RESEARCH MEMO
JULY 2024

AI GOVERNANCE INITIATIVE | OXFORD MARTIN SCHOOL | UNIVERSITY OF OXFORD

AISIs' Roles in Domestic and International Governance

Authors: Marta Ziosi, Claire Dennis, Robert Trager, Simeon Campos, Ben Bucknall, Charles Martinet, Adam L. Smith, Merlin Stein

AISIs' Roles in Domestic and International Governance

Marta Ziosi,[1] Claire Dennis,[1,2] Robert Trager,[1,2] Ben Bucknall,[1,2] Simeon Campos,[3] Charles Martinet,[1] Adam L. Smith,[4] Merlin Stein[5]

[1] Oxford Martin AI Governance Initiative (AIGI), [2] Centre for the Governance of AI (GovAI), [3] SaferAI, [4]Independent Standards Researcher, [5] University of Oxford

Executive Summary

The AI Safety Institutes (AISIs) are poised to assume an increasingly significant role in the governance of advanced AI. The Oxford Martin AI Governance Initiative recently held an expert workshop to explore the roles AISIs can play both domestically and internationally, with attendees suggesting a variety of functions for these organisations:

1. **Conducting Evaluations:** AISIs should develop and conduct rigorous safety evaluations and coordinate efforts internationally to track and forecast risks, addressing challenges in mutual recognition of results, and the risk of evaluation gaming.
2. **Standards Development:** AISIs can enhance AI safety standards by promoting their internationalisation, building consensus, bridging gaps from principles to practice, and strategically leveraging traditional standard development organisations (SDOs).
3. **Regulatory vs. Technical Function:** AISIs need to maintain independence from regulatory bodies to foster cooperation with non-government stakeholders while supporting regulatory capacity-building and providing an apolitical, scientific basis for AI governance.
4. **Classified Material and Information Sharing:** AISIs can act as intermediaries between defence sectors and AI safety efforts, developing dual-track evaluations and best practices for information sharing, especially regarding classified material.
5. **International Coordination Among AISIs:** International coordination is essential, with AISIs collaborating on joint testing, mutual recognition of certifications, aligning mandates, and enhancing horizontal and vertical coordination to address global AI safety priorities.
6. **Regional AISIs:** Regional AISIs can provide a single point of contact for a region, focus on capacity-building, communicate local risks globally, prioritise region-specific concerns, and potentially adopt work from other AISIs to support countries with less technical capabilities.
7. **State of the Science Report:** AISIs should lead or contribute to periodic "state of the science reports," enhancing their global legitimacy and coordination with the UN or other institutions to build international scientific consensus on AI risks.

We are grateful to Joslyn Barnhart,[6] Janet Egan,[7] Owen Larter,[8] and Var Shankar[9] for helpful comments.

[6]Google DeepMind, [7]Harvard Kennedy School, [8]Microsoft, [9]Enzai

Introduction

The AI Safety Institutes (AISIs) - and other institutions that may play similar roles, such as the Chinese AI Safety Network and Unit A3 of the EU AI Office - are increasingly important actors in the governance of advanced AI. The Oxford Martin AI Governance Initiative (AIGI) recently convened a small expert workshop – held under the Chatham House rule – to explore the role that AISIs could play both domestically and internationally. Attendees suggested a wide variety of functions AISIs can fulfil, including roles as evaluators of models, developers of standards, and coordinators of third parties, and noted that a network of AISIs might coordinate these efforts internationally. This memo summarises the key takeaways, questions, and tradeoffs which emerged from the workshop about AISIs' structure and functions. These include: conducting evaluations, participation in standards development, coordination with regulatory bodies, classified material and information-sharing, ability to coordinate internationally, the potential formation of regional AISIs, and AISI contributions to the international scientific report on AI safety.

