3 October 2024

BSA COMMENTS ON PROPOSALS PAPER FOR INTRODUCING MANDATORY GUARDRAILS FOR AI IN HIGH-RISK SETTINGS

Submitted Electronically to the Department of Industry, Science and Resources

BSA | The Software Alliance (**BSA**)[1] welcomes the opportunity to submit comments to the Department of Industry, Science and Resources (**DISR**) on its Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings (**Proposals Paper**).[2]

BSA is the leading advocate for the global software industry. BSA members create technology solutions that power other businesses, including cloud storage services, customer relationship management software, human resources management programs, identity management services, security solutions, and collaboration software. Our members are on the leading edge of providing AI-enabled products and services, and tools used by others in the development of AI systems and applications. As a result, they have unique insights into the technology's tremendous potential to spur digital transformation and the policies that can best support the responsible use of AI.

We welcome DISR's efforts to set clear expectations on how to use AI safely and responsibly. Unlocking the full potential of AI will require a dynamic and flexible policy framework that spurs responsible AI innovation while providing clear guidance to organisations on how they can implement robust safeguards.

Summary of BSA's Recommendations

BSA offers the following recommendations, which respond to the specific policy issues highlighted in the Proposals Paper.

Definitions

1. DISR should identify a specific and defined set of uses cases that are considered high-risk. That defined set of use cases should be grounded in the proposed principles, but will give companies specific guidance about when a system is high-risk or not. For example, AI systems may be high-risk when they are specifically developed to make consequential decisions, which determine an individual's eligibility for and result in the provision or denial of housing, employment, credit, education, access to physical places of public accommodation, healthcare, or insurance. A principles-only approach is too vague and does not provide

[1] BSA's members include: Adobe, Alteryx, Altium, Amazon Web Services, Asana, Atlassian, Autodesk, Bentley Systems, Box, Cisco, Cloudflare, CNC/Mastercam, Cohere, Dassault, Databricks, DocuSign, Dropbox, Elastic, ESTECO SpA, EY, Graphisoft, Hubspot, IBM, Informatica, Kyndryl, MathWorks, Microsoft, Nikon, Notion, Okta, OpenAI, Oracle, PagerDuty, Palo Alto Networks, Prokon, Rockwell, Rubrik, Salesforce, SAP, ServiceNow, Shopify Inc., Siemens Industry Software Inc., Splunk, Trend Micro, Trimble Solutions Corporation, TriNet, Twilio, Workday, Zendesk, and Zoom Video Communications, Inc.

[2] Introducing Mandatory Guardrails for AI in High-Risk Settings: Proposals Paper, September 2024, https://storage.googleapis.com/converlens-au-industry/industry/p/prj2f6f02ebfe6a8190c7bdc/page/proposals_paper_for_introducing_mandatory_guardrails_for_ai_in_high_risk_settings.pdf.

300 Beach Road
#30-06 The Concourse
Singapore 199555

P: +65 6292 2072
F: +65 6292 6369
W: bsa.org

Regional Representative Office
UEN: S97RF0005K

sufficient guidance to companies on the scenarios in which new safeguards should be adopted and implemented.

2. DISR should not adopt the definition of General Purpose AI (**GPAI**) model in the Proposals Paper, as it is overly broad and provides little guidance to organisations. DISR should also avoid treating all GPAI as high-risk. As it stands, the proposed definition of GPAI would include GPAI models used for R&D, which does not reflect a risk-based approach.

3. DISR should not apply the mandatory guardrails to all GPAI models. The guardrails appropriate for high-risk uses of AI should be distinct from any guardrails imposed on GPAI models, since they present different types of risks.

Guardrails for Ensuring Testing, Transparency and Accountability of AI

4. BSA supports designing the mandatory guardrails to be interoperable with those of other comparable jurisdictions. BSA also encourages Australia to map the guardrails to existing frameworks of other jurisdictions to guide users on how adopting one framework can be used to meet the criteria of the other,[3] which will facilitate international harmonisation and interoperability.

5. DISR should limit the definition of developer to entities which "design, code or produce" an AI system and the definition of deployer to entities which "use" an AI system.

6. The allocation of responsibilities between developers and deployers in Attachment E of the Proposals Paper is helpful, but BSA is concerned about certain assumptions made regarding the developer-deployer relationship, most notably the assumption that the deployer will and should always maintain an ongoing relationship and share data with the developer post-deployment.

