BSA Comments on the Request for Comments on the US AI Safety Institute's Draft Document: Managing Misuse Risk for Dual-Use Foundation Models
September 6, 2024

BSA | The Software Alliance appreciates the opportunity to provide comments on the US AI Safety Institute's (AISI) draft guidance on Managing Misuse Risk for Dual-Use Foundation ("Draft Guidance'). BSA is the leading advocate for the global software industry.[1] BSA members are at the forefront of developing cutting-edge services — including AI — and their products are used by businesses across every sector of the economy.[2] For example, BSA members provide tools including cloud storage and data processing services, customer relationship management software, human resource management programs, identity management services, cybersecurity services, and collaboration software.

BSA members are on the leading edge of providing AI-enabled products and services and have unique insights into the technology's tremendous potential to spur digital transformation and policies that can support the responsible use of AI. BSA's views are informed by our experience working with member companies to develop the BSA Framework to Build Trust in AI,[3] a risk management framework we published three years ago to help companies mitigate the potential for unintended bias in AI systems.[4]

We appreciate the National Institute of Standards and Technology's (NIST) work to identify, evaluate, and address AI risks, including its AI Risk Management Framework and accompanying playbook and operational profiles. AISI, in just a few short months, has also expanded these efforts, including by facilitating implementation of President Biden's AI Executive Order and engaging with stakeholders through its consortium, in which BSA is pleased to participate.

BSA welcomes the opportunity to provide input on the Draft Guidance, which can help companies assess and address AI risks. We appreciate the Draft Guidance's identification of potential risks that may arise in addressing misuse of foundation models, ranging from creation of biological weapons to non-consensual intimate imagery. We note that, in

---

[1] BSA's members include: Adobe, Alteryx, Asana, Atlassian, Autodesk, Bentley Systems, Box, Cisco, CNC/Mastercam, Cohere, Databricks, DocuSign, Dropbox, Elastic, EY, Graphisoft, Hubspot, IBM, Informatica, Kyndryl, MathWorks, Microsoft, Notion, Okta, OpenAI, Oracle, PagerDuty, Palo Alto Networks, Prokon, 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] *See* BSA | The Software Alliance, Artificial Intelligence in Every Sector, *available at* https://www.bsa.org/files/policy-filings/06132022bsaaieverysector.pdf.
[3] *See* BSA | The Software Alliance, Confronting Bias: BSA's Framework to Build Trust in AI, *available at* https://www.bsa.org/reports/confronting-bias-bsas-framework-to-build-trust-in-ai.
[4] BSA has testified before the United States Congress and the European Parliament on the Framework and its approach to mitigating AI-related risks. *See, e.g.,* Testimony of Victoria Espinel, Public Hearing on AI & Bias, Special Committee on Artificial Intelligence in a Digital Age, European Parliament, Nov. 30, 2021, *available at* https://www.europarl.europa.eu/cmsdata/244265/AIDA_Verbatim_30_November_2021_EN.pdf; Testimony of Victoria Espinel, The Need for Transparency in Artificial Intelligence, Before the Senate Committee on Commerce, Science, and Transportation Subcommittee on Consumer Protection, Product Safety, and Data Security, *available at* https://www.bsa.org/files/policy-filings/09122023aitestimonyoral.pdf.

200 Massachusetts Avenue, NW
Suite 310
Washington, DC 20001

P 202-872-5500
W bsa.org

addition to these important issues, stakeholders have also encouraged policymakers to address other issues that could cause immediate real-world harm. We also appreciate that the Draft Guidance identifies challenges associated with managing the risks it highlights, describes objectives for overcoming those challenges, recommends practices to help achieve those objectives, and identifies documentation that can help identify how companies have managed such risks. The Draft Guidance solicits information on, among other things, practical challenges with implementing the objectives, monitoring actions, confidentiality issues associated with sharing documentation, and collaboration among actors in the AI supply chain. Our comments below focus on these issues, as well as some additional issues, related to the Draft Guidance's recommended practices. Specifically, we highlight that:

