POLICY GUIDE

FEBRUARY 2025

International Dialogues on AI Safety

About Us

International Dialogues on AI Safety

The International Dialogues on AI Safety (IDAIS) bring together leading scientists from around the world to collaborate on mitigating risks from AI. The inaugural IDAIS event in October 2023 was convened by Turing Award winners Yoshua Bengio and Andrew Yao, UC Berkeley professor Stuart Russell, OBE, and founding Dean of the Tsinghua Institute for AI Industry Research Ya-Qin Zhang.

Safe AI Forum

IDAIS is supported by the Safe AI Forum, an organization co-founded by Fynn Heide and Conor McGurk. The Safe AI Forum does not receive funding from any corporate donors.

If you are keen to push forward research and action in line with the policy areas mentioned in this document, please get in touch with us at: idais@saif.org

Introduction

Starting in late 2023, the International Dialogues on AI Safety has convened leading scientists and academics from around the world to build consensus on risks from frontier AI systems and the governance interventions needed to tackle these risks. Based on this consensus, IDAIS participants have signed public statements outlining a set of goals for AI safety and governance.

This policy guide aims to connect the goals recommended in the consensus statements to direct policy action that policymakers, philanthropists, companies, and researchers can consider taking to improve the state of AI safety and governance.

This guide is split into four key policy areas, based on the goals recommended in the statements:

- AI safety research
- Testing and evaluation
- Domestic governance
- International governance

Within each area, there are a set of recommended goals, drawn directly from the consensus statements. For each goal, the guide provides an assessment of current progress towards the goal, as well as policies that can be considered to take further action towards the goal.

This guide is meant to be a living document that is updated after each successive dialogue. It is also modular, and readers should use the overview to identify areas of key interest and relevance, where they can dive deeper.

Readers should note that while the high-level goals drawn from the statements are endorsed by all signatories of the statements, the policy actions are suggestions drawn from a range of sources, including discussions at dialogues, as well as further research and analysis. Specific policy recommendations therefore do not necessarily reflect the views of all signatories.

Overview

Readers can use the table below to go through to areas of interest via the 'goals from consensus statements'.

Legend: - *Areas of Limited Progress* - *Areas of Moderate Progress*

Category	Goals from Consensus Statements	Potential Policies
AI Safety Research	Ensuring necessary resourcing of AI safety R&D	• Fund AI Safety Institutes, or related national/state-level safety organizations, through a wide range of sources. • Introduce tax credits for spending on AI safety research. • Mandate minimum safety investment levels (or other safety-relevant criteria) in public procurement guidelines. • Support public universities to conduct AI research by providing financial and in-kind resources. • Issue specific requests for proposals or grants for 3rd party safety research. • Launch competitions, incubators, and innovation prizes to accelerate safety research. • Set up global collaborative funds dedicated to funding AI safety R&D.
	Enabling international collaboration on AI safety	• Establish joint global research programs aimed at tackling the most significant safety challenges. • Set up research networks and high-impact conferences for AI safety. • Create channels for talent development, exchange, and mutual learning. • Set up jointly taught university courses, seminars, and modules.
	Supporting research into verification methods	• Advance high priority research identified in existing research agendas. • Identify specific verification projects that should be collaborated on internationally. • Collaboratively stress-test emerging hardware-enabled verification mechanisms. • Organize global competitions for verification research.

Category	Goals from Consensus Statements	Potential Policies
Testing and Evaluation	Defining and evaluating red lines and early-warning thresholds	• Develop a common taxonomy for red lines, early warning thresholds, and other key thresholds. • Develop in-house technical capacity with security clearances to conduct capability evaluations relevant to weapons development and cyberattack red lines. • Require companies to conduct evaluations on frontier models and share results with relevant regulators to build state capacity and visibility. • Establish a clear end-to-end process for domestic and international information disclosure related to red lines and early warning thresholds. • Develop methods to verify that necessary mitigations are in place. • Conduct research into red line/threshold violation protocols. • Research and develop methods to reliably shut down AI systems that have crossed or are in danger of crossing red lines.
	Setting up an ecosystem of 3rd party audits and evaluations	• Develop a comprehensive evaluation framework, which provides clarity on the role of 3rd party evaluators. • Develop a comprehensive auditing framework, which provides clarity on the role of 3rd party auditors. • Develop and share evaluation infrastructure. • Provide resources such as compute or API credits to support 3rd party evaluators. • Establish grants and prizes to incentivize R&D on evaluations of concerning capabilities.
	Requiring safety cases and/or guarantees	• Direct funding towards research into ‘safety cases’. • Direct funding towards alternative scientific or engineering paradigms which may be safe by design.
	Implementing domestic model registration	• Ensure that model registration policies provide wider and deeper visibility into emerging AI risks. • Conduct research and build consensus on refined compute, capability, and risk thresholds that should trigger greater regulatory scrutiny.

Category	Goals from Consensus Statements	Potential Policies
Domestic Governance	Monitoring large scale data centers	• Leverage cloud compute providers as intermediaries to gain better visibility into AI development trends and enforce regulations more effectively on model developers. • Develop privacy-preserving tools to enable a wider range of governance activity. • Develop robust compute workload classification methods.
	Putting in place AI incident tracking and reporting	• Establish a comprehensive hybrid approach to incident tracking and reporting. • Encourage more proactive documentation of AI systems to enableeee better analysis of incidents. • Research and develop automated metadata collection mechanisms.
	Adopting risk management and risk assessment practices	• Conduct research on AI risk management systems for frontier AI labs, as adapted from other safety-critical industries. • Establish regulatory requirements for risk management systems. • Assess whether company practices follow existing government guidance and recommend improvements.
	Requiring post-deployment monitoring	• Require companies to carry out and share interconnected post-deployment monitoring data. • Collect data about the integration of AI systems into critical infrastructure. • Invest in technical governance approaches that increase visibility into agentic systems.
	Instituting a professional code of ethics for AI practitioners	• Develop and require AI practitioners to take a 'Hippocratic Oath' for AI. • Enshrine a 'Right to Warn' that allows AI practitioners to exercise professional ethical responsibility.
	Establishing AI safety as a global public good	• Ensure international collaboration on AI safety is protected from broader geopolitical competition on AI. • Identify, develop, and share safety technologies, such as 'PALs for AI'. • Conduct further research to clarify the 'AI safety as a global public good' concept.

Category	Goals from Consensus Statements	Potential Policies
International Governance	Establishing conditional market access through globally aligned standards on red lines	• Identify a suitable channel for red lines standards development within existing coordination mechanisms. • Establish regulatory markets as a mechanism for coordinating regulatory goals and market access. • Adopt a jurisdictional certification process and a multilateral export control regime to manage conditional market access.
	Designing and negotiating international agreements and/or setting up an international organization	• Build consensus around which elements of frontier AI governance should be internationalized. • Align on whether key international frontier risk governance functions should be established through existing institutions or new processes. • Organize confidence-building measures to ensure necessary international trust to build more extensive governance processes. • Coordinate on key emerging safety standards.

AI Safety Research

Ensuring necessary resourcing of AI safety R&D

Goals

*We call on **leading AI developers to make a minimum spending commitment of one third of their AI R&D on AI safety** and for government agencies to fund academic and non-profit AI safety and governance research in at least the same proportion. - IDAIS-Oxford, 2023*

*Additional funding will be required to support the growth of this field: we **call for AI developers and government funders to invest at least one third of their AI R&D budget in safety.***

- IDAIS-Beijing, 2024

***States must sufficiently resource AI Safety Institutes,** continue to convene summits and support other global governance efforts. - IDAIS-Venice, 2024*

*States, philanthropists, corporations and experts should **enable global independent AI safety and verification research through a series of Global AI Safety and Verification Funds.** These funds should scale to a significant fraction of global AI research and development expenditures to adequately support and grow independent research capacity. - IDAIS-Venice, 2024*

Progress

Limited

⅓ investment of AI R&D into safety research would be approximately US$40B. It appears unlikely that global AI safety funding even reaches US$1B today, suggesting that less than 0.5% of overall AI R&D is spent on safety.[1]

Potential Policies

- **Fund AI Safety Institutes, or related national/state-level safety organizations, through a wide range of sources.** There is wide variance on this front—the UK has resourced its AI Safety Institute with over US$100M, while the US AI Safety Institute has received approximately 10% of that amount. States may consider drawing in diverse sources of funding ranging from temporary or permanent

[1] Given the lack of standardized data on this subject, any estimation will be imprecise. The goal of this estimate is to provide a directional sense, using the following method: Most AI R&D is private investment, and excluding M&A activity comes up to US$110B, as estimated by the Stanford AI Index 2024. We assume government investment is about 5% of private investment, leading to approximately US$120B in total AI investment, 33% of which is US$40B. AI safety commitments today are mostly government-led, and amount to little more than US$200M.

funding from within government, implementing targeted taxation and redistribution, and/or sourcing voluntary contributions from philanthropists and companies.[2] [3]