7. **Guardrails 1 and 2:** BSA supports the establishment of accountability processes and risk management programs for high-risk uses of AI, and encourage the mandatory guardrails to recognise the importance of conducting impact assessments for high-risk uses of AI. However, we also recommend that accountability processes and impact assessments remain confidential to preserve incentives for organisations to implement them. BSA is also concerned that the GPAI developers are expected to address risks against all foreseeable use cases by their clients and reiterates the need to define a clear set of high-risk use cases and obligations for developers. Relatedly, BSA published a set of Best Practices for AI Governance,[4] which outlines comprehensive strategies for establishing AI governance programs.

8. **Guardrail 3:** BSA agrees with the importance of having safeguards in place to securely store and manage data, and recommends making this the core of Guardrail 3. In this regard, DISR should also emphasise the importance of cybersecurity measures for securing AI systems. BSA also agrees that training data should be legally obtained but cautions against imposing overly prescriptive requirements. BSA is also concerned that imposing a requirement to disclose data sources could lead to the disclosure of trade secrets or confidential information about the design or use of an AI system.

9. **Guardrail 4:** BSA supports robust internal testing requirements, but advises against imposing requirements for organisations to conduct external testing (i.e., testing by third parties) given

[3] One example of such a crosswalk was published by the US National Institute of Standards and Technology. See Crosswalk of NIST AI Risk Management Framework and IMDA AI Verify Testing Framework, October 2023, https://aiverifyfoundation.sg/downloads/AI_RMF_and_AI_Verify_Crosswalk.pdf

[4] BSA Best Practices for AI Governance, July 2024, https://www.bsa.org/policy-filings/best-practices-for-ai-governance.

concerns about inadvertently disclosing trade secrets or similar sensitive proprietary information, as well as the nascency of existing technical standards for AI testing.

10. **Guardrail 6:** BSA supports the development and use of content authentication and provenance mechanisms to help users identify AI-generated content.

11. **Guardrail 7:** BSA agrees with the importance of ensuring consumers can exercise existing rights when those rights apply to decisions made by AI. BSA also supports limiting this Guardrail to consumer-facing deployers, who are best placed to discharge the obligations therein.

12. **Guardrail 8:** BSA supports encouraging transparency and information sharing across the AI value chain, but any guidelines or requirements on information sharing should reflect that different entities possess distinct knowledge and abilities to share information, and must respect the need to protect confidential information, including trade secrets. BSA also cautions against legislating a general requirement for incident reporting between different organisations, as it may not always promote effective risk management and can be arranged through contract where appropriate.

13. **Guardrails 9 and 10:** BSA reiterates its support for measures that promote accountability, robust testing and transparency while remaining interoperable with other markets, but advise against imposing requirements for organisations to conduct external testing (i.e., testing by third parties).

Regulatory Options to Mandate Guardrails

14. BSA recommends that if Australia adopts mandatory guardrails, it does so through clear legislation which applies across all industry sectors and the Government, ensuring consistency in how individual regulators implement the guardrails. This will resolve concerns regarding regulatory fragmentation and create a clear and certain operating environment for organisations and businesses across all sectors. BSA also supports continued engagement with industry stakeholders and, to the extent that another committee or expert group is formed, BSA encourages DISR to include more representatives from the industry as they can provide more practical and business-oriented viewpoints.

Definitions

Defining AI

Given that AI systems are developed and deployed in an international context, guidelines and standards that apply to AI should operate across different jurisdictions to facilitate and promote further adoption and use of AI technologies. Definitions pertaining to AI should be aligned across jurisdictions, and indeed within the jurisdiction, to ensure that all stakeholders have a common understanding of AI. In this regard, BSA supports the definition of AI used in the Proposals Paper,[5] which is the same definition used by the Organisation for Economic Co-operation and Development (**OECD**).[6]

Defining High-Risk AI

The Proposals Paper suggests two broad categories of high-risk AI: a) high-risk AI based on intended and foreseeable uses; and b) all GPAI models, where the possible risks and applications cannot be

[5] Proposals Paper (2024), p. 8. AI is defined as "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment".

[6] Updates to the OECD's definition of an AI system explained, November 2023, https://oecd.ai/en/wonk/ai-system-definition-update.

foreseen. We raise concerns with both definitions, as well as the Proposals Paper's assumption that the same set of mandatory guardrails would apply to these two very different types of AI.