- Developers should be encouraged to conduct internal testing to assess the threat profile of a foundation model (Practice 1.1);
- AISI should continue to encourage foundation model developers to consider security issues more broadly and identify tools that can help developers address the unique security issues that arise in the large language model (LLM) context (Practice 3.1);
- AI red teaming is an important practice, but stakeholders should encourage appropriate adoption by considering the impact of resource-intensive processes and use for appropriate threat profiles, the reliability of AI red teaming conducted by in-house teams, and the need for standardization of AI red teaming practices (Practice 4.2);
- Post-deployment monitoring may not always be feasible for developers, and neither is requesting that third-party distribution channels monitor misuse and report this information back to developers (Practice 6.1);
- Transparency is an important goal, but some of the information AISI recommends sharing, such as test results and data sources, should be protected from disclosure (Practice 7.1);
- AISI should leverage BSA's recent industry guidance on information sharing along the general purpose AI value chain (Practice 7.1); and
- AI incident reporting raises important questions about how it should be implemented because of the complexity and ubiquity of AI, which affects who will know about incidents; how to define the scope of an AI incident, including causation and overlap with other reporting regimes, such as cybersecurity, privacy, and critical infrastructure; and where incidents should be reported (Practice 7.3).

I. Objective 1: Anticipate Potential Misuse Risk

The Draft Guidance encourages foundation model developers to anticipate potential misuse. To achieve this objective, Practice 1.1 recommends that developers identify threat profiles for the most significant ways bad actors could misuse the model. To implement this practice, the Draft Guidance recommends that developers consider consulting external experts, including potentially granting them model access to conduct "open-ended experimentation" that could uncover potential misuses.

Although understanding model capabilities and establishing threat profiles can help developers understand potential misuse, the Draft Guidance does not recognize the concerns that may arise from providing open-ended third-party access to AI models. Model development is an extremely sensitive process, and developers take important steps to protect both the model itself (including securing model weights) and information associated with the process of developing the model, including design and training data. Opening the

model to third parties may create new vulnerabilities and can put proprietary technology at risk. The Draft Guidance should account for these concerns, if it retains these recommendations. Alternatively, the Draft Guidance could emphasize the importance of internal testing to help identify threats in gap profiles, which can achieve the same objective without presenting these concerns.

Recommendation: Practice 1.1 should recognize that developers can use best practices to assess potential capabilities and misuse internally. To the extent that the Draft Guidance continues to encourage developers to consider consulting external experts, it should also encourage companies to account for risks that such activities may pose.

II. Objective 3: Manage the Risks of Model Theft

The Draft Guidance also focuses on the importance of managing the risks of model theft. Specifically, Objective 3 encourages foundation model developers to take steps to prevent the theft of information and assets that would allow bad actors to recreate the foundation model. To implement this objective, Practice 3.1 advises developers to consider the company's compliance with cybersecurity best practices when assessing the risk of model theft.

BSA strongly supports robust cybersecurity practices that build trust in digital technologies and protect consumers and businesses from harm. Implementation of cybersecurity best practices is critical for all software, including AI. Responsible cybersecurity practices can address vulnerabilities that arise in the AI context, such as prompt injections, data poisoning, or unauthorized code execution, that bad actors can exploit. Companies assessing the risk of misuse should identify circumstances that led to previous model thefts to inform their assessment of misuse risk and their capability of mitigating similar threats. In addition to focusing on model theft, the guidance should encourage foundation model developers to consider their overall security posture, how common vulnerabilities may be exploited in new ways within LLMs, and how traditional remediation strategies can be adapted to this context.[5]

Recommendation: AISI should continue to encourage foundation model developers to consider security issues more broadly and identify tools that can help developers address the unique security issues that arise in the LLM context.

III. Objective 4: Measure the Risk of Misuse

The Draft Guidance also highlights the objective of measuring the risk of misuse. Practice 4.2 recommends that developers implement this objective by conducting red team testing to assess whether bad actors can bypass safeguards or misuse particular capabilities, including by using external experts.

---

[5] In addition to the valuable resources that NIST has provided on cybersecurity, AI risk management, and secure software development for AI, another resource with practical guidance for addressing LLM security issues that some companies have consulted is the Open Web Application Security Project, of OWASP, Top 10 for LLM list, which identifies common vulnerabilities, attack scenarios, and steps to prevent security breaches in the unique LLM context based on guidance from an international group of nearly 500 leading experts from a diverse cross-section of companies, including AI, security, and hardware companies, and academia. *See* OWASP Top 10 for LLM 1.0, available at https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_0.pdf.