- **Introduce tax credits for spending on AI safety research.** A draft AI model law led by Zhou Hui from the Chinese Academy of Social Sciences Legal Research Center, introduced an idea for safety tax credit. This credit would amount to at least 30% the amount invested by AI developers and providers, in the R&D or procurement of dedicated equipment used for safety and governance purposes.[4]
- **Mandate minimum safety investment levels (or other safety-relevant criteria) in public procurement guidelines.** States possess significant levers in shaping the behavior of companies by including a range of criteria as part of public procurement requirements. For example, the Office of Management and Budget in the US issued policies that require federal agencies to responsibly procure AI. These policies include requiring adequate testing and safeguards and external AI red-teaming.[5]
- **Support public universities to conduct AI research by providing financial and in-kind resources.** Endowed chairs in AI safety research at universities would help to build AI safety up as an academic field, while initiatives such as the National AI Research Resource in the US could help academics access scarce resources (e.g., compute) that would otherwise be challenging to obtain.[6]
- **Issue specific requests for proposals or grants for 3rd party safety research.** For example, the UK AI Safety Institute set aside ~US$10M for systemic AI safety grants, which aim to support research into mitigating the impacts of AI at a societal level.[7]
- **Launch competitions, incubators, and innovation prizes to accelerate safety research.** Targeted and accelerated programs in this vein include XPrize and Entrepreneur First's defensive acceleration program.[8]
- **Set up global collaborative funds dedicated to funding AI safety R&D.** Collaborative funds pool resources from multiple organizations allowing for larger-scale funding. They can be set up as public-private partnerships, allowing government funding to go further by drawing in private donations. These funds could disburse grants and/or invest directly in AI safety projects. Examples of existing collaborative funds include The Audacious Project, housed at TED, and the ClimateWorks Foundation.[9]

---

[2] Selective taxation is used to fund public programs that address some of the public safety risks brought by a given industry. For example, in California, tobacco taxes are used to fund anti-smoking campaigns.

[3] In countries such as the US, government agencies can accept donations from private foundations. For further information, see: https://exponentphilanthropy.org/qa/can-a-private-foundation-make-a-grant-to-a-government-agency/

[4] ZHOU Hui et al., 'The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version', 16 April 2024, https://doi.org/10.5281/ZENODO.10974162.zh.

[5] Shalanda Young, 'Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence' (Executive Office of the President, Office of Management and Budget, 1 November 2023), https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf.

[6] 'National Artificial Intelligence Research Resource Pilot', n.d., https://new.nsf.gov/focus-areas/artificial-intelligence/nairr.

[7] 'Tech Secretary Unveils £8.5 Million Research Funding Set to Break New Grounds in AI Safety Testing', 22 May 2024, https://www.gov.uk/government/news/tech-secretary-unveils-85-million-research-funding-set-to-break-new-grounds-in-ai-safety-testing.

[8] 'XPrize', n.d., https://www.xprize.org/; Matt Clifford, 'Introducing Def/Acc at EF', 20 May 2024, https://www.joinef.com/posts/introducing-def-acc-at-ef/.

[9] 'The Audacious Project', n.d., https://www.audaciousproject.org/faq; 'ClimateWorks Foundation', n.d., https://www.climateworks.org/.

Enabling international collaboration in AI safety

Goals

> *Concerted effort by the global research community in both AI and other disciplines is essential; we* ***need a global network of dedicated AI safety research and governance institutions*** *-IDAIS-Oxford, 2023*

> ***We encourage building a stronger global technical network to accelerate AI safety R&D and collaborations through visiting researcher programs and organizing in-depth AI safety conferences and workshops*** *- IDAIS-Beijing, 2024*

> *Eventually, comprehensive verification could take place through several methods, including third party governance (e.g., independent audits), software (e.g., audit trails) and hardware (e.g., hardware-enabled mechanisms on AI chips).* ***To ensure global trust, it will be important to have international collaborations developing and stress-testing verification methods*** *- IDAIS-Venice, 2024*

Progress

Moderate
There has been moderate progress towards building a global network of dedicated AI safety research institutions. A network of AI safety institutes (AISIs) has been set up to coordinate between nationally-representative AI safety-focused entities. Moreover, bilateral talent sharing agreements, such as the Memorandum of Understanding between the UK and US AISIs, have also been set up.

Safety has also become a part of several AI conferences. For example, safety and alignment for a have been organized as part of conferences such as ICML, the Beijing Academy of AI annual conference, and the World AI Conference.

However, there are still there are still very few conferences dedicated to AI safety and limited institutionalized partnerships between AI safety labs across the world.

Potential Policies

- **Establish joint global research programs aimed at tackling the most significant safety challenges.** Examples from other fields include the Human Genome Project, which was an international effort that generated the first sequence of the human genome over the course of 13 years. It brought together researchers from 20 institutions across the US, UK, France, Germany, Japan, and China.[10]
- **Set up research networks and high-impact conferences for AI safety.** There has been some progress towards building networks and conferences around existing machine learning ecosystems.

[10] 'Human Genome Project', n.d., https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project.

However, there remains space to set up more dedicated infrastructure for safety research to accelerate its growth as an academic discipline. Potential models for high-impact conferences include the American Federal Reserve's annual Jackson Hole Economic Symposium, which is attended by academics and central bank leaders from across the world, and often has significant impact on monetary policy and the global economy.[11]

- **Create channels for talent development, exchange, and mutual learning.** Specific AI-safety focused exchange programs, scholarships, and fellowships, stewarded by leading AI safety scientists, could pave the way for a more coordinated global field. These could be modeled after a range of broader existing programs such as Schwarzman Scholars, Fulbright, or the Luce scholarship.
- **Set up jointly taught university courses, seminars, and modules.** While courses on AI safety are taught in some universities (e.g., Intro to AI Safety at UC Berkeley), more of such efforts—especially if jointly organized by leading academics from around the world—could help to rapidly build up global talent that have a shared understanding of the field.

Supporting research into verification methods

Goals

*In addition to foundational AI safety research, these funds would focus on **developing privacy-preserving and secure verification methods, which act as enablers for domestic governance and international cooperation.** These methods would allow states to credibly check an AI developer's evaluation results, and whether mitigations specified in their safety case are in place. In the future, these methods may also allow states to verify safety-related claims made by other states, including compliance with the Safety Assurance Frameworks and declarations of significant training runs - IDAIS-Venice, 2024*

Progress

Limited
There has been limited research into verification methods that would allow one actor to check AI safety-related claims made by another actor. While there has been some initial work into some topics such as hardware-enabled mechanisms, many verification questions have received little research, even though trusted assurances may require extensive research and stress-testing.

[11] 'How Jackson Hole Became an Economic Obsession', *The New York Times*, 25 August 2023, https://www.nytimes.com/2023/08/24/business/economy/jackson-hole-economic-conference.html.

Potential Policies

- **Advance high priority research identified in existing research agendas.** There are a range of in-progress or published research agendas which identify critical research questions that need to be addressed to advance the state of verification research.[12] [13] [14] [15] [16] [17] Collectively, these research agendas provide a roadmap for philanthropists, states, and researchers to direct further resources to make progress in this critical area.
- **Identify specific verification projects that should be collaborated on internationally.** Some verification projects are either stronger or only possible when they are produced internationally. A case in point is the US-Soviet Joint Verification Experiment of 1988, which allowed both states to test the explosive yields of the other party's underground detonations at close range to enable verification for the Threshold Test Ban Treaty, which had been at an impasse since 1974.[18]
- **Collaboratively stress-test emerging verification mechanisms.** Many proposals for verification include components that are open-sourced, which would allow global stress-testing of protocols and build trust. Researchers from around the world can attempt to stress-test and improve this open-source technology to build collaborative verification mechanisms.[19]
- **Organize global competitions for verification research.** States can launch global competitions on developing secure verification methods, similar to global competitions on post-quantum cryptography standardization launched by the US National Institute for Standards and Technology.[20]

[12] Mauricio Baker et al. (Forthcoming). RAND.
[13] Ben Harack et al. (Forthcoming). Oxford University
[14] Scher & Thiergart, 'Mechanisms to Verify International Agreements About AI Development', November 2024, https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development
[15] Anka Reuel et al., 'Open Problems in Technical AI Governance' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.14981; Miles Brundage et al., 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims' (arXiv, 20 April 2020), http://arxiv.org/abs/2004.07213.
[16] Akash Wasil et al. (Forthcoming).
[17] James Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees', 23 August 2024, https://yoshuabengio.org/wp-content/uploads/2024/09/FlexHEG-Interim-Report_2024.pdf.
[18] '30th Anniversary of JVE: Snezhinsk Scientists Reach out to Congratulate the US Colleagues', n.d., https://nonproliferation.org/lab-to-lab-joint-verification-experiment.
[19] Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees'.
[20] 'Post-Quantum Cryptography', n.d., https://www.nist.gov/pqcrypto.