High-Risk AI based on intended and foreseeable uses

BSA supports a risk-based approach that focuses on use cases that create high risks to individuals. As such, we agree with the intent of the Principles set out in the Proposals Paper, which consider whether an AI system is high-risk based on its application and use, rather than the sector that it is used or deployed in.

It is important that AI legislation clearly define a set of high-risk uses of AI to provide clarity for both companies and individuals about when new guardrails are required. In this regard, we note that the Proposals Paper did not provide a definition of what would constitute a high-risk use of AI, and only set out principles which are to be referred to when assessing whether an AI system is high-risk.

We therefore urge you to identify a specific and defined set of use cases that are considered high-risk. For example, BSA has encouraged policymakers to focus on AI systems specifically developed to make consequential decisions, which determine an individual's eligibility for and result in the provision or denial of housing, employment, credit, education, access to physical places of public accommodation, healthcare, or insurance. In this regard, Table 1 of the Proposals Paper is a good reference point.[7] A principles-only approach leaves significant ambiguity for companies trying to adopt and implement new safeguards. In several places, the Proposals Paper provides few specifics about how to apply these concepts of high-risk. Notably, Principle (e) requires organisations to consider "the risk of adverse impacts to the broader Australian economy, society, environment and rule of law". As currently drafted, the scope of this Principle is too wide and would not be helpful in guiding organisations. Based on the examples provided in the Proposals Paper to explain Principle (e), the risks that Principle (e) seek to address primarily relate to the creation, amplification, and use of harmful or misleading content. This more specific and concrete scope can be better reflected in clearer language in a revised Principle (e).

Recommendation: DISR should identify a specific and defined set of uses cases that are considered high-risk. That defined set of use cases should be grounded in the proposed principles, but will give companies specific guidance about when a system is high-risk or not. For example, AI systems may be high-risk when they are specifically developed to make consequential decisions, which determine an individual's eligibility for and result in the provision or denial of housing, employment, credit, education, access to physical places of public accommodation, healthcare, or insurance. A principles-only approach is too vague and does not provide sufficient guidance to companies on the scenarios in which new safeguards should be adopted and implemented.

High-Risk AI: GPAI Models

The Proposals Paper proposes defining GPAI models as "an AI model that is capable of being used, or capable of being adapted for us, for a variety of purposes, both for direct use as well as for integration in other systems".[8]

BSA has substantial concerns with the proposed definition of GPAI models and the assumption that *all* GPAI models are high-risk. GPAI models are not high-risk by default – whether a GPAI model is high-risk depends on what the GPAI model is used for. The proposed definition of GPAI is extremely broad and would sweep in an enormous range of AI models that are used and applied in routine, low-risk settings by individuals and businesses. This overbroad definition risks regulating a wide variety of AI models regardless of their actual risk profile. Many AI models used for everyday tasks, such as automating email sorting, personalizing user interfaces, or generating routine business analytics, would be categorised as GPAI under this definition, despite posing little to no risk to individuals or society.

[7] Proposals Paper (2024), p. 26.

[8] Proposals Paper (2024), p. 28.

One alternative is to define GPAI Models in line with the EU AI Act's definition of GPAI model and only including a subset of such models in the scope of the proposal. The EU AI Act defines a GPAI model as "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are released on the market".[9] The EU AI Act does not treat all GPAI models meeting this definition as high-risk; rather, it applies additional safeguards to a subset of GPAI models deemed to have "systemic" risk. This approach is more nuanced than the Proposals Paper's and presents a lower risk of inadvertently capturing low-risk AI routinely used for benign purposes. Additionally, the exclusion of AI models used for research, development, or prototyping prior to market release ensures that innovation is not stifled during the early stages of AI development.

Recommendation: DISR should not adopt the definition of GPAI model in the Proposals Paper, as it is overly broad and provides little guidance to organisations. DISR should also avoid treating all GPAI as high-risk. As it stands, the proposed definition of GPAI would include GPAI models used for R&D, which does not reflect a risk-based approach.