Red teaming is a critical cybersecurity tool that can test whether a company's security posture can withstand attacks in the real world and whether additional safeguards are necessary. There are few disadvantages of red teaming in the cybersecurity context, but one drawback is that coverage is not necessarily comprehensive, as red teams targeting a way to penetrate systems to access sensitive data may achieve this goal without triggering all defenses, so it may not shed light on the effectiveness of all safeguards. In addition, red teaming is costly and requires significant resources, creating challenges for smaller companies.

Adapting the practice of red teaming to the AI context could significantly enhance AI testing and the ability to address the risk of misuse of AI models before they are deployed in the real world. However, stakeholders should work together to ensure red teaming is adapted appropriately in the AI context by accounting for three relevant considerations:

- First, the resource challenges in the cybersecurity context apply equally here, and indeed may be heightened, given the more limited number of start-ups and SMEs that develop AI foundation models. In addition, in some cases, even with fewer resource constraints, the AI system's threat profile may not warrant these additional investments in every instance.

- Second, where developers do engage in AI red teaming, they may opt to use knowledgeable in-house teams to perform the testing. In some contexts, such as the creation of biological weapons, developers may also find it useful to consult domain experts, as Practice 4.2 suggests. However, that may not be the case more broadly with respect to other types of risks, which may be appropriately assessed by internal teams.

- Third, while red teaming is an established cybersecurity best practice, the red teaming of AI models is more nascent and, as a result, there is a lack of standardized techniques to evaluate the same AI threats. Further, even where common AI red teaming techniques are used, companies may implement them differently. This variation among AI red team tools and processes makes it difficult for companies to leverage AI red team results to assess the safety of their products in comparison to other systems.

Recommendation: As AI foundation model developers increasingly leverage red teaming to identify vulnerabilities, stakeholders should consider how to account for resource-intensive processes, recognize the reliability of AI red teaming performed by in-house teams, and develop standardized practices that enable objective safety measurement across AI systems.

IV. Objective 6: Collect and respond to information about misuse after deployment

Objective 6 of the Draft Guidance advises foundation model developers to collect information after a model has been deployed to understand misuse risk to adjust deployments and inform future risk management. To aid in this effort, Practice 6.1

recommends that developers monitor distribution channels, including requesting that those distribution channels monitor and share relevant information with the developer.

At the outset, the Draft Guidance briefly acknowledges that multiple companies comprise the AI supply chain, although the intended focus of the guidance is on initial foundation model developers. However, the complexity of that supply chain and distribution models and appropriate role that different actors should play are still relevant to the Draft Guidance, because different actors along the supply chain will have access to different types of information and be positioned to take different actions to mitigate potential risks.

Objective 6, and Practice 6.1 in particular, ignore the differences between different companies along the AI supply chain and, as a result, propose measures that may not be workable in practice. For example, Practice 6.1 recommends that a developer build or procure systems to enable automated detection of misuse — even though the developer's ability to use these tools may be affected by its distribution method and lack of information about how the model is being used.

Generally, foundation model developers may distribute their AI models through several different methods, such as an API, a Software-as-a-Service (SaaS) provider, on-premise, or open source — with varying levels of insight into how the model is being used. As a result, these distribution methods will affect the developer's ability to monitor potential misuse. For example, in a model hosted through an API, foundation model developers may have some visibility into the model's operation, though the context of use may not be apparent. In the SaaS context, the model may still be hosted via an API, but the information the foundation model developer has is more limited, as it typically won't know the identity of the SaaS provider's customers deploying the system and, as a result, is not in a position to know which customers generated the relevant data or the context of use. In addition, the foundation model developer's access may be further limited by contractual information or technical measures, such as obfuscation, that the SaaS provider implements. In the on-premise context, foundation model developers lack any information about its use once the model is delivered, as the customer has complete control of the model customization and physical IT infrastructure it uses to operate the software. In the open source context, the foundation model developer not only lacks information about its subsequent use, but it also doesn't know the identity of the entity that integrated or deployed the model.