Testing and Evaluation

Defining and evaluating red lines and early-warning thresholds

Goals

We also recommend ***defining clear red lines that, if crossed, mandate immediate termination of an AI system — including all copies — through rapid and safe shut-down procedures.*** *Governments should cooperate to instantiate and preserve this capacity. Moreover, prior to deployment as well as during training for the most advanced models,* ***developers should demonstrate to regulators' satisfaction that their system(s) will not cross these red lines*** *- IDAIS-Oxford, 2023*

We propose ***red lines in AI development as an international coordination mechanism...***

- *Autonomous Replication or Improvement: No AI system should be able to copy or improve itself without explicit human approval and assistance. This includes both exact copies of itself as well as creating new AI systems of similar or greater abilities.*
- *Power Seeking: No AI system should take actions to unduly increase its power and influence.*
- *Assisting Weapon Development: No AI systems should substantially increase the ability of actors to design weapons of mass destruction, or violate the biological or chemical weapons convention.*
- *Cyberattacks: No AI system should be able to autonomously execute cyberattacks resulting in serious financial losses or equivalent harm.*
- *Deception: No AI system should be able to consistently cause its designers or regulators to misunderstand its likelihood or capability to cross any of the preceding red lines. - IDAIS-Beijing, 2024*

We should set early-warning thresholds: levels of model capabilities indicating that a model may cross or come close to crossing a red line. *This approach builds on and harmonizes the existing patchwork of voluntary commitments such as responsible scaling policies. Models whose capabilities fall below early-warning thresholds require only limited testing and evaluation, while more rigorous assurance mechanisms are needed for advanced AI systems exceeding these early-warning thresholds. - IDAIS-Venice, 2024*

Progress

Moderate
While there has been limited direct work on the red lines concept, there has been continued progress on related work. For example, several leading AI companies have signed the Frontier Safety Commitments at the AI Seoul Summit, committing to defining thresholds at which risks posed by a model or system would be deemed intolerable. AI companies and AI safety institutes have also made progress on designing and running concerning capability evaluations to test whether models are meeting these predefined thresholds.

Potential Policies

Broader actions that aim to ensure necessary resourcing of AI safety R&D and build state visibility into the activities of model developers through domestic regulation, could also contribute towards this goal.

- **Develop a common taxonomy for red lines, early warning thresholds, and other key concepts.** While the IDAIS-Beijing consensus statement provides a foundational set of red lines, significant further work is required to build consensus across a broader set of stakeholders on the definition of red lines and related concepts such as early warning thresholds.[21] Further work is also required to determine how these concepts map on to each other. The OECD carried out a similar multi-stakeholder consultation process to build a taxonomy for AI incidents and hazards through its AI Expert Group on AI Incidents.[22]
- **Develop in-house technical capacity with security clearances to conduct capability evaluations relevant to weapons development and cyberattack red lines.** National security actors have expertise in evaluating cyber and Chemical, Biological, Radiological, and Nuclear (CBRN) threats. Developing a deep understanding of AI-enabled risks in these domains requires access to classified information and national security experts. States are uniquely positioned to develop in-house technical capacity that can design and run model evaluations, while being connected to secret and top-secret information sources and experts.[23]
- **Require companies to conduct evaluations on frontier models and share results with relevant regulators.** Broad requirements for companies to conduct evaluations on frontier models and share results will help build state capacity and visibility, which will be needed to develop specific future requirements on red lines-related evaluations. There has already been progress in this area—several American companies have voluntarily committed to conducting internal and external testing and evaluation, and are required by an executive order to share these results with the US government for models above a specific compute threshold.[24] [25] The EU AI Act also requires that developers of general purpose AI models that pose systemic risks conduct model evaluations.[26]
- **Establish a clear end-to-end process for domestic and international information disclosure related to red lines and early warning thresholds.** Currently, there is some ambiguity around the types of information states require companies to report with respect to both safety and security testing, how different actors within government should coordinate that information, and which parts of a given government should respond to emerging threats. States should work with companies and academics to clarify which actors (contractors, 3rd parties, etc.) should report to which government actor, and to

[21] Early warning thresholds, when passed, could indicate that critical model capability or risk levels have been reached and predefined actions should be taken. These thresholds build towards but are less severe than red lines, providing advance warning that models are potentially on track to cross them.
[22] 'Defining AI Incidents and Related Terms', OECD Artificial Intelligence Papers, vol. 16, OECD Artificial Intelligence Papers, 6 May 2024, https://doi.org/10.1787/d1a8d965-en.
[23] Akash Wasil et al., 'AI Emergency Preparedness: Examining the Federal Government's Ability to Detect and Respond to AI-Related National Security Threats' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.17347.
[24] 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', § 4 (2023).
[25] 'White House Voluntary AI Commitments' (White House, n.d.), https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.
[26] 'EU AI Act', § 55 (2024).

specify what types of information would be most valuable to provide.[27] Further work in this area could also identify what types of information should be shared internationally.

- **Develop methods to verify that necessary mitigations are in place.** As AI systems pass different early warning threshold and potentially pose greater risks, mitigations are necessary to manage rising risks, including security measures (e.g., hardened cybersecurity) to prevent the theft of model weights and safety measures (e.g., red-teaming, content filters) to ensure safe deployment. Currently such mitigations are voluntarily implemented by AI companies, but in the future, states may need to require companies to both implement and prove that these mitigations are in place. Further research is required to develop methods for such verification to take place.
- **Conduct research into red line/threshold violation protocols.** Limited research has been dedicated to determining how states, corporations, and other key actors should act if a model evaluation suggests that a red line has been or may be crossed. Further work is required to understand what response protocols should look like both domestically and internationally. An example of a potential protocol in this vein is Coordinated Pausing, which is a proposal for AI corporations to agree to collectively pause research and development in specific areas when model evaluations suggest certain unacceptable risk thresholds have been crossed, and only resume development in those areas when certain safety criteria have been met.[28]
- **Research and develop methods to reliably shut down AI systems that have crossed or are in danger of crossing red lines.** Research in this area is nascent, but it may be necessary to develop ways to safely shut down AI systems, once it is discovered that this system is crossing red lines. Proposed solutions include an 'off-switch' installed onto AI hardware (e.g., semiconductor chips).[29]

Setting up an ecosystem of 3rd party audits and evaluations

Goals

> *[Governments] should require that AI developers of frontier models be* ***subject to independent third-party audits evaluating their information security and model safety*** *- IDAIS-Oxford, 2023*

> *States should mandate that developers conduct regular testing for concerning capabilities, with* ***transparency provided through independent pre-deployment audits by third parties granted sufficient access to developers' staff, systems and records necessary to verify the developer's claims.*** *Additionally, for models exceeding early-warning thresholds, states could require that* ***independent experts approve a developer's safety case prior to further training or deployment.*** *- IDAIS-Venice, 2024*

[27] Joe O'Brien et al., 'Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.01420.
[28] Jide Alaga and Jonas Schuett, 'Coordinated Pausing: An Evaluation-Based Coordination Scheme for Frontier AI Developers' (arXiv, 2023), https://doi.org/10.48550/ARXIV.2310.00374.
[29] 'Considerations and Limitations for AI Hardware-Enabled Mechanisms', 10 March 2024, https://blog.heim.xyz/considerations-and-limitations-for-ai-hardware-enabled-mechanisms/.

*States should mandate that developers conduct regular testing for concerning capabilities, with **transparency provided through independent pre-deployment audits by third parties granted sufficient access to developers' staff, systems and records necessary to verify the developer's claims**. Additionally, for models exceeding early-warning thresholds, states could require that **independent experts approve a developer's safety case prior to further training or deployment**. - IDAIS-Venice, 2024*

Progress

Moderate
There has been moderate progress in the development of a 3rd party auditing and evaluation ecosystem. Evaluations tend to include tests run in order to understand the risk a model or system poses. Across some countries, AI safety institutes, national research institutions, nonprofits, and for-profit companies have been set up to play the role of 3rd party evaluators. However, there has been slower progress on 3rd party evaluation becoming required by governments. Some AI companies are providing pre-deployment access to 3rd party evaluators; however, such access is voluntary and provided on terms dictated by the AI companies.

Audits are a broader type of activity, and may include the governance systems of companies, product safety, and the impact of products on downstream users. There has been more progress on audits becoming required by governments. For example, by 2026, the EU AI Act will require providers of some types of high-risk AI systems to undergo third-party conformity assessments, which are a type of audit.

Potential Policies

Broader actions that aim to ensure necessary resourcing of AI safety R&D could also contribute towards this goal.

- **Develop a comprehensive evaluation framework, which provides clarity on the role of 3rd party *evaluators*.** 3rd parties will increasingly play a role in either directly evaluating models, or in auditing and verifying that certain evaluations have been run by companies. States, in collaboration with researchers and companies, need to clarify where 3rd party testing will be most necessary, and what type of 3rd party testers (e.g., government-linked or private) will be needed for different types of risks.[30] This will also require building consensus on issues such as pre-deployment model access, and use of trusted evaluation infrastructure.
- **Develop a comprehensive auditing framework, which provides clarity on the role of 3rd party *auditors*.** Audits may target governance systems of technology providers (e.g., risk management systems of AI developers), verification of model characteristics (e.g., robustness), reviewing documentation of model limitations (e.g., model cards), and/or review of downstream impacts of models on users.[31] Consensus-building is also required on the distribution of responsibilities between the public and private sector for external audits. Some audits, in particular white- or gray-box audits

[30] 'Third-Party Testing as a Key Ingredient of AI Policy', 25 March 2024, https://www.anthropic.com/news/third-party-testing.
[31] Jakob Mökander et al., 'Auditing Large Language Models: A Three-Layered Approach', *AI and Ethics*, 30 May 2023, https://doi.org/10.1007/s43681-023-00289-2.

involving sensitive information, may require the involvement of public bodies, as is the case in nuclear safety and life sciences. On the other hand, black box model audits and governance audits involving less sensitive information may be more effectively carried out by a regulated market of private auditors.[32]