The Proposals Paper also proposes to apply mandatory guardrails to all GPAI models. As observed in the Proposals Paper, this approach "assumes that all GPAI are inherently high-risk because of the nature and ability of such models".[10]

BSA strongly disagrees with the application of the mandatory guardrails to all GPAI models, The Proposals Paper assumes that the guardrails appropriate for high-risk uses of AI will also work in the context of GPAI models. However, the guardrails appropriate for high-risk uses of AI (e.g., an AI system specifically developed to determine an individual's eligibility of housing) should be distinct from any guardrails applied to GPAI models, since they present different types of risks.

Relatedly, treating all GPAI as high-risk is fundamentally inconsistent with adopting a risk-based approach to AI policy and governance. As a general principle, the scope of any regulatory obligations should be a function of the degree of risk and the potential scope and severity of harm.

Recommendation: DISR should not apply the mandatory guardrails to all GPAI models. The guardrails appropriate for high-risk uses of AI should be distinct from any guardrails imposed on GPAI models, since they present different types of risks.

Guardrails for Ensuring Testing, Transparency and Accountability of AI

Interoperability and Standards

As policymakers around the world are developing regulatory approaches to AI, the global nature of today's technology ecosystem demands coordinated policy responses to foster innovation. In this regard, BSA supports designing the mandatory guardrails to be "interoperable with those that other comparable jurisdictions have developed and adopted", and "align[ed] with national and international standards", including those set out in the EU AI Act and *ISO/IEC 42001:2023 Artificial Intelligence Management System*.[11]

As one example of the importance of mapping guidance across jurisdictions, BSA notes that the US National Institute of Standards and Technology (**NIST**) and Singapore's Infocomm Media Development Authority (**IMDA**) jointly published a "crosswalk"[12] of NIST's AI Risk Management

[9] EU AI Act, Article 3(56).

[10] Proposals Paper (2024), p. 29.

[11] Proposals Paper (2024), p. 31.

[12] Crosswalk of NIST AI Risk Management Framework and IMDA AI Verify Testing Framework, October 2023, https://aiverifyfoundation.sg/downloads/AI_RMF_and_AI_Verify_Crosswalk.pdf.

Framework[13] and IMDA's AI Verify Testing Framework.[14] This crosswalk serves as a mapping document that guides users on how adopting one framework can be used to meet the criteria of the other. When publishing the crosswalk, NIST and IMDA highlighted that this "is an important step towards harmonisation of international AI governance frameworks to reduce industry's cost to meet multiple requirements". Indeed, as Australia continues to develop guardrails for high-risk AI, we urge Australia to explore arrangements like the NIST-IMDA crosswalk to facilitate international harmonisation and interoperability.

Recommendation: BSA supports designing any mandatory guardrails to be interoperable with those of other comparable jurisdictions. BSA also encourages Australia to map those guardrails to existing frameworks of other jurisdictions to guide users on how adopting one framework can be used to meet the criteria of the other, which will facilitate international harmonisation and interoperability.

Defining Developers and Deployers and Allocating Responsibilities

The Proposals Paper defines both developers[15] and deployers[16] of AI. BSA is encouraged to see that there is a policy intention to reflect the distinct roles and responsibilities of different organisations in the AI supply chain. As the Proposals Paper notes, "AI systems often involve a complex network of actors responsible for different aspects of the system's development and deployment", and as such, "it will be important to clarify the roles and specific obligations" of actors across the AI supply chain.[17].

However, we have the following concerns regarding the proposed definitions:

- Developers are defined as "organisations or individuals who design, build, train, adapt or combine AI models and applications". This definition is overly broad, and we are concerned that entities who "train, adapt or combine" are considered AI developers. This broad definition could cover organisations that are merely adapting or customising AI models for low-risk, specialized applications, including those that merely decide to use more than one model developed by another company in its own existing software application, which may not be positioned to carry out the obligations of developers. Instead, the definition should clearly focus on those engaged in the actual development of AI systems. We suggest defining AI developers as entities which "design, code or produce an AI system".
- Deployers are defined as "any individual or organisation that supplies or uses an AI system to provide a product or service".[18] This definition is also overly broad. The use of the term "supplies" may inadvertently include companies that do not actually use an AI system in practice. The definition of deployer should narrowed, to focus on those who actually use the AI system in delivering a product or service. We strongly recommend limiting the definition of AI developers to entities which "use" AI systems.
- While it is important to recognize the different roles of developers and deployers, the AI ecosystem is complex and often involves multiple companies that may develop an AI model, integrate that model into a particular AI system, and use the AI system for a specific task. There are two general and important touchpoints where policymakers have focused efforts to manage AI risks: AI developers, or organisations that design, code, or produce AI systems, and AI deployers, organisations that use AI systems. These entities can each advance responsible AI, but they do so in very different ways based on their roles. These roles are not

[13] AI Risk Management Framework, January 2023, https://doi.org/10.6028/NIST.AI.100-1.