In light of these considerations, the Draft Guidance's recommendation to monitor misuse is not practical in a range of scenarios. Nor does requesting that distributors monitor use and share the information with the foundation model developer avoid the issue. For example, SaaS providers that integrate an AI model into their applications often enable customers to further train the model on their own data in their customer instance. Importantly, many SaaS providers do not have access to customer data and, indeed, their contractual obligations, as well as privacy and security policies, prevent them from looking at customer data. As a result, they often have no visibility into how an AI system is being used by their business customers. In addition, there may be several additional layers in the supply chain between the foundation model developer and company distributing the system to the deployer and, therefore, no direct line of communication for sharing information to the original developer on misuse in deployment.

In contrast, deployers using AI systems may be best positioned to monitor misuse. Indeed, they are the entities in the AI supply chain that have the most information about the context in which a system is used and details about how the AI system is working in the real world, including how the results of the AI system were leveraged, such as whether they were used to make important decisions about consumers and other factors that influenced those

decisions. Foundation model developers, by contrast, can build in safeguards during the AI model's initial development that may help limit subsequent misuse and adjust the model after it is placed on the market to improve functionality or address problems it discovers during its own continued testing and analysis. However, to the extent that foundation model developers have visibility into hosted models, the customer and context of use may not be apparent, and privacy considerations may affect how they handle the data to which they have access. Moreover, terms of service may limit their access to and use of this information. As a result, foundation model developers generally are not positioned to know about or track how other companies are using AI systems.

Practice 6.1 appears to recognize at least some of these limitations, because it suggests that these considerations only apply "where possible." We strongly recommend recognizing the wide range of scenarios in which such monitoring will not be possible.

Recommendation: Practice 6.1 should be revised to avoid suggesting that developers can effectively monitor misuse of an AI model after it is deployed by another company. Instead, it should recognize other actors in the ecosystem – AI deployers – may be better placed to monitor misuse of AI systems in deployment.

V. Objective 7: Provide appropriate transparency about misuse risk

A. Information Sharing

1. Practice 7.1 should be revised to avoid implicating the disclosure of proprietary or confidential business information.

Objective 7 in the Draft Guidance advises foundation model developers to provide appropriate transparency about misuse risk to facilitate understanding, accountability, and scientific development related to model misuse. A key component of an effective AI policy framework is ensuring the accountability of AI actors. An essential part of achieving this goal is increasing transparency – not only about companies' AI products and services, but about how they build and use them to ensure that they are trustworthy.

There are different audiences for transparency — customers, regulators, and different members of the public — and the amount, nature, and method of information, as well as the protections associated with it, may vary based on the audience. Generally, in all three circumstances, it is important for companies to share the key capabilities, limitations, and features of their AI systems, but this may be accomplished in different ways. For example, companies may provide model cards or data sheets for use by business customers, regulators, and technology experts, but the wider public may benefit from plain language explanations on public-facing websites. Companies may also supplement these tools with contracts and industry templates to provide additional information that may not be public — for proprietary reasons, relevance, or technical complexity — but is nonetheless important for other companies in the supply chain to identify and manage risks.

Against this backdrop, one key overarching consideration is the importance of keeping trade secrets or other sensitive business information confidential to protect intellectual property, privacy, and security interests and ensure businesses' continued competitiveness. Companies should be able to protect this information in all circumstances. In the regulatory context, some sensitive information may be required to be disclosed to a regulator, for example, in the context of an investigation, but government entities that access information in those scenarios must still protect its confidentiality and withhold it from public disclosure. In the commercial context, companies are entitled to withhold trade secrets. Because public

disclosures do not have a mechanism to protect trade secrets or other confidential business information, the amount and nature of information provided must be more limited. In short, the nature of the content that is shared, as well as the audience, matters.

Practice 7.1 encourages developers to publish regular transparency reports — without addressing the need to protect sensitive information or trade secrets. We strongly recommend revising this practice, to avoid suggesting that public transparency reports include such information. For example, Practice 7.1 suggests that foundation model developers provide public transparency reports that include AI model evaluation processes and results and data sources. However, companies make significant investments in training data and processes to be competitive in the AI marketplace and, as a result, this information is very sensitive. While companies may provide a general overview of training data, including its curation and provenance, they should not be expected to provide more granular details that could not only disclose proprietary information that undermines their business operations but, in the cybersecurity context, could also provide a roadmap to bad actors on how to undo built-in security protections. Testing and evaluation information, if disclosed publicly, could also raise concerns and, in some cases, may be irrelevant where mitigations were made to address issues that arose in earlier stages.