- **Develop and share evaluation infrastructure.** Evaluation infrastructure, such as agent scaffolding, capability elicitation, and grading tools, would enable 3rd party evaluators to evaluate models more effectively and cheaply.[33] The UK AISI, for example, has open-sourced its evaluation framework, Inspect, while Singapore's AI Verify Foundation has open-sourced Project Moonshot, a testing platform that allows companies to run pre-existing tests and automated red-teaming on their models more easily.[34] Corporations could be encouraged to do the same.
- **Provide resources such as compute or API credits to support 3rd party evaluators.** States, corporations and philanthropists could set up common resource pools (e.g., a national research cloud, API credits) to support 3rd party evaluators. For example, the UK's AI Research Resource, a cluster of advanced computers for AI research, is receiving US$1 billion to support R&D.[35]
- **Establish grants and prizes to incentivize R&D on evaluations of concerning capabilities.** An example is Anthropic's funding initiative for third-party organizations to develop evaluations that effectively measure advanced capabilities in AI models, including cybersecurity, CBRN capabilities, and model autonomy.[36] Promising research directions include improving reliability, scalability, comprehensiveness of model evaluations, as well as defining best practices for monitoring agentic systems.[37]

Requiring safety cases and/or guarantees

Goals

*AI developers should also be required to share comprehensive ... **predictions about their systems' behaviour in third party evaluations and post-deployment** with relevant -IDAIS-Oxford, 2023*

*The onus should be on developers to convincingly demonstrate that red lines will not be crossed such as through rigorous empirical evaluations, **quantitative guarantees or mathematical proofs** - IDAIS-Beijing, 2024*

---

[32] Merlin Stein et al., 'Public vs Private Bodies: Who Should Run Advanced AI Evaluations and Audits? A Three-Step Logic Based on Case Studies of High-Risk Industries' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.20847.
[33] Gabriel Mukobi, 'Reasons to Doubt the Impact of AI Risk Evaluations' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2408.02565.
[34] 'Project Moonshot: An LLM Evaluation Toolkit', n.d., https://aiverifyfoundation.sg/project-moonshot/; 'Inspect: An Open-Source Framework for Large Language Model Evaluations', n.d., https://inspect.ai-safety-institute.org.uk/.
[35] 'AI Research Resource Funding Opportunity Launches', 24 January 2024, https://www.ukri.org/news/ai-research-resource-funding-opportunity-launches/.
[36] 'A New Initiative for Developing Third-Party Model Evaluations', 1 July 2024, https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations.
[37] Anka Reuel et al., 'Open Problems in Technical AI Governance' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.14981.

> *Developers should* ***submit a high-confidence safety case, i.e., a quantitative analysis that would convince the scientific community that their system design is safe****, as is common practice in other safety-critical engineering disciplines. Additionally, safety cases for sufficiently advanced systems should discuss organizational processes, including incentives and accountability structures, to favor safety. - IDAIS-Venice, 2024*

Progress

Limited
There has been limited progress on developers being required to provide more comprehensive safety guarantees. However, there has been more work on safety cases, provable safety and other approaches which could provide greater certainty about the safety of AI systems.

Potential Policies

- **Direct funding towards alternative scientific or engineering paradigms which may be safe by design.** Most AI development today is focused on the existing Transformer-based deep learning paradigm. It is possible that systems built in this paradigm may never be sufficiently safe to deploy AI systems in important safety-critical domains (e.g., aviation). As such states and philanthropists may want to direct funding towards other research approaches, such as Guaranteed Safe AI.[38]
- **Direct funding towards research into 'safety cases'.** A safety case is a structured argument that an AI system is safe within a particular training or deployment context. Positively arguing that a system is safe may be necessary for safety-critical deployment or if frontier AI models show signs of potentially posing catastrophic risk. A safety case may include evidence of sufficiently strong control measures and trustworthiness despite capability to cause harm, or other methods of demonstrating safety.[39] Some AI companies, such as Anthropic, have started hiring researchers for such roles, while other organizations such as Apollo Research have begun to build domain-specific approaches to fleshing out safety cases for AI 'scheming'. [40] [41] [42]

[38] David 'davidad' Dalrymple et al., 'Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2405.06624.
[39] Joshua Clymer et al., 'Safety Cases: How to Justify the Safety of Advanced AI Systems' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2403.10462.
[40] 'Safety Case Specialist (Safety Mitigations) (London)', n.d., https://www.linkedin.com/jobs/view/safety-case-specialist-safety-mitigations-london-at-anthropic-3956180495/?originalSubdomain=uk.sa.
[41] Mikta et al., 'Towards Evaluations-Based Safety Cases for AI Scheming' (Apollo Research, 1 November 2024),.https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/6724bd285d993323e03b89d6/1730460969317/Toward_evaluations_based_safety_cases_for_AI_scheming.pdf.
[42] Marie Davidsen Buhl et al., 'Safety Cases for Frontier AI' (arXiv, 28 October 2024), http://arxiv.org/abs/2410.21572.

Domestic Governance

Domestic governance actions in this section exclude testing and evaluation, which is covered separately.

Implementing domestic model registration

Goals

> *In domestic regulation, we* ***recommend mandatory registration for the creation, sale or use of models above a certain capability threshold, including open-source copies and derivatives****, to enable governments to acquire critical and currently missing visibility into emerging risks. - IDAIS-Oxford, 2023*
>
> *We should immediately* ***implement domestic registration for AI models and training runs above certain compute or capability thresholds.*** *Registrations should ensure governments have visibility into the most advanced AI in their borders and levers to stem distribution and operation of dangerous models. - IDAIS-Beijing, 2024*

Progress

Moderate
There has been moderate progress towards model and training run registration. In both the EU and US, model developers are required or will soon be required to notify governments if they are developing AI models above a specified compute threshold, which serves as a proxy for concerning capabilities and risk. In China, model developers are required to register all models that are public-facing with the Cyberspace Administration of China. Compute thresholds have also been referred to in two draft AI law proposals written by leading Chinese legal scholars and experts.

Potential Policies

- **Ensure that model registration policies provide wider and deeper visibility into emerging AI risks.** States have already taken initial steps in requiring developers to register certain types of AI models.[43] Beyond existing registration requirements, states may want to consider signaling that a wider range of documentation and data may be required for model registration in the future, creating incentives for this data to be collected today. This could include data beyond what is already publicly disclosed through model cards, such as detailed information about the model's training process, including the datasets used, data sanitization practices, hardware and software components involved, and known vulnerabilities in these components.[44]

[43] For example, the Biden Administration's executive order on AI requires that developers of dual-use foundation models notify the government if they are developing or planning to develop a model above a compute threshold. Chinese model registration regulations apply to all public-facing generative AI models, but not to models used for research and development. In the EU, AI systems that are classified as high-risk will need to be registered and developers of general purpose AI models which pose systemic risks will also need to notify the EU AI Office.
[44] Noam Kolt et al., 'Responsible Reporting for Frontier AI Development' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2404.02675.

- **Conduct research and build consensus on refined compute, capability, and risk thresholds that should trigger greater regulatory scrutiny**. Current model registration requirements in the EU and US broadly use compute thresholds as a proxy for risk.[45] This is a blunt approach that provides an initial and tentative line in the sand for regulators. Further refinement of thresholds and related evaluation and testing methods will ensure that greater scrutiny of AI developers is tied to fine-grained indicators of concerning capabilities and risks.[46]

Monitoring large-scale data centers

Goals

Governments should monitor large-scale data centers... - IDAIS-Oxford, 2023

Progress

Limited
There has been limited progress towards monitoring large-scale data centers. The US government has issued proposed rules that would require cloud compute providers (who often run their own large-scale data centers) to verify the identity of foreign customers, and report transactions with foreign persons that would result in training large AI models that could be used for cyberattacks.

Although other jurisdictions may have existing laws that could be applied to frontier AI-relevant data center monitoring, there has been limited work exploring this.

Potential Policies

- **Leverage cloud compute providers as intermediaries to gain better visibility into AI development trends and enforce regulations more effectively on model developers.** Given the concentration of the cloud compute provider market, regulating cloud compute providers would provide governments with a higher leverage target for governance.[47] For example, governments could license domestic cloud providers who meet certain infrastructure security requirements and agree to verify that model developers using their cloud compute services are complying with domestic requirements.[48] More broadly, compute providers could be a promising intermediary for governance by:
    - Providing additional physical security and cybersecurity for AI models.
    - Keeping records of high-level information such as a customer's compute usage.

[45] Sara Hooker, 'On the Limitations of Compute Thresholds as a Governance Strategy' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.05694.
[46] Leonie Koessler, Jonas Schuett, and Markus Anderljung, 'Risk Thresholds for Frontier AI' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2406.14713.
[47] Lennart Heim et al., 'Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2403.08501.
[48] 'Global Governance: Goals and Lessons for AI' (Microsoft, n.d.), https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lhQ0.