[14] AI Verify Testing Framework, May 2022, https://file.go.gov.sg/aiverify.pdf.

[15] Proposals Paper (2024), p.32. Developer is defined as "organisations or individuals who design, build, train, adapt, or combine AI models and applications."

[16] Proposals Paper (2024), p. 32. Deployer is defined as "any individual or organisation that supplies or uses an AI system to provide a product or service."

[17] Proposals Paper (2024), p. 31-32.

[18] Proposals Paper (2024), p. 32.

necessarily static to an organisation and may change depending on the context. Each type of company has access to different types of information and can take different actions to mitigate risks associated with the AI system. AI legislation must account for these differences to create effective and workable obligations.

Recommendation: DISR should limit the definition of developer to entities which "design, code or produce" an AI system and the definition of deployer to entities which "use" an AI system.

BSA also notes that Attachment E of the Proposals Paper sets out how each of the proposed mandatory guardrails will impose distinct obligations depending on whether an entity is a developer or deployer. This is a helpful addition to the Proposals Paper and serves to clearly allocate responsibilities, promoting certainty.

However, BSA is concerned about certain assumptions when allocating responsibilities. Notably, under Guardrail 4, developers are obliged to "monitor and refine their AI system once deployed based on feedback and data provided by deployers where appropriate".[19] This assumes that: a) the deployer will maintain an ongoing relationship and share data with the developer post-deployment; and b) maintaining this relationship and sharing data is desirable. However, this is sometimes not the case – deployers often run sensitive or commercially valuable data through an AI system after deployment, and as such sharing the data with the developer might import substantial commercial risk. Furthermore, the data may also be subject to confidentiality requirements or other restrictions. In such cases, sharing data with the developer may not be feasible or desirable due to security, privacy, or commercial considerations.

Recommendation: The allocation of responsibilities between developers and deployers in Attachment E of the Proposals Paper is helpful, but BSA is concerned about certain assumptions made regarding the developer-deployer relationship, most notably the assumption that the deployer will and should always maintain an ongoing relationship and share data with the developer post-deployment.

Proposed Mandatory Guardrails

BSA appreciates the Proposals Paper's work to develop proposed mandatory guardrails that will support the responsible development and deployment of AI. Our comments on specific mandatory guardrails are as follows:

Guardrails 1 and 2:

BSA supports the establishment and implementation of accountability processes and risk management programs for high-risk uses of AI. These guardrails enable organisations to identify the personnel, policies and processes necessary to manage AI risks. Key elements may include clearly assigning roles and responsibilities, establishing formal policies, using evaluation mechanisms, ensuring executive oversight, performing impact assessments for high-risk AI, and having internal independent review mechanisms, such as interdepartmental governance or ethics committees, to evaluate and address AI issues that pose high risks. Organisations can incorporate these practices into a broader corporate risk management program or establish them in a separate AI program.

We also encourage the mandatory guardrails to recognise that a key part of an effective risk management program is conducting impact assessments for high-risk AI. Impact assessments enable organisations that develop or deploy high-risk AI to identify and mitigate risks. By allowing personnel across the organisation to examine the objectives, data preparation, design choices, and testing results, these assessments help refine AI products and services and drive internal changes to an organisation's risk management program. Implementing these changes enables organisations to better address existing concerns and adapt to new risks as they emerge.

[19] Proposals Paper (2024), p. 66.

However, we are concerned that Guardrail 1 refers to "making accountability processes publicly available and accessible to improve public confidence in AI products and services".[20] Rather than requiring public disclosure of these measures, we encourage DISR to recognise that accountability processes and impact assessments should be treated as confidential to preserve the incentives for organisations to implement them through rigorous processes that identify and mitigate a wide range of potential risks. The fact that assessments are being performed for high-risk uses of AI systems promotes trust for external stakeholders because they will know that an organisation is conducting a thorough examination of AI systems; those assessments should also be available to regulators in the course of an investigation, under existing domestic laws.