Recommendation: Practice 7.1 should be revised to avoid recommending public disclosure of sensitive proprietary information.

2. The Draft Guidance should not recommend that developers disclose all of the documentation supporting the recommended practices because it could include sensitive proprietary information.

We also note that each objective throughout the Draft Guidance includes suggested documentation for developers to maintain to support implementation of the best practices to enhance transparency. The guidance states generally that developers should share the documentation with the public or relevant parties, but it doesn't specify which set of documents are appropriate to share with each audience. While we support measures that enable companies to demonstrate accountability, we note that the documentation may also include sensitive or confidential information and, as a result, should not be publicly disclosed. In many circumstances, it also may not be appropriate for sharing with third parties. However, we also emphasize that accountability tools still exist that ensure companies are implementing responsible practices and keeping appropriate records, as regulators could request them in connection with a law enforcement investigation.

Recommendation: The Draft Guidance should be revised to clarify that developers do not need to share the underlying documentation supporting implementation of the best practices with the public or third parties.

3. The Draft Guidance should leverage BSA's recently-published best practices on sharing information along the general purpose AI value chain.

BSA has previously highlighted the critical role that information sharing along the AI supply chain has to enhancing accountability across the ecosystem. Notably, BSA recently published best practices that, among other things, describes the information that different AI actors should share with other companies in the AI supply chain and provides a template

that companies can use for this purpose.[6] These best practices, which are attached to this submission, can help downstream AI actors assess and address risks. The BSA best practices apply to actors across various stages of the development and distribution of general purpose AI: developers of general purpose AI models, such as generative AI models, companies that integrate general purpose AI models into AI systems, and companies that make the original general purpose models available to business customers without integrating the model into an AI system. BSA's best practices anticipate that these AI actors will provide the pertinent information about AI models and systems to the next company in the chain, ending with deployers, who will use AI systems for a specific purpose.

As a general matter, information sharing should depend on what is reasonable and appropriate based on the risks and capabilities, the different roles of actors in the AI supply chain, and the relationship between the various actors. Generally, the BSA best practices reflect that general purpose model developers should share information with downstream providers including the release and revision dates for the AI model, modality, information necessary to integrate the model into a system, an overview of training data, and an acceptable use policy, if applicable. The BSA best practices also reflect that downstream providers should share information about the AI system into which a model is integrated, and any relevant changes they made to the original model. Finally, for companies simply making the AI model available to their customers without integrating it into a system, the BSA best practices recognize it is appropriate to share information received from the model developer and the fact that no changes were made.

Recommendation: We encourage AISI to revise the Draft Guidance to leverage BSA's best practices on information sharing in addressing how developers of foundation models can manage the risk of misuse.

B. Incident Reporting

Practice 7.3 recommends that developers report incidents and hazards related to the foundation model to AI incident databases. It advises developers to define the category of misuse events to report and share verified reports to relevant third parties.

Generally, incident reporting in the AI context raises several important questions, as there are an array of issues to address — such as who should report, when to report, and where to report — that are complicated by a range of factors, including both the complexity and ubiquity of AI.

Who should report an AI incident?
Practice 7.3 in the Draft Guidance places responsibility for reporting misuse of foundation models on developers. As discussed above, AI's supply chain is complex, and multiple companies play several different roles in how AI systems are built and used. Companies' ability to identify and address risks vary based on their role and, for foundation model developers, how the product was distributed. In some cases, foundation model developers could receive information from third parties alerting them of problems, including through contractual obligations with companies with whom they have a direct relationship, voluntary mechanisms for receiving information about incidents, or information they are able to

[6] *See* BSA | The Software Alliance, Best Practices for Information Sharing Along the General Purpose AI Value Chain, *available at* https://www.bsa.org/policy-filings/best-practices-for-information-sharing-along-the-general-purpose-ai-value-chain.

observe directly about the model where, for example, it is hosted via an API and access is not otherwise limited through contractual or technical means. In most cases, however, foundation model developers will not be the entities aware of misuse, nor would they be positioned to address misuse where the model has been deployed downstream. Instead, deployers using the system may be better positioned than other companies in the supply chain to identify and report misuse incidents, though even deployers could encounter difficulties in detecting misuse despite monitoring how the system is working in the real world. In light of these considerations, AISI should consider the roles of various AI actors further to develop more tailored recommendations about who may be the appropriate entity to report misuse to ensure the Draft Guidance works in practice.