  - Verifying key properties such as customer identities and AI system properties.
  - Enforcing compute access restrictions for non-compliant customers.
- **Develop privacy-preserving tools to enable a wider range of governance activity.** Privacy-preserving tools could allow cloud compute providers to verify more fine-grained information, such as detailed workloads, which are relevant to governance. Detailed workload verification is currently not possible as compute providers would require direct access to models or code to verify certain governance-relevant properties (e.g., that a model evaluation had been run), which would be a breach of the confidentiality and privacy agreements they have with customers. Further research into confidential computing and other privacy-preserving approaches would allow for more detailed workload verification without exposing sensitive data.[49]
- **Develop robust compute workload classification methods.** Compute workload classification would allow compute providers to determine whether the type of computation carried out by a customer should be part of an oversight regime (e.g., a large AI training run). Research is needed to create methods for compute workload classification that are resistant to adversarial gaming. Currently, adversarial actors may be able to change the computational pattern of their workload to evade detection.[50]

Putting in place AI incident tracking and reporting

Goals

Governments should ... track AI incidents - IDAIS-Oxford, 2023

Progress

Moderate
There has been moderate progress towards incident tracking. The Partnership on AI and the OECD both maintain different databases of AI incidents. The EU AI Act also requires the providers of high-risk AI systems and general purpose AI systems that pose systemic risks to report serious incidents to national market surveillance authorities or the EU AI Office. Beyond these examples, there remains limited comprehensive AI incident tracking.

Potential Policies

- **Establish a comprehensive hybrid approach to incident tracking and reporting.** Such a system could combine mandatory reporting for major incidents, voluntary reporting for minor incidents, and a public portal for public reporting. This system could be supported by a standardized

[49] Heim et al., 'Governing Through the Cloud'.
[50] Heim et al.

framework for data collection to ensure comparability of data collected from various sources, and an incident investigation team that would look into root causes of major incidents.[51]

- **Encourage more proactive documentation of AI systems to enable better analysis of incidents.** Current AI incident practices tend to be reactive, in that information about a system and incident are logged after the fact. More proactive documentation of AI systems before AI incidents have even occurred would make it easier to track the full lifecycle of the AI system should an incident occur. This would also help investigators determine what early signs of system failure look like.[52]
- **Research and develop automated metadata collection mechanisms.** Mechanisms such as flight recorders—also known as 'black boxes'—are used to provide critical contextual and technical metadata around an incident. They are not examined during routine operations, but are helpful during investigation. Comparable mechanisms for AI systems have not been developed, but may include 'snapshots' of model's technical data (e.g., model weights, input and output logs) at regular checkpoints.[53]

Adopting risk management and risk assessment practices

Goals

*developers should also be required to **share comprehensive risk assessments, [and] policies around risk management** - IDAIS-Oxford, 2023*

Progress

Moderate
There has been moderate progress towards risk assessments and risk management. Several leading AI companies signed the Frontier Safety Commitments at the AI Seoul Summit, which require them to produce risk assessment and mitigation plans. Some companies have already published such plans, variously referred to as responsible scaling, frontier safety, or preparedness policies, and also set up dedicated teams focused on the safety of current and future systems.

Some jurisdictions, such as the EU, have also begun to require developers to implement risk management systems, while other jurisdictions such as the US, have put forward voluntary risk management frameworks that companies can choose to adopt.

However, there is limited evidence that more comprehensive risk management practices have been adopted, such as a three lines of defense model, which would include several layers of risk monitoring and independent audits housed within a single company.

[51] Ren Bin Lee Dixon and Heather Frase, 'An Argument for Hybrid AI Incident Reporting' (Center for Security and Emerging Technology, March 2024), https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/.
[52] Violet Turri and Rachel Dzombak, 'Why We Need to Know More: Exploring the State of AI Incident Documentation Practices', in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '23: AAAI/ACM Conference on AI, Ethics, and Society, Canada: ACM, 2023), 576–83, https://doi.org/10.1145/3600211.3604700.
[53] Dixon and Frase, 'An Argument for Hybrid AI Incident Reporting'.

Potential Policies

- **Conduct research on AI risk management systems for frontier AI labs, as adapted from other safety-critical industries.** Risk management is a continuous iterative process run throughout the entire model lifecycle, consisting of risk identification, analysis and evaluation. Key tools include risk taxonomies for structuring the risk universe, the Delphi technique for collating expert forecasts of the likelihood of different risk scenarios, and risk matrices which help decision makers determine when specific risks need to be addressed based on likelihood and potential severity. While initial research has been carried out to develop some methods, further work is needed to determine other key issues, such as risk appetites and risk tolerances.[54]
- **Establish regulatory requirements for risk management systems**. States can play a role in providing clearer guidelines on how companies should carry out risk assessment and what of that should be reported to regulators or 3rd party auditors. For example, Article 9 of the EU AI Act specifies that risk management systems should include detailed information about the types of risk analysis, identification, and evaluation companies must conduct, (e.g., evaluating risks that may arise after deployment, assessed via a post-market monitoring system).
- **Assess whether company practices follow existing government guidance and recommend improvements.** Researchers and regulators can evaluate the extent to which companies are adhering to established guidance and following through on their voluntary commitments. An example of this is a report from the Institute for AI Policy and Strategy, which assesses Anthropic's responsible scaling policy against guidance issued by the UK government.[55]

Requiring post-deployment monitoring

Goals

> *Advanced AI systems may increasingly engage in complex multi-agent interactions with other AI systems and users. This interaction may lead to emergent risks that are difficult to predict.* ***Post-deployment monitoring is a critical part of an overall assurance framework, and could include continuous automated assessment of model behavior, centralized AI incident tracking databases, and reporting of the integration of AI in critical systems. Further assurance should be provided by automated run-time checks, such as by verifying that the assumptions of a safety case continue to hold*** *and safely shutting down a model if operated in an out-of-scope environment - IDAIS-Venice, 2024*

[54] Leonie Koessler and Jonas Schuett, 'Risk Assessment at AGI Companies: A Review of Popular Risk Assessment Techniques from Other Safety-Critical Industries' (arXiv, 2023), https://doi.org/10.48550/ARXIV.2307.08823.
[55] Bill Anderson-Samways et al., 'Responsible Scaling: Comparing Government Guidance and Company Policy' (Institute for AI Policy and Strategy, 11 March 2024), https://www.iaps.ai/research/responsible-scaling.

Progress

Moderate
There has been moderate progress towards some forms of post-deployment monitoring. AI developers such as OpenAI monitor usage patterns to detect misuse of AI systems for activities such as cyberattacks.[56] Chinese companies such as Tencent and Alibaba run comprehensive input and output filters when a model is deployed, to ensure that malicious prompts and output are filtered away.[57] Moreover, some of the progress mentioned under incident reporting also contributes towards robust post-deployment monitoring. Finally, the EU AI Act requires that providers of high-risk AI systems put in place a post-market monitoring which collects relevant data on the performance of the system and monitor compliance with other EU AI Act requirements.

AI companies and AI Safety Institutes also are building up capacity to understand the impact of AI at a societal level, by hiring researchers to focus on assessing societal impacts of AI systems.[58]

However, most post-deployment monitoring is done by model developers, with limited visibility provided to governments, including in safety-critical industries.

Note: Incident tracking is discussed in detail in a separate section.

Potential Policies

- **Require companies to collect and share interconnected post-deployment monitoring data.** Currently governments have very limited visibility into post-deployment data. Interconnected post-deployment monitoring refers to linking different kinds of post-deployment information (e.g., model integration and usage, application usage, incident information) to each other for better comprehensive risk assessment, as well as linking post-deployment information to specific risk mitigation. Such an approach has already been partly effective in other industries. For example, the US Food and Drug Administration monitors and connects population-level impacts of drugs to the observations of doctors, informing decisions such as whether to apply new warning labels.[59]
- **Monitor the activation of concerning features that should trigger automatic shutdown.** Advances in some AI safety research directions, such as mechanistic interpretability and representation engineering, are making it possible for AI developers to isolate features (i.e., higher-level concepts) that correspond to undesirable or harmful behaviors (e.g., deception).[60] Anthropic, in a recent

[56] 'Influence and Cyber Operations: An Update' (OpenAI, 9 October 2024), https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf.
[57] '大模型安全与伦理报告 / Tencent Large Model Safety-Security and Governance Report' (Tencent Research Institute, 24 January 2024), https://mp.weixin.qq.com/s/KCWw9gBwUnzywyNW_K8-4A.l.
[58] 'Psychological and Social Risks (Societal Impacts) Workstream Lead (UK AISI)', n.d., https://boards.eu.greenhouse.io/aisi/jobs/4399639101; 'Research Engineer, Societal Impacts (Anthropic)', n.d., https://boards.greenhouse.io/anthropic/jobs/4251453008.
[59] Merlin Stein, Jamie Bernardi, and Connor Dunlop, 'The Role of Governments in Increasing Interconnected Post-Deployment Monitoring of AI' (arXiv, 7 October 2024), http://arxiv.org/abs/2410.04931.
[60] Andy Zou et al., 'Representation Engineering: A Top-Down Approach to AI Transparency' (arXiv, 10 October 2023), http://arxiv.org/abs/2310.01405.

exploratory blog post, has argued that it may be possible to automatically detect features that should result in immediate shutdown at runtime.[61]

- **Collect data about the integration of AI systems into critical infrastructure.** The Biden Administration 2023 AI Executive Order calls for Sector Risk Management Agencies to assess potential risks related to the use of AI in critical infrastructure sectors.[62] While such agencies can and should play a role in assisting critical infrastructure operators to responsibly adopt, they should minimally require data from AI developers and/or critical infrastructure operators about their use of AI systems.[63]
- **Invest in technical governance approaches that increase visibility into agentic systems**. Visibility, particularly post-deployment visibility into the activities of increasing agentic AI systems, is extremely limited today. States can resource technical governance approaches focused on building visibility such as agent identifiers (watermarks, IDs), real-time monitoring, and activity logging to ensure that users and governments can have visibility into how AI agents are deployed in society.[64]

Instituting a professional code of ethics for AI practitioners

Goals

*Moreover, **states can help institute ethical norms for AI engineering**, for example by stipulating that engineers have an individual duty to protect the public interest similar to those held by medical or legal professionals. - IDAIS-Venice, 2024*

Progress

Limited
While some industry associations such as the British Computer Society argued that technologists working in high-stakes IT roles, particularly in AI, should be registered professionals that meet independent standards of ethical practice, states do not seem to be considering instituting such norms.