BSA is also concerned that, as an example of risk mitigation under Guardrail 2, the Proposal Paper states that "GPAI developers should take responsibility for addressing risks against all foreseeable use cases by their clients".[21] This assumes that GPAI developers can identify and address all foreseeable use cases. However, GPAI models are, by definition, capable of very wide range of uses. In this regard, we reiterate the need to define a clear set of high-risk use cases and obligations for developers when their models are put to such uses.

Recommendation: BSA supports the establishment of accountability processes and risk management programs for high-risk uses of AI and encourages the mandatory guardrails to recognise the importance of conducting impact assessments for high-risk uses of AI. However, we also recommend that accountability processes and impact assessments remain confidential to preserve incentives for organisations to implement them. BSA is also concerned that the GPAI developers are expected to address risks against all foreseeable use cases by their clients and reiterates the need to define a clear set of high-risk use cases and obligations for developers. Relatedly, BSA published a set of Best Practices for AI Governance,[22] which outlines comprehensive strategies for establishing AI governance programs.

Guardrail 3:

BSA agrees that "[o]rganisations must ensure that they have appropriate data governance, privacy and cybersecurity measures in place"[23] to manage data quality and provenance. The data-intensive nature of AI underscores the importance of having safeguards in place to securely store and manage data. We recommend making this the core of Guardrail 3.

While Guardrail 3 appears to focus on data quality and provenance, we urge DISR to also recognise and emphasise the importance of cybersecurity measures for securing AI systems. As enterprises increasingly use AI, they should revisit the practices they use to protect the confidentiality, integrity, and availability of their information and information systems from threats and vulnerabilities. In this regard, the NIST AI Risk Management Framework is a useful reference point.[24] Similarly, non-profit organisations like MITRE have developed an AI security resource, the Adversarial Threat Landscape for AI Systems (**ATLAS**) database,[25] which catalogs the tactics and techniques used by adversaries to attack AI systems.

[20] Proposals Paper (2024), p. 35-36.

[21] Proposals Paper (2024), p. 36.

[22] BSA Best Practices for AI Governance, July 2024, https://www.bsa.org/policy-filings/best-practices-for-ai-governance.

[23] Proposals Paper (2024), p. 37.

[24] AI Risk Management Framework, January 2023, https://doi.org/10.6028/NIST.AI.100-1. We further note that in July 2024, NIST released the NIST-AI-600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (https://doi.org/10.6028/NIST.AI.600-1). Developed in part to fulfill the President's Executive Order (EO) on Safe, Secure, and Trustworthy AI, the profile can help organisations identify unique risks posed by generative AI and proposes actions for generative AI risk management that best aligns with their goals and priorities.

[25] MITRE ATLAS Database, https://atlas.mitre.org/.

BSA also agrees that AI training data should be legally obtained.[26] Organisations that develop and train AI systems remain subject to other laws, including laws for training material to be accessed legally, as well as prohibitions against accessing illegal and harmful material. However, we caution against imposing prescriptive requirements that govern training data, given the wide variety of contexts in which AI systems will be developed. Availability of training data, particularly high-quality datasets, is especially important for developing AI in a safe and responsible way – poorly trained AI systems can present significant risks, such as biased outcomes and inaccurate analysis, which misinform rather than value-add. Prescriptive requirements on training data can also create an uneven playing field in AI development - established companies with more financial capabilities can better access large data sets, marginalising new entrants and stifling open-source AI development.

We note that there is also a proposed requirement to disclose data sources.[27] We are concerned that this requirement could lead to the disclosure of trade secrets or confidential or proprietary information about the design or use of an AI system. This would discourage investment in AI. There may be situations where disclosing a summary of the training data sources is appropriate, but a detailed accounting would be impractical and should not be required.

Recommendation: BSA agrees with the importance of having safeguards in place to securely store and manage data, and recommends making this the core of Guardrail 3. In this regard, DISR should also emphasise the importance of cybersecurity measures for securing AI systems. BSA also agrees that training data should be legally obtained but cautions against imposing overly prescriptive requirements. BSA is also concerned that imposing a requirement to disclose data sources could lead to the disclosure of trade secrets or confidential information about the design or use of an AI system.

