.5.1.7) [144].

The second component of deployment stage monitoring is periodic testing of the deployed AI system itself against new jailbreaks [94,95] or red teaming techniques that fit the deployment context, but may not yet have been observed "in the wild". The intent of this component is to verify that the deployed AI system remains robust to novel attacks as they are developed, and if necessary is hardened or taken offline in the event that a novel attack succeeds in eliciting harmful behavior.

We list below some key considerations for effective deployment stage monitoring.

Q.1 Usage pattern reporting

The AIMD could provide the regulator with periodic reports of model input-outputs in each deployment context, including any detected high-risk interactions. High-risk interactions could be investigated, and potentially incorporated into the public evaluation set. High-risk interactions could also be shared with vetted third party evaluators, under a privacy agreement, to support their development of further private evaluations (LOE3, 3.2.2).

- If the actual usage pattern of the deployment context diverges sharply from the anticipated deployment context, the regulator could pause deployment of the model and investigate. If the actual usage pattern of the deployment context begins to drift away from the anticipated deployment context over time, the regulator could require that the AIMD put the model through a new deployment approvals process under the updated deployment context.

- Users responsible for high-risk interactions could be subjected to rigorous KYC. In the event that it is impossible to confirm the identity of a high-risk user, that user could be banned from use of the model.

- Unrestricted API access to a model can be used to construct synthetic datasets that let an attacker train small models that approach the performance of a larger model served via API, more cheaply than training the original larger model (Introduction, 0.5.1.7). High-volume users of a public high-risk model API could therefore be subjected to rigorous KYC procedures [337]. KYC could initially be based on existing best practices in the financial sector [147].

- For very large inference customers, the AIMD could directly track usage and interview the customer to ensure that they have an awareness of the large critical use cases that their AI systems support.

- KYC thresholds could include industry-wide blacklists and whitelists of specific users, maintained by the regulator.

- Additionally, the AIMD could randomly interview a small subset of customers that do not raise ordinary KYC flags, in order to build an institutional understanding of typical usage patterns in a deployment context and their changes over time.

- The AIMD could report incidents associated with its deployed models via an incident database available to the regulator when they meet a predetermined severity threshold (LOE4, 4.1.3.5). This severity threshold could be determined by the regulator.

Q.2 Red teaming

Deployment stage monitoring could include ongoing red teaming by vetted third parties. If a high-risk AI model is deployed via public API, this could also include bug bounties and periodic hackathons to test the monitoring regime's ability to detect and defuse interactions that could signal weaponization risk, prosaic accident risk, or catastrophic risk from loss of control.

Government CBRN experts with access to classified information could also be involved in red teaming high-risk AI models and performing CBRN, cyber, and other dangerous capability evaluations (LOE3, 3.1.1.3).

Q.3 Failed evaluation protocol

If, during deployment, a high-risk AI model fails any evaluation, the regulator could require all AIMDs to pause some ongoing training runs, or to suspend some existing model deployments, depending on the nature of the failure. If this occurs:

- The AIMD should generally cut off public access to the model. There should be a contingency plan ensuring that a fallback model is in place for such situations, to be activated temporarily during the pause to support critical customer use cases. For example, if a hospital is using the model for aspects of its care, there should be a fallback model available with the same user interface or the same API signature to ensure some level of uninterrupted service.

- The AIMD should take strong steps to sandbox the model and do a comprehensive assessment of it, including developing and running more internal evaluations.

- The regulator could mandate that other AIMDs run additional evaluations on their own models, even if those models are already deployed. Regulator disclosure to other AIMDs at this stage could be on an anonymized basis, and similar to reporting a software security vulnerability.

- The regulator could consult with key subject-matter experts relating to the failed evaluation, such as CBRN or cybersecurity experts if the failed evaluation involves a dangerous capability in those domains.

- If safety is not proven to the regulator’s satisfaction, the tiered model licensing thresholds (see LOE4, 4.1.3.4) could be re-evaluated.