I. Conducting Evaluations

Participants noted that a major contribution of the AISIs could be on frontier technical research for model evaluations, enabled by access to leading models. The UK AISI has significant technical capacity to develop evaluations and standards, and has been doing so. Other AISIs, among which Japan, US, Singapore and Canada, might begin to follow suit. There is a question, however, of whose or which AI systems each AISI should evaluate, and whether an evaluation from one single AISI can be valid and recognised by others. Additionally, the policy and regulatory significance of AISI test results is not clear, nor how evaluation results should inform or trigger risk management decisions, such as whether training or deployment of a new model should be allowed to proceed or whether additional safeguards are required. Participants recommended AISIs:

- **Develop and conduct safety evaluations.** Coordinate efforts to develop robust methods for verifying and validating the safety and reliability of AI systems, including automated capability assessments, red-teaming, human uplift testing, AI agents evaluations, and certification processes. A differentiation could be made between capability evaluations for advanced AI models and safety assurance evaluations for specific contexts. Examples could include 1) conducting model testing to confirm claimed capabilities, 2) red teaming to stress test applications, and 3) field testing to investigate how people engage with AI in regular use.
- **Establish clear policies on which AI systems AISIs should evaluate and mutual recognition of evaluation results.** Given that the quality of evaluations depends on levels of access and transparency, and that frontier developers heavily restrict in-depth access to their models, participants noted that AISIs may be limited to evaluating models within their jurisdictions. This may require a system for mutual recognition of evaluations results between AISIs internationally (more in Section V). There are also limitations on information sharing between AISIs for certain evaluations related to

national security, which require access to classified information (more in section IV).

- **Establish "rules of engagement" for evaluations (i.e. modalities, terms and conditions of access).** Participants noted that there is an important risk of "evaluation gaming" if frontier model providers have access to the details and methodology of the evaluation being performed and can optimise aggressively against it. This could result in models being evaluated as safe in the testing environment, without this being a good benchmark of their safety in the real world. In order to avoid this, attendees noted it seems important to find ways to legally protect evaluation methods from being used (not only for training or finetuning but also simply for screening) by the AI providers. Additionally, it was noted that AISIs should conduct red teaming, using methods that are not specified in advance, to minimise the risk of gaming and responding to unique aspects of systems.
- **Advance the "science of evaluations."** While current technical evaluation and metrology practices[1] are still nascent and insufficiently effective, AISIs could contribute to a "science of evaluations" in order to advance methods to assess advanced AI systems' safety-relevant properties such as potential for misuse, societal harms, safeguards, system security, controllability, and robustness.
- **Track and forecast AI capabilities and use.** AISIs are well positioned to track and forecast AI capabilities and usage at an ecosystem level, which participants noted seems equally as important as tracking individual AI system capabilities. This could be especially true for systemic and emergent sources of risk, which may not otherwise have natural parties to track or take ownership of them.
- **Expand beyond pre-deployment evaluations.** While this memo emphasises AISIs' role in developing model evaluations, AISIs should also develop mechanisms for post-deployment monitoring and incident reporting. This could include rapid response protocols, risk thresholds and frameworks, incident monitoring systems, and auditing or other forms of external oversight.

[1] Practices regarding the science of measurement.

II. Standards Development

Traditional standard-setting processes may not be suited to developing AI safety standards for two reasons: (1) they distill known fields of information into standards rather than investigate the frontier of knowledge, (2) they largely do not draw upon information from national security communities. Even when they do incorporate such information, the rapid progression and high stakes of advanced AI systems will likely demand a more agile approach to integrating cutting-edge insights into standards. Concurrently, alternative standard bodies and processes, although potentially faster, may lack the legitimacy, the recognition and the knowledge about standards setting processes. Participants noted several ways in which AISIs can contribute and improve standards development:

- **Internationalise AI safety standards.** AISIs should identify or develop robust standard practices on AI safety and encourage their harmonisation across jurisdictions. An approach consistent with common international standards practice is for individual AISI personnel to engage with - or even become members of - their local standards organisations and through them with international standards bodies. Increased access to standardisation processes through regional AISIs (section VI) or ad-hoc procedures (e.g., regional experts) should be considered to ensure participation from all jurisdictions, with the aim to ensure legitimacy through a comprehensive multi-stakeholder process.
- **Create international standards consensus.** Consensus on standards is facilitated by consensus on risks. AISIs can build international scientific consensus around AI risks by publishing periodic "state of the science reports," as the UK AISI is doing through its International Scientific Report on Advanced AI Safety (see Section VII). Additionally, AI Safety summits could be used as a platform for periodic reports on best practices for evaluations and industry commitments towards standardisation. Such outputs could feed into wider UN processes to enhance global credibility and drive international standardisation efforts.
- **Bridge the gap between principles and standard practices.** AISIs could provide technical reports, guidance or support on how to proceed from abstract "principles" (e.g. voluntary commitments) to standards "practice" (e.g., technical specifications on AI safety) for individual actors. They could, for example, create crosswalks[2] between voluntary commitments related to AI safety and existing standards on risk management or product safety.[3]
- **Strategically leverage traditional standards development organisations (SDOs) structures.** AISIs could leverage traditional SDOs paths and structures to fast-track standard development by, for example, utilising the fast-track option for ISO standards. This could be done for urgent topics on which consensus can be more