Guardrail 4:

BSA encourages measures that incentivise safety and security. Robust testing and evaluation of high-risk AI systems for safety, security, accuracy, and fairness is critical and is prioritised in the US NIST AI Risk Management Framework, which BSA supports. However, we advise against imposing requirements for organisations to conduct external testing (i.e., testing by third parties). Internal testing — which can be performed by a team of employees that is independent from the team tasked with developing an AI system — can identify and mitigate risks without creating concerns about sharing trade secrets, information that could jeopardise information or network security, and other proprietary information that will arise in external testing. Furthermore, existing technical standards for AI testing are nascent and should be developed consistently with longstanding voluntary, market-driven, and consensus-based approaches to standards development.

Recommendation: BSA supports robust internal testing requirements, but advises against imposing requirements for organisations to conduct external testing (i.e., testing by third parties) given concerns about inadvertently disclosing trade secrets or similar sensitive proprietary information, as well as the nascency of existing technical standards for AI testing.

Guardrail 6:

BSA supports the development and deployment of reliable content authentication and provenance mechanisms (e.g., watermarking) that can help users identify AI-generated content. We support efforts by the Content Authenticity Initiative (**CAI**) to promote the open Coalition for Content Provenance and Authenticity (**C2PA**) standard for content authenticity and provenance. This standard will help consumers decide what content is trustworthy and promote transparency around the use of AI. In conjunction with watermarking, the CAI approach provides secure, indelible provenance. Embracing open standards like that developed by C2PA facilitates interoperability and enhances the integrity of digital content ecosystems. We also note that what constitutes state of the art in ensuring

[26] Proposals Paper (2024), p. 37.

[27] Proposals Paper (2024), p. 37.

solutions for content provenance will evolve over time and encourage DISR to ensure that any governance framework accommodates such developments.

Recommendation: BSA supports the development and use of content authentication and provenance mechanisms to help users identify AI-generated content.

Guardrail 7:

The Proposals Paper requires deployers,[28] to "establish internal organisational avenues for people negatively affected by AI systems to raise concerns or complaints".[29] As a general principle, BSA agrees that people should be notified when they are interacting with a high-risk AI system, and that it is important to ensure that consumers can exercise existing rights — where these existing consumer rights apply to AI-related decisions. BSA also supports limiting this requirement to consumer-facing deployers, as they are the entities interfacing with end-users, and are thus best placed to discharge such obligations.

Recommendation: BSA agrees with the importance of ensuring consumers can exercise existing rights when those rights apply to decisions made by AI. BSA also supports limiting this Guardrail to consumer-facing deployers, who are best placed to discharge the obligations therein.

Guardrail 8:

BSA agrees that transparency and information sharing across the AI supply chain will "help organisations meet their legal obligations and enable them to effectively identify and mitigate risks".[30] In this regard, BSA has developed a set of Best Practices for Information Sharing Along the General Purpose AI Value Chain,[31] which set out the recommended types of information that different entities along the AI value chain should share with each other, as well as templates for use by these entities to share information. The central principle underlying our Best Practices document is that different entities in the AI value chain possess distinct knowledge and abilities to share information, and that AI policies and corporate practices must take this into consideration when reflecting the varying roles and responsibilities of different entities within the AI ecosystem.

However, any guardrails addressing these issues must retain flexibility given the different roles that different organisations may play in developing or using an AI system, and the different abilities they may have to share information with others. For example, deployers often run sensitive or commercially valuable data through an AI system after deployment, and sharing that data with the developer might import substantial commercial risk. Furthermore, the data may also be subject to confidentiality requirements or other restrictions. In such cases, sharing data with the developer may not be feasible or desirable due to security, privacy, or commercial considerations. There are also situations where it is simply not useful to report an incident, e.g., where the model is an open-source model. To the extent that information sharing and incident reporting are necessary, organisations can (and already do) identify appropriate feedback channels via contractual arrangements.

Further, the suggestion in Guardrail 2, which refers to Guardrail 8, that recommends deployers "give feedback to developers and contribute to the risk management discussion and design," ignores the practical reality of the AI supply chain and the varied roles of the actors, as deployers are not involved and are not positioned to provide input in the initial design of an AI system. Moreover, depending on how the original AI model is distributed, such as third-party models distributed via SaaS providers to their business customers. the developer may not even know the identity of the deployer.

---

[28] Proposals Paper (2024), p.40 and p. 67.

[29] Proposals Paper (2024), p. 67.