What Is the scope of an AI incident?
Another significant issue is when to report or, in other words, defining what constitutes the scope of an AI incident. Several factors complicate where to draw the line – difficulty in categorizing misuse, inability to determine causation, and the need to avoid duplicative reporting requirements.

First, the inherent functionality of foundation models intentionally allows them to be used in unpredictable ways, and there is a lack of objective benchmarks about what harms to protect against, making it difficult to identify what should be categorized as misuse.

Second, a host of issues related to causation affect the ability to determine when an AI incident occurs, including with respect to integration and combination of AI models. For example, foundation models are integrated into software applications and, in some cases, combined with other models in the same system – ensemble models – to improve accuracy and reliability of predictions. It may be difficult to discern whether misuse involves manipulation of the AI components of the application and, if so, the specific foundation model that was misused. Further, the widespread adoption of AI in consumer and business products has made AI nearly ubiquitous, which means that an AI incident rarely occurs in isolation, and causation may be difficult to determine because it is intertwined with other issues. For example, when an AI-enabled application inadvertently discloses a consumer's personal information, is it an AI incident or a privacy incident? If a company's AI-enabled cybersecurity defenses are compromised, is it an AI incident or a security incident? If AI-enabled software is helping to operate an electrical grid that loses power, is it an AI incident or a critical infrastructure incident? As discussed more fully below, it also raises the question of whether companies must report incidents under multiple overlapping frameworks.

Third, in addition to the causation issues implicated when AI is used across diverse use cases, there are also policy considerations that weigh in favor of excluding those AI incidents that implicate other policy issues from the scope of AI incidents. The overlapping issues highlighted above implicate other incident reporting frameworks, such as the Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA) and forthcoming implementing rules, the Securities and Exchange Commission's rules requiring publicly-traded companies to disclose material cybersecurity incidents, data breach notification requirements in state and federal sector-specific laws, and foreign laws, such as the EU's GDPR and NIS2 Directive, which cover similar issues. Even where AI is involved, it is imperative that policymakers avoid duplicative reporting requirements, as they impose significant compliance burdens and impact competitiveness, particularly for startups and SMEs and for entities whose business operations are at the intersection of issues covered by different reporting frameworks, such as a global, publicly-traded company that provides AI-enabled cybersecurity services for critical infrastructure.

Practice 7.3 of the Draft Guidance allows developers to define a misuse incident, which enables them to avoid areas that intersect with other reporting frameworks that may govern their business. However, it doesn't address the broader concern of creating a reliable reporting ecosystem with a common baseline for what a covered incident is, which is necessary to assess the extent of misuse of the foundation model and how it compares across AI systems. Nor does it address the significant issue of determining causation – whether the misuse incident involved AI and, if so, which foundation model was misused. We recommend that AISI consider these issues further as it finalizes the Draft Guidance.

Where should companies report AI incidents?
The issue of where to report misuse of foundation models is similarly affected by the distributed structure of the AI supply chain, as companies involved in the development and use of AI systems may not have relationships with, or even know, each other. As a result, there may be challenges in how misuse incidents are communicated either upstream or downstream in the supply chain. Nonetheless, as discussed above, there are discrete, but not comprehensive, efforts to improve information sharing among AI actors, and some companies may have mechanisms in place to receive third-party reports. In many cases, government entities are not ideal recipients of the reports for several reasons. Indeed, it would consume scarce public resources to administer a reporting system apart from existing reporting infrastructures that already exist, require disclosure of sensitive information to be stored in government systems that may lack the same security infrastructure of private companies, and have limited utility, particularly where both "hazards" and "incidents" are reported, as AISI's Draft Guidance suggests, because the reportable events would include threats that did not mature into actual harms and, as a result, do not warrant further corrective action. In other circumstances, public disclosures may also not be a good solution, particularly where a vulnerability could still be exploited to cause additional harm and, as noted above, where there is no practical guidance for a consumer to take action to address an issue, unlike software security incidents where they can download a patch to fix a vulnerability.