[2] Crosswalks map the provisions of laws, regulations, standards, and frameworks to subcategories helping organisations prioritise activities or outcomes to facilitate conformance. A set of sample crosswalks between the NIST Risk Management Framework and a set of standards and regulations can be found here: https://airc.nist.gov/AI_RMF_Knowledge_Base/Crosswalks

[3] Other organisations which can participate in these processes include the OECD, which was recently tasked by the G7 with creating a voluntary reporting regime for AI developers in line with the Hiroshima Process Code of Conduct.

easily achieved (e.g., terminology standards for advanced AI). They could also turn technical specifications into international standards by becoming a Publicly Available Specification submitter under ISO. This option could be used, for example, once AISIs have developed best practices for safety evaluations. Additionally, AISIs could provide specialised expertise through the inclusion of expert bodies in AI as liaison members in ISO/IEC committees. In particular, they could provide a "dose of technical clarity" when certain standards proposals ask for desiderata that are unachievable using current methods.

III. Regulatory vs. Technical Function

To maintain a cooperative relationship with non-government stakeholders, including private industry, workshop participants noted that it is crucial for AISIs to maintain independence from regulatory bodies. However, if AISIs' expertise and judgment influence regulations, these stakeholders might still view AISIs as an enforcement agency, potentially undermining cooperative efforts. AISIs could also be immensely beneficial to other government agencies, pooling technical expertise and building regulatory capacity on AI that otherwise would not exist. Additionally, assuming a regulatory function could allow AISIs to compel access to models more easily. It is thus important to strike a sensible balance between collaboration and coordination between AISIs and regulatory bodies. This balance includes considering whether AISIs should maintain their current status and position within the civil service, or explore alternative structures (e.g., evolving towards a more independent, yet state-backed model similar to the UK Space Agency). Attendees recommended AISIs:

- **Define a clear separation of roles between AISIs and institutional/regulatory bodies**. AISIs could promote their role as a "neutral" or trusted intermediary by, for example, strategically building infrastructure for a third-party ecosystem of evaluators or of accreditation of third-party auditors. This could be done by including structured research access or funding pools based on centrally collected fees. Outsourcing certification to third parties could enhance AISIs' perceived neutrality, but would reduce their control over the process and outcomes.
- **Support capacity-building for and with other relevant agencies.** Provide support and help upskill sector-specific regulators on technical issues. In the US context, for example, agencies such as FEMA/CISA will struggle to have the ability to deal with rogue agents, unless being taught and resourced at least in part by the pool of experts from AISIs. At the same time, these external agencies will possess relevant know-how useful for AISI work; for example, what the current response authorities and capacities are to respond to incidents, and how societal resilience might change over time.
- **Provide an apolitical, scientific basis for governance.** As part of the civil service of many countries, AISIs may be poised to be scientific institutions which are less responsive to political changes. To play this role effectively, AISIs should engage in bi/nonpartisan public outreach. Broad popular support for the institutions will help them weather electoral cycles. Additionally, while regulatory powers could help AISIs compel access to AI models, non-regulatory approaches also exist to secure access

while maintaining AISIs' neutrality. For instance, the US mandated reporting above a compute threshold through executive action relying on pre-existing law.