[30] Proposals Paper (2024), p. 41.

[31] BSA Best Practices for Information Sharing Along the General Purpose AI Value Chain, September 2024, https://www.bsa.org/policy-filings/best-practices-for-information-sharing-along-the-general-purpose-ai-value-chain

Finally, BSA notes that DISR has not provided a comprehensive list of the types of information that different entities should share with each other, unlike in the EU AI Act.[32]

Recommendation: BSA supports encouraging transparency and information sharing across the AI value chain, but any guidelines or requirements on information sharing should reflect that different entities possess distinct knowledge and abilities to share information, and must respect the need to protect confidential information, including trade secrets. BSA also cautions against legislating a general requirement for incident reporting between different organisations, as it may not always promote effective risk management and can be arranged through contract where appropriate.

Guardrails 9 and 10:

As highlighted previously in our comments on Guardrails 1-2, 3 and 4, BSA supports measures that promote accountability, robust testing and transparency while remaining interoperable with other markets. However, we reiterate our concerns regarding requirements for testing and assessments by third parties. Requiring assessments by third parties imports concerns about sharing trade secrets, information that could jeopardise information or network security, and other proprietary information that will arise in external testing. Furthermore, existing technical standards for AI testing are nascent and should be developed consistently with longstanding voluntary, market-driven, and consensus-based approaches to standards development.

Recommendation: BSA reiterates its support for measures that promote accountability, robust testing and transparency while remaining interoperable with other markets, but advise against imposing requirements for organisations to conduct external testing (i.e., testing by third parties).

Regulatory Options to Mandate Guardrails

BSA appreciates DISR's comprehensive consideration of three separate regulatory options. Our recommendation is that if Australia adopts mandatory guardrails, it should do so through clear legislation which applies across all industry sectors and the Government. This creates a consistent approach to AI governance rather than leading to potentially fragmented implementation across individual sectoral regulators.

In considering how to adopt AI guardrails, a key consideration is the likelihood of regulatory fragmentation. Given the cross-cutting nature of AI, we agree with DISR's analysis that an approach that provides regulators with substantial leeway and powers to adopt and implement guardrails "is likely to exacerbate gaps and inconsistencies within the current regulatory system".[33] We also agree with the observation that regulators may have competing regulatory priorities, or may lack the resources and technical capability of implementing the guardrails appropriately. This is evident from the examples provided in the Proposals Paper of how regulators, such as the eSafety Commissioner, have already included AI-related provisions and obligations in the regulations under their domain.[34] In the long run, such fragmentation will lead to inconsistent compliance obligations for organisations and create uncertainty across sectors, undermining the effectiveness of AI governance.

Finally, BSA recalls that an AI Expert Group was formed to advise DISR on options for mandatory high-risk AI guardrails, which culminated in the Proposal Paper. Regardless of which regulatory approach is taken, BSA encourages DISR to continue engaging with industry stakeholders in similar committees and expert groups. Relatedly, BSA also notes that majority of the appointees to the AI Expert Group advising on the guardrails are from academia. While academic perspectives are invaluable, a more diverse range of appointees, including representatives from the industry, can present more practical and business-oriented viewpoints. As enterprise software companies, BSA

[32] Proposals Paper (2024), p. 41, c.f., EU AI Act, Article 13.

[33] Proposals Paper (2024), p. 47.

[34] Proposals Paper (2024), p. 44.

members have extensive experience in developing, deploying and managing AI systems at scale and are therefore well-placed to offer critical insights into the real-world impacts of AI regulation.

Recommendation: BSA recommends that if Australia adopts mandatory guardrails, it do so through clear legislation which applies across all industry sectors and the Government, ensuring consistency in how individual regulators implement the guardrails. This will resolve concerns regarding regulatory fragmentation and create a clear and certain operating environment for organisations and businesses across all sectors. BSA also supports continued engagement with industry stakeholders and, to the extent that another committee or expert group is formed, BSA encourages DISR to include more representatives from the industry as they can provide more practical and business-oriented viewpoints.

Conclusion

We hope that our comments will assist DISR. We look forward to serving as a resource as you continue to engage in policy discussions on this issue. Please do not hesitate to contact me if you have any questions regarding this submission or if I can be of further assistance.

Sincerely,

Tham Shen Hong
Senior Manager, Policy – APAC