Potential Policies

- **Develop and require AI practitioners to take a 'Hippocratic Oath' for AI.** The Hippocratic Oath acts as an individualized precautionary principle for doctors, by which they pledge to 'first, do no harm', put the interests of patients first, and uphold high standards of professional integrity. AI practitioners, particularly those working on frontier AI systems which may pose catastrophic risks to humanity,

[61] Roger Grosse, 'Three Sketches of ASL-4 Safety Case Components', 5 November 2024, https://alignment.anthropic.com/2024/safety-cases/.
[62] 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', § 4 (2023).
[63] Kyle Crichton et al., 'Securing Critical Infrastructure in the Age of AI' (Center for Security and Emerging Technology, 15 October 2024), https://cset.georgetown.edu/publication/securing-critical-infrastructure-in-the-age-of-ai/.
[64] Alan Chan et al., 'Visibility into AI Agents' (arXiv, 17 May 2024), http://arxiv.org/abs/2401.13138.

could be required to take such an oath. While penalties for going against an oath may be difficult to enforce, the principles in the oath can be reflected in specific professional standards which AI practitioners can be held to.

- **Enshrine a 'Right to Warn' that allows AI practitioners to exercise professional ethical responsibility**. As with doctors, professional responsibility may conflict with corporate incentives.[65] However, many potential risks from AI are not covered by existing law. An open letter signed by several former employees of frontier AI labs has called for enhanced whistleblower protections to be enshrined under a 'Right to Warn', with other AI practitioners drawing attention to this issue in Senate hearings.[66]

[65] Sharon A. Clark, 'The Impact of the Hippocratic Oath in 2018: The Conflict of the Ideal of the Physician, the Knowledgeable Humanitarian, Versus the Corporate Medical Allegiance to Financial Models Contributes to Burnout', *Cureus* 10, no. 7 (30 July 2018): e3076, https://doi.org/10.7759/cureus.3076.
[66] 'A Right to Warn about Advanced Artificial Intelligence', 4 June 2024, https://righttowarn.ai/; Sophie Luskin, 'Need for Whistleblower Protections in Artificial Intelligence Industry Discussed in Senate Judiciary Hearing', *Whistleblower Network News*, 24 September 2024, https://whistleblowersblog.org/corporate-whistleblowers/need-for-whistleblower-protections-in-artificial-intelligence-industry-discussed-in-senate-judiciary-hearing/.

International Governance

Domestic governance actions, actions on AI safety funding and collaboration, as well as testing and evaluation, will be reflected in their respective linked sections. This section will focus specifically on the international dimensions of the specified goals.

Establishing AI safety as a global public good

Goals

> ***AI safety is a global public good** that should be supported by public and private investment, with advances in safety shared widely - IDAIS-Oxford, 2023*
>
> *The global nature of these risks from AI makes it necessary to **recognize AI safety as a global public good**, and work towards global governance of these risks... we call on states to carve out AI safety as a cooperative area of academic and technical activity, distinct from broader geo-strategic competition on development of AI capabilities - IDAIS-Venice, 2024*

Progress

Moderate
The AI Summits have brought together key AI powers across different cultural contexts and political systems to discuss AI safety issues. AI safety and risk has also continued to be a key topic for intergovernmental dialogue between the US and China. At the same time, various leading Chinese experts have begun expanding on the idea of AI safety as a global public good.

Nonetheless, the current situation is tenuous, and there is a real risk that AI safety becomes linked with broader geopolitical tensions around AI competition.

Potential Policies

- **Ensure international collaboration on AI safety is protected from broader geopolitical competition on AI.** States can take steps to ensure that restrictive policies such as export controls do not inadvertently impact AI safety collaboration. A case in point—American export controls on Huawei in 2019, led to the inadvertent exclusion of Huawei from certain standards-setting bodies. Organizations such as the Standard Performance Evaluation Corporation (SPEC), which sets standards for server energy efficiency and maintains a software toolkit to determine server energy efficiency levels, were not able to engage with Huawei, leading to a split in global standards-setting on this topic. Adjustments to export controls were eventually made after more than two years, but have thus far not

been able to repair the lasting damage to international coordination on server energy efficiency standards.[67]

- **Identify, develop, and share safety technologies, such as 'PALs for AI'.** Permissive action links (PALs) are a type of electromechanical lock that improved nuclear security during the Cold War by preventing unauthorized and accidental deployment of nuclear weapons. Core elements of this technology were developed by the US and shared with the USSR, in order to improve global nuclear security. Key American national security figures such as Jason Matheny, former Deputy Director for National Security at the U.S. Office of Science and Technology and current CEO of RAND, have called for PALs for AI to be developed and shared even amongst competitors.[68] Examples of PALs for AI may include safety techniques and approaches that are strictly safety-enhancing, such as anomaly detection to identify when AI systems are behaving in potentially hazardous ways.[69]
- **Conduct further research to clarify the 'AI safety as a global public good' concept.** The International Monetary Fund defines a global public good as a good which individuals cannot be charged to use, where benefit to each individual is small and where the benefits are realized in the far future.[70] Parts of this definition and other analogous global public goods (e.g., the environment) may apply to AI safety, but novel intellectual progress may be required to identify what AI safety as a global public good would look like. While there has been early work in this direction by Chinese AI governance researchers, further work is still required.[71]

Establishing conditional market access through globally aligned standards on red lines

Goals

> ***Domestic regulators ought to adopt globally aligned requirements to prevent crossing these red lines. Access to global markets should be conditioned on domestic regulations meeting these global standards*** *as determined by an international audit, effectively preventing development and deployment of systems that breach red lines. - IDAIS-Beijing, 2024*

[67] Nigel Cory, 'The U.S.-China Tech Conflict Fractures Global Technical Standards: The Example of Server and Datacenter Energy Efficiency', 22 August 2023, https://itif.org/publications/2023/08/22/the-us-china-tech-conflict-fractures-global-technical-standards-the-example-of-server-and-datacenter-energy-efficiency/.
[68] Jeffrey Ding, 'Keep Your Enemies Safer: Technical Cooperation and Transferring Nuclear Safety and Security Technologies', *European Journal of International Relations*, 27 April 2024, 13540661241246622, https://doi.org/10.1177/13540661241246622.
[69] Oliver Guest and Zoe Williams, 'Topics for Track IIs: What Can Be Discussed in Dialogues about Advanced AI Risks without Leaking Sensitive Information?' (Institute for AI Policy and Strategy, 2 May 2024), https://www.iaps.ai/research/dialogue-topics.
[70] Moya Chin, 'What Are Global Public Goods?', December 2021, https://www.imf.org/en/Publications/fandd/issues/2021/12/Global-Public-Goods-Chin-basics.
[71] 王迎春 et al., '人工智能安全作为 全球公共产品 研究报告 / AI Safety as Global Public Goods Working Report', 5 July 2024, https://www.sipa.sjtu.edu.cn/Kindeditor/Upload/file/20240704/%E7%A0%94%E7%A9%B6%E6%8A%A5%E5%91%8A%E6%89%8B%E5%86%8C-04.pdf.

Progress

Limited
There has been no progress to date on establishing conditional market access through globally aligned standards on red lines. Further work is required to detail and operationalize red lines-related standards, and to link these to conditional market access.

Potential Policies

- **Identify a suitable channel for red lines standards development within existing coordination mechanisms.** A range of existing processes—the network of AI safety institutes, AI Summits, ISO/IEC—could provide avenues for states to align their domestic red lines standards once they are developed. Given that frontier models are only being developed in a few countries, it may be more effective for the initial standards alignment approach to involve key AI powers.
- **Establish regulatory markets as a mechanism for coordinating regulatory goals and market access.** The pace of technological development and scope of technological transformation may exceed the ability of states, with large and slow-moving bureaucracies, to adapt and respond directly. Instead, states could consider outsourcing technical aspects of governance to licensed private regulators.[72] Under this model, states would specify high-level goals for private regulators to achieve. States across the world could come together to establish minimal global regulatory goals, conditioning market access for AI developers on being certified by recognized private regulators.
- **Adopt a jurisdictional certification process and a multilateral export control regime to manage conditional market access.** A proposal in this vein is for an International AI Organization, roughly analogous to the IAEA. The IAIO would be responsible for setting standards, certifying that states have sufficient capacity to monitor and regulate frontier AI developers, and coordinating enforcement. In particular, it could work with states to deny non-compliant states access to relevant AI inputs (e.g., semiconductors) and ability to export AI outputs (e.g., models) to key markets, through a multilateral export control regime.[73]

Designing and negotiating international agreements and/or setting up an international organization

Goals

> *Governments around the world — especially of leading AI nations — have a responsibility to develop measures to prevent worst-case outcomes from malicious or careless actors and to rein in reckless competition.* ***The international community should work to create an international coordination process for advanced AI in this vein****. - IDAIS-Oxford, 2023*

[72] Gillian K. Hadfield and Jack Clark, 'Regulatory Markets: The Future of AI Governance' (arXiv, 2023), https://doi.org/10.48550/ARXIV.2304.04914.
[73] Robert Trager et al., 'International Governance of Civilian AI: A Jurisdictional Certification Approach' (arXiv, 2023), https://doi.org/10.48550/ARXIV.2308.15514.

We should take measures to prevent the proliferation of the most dangerous technologies while ensuring broad access to the benefits of AI technologies. To achieve this we should ***establish multilateral institutions and agreements to govern AGI development safely and inclusively with enforcement mechanisms to ensure red lines are not crossed and benefits are shared broadly****.*
- IDAIS-Beijing, 2024

States should agree on technical and institutional measures required to prepare for advanced AI systems, regardless of their development timescale. To facilitate these agreements, we ***need an international body to bring together AI safety authorities****, fostering dialogue and collaboration in the development and auditing of AI safety regulations across different jurisdictions. This* ***body would ensure states adopt and implement a minimal set of effective safety preparedness measures****, including model registration, disclosure, and tripwires...Over time, this body could also set standards for and commit to using verification methods to enforce domestic implementations of the Safety Assurance Framework. These methods can be mutually enforced through incentives and penalty mechanisms, such as conditioning access to markets on compliance with global standards*
- IDAIS-Venice, 2024

Progress

Limited
There has been limited progress towards creating various international coordination processes that are linked to red lines and focused on risks from advanced AI, despite a proliferation of international governance efforts.

Several international coordination processes have been established, including the UN High-Level Advisory Body (HLAB), the AI Summit process, the G7 Hiroshima Process, the International Scientific Report on the Safety of Advanced AI, and China's Global AI Governance Initiative.

With respect to international institutions, the interim report issued by the UN HLAB raises several key functions that would need to be performed for global AI governance. Several experts have also suggested various new models for multilateral organizations that could focus on AI governance.

However, there is very limited international coordination on catastrophic risks from AI, and little focus on enforcement mechanisms linked to red lines.

Potential Policies

- **Build consensus around which elements of frontier AI governance should be internationalized[74].** Internationalization tends to be required in limited circumstances, when domestic governance alone

[74] Dennis et al., 'What Should be Internationalised in AI Governance', November 2024, http://oxfordmartin.ox.ac.uk/publications/what-should-be-internationalised-in-ai-governance

is unable to adequately address risks. Amongst other reasons, internationalization may be required if risks are transnational in nature (i.e., risks originating from one jurisdiction can impact others, such as extremist terrorism), and in situations where states are unlikely to take costly unilateral actions to manage risks (e.g., climate commitments). Internationalization also sits on a spectrum, from standards being set entirely by international bodies, to simply ensuring interoperability of domestic standards. Consensus on what needs to be internationalized, for what reasons, and to what extent is required to build an international AI governance system.[75]

- **Align on whether key international frontier risk governance functions should be established through existing institutions or new processes**. New processes and functions allow for clearer focus without prior institutional baggage, but are also less embedded in broader structures, facing potential issues of resourcing and legitimacy. States should align on whether frontier AI governance functions should be set up as a new process (e.g., the AI Summit process) or go through existing channels (e.g., the UN High-Level Advisory Body on AI). In some cases, it may be necessary to split up a function across several institutions to serve distinct but related goals. For example, it may be be necessary to split the function of building global scientific consensus on AI risks between a broader UN-led process that involves deep member state engagement, similar to the IPCC, and a separate independent report that focuses on the safety of advanced AI, continuing the process started by the UK-commissioned 'International Scientific Report on the Safety of Advanced AI'.[76]
- **Organize confidence-building measures to ensure necessary international trust to build more extensive governance processes.** Confidence-building measures (CBMs) are procedures and measures that aim to reduce uncertainties and likelihoods of escalation resulting from ambiguities and opaqueness, and may also help to defuse crisis situations. They serve three key functions: promoting transparency; providing avenues for communication and coordination; and offering pathways for cooperation, collaboration, and integration. In the context of AI, CBMs could include crisis hotlines, incident sharing, model cards, content provenance, collaborative red-teaming exercises, table-top exercises, as well as dataset and evaluation sharing.[77]
- **Coordinate on key emerging safety standards.** Through bilateral working groups, existing standards-setting bodies or emerging institutions such as the network of safety institutes, states could set standards for key areas such as identifying key risks, and determining what types of model or training run registration information may need to be shared internationally. An example of a bilateral risk mapping exercise is the crosswalk linking Singapore's risk framework to the US's.[78]

---

[75] For a taxonomy of international governance functions, see Matthijs M. Maas and José Jaime Villalobos Ruiz, 'International AI Institutions: A Literature Review of Models, Examples, and Proposals', *SSRN Electronic Journal*, 2023, https://doi.org/10.2139/ssrn.4579773.
[76] Claire Dennis et al., 'The Future of International Scientific Assessments of AI's Risks' (Carnegie Endowment for International Peace, 27 August 2024), https://carnegieendowment.org/research/2024/08/the-future-of-international-scientific-assessments-of-ais-risks?lang=en.
[77] Sarah Shoker et al., 'Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings' (arXiv, 2023), https://doi.org/10.48550/ARXIV.2308.00862.
[78] 'Joint Mapping Exercise between Singapore IMDA and the US NIST', 13 October 2023, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/nist-imda-joint-mapping-exercise.

References

'30th Anniversary of JVE: Snezhinsk Scientists Reach out to Congratulate the US Colleagues', n.d. https://nonproliferation.org/lab-to-lab-joint-verification-experiment.

'A New Initiative for Developing Third-Party Model Evaluations'. Anthropic, 1 July 2024. https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations.

'A Right to Warn about Advanced Artificial Intelligence', 4 June 2024. https://righttowarn.ai/.

'AI Research Resource Funding Opportunity Launches', 24 January 2024. https://www.ukri.org/news/ai-research-resource-funding-opportunity-launches/.

Alaga, Jide, and Jonas Schuett. 'Coordinated Pausing: An Evaluation-Based Coordination Scheme for Frontier AI Developers'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2310.00374.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, et al. 'Frontier AI Regulation: Managing Emerging Risks to Public Safety'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2307.03718.

Anderson-Samways, Bill, Shaun Ee, Joe O'Brien, Marie Buhl, and Zoe Williams. 'Responsible Scaling: Comparing Government Guidance and Company Policy'. Institute for AI Policy and Strategy, 11 March 2024. https://www.iaps.ai/research/responsible-scaling.

Baseni, Mikta, Marius Hobbhahn, David Lindner, Alex Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, et al. 'Towards Evaluations-Based Safety Cases for AI Scheming'. Apollo Research, 1 November 2024. https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/6724bd285d993323e03b89d6/1730460969317/Toward_evaluations_based_safety_cases_for_AI_scheming.pdf.

Buhl, Marie Davidsen, Gaurav Sett, Leonie Koessler, Jonas Schuett, and Markus Anderljung. 'Safety Cases for Frontier AI'. arXiv, 28 October 2024. http://arxiv.org/abs/2410.21572.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims'. arXiv, 20 April 2020. http://arxiv.org/abs/2004.07213.

Chan, Alan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, et al. 'Visibility into AI Agents'. arXiv, 17 May 2024. http://arxiv.org/abs/2401.13138.

Chin, Moya. 'What Are Global Public Goods?', December 2021. https://www.imf.org/en/Publications/fandd/issues/2021/12/Global-Public-Goods-Chin-basics.

Clark, Sharon A. 'The Impact of the Hippocratic Oath in 2018: The Conflict of the Ideal of the Physician, the Knowledgeable Humanitarian, Versus the Corporate Medical Allegiance to Financial Models Contributes to Burnout'. *Cureus* 10, no. 7 (30 July 2018): e3076. https://doi.org/10.7759/cureus.3076.

Clifford, Matt. 'Introducing Def/Acc at EF', 20 May 2024. https://www.joinef.com/posts/introducing-def-acc-at-ef/.

'ClimateWorks Foundation', n.d. https://www.climateworks.org/.

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

'Considerations and Limitations for AI Hardware-Enabled Mechanisms', 10 March 2024. https://blog.heim.xyz/considerations-and-limitations-for-ai-hardware-enabled-mechanisms/.

Cory, Nigel. 'The U.S.-China Tech Conflict Fractures Global Technical Standards: The Example of Server and Datacenter Energy Efficiency', 22 August 2023. https://itif.org/publications/2023/08/22/the-us-china-tech-conflict-fractures-global-technical-standards-the-example-of-server-and-datacenter-energy-efficiency/.

Crichton, Kyle, Jessica Ji, Kyle Miller, and John Bansemer. 'Securing Critical Infrastructure in the Age of AI'. Center for Security and Emerging Technology, 15 October 2024. https://cset.georgetown.edu/publication/securing-critical-infrastructure-in-the-age-of-ai/.

Dalrymple, David 'davidad', Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, et al. 'Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2405.06624.

'Defining AI Incidents and Related Terms'. OECD Artificial Intelligence Papers. Vol. 16. OECD Artificial Intelligence Papers, 6 May 2024. https://doi.org/10.1787/d1a8d965-en.

Dennis, Claire, Hadrien Pouget, Robert Trager, Jon Bateman, Renan Araujo, Belinda Cleeland, Malou Estier, et al. 'The Future of International Scientific Assessments of AI's Risks'. Carnegie Endowment for International Peace, 27 August 2024. https://carnegieendowment.org/research/2024/08/the-future-of-international-scientific-assessments-of-ais-risks?lang=en.