IV. Classified Material and Information Sharing:

Some information required for evaluations is classified. Currently, AISIs coordinate with teams in defence establishments to perform evaluations, but AISIs may not themselves possess all the information required to administer evaluations in safety-critical or classified domains. Thus, all elements of model evaluations may not be shareable internationally. AISIs could coordinate to develop best practices for sharing information, including provisions in MoUs for information exchange. Participants agreed AISIs could:

- **Serve as a trusted intermediary between defence sectors and other AI safety efforts.** This could be done by developing evaluation methods in coordination with defence personnel, testing them against AI models, and sharing outputs for comparison against classified data.
- **Develop a dual-track for evaluations.** This could entail (1) standardisation of a large number of evaluations, some of which are performed by a set of mutually recognised AISIs (when evaluations deal with critical safety issues), and some of which are performed by AISI-certified private evals firms; (2) a set of evaluations that rely on classified material, which can only be performed by an AISI with access to that material, and are required for access to markets that coordinate with that AISI. In other industries, the decision to outsource evaluations to certified third-party auditors or keep them in-house depends on factors such as the feasibility of standardising the evaluations, market concentration and the criticality of safety concerns[4].
- **Report information internationally.** AISIs can collect information domestically from AI developers and compute providers, and report some information (potentially metadata) internationally to an international body or through bilateral networks. As in areas of finance, international reporting prevents actors from avoiding oversight regimes, for instance assisting in detection of attempts to structure or split large-model training across jurisdictions. While the focus of industry commitments to date has been on AI developers, consideration should be given to how to best leverage compute providers.

V. International Coordination Among AISIs:

International coordination among AISIs is important for international standards development. Participants noted that the roles and functions of AISIs should likely be differentiated based on the AI capabilities being developed and used in their respective jurisdictions. They emphasised the importance of AISIs recognising each other's certification decisions to streamline the process for firms to have fewer points of contact and promote interoperability, a core competency of international standards developing organisations (SDOs). Participants noted that the UK AISI, for example, is meant to be participating in the evaluation of

[4] For example, in the automotive industry, some safety-critical evaluations, like crash testing, are typically performed by recognised internal bodies, while less critical evaluations, such as emissions testing, can be outsourced to certified third-party facilities.

US-based AI systems - and has the unusual asset of a formal UK/US AISI MoU, first-mover advantage, and comparatively high level of technical talent density. At the same time, AISIs in other countries may lack technical resources, lack model access, or be limited to evaluating AI systems developed in their own countries. The network of AISIs should be modulated to account for and take advantage of this differential capacity, while still maintaining baseline structural (e.g., relation to government or other regulatory agencies) and functional (e.g., regulatory vs scientific) similarities to coordinate easily and avoid misaligned objectives. AISIs should:

- **Enhance interagency collaboration and coordination.** Coordinate horizontally across AISIs to gather evidence-based consensus on advanced AI issues and lay the groundwork for standardisation. As mentioned above, liaise vertically with their respective national standard bodies and, through them, with international standards bodies to ensure alignment of standards initiatives.
- **Conduct joint testing for AI safety and sharing of technical methods.** This involves closely collaborating to iteratively develop robust suites of evaluations for AI models, systems, and agents. The US and UK, for example, already plan to conduct joint testing on a publicly accessible model. AISIs should leverage this process to make progress on technical methods to secure model access. AISIs should also share model information and technical tools (such as evaluation repositories, and task standards) with partner AISIs.
- **Mutually recognise certification between AISIs.** Different AISIs will have differential access to models, depending on their geographical location. Additionally, AISIs will differ in their technical capacity. However, developers are likely to prefer a simple evaluations ecosystem. Ideally, one entity would certify a system as safe for all jurisdictions, or at least a variety of evaluations would only need to be performed once. Mutual recognition agreements between AISIs could be put in place for this. One resulting structure could be a network of AISIs where companies get tested by their local AISIs. These AISIs can then share some information in the network internationally to ensure compliance with mutually agreed-upon standards.
- **Align AISIs' mandates in a complementary fashion to address global AI safety priority topics.** AISIs won't all have the same remit; for example, some don't seem to focus on national security at all. Each institute could specialise in a specific area of expertise within a coordinated global network. Specialisation could also allow AISIs to inform and support each other without competing for talent or resources. Such diversification could require broadening the definition of "safety" to collectively address the full range of global AI safety priorities among AISI member states. Mandates could be kept flexible or updated periodically, using the AI safety summits as a platform to report on mandates deliverables and assess room for collaboration. However, specialisation should not come at the cost of excessive divergence that could hinder collaboration and effectiveness between AISIs. All AISIs could agree on certain baseline functions and structures through a shared charter.