Practice 7.3 sidesteps concerns related to specific options for where incidents should be reported, suggesting generally that foundation model developers share them with "relevant third parties, such as AI incident databases." It is unclear what entities are considered relevant third parties, but if it includes other actors involved in the development or use of the foundation model, the supply chain information sharing issues described above would need to be addressed. Practice 7.3 does explicitly refer to AI incident databases. Notably, the OECD has an AI Incidents Monitor aimed at showing global AI risk patterns, which relies on news articles.[7] AISI's guidance refers to "verified reports" of misuse, and it is unclear what this means, or who verifies the reports. But to the extent it pertains to news articles, reporting to services like the OECD AI Incidents Monitor could be crowdsourced, and not necessarily a responsibility of the foundation model developer. We recommend that AISI further consider the appropriate entity that should receive AI incident reports and whether foundation model developers are best suited to share that information with the designated entity.

[7] *See* OECD AI Incidents Monitor, *available at* https://oecd.ai/en/incidents?search_terms=%5B%5D&and_condition=false&from_date=2014-01-01&to_date=2024-08-24&properties_config=%7B%22principles%22:%5B%5D,%22industries%22:%5B%5D,%22harm_types%22:%5B%5D,%22harm_levels%22:%5B%5D,%22harmed_entities%22:%5B%5D%7D&only_threats=false&order_by=date&num_results=20.

Recommendation: Practice 7.3 should acknowledge that there are several issues to resolve before an effective AI incident reporting ecosystem can be established, including: (1) assigning responsibility to the appropriate actors in the AI supply chain; (2) defining the scope of an AI incident in a way that addresses causation issues and avoids duplication with other reporting frameworks; and (3) identifying the appropriate parties who should receive the incident reports.

* * *

Thank you for the opportunity to provide comments on AISI's guidance. We look forward to continuing to participate in the AISI consortium and serving as a resource as you, as well as other organizations within NIST, consider mechanisms for advancing responsible AI.

Respectfully submitted,

Shaundra Watson
Senior Director, Policy
BSA | The Software Alliance

200 Massachusetts Avenue, NW
Suite 310
Washington, DC 20001

P 202-872-5500
W bsa.org

Best Practices for Information Sharing Along the General Purpose AI Value Chain

BSA members are advancing trust and ethics in artificial intelligence (AI) and we support policies that ensure AI systems are developed and used responsibly. To be effective, AI policies and corporate practices must reflect that different entities have different roles in the AI ecosystem, and therefore will have different obligations based on those roles. For example, an AI developer, an AI deployer, and other parties within the value chain will have different information about characteristics, capabilities, and limitations of an AI system and how an AI system was developed or operates. As a result, these different companies will have different abilities to share information with other actors.

BSA's best practices for information sharing along the general purpose AI (GPAI) value chain are intended to support transparency between these different actors.[1]

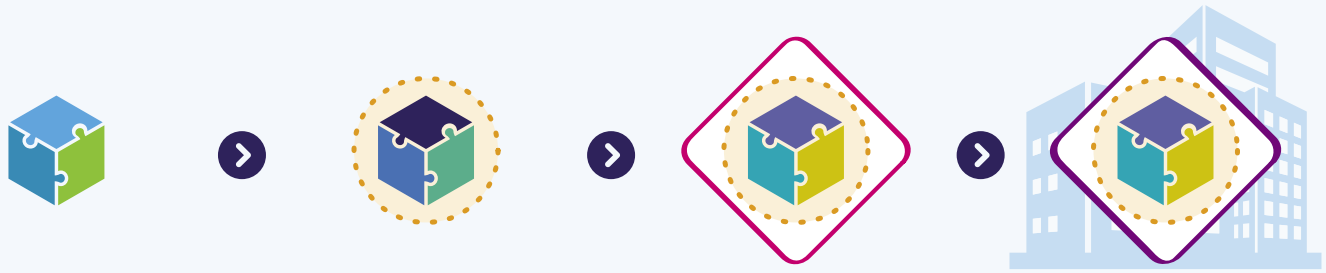
BSA'S BEST PRACTICES:

1. **Identify the entities that should share information.**
2. **Identify the audience for such information.**
3. **Identify