Dennis et al., 'What Should be Internationalised in AI Governance', November 2024, http://oxfordmartin.ox.ac.uk/publications/what-should-be-internationalised-in-ai-governance

Ding, Jeffrey. 'Keep Your Enemies Safer: Technical Cooperation and Transferring Nuclear Safety and Security Technologies'. *European Journal of International Relations*, 27 April 2024, 13540661241246622. https://doi.org/10.1177/13540661241246622.

Dixon, Ren Bin Lee, and Heather Frase. 'An Argument for Hybrid AI Incident Reporting'. Center for Security and Emerging Technology, March 2024. https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/.

EU AI Act, § 55 (2024).

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, § 4 (2023).

'Global Governance: Goals and Lessons for AI'. Microsoft, n.d. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lhQ0.

Guest, Oliver, and Zoe Williams. 'Topics for Track IIs: What Can Be Discussed in Dialogues about Advanced AI Risks without Leaking Sensitive Information?' Institute for AI Policy and Strategy, 2 May 2024. https://www.iaps.ai/research/dialogue-topics.

Grosse, Roger. 'Three Sketches of ASL-4 Safety Case Components', 5 November 2024. https://alignment.anthropic.com/2024/safety-cases/.

Hadfield, Gillian K., and Jack Clark. 'Regulatory Markets: The Future of AI Governance'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2304.04914.

Heim, Lennart, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A Osborne, and Noa Zilberman. 'Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2403.08501.

Hooker, Sara. 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.05694.

'Human Genome Project', n.d. https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project.

'Influence and Cyber Operations: An Update'. OpenAI, 9 October 2024. https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf.

'Inspect: An Open-Source Framework for Large Language Model Evaluations', n.d. https://inspect.ai-safety-institute.org.uk/.

'Joint Mapping Exercise between Singapore IMDA and the US NIST', 13 October 2023. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/nist-imda-joint-mapping-exercise.

Kim, Jerome H, Peter Hotez, Carolina Batista, Onder Ergonul, J Peter Figueroa, Sarah Gilbert, Mayda Gursel, et al. 'Operation Warp Speed: Implications for Global Vaccine Security'. *The Lancet Global Health* 9, no. 7 (July 2021): e1017–21. https://doi.org/10.1016/S2214-109X(21)00140-6.

Koessler, Leonie, and Jonas Schuett. 'Risk Assessment at AGI Companies: A Review of Popular Risk Assessment Techniques from Other Safety-Critical Industries'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2307.08823.

Koessler, Leonie, Jonas Schuett, and Markus Anderljung. 'Risk Thresholds for Frontier AI'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2406.14713.

Kolt, Noam, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin Esvelt, Gillian K. Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B. Sandbrink, and Thomas Woodside. 'Responsible Reporting for Frontier AI Development'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2404.02675.

'Living with AI and Emerging Technologies: Meeting Ethical Challenges through Professional Standards'. BCS, Chartered Institute for IT, 15 February 2024. https://www.bcs.org/media/jgmfqo2i/living-with-ai-and-emerging-technologies.pdf.

Luskin, Sophie. 'Need for Whistleblower Protections in Artificial Intelligence Industry Discussed in Senate Judiciary Hearing'. *Whistleblower Network News*, 24 September 2024. https://whistleblowersblog.org/corporate-whistleblowers/need-for-whistleblower-protections-in-artificial-intelligence-industry-discussed-in-senate-judiciary-hearing/.

Maas, Matthijs M., and José Jaime Villalobos Ruiz. 'International AI Institutions: A Literature Review of

Models, Examples, and Proposals'. *SSRN Electronic Journal*, 2023. https://doi.org/10.2139/ssrn.4579773.

Mökander, Jakob, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 'Auditing Large Language Models: A Three-Layered Approach'. *AI and Ethics*, 30 May 2023. https://doi.org/10.1007/s43681-023-00289-2.

Mukobi, Gabriel. 'Reasons to Doubt the Impact of AI Risk Evaluations'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2408.02565.

'National Artificial Intelligence Research Resource Pilot', n.d. https://new.nsf.gov/focus-areas/artificial-intelligence/nairr.

'NTU Serving up New Undergrad Course on Meat Alternatives', 25 June 2021. https://www.edb.gov.sg/en/business-insights/insights/ntu-serving-up-new-undergrad-course-on-meat-alternatives.html.

O'Brien, Joe, Shaun Ee, Jam Kraprayoon, Bill Anderson-Samways, Oscar Delaney, and Zoe Williams. 'Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.01420.

Petrie, James, Onni Aarne, Nora Ammann, and David 'davidad' Dalrymple. 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees', 23 August 2024. https://yoshuabengio.org/wp-content/uploads/2024/09/FlexHEG-Interim-Report_2024.pdf.

'Post-Quantum Cryptography', n.d. https://www.nist.gov/pqcrypto.

'Project Moonshot: An LLM Evaluation Toolkit', n.d. https://aiverifyfoundation.sg/project-moonshot/.

'Psychological and Social Risks (Societal Impacts) Workstream Lead (UK AISI)', n.d. https://boards.eu.greenhouse.io/aisi/jobs/4399639101.

'Research Engineer, Societal Impacts (Anthropic)', n.d. https://boards.greenhouse.io/anthropic/jobs/4251453008.

Reuel, Anka, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, et al. 'Open Problems in Technical AI Governance'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.14981.

'Safety Case Specialist (Safety Mitigations) (London)', n.d. https://www.linkedin.com/jobs/view/safety-case-specialist-safety-mitigations-london-at-anthropic-3956180495/?originalSubdomain=uk.

Scher & Thiergart, 'Mechanisms to Verify International Agreements About AI Development', November 2024, https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development

Shoker, Sarah, Andrew Reddie, Sarah Barrington, Ruby Booth, Miles Brundage, Husanjot Chahal, Michael Depp, et al. 'Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2308.00862.

Stein, Merlin, Jamie Bernardi, and Connor Dunlop. 'The Role of Governments in Increasing Interconnected Post-Deployment Monitoring of AI'. arXiv, 7 October 2024. http://arxiv.org/abs/2410.04931.

Stein, Merlin, and Connor Dunlop. 'Safe beyond Sale: Post-Deployment Monitoring of AI'. Ada Lovelace Institute, 28 June 2024. https://www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/.

Stein, Merlin, Milan Gandhi, Theresa Kriecherbauer, Amin Oueslati, and Robert Trager. 'Public vs Private Bodies: Who Should Run Advanced AI Evaluations and Audits? A Three-Step Logic Based on Case Studies of High-Risk Industries'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.20847.

'Tech Secretary Unveils £8.5 Million Research Funding Set to Break New Grounds in AI Safety Testing'. UK Government, 22 May 2024. https://www.gov.uk/government/news/tech-secretary-unveils-85-million-research-funding-set-to-break-new-grounds-in-ai-safety-testing.

'The Audacious Project', n.d. https://www.audaciousproject.org/faq.

Smialek, Jeanna. 'How Jackson Hole Became an Economic Obsession'. *The New York Times*, 25 August 2023. https://www.nytimes.com/2023/08/24/business/economy/jackson-hole-economic-conference.html.

'Third-Party Testing as a Key Ingredient of AI Policy'. Anthropic, 25 March 2024. https://www.anthropic.com/news/third-party-testing.

Trager, Robert, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, et al. 'International Governance of Civilian AI: A Jurisdictional Certification Approach'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2308.15514.

Turri, Violet, and Rachel Dzombak. 'Why We Need to Know More: Exploring the State of AI Incident Documentation Practices'. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 576–83. Canada: ACM, 2023. https://doi.org/10.1145/3600211.3604700.

Wasil, Akash, Tom Reed, Jack Miller, and Peter Barnett. 'Verification Methods for International AI Agreements', 2024. https://doi.org/10.2139/ssrn.4938419.

Wasil, Akash, Everett Smith, Corin Katzke, and Justin Bullock. 'AI Emergency Preparedness: Examining the Federal Government's Ability to Detect and Respond to AI-Related National Security Threats'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.17347.

'White House Voluntary AI Commitments'. White House, September 2023. https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

'XPrize', n.d. https://www.xprize.org/.

Young, Shalanda. 'Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence'. Executive Office of the President, Office of Management and Budget, 1 November 2023. https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf.

Zou, Andy, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, et al. 'Representation Engineering: A Top-Down Approach to AI Transparency'. arXiv, 10 October 2023. http://arxiv.org/abs/2310.01405.

ZHOU Hui, ZHU Yue, ZHU Lingfeng, SU Yu, YAO Zhiwei, WANG Jun, CHEN Tianhao, et al. 'The Model

Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version', 16 April 2024. https://doi.org/10.5281/ZENODO.10974162.

'大模型安全与伦理报告 / Tencent Large Model Safety-Security and Governance Report'. Tencent Research Institute, 24 January 2024. https://mp.weixin.qq.com/s/KCWw9gBwUnzywyNW_K8-4A.

王迎春, 贾 开, 陈 玲, 赵 静, 秦川申, 袁 媛, 傅宏宇, and 梁兴洲. '人工智能安全作为 全球公共产品 研究报告 / AI Safety as Global Public Goods Working Report', 5 July 2024. https://tinyurl.com/AIGPGWP