VI. Regional AISIs:

For countries lacking the capacity to establish national AISIs, regional AISIs can aggregate expertise and infrastructure, supporting global standards and local enforcement. For example, ASEAN could establish a regional AISI to pool resources and provide a technical delegation for the region. This model has worked well in the Financial Action Task Force (FATF), which includes nine independent regional bodies with their own charters to implement FATF standards in their respective regions, concentrate on issues specific to the area, and provide technical assistance to their member states. Countries in a particular region can be both direct members of the FATF and the regional body.

These regional AISIs could allow countries with less technical capabilities to benefit from pooled local talent, computing power, data, and hardware, while also boosting their voice and inclusion in international governance decisions. For example, there is a need for many states to have input into standardising processes, and since all states likely won't develop an AISI, it implies a need for regional AISIs or similar institutions. However, it may be important for some AISIs to have distinct functions or mandates, given that companies may limit model access to the AISI in which they are based.

- **Provide a single point of contact for a region.** Countries with high trust may be able to coordinate to provide a single point of contact for a region. "High trust" could imply countries with strong, positive foreign diplomatic relations, stable domestic political institutions and solid technical capacity. This could allow foreign governments to trust that evaluations carried out elsewhere are effective and help overcome classified information challenges.
- **Focus on capacity-building mandates.** Regional or national AISIs in lower-capacity contexts could have more targeted mandates focused on capacity-building, technology transfer, and specific areas of research. Such AISIs could form partnerships with universities and industry partners to develop local talent. They could also focus on identifying promising AI technologies developed elsewhere and adapting them to the local context, as well as springboarding local innovations through the AISI network.
- **Communicate local and regional risks globally.** AISIs with less technical capabilities should maintain open communication with AISIs where advanced AI is developed. This would allow them to convey potential and realised risks associated with the local deployment of advanced models in their specific contexts. They can also help assess the impact of these models on less digitised communities by developing contextual socio-technical evaluations.
- **Focus on region-specific concerns**. Regional AISIs should prioritise issues that are of primary importance to countries in their region. For example, some countries may be more concerned about the risks of disinformation and its impact on society, while others may focus on the potential loss of control over advanced AI systems.
- **Adopt work of other AISIs.** One participant noted that an alternative to establishing regional AISIs would be for countries without AISIs to simply recognise and adopt the evaluation results and recommendations of other AISIs, eliminating the need to establish their own institution.

VII. State of the Science Report:

As mentioned above, AISIs can build international scientific consensus around AI risks by publishing periodic "state of the science reports," as the UK AISI has done through its International Scientific Report on Advanced AI Safety. Participants noted that AISIs are well-positioned to produce authoritative science reports, and that the UK AISI has demonstrated its capability to deliver a substantial report in a short time period. However, AISIs currently lack widespread geographical representation, which is critical for the report to serve as a foundation on international AI safety agreements between countries. To improve their global legitimacy to lead this report, participants emphasised that AISIs should improve stakeholder engagement, coordination, and communication of results.

- **Formalise input from AISIs into ongoing updates to the "State of the Science" report, with coordinated check-ins at the Safety Summits.** AISIs already have the ambition to conduct foundational AI safety research, as well as access to models from leading AI developers. This makes them perhaps best-positioned to conduct cutting-edge AI safety research and feed any timely and relevant findings directly into a formal report process led by the UN or other international institutions, or to lead the safety report writing entirely.
- **Participants noted the importance of maintaining the State of the Science report as an independent, scientist-led publication, released every six months or annually at the AI safety summits.** Given the scarcity of technical expertise, one participant noted it is important to avoid competing reports and maintain a single internationally framed report, rather than "national state of the science reports." Pooling the limited existing expertise within AISIs could help reduce duplicative efforts and compound investments already made in these science-leaning institutions.
- **Formalise AISI coordination with the UN or other institutions to lead a broad political process for global adoption of the report.** Given that not all countries have AISIs, AISI research to support international scientific consensus on AI safety should likely be integrated into a formal UN-led process or other institution with global reach and credibility. However, as AI development is primarily driven by the private sector, AISIs with access to cutting-edge models have a crucial role to play in providing insights that cannot be obtained elsewhere. Regardless of where the report is ultimately housed, participants emphasised the critical role of AISIs in supporting frontier scientific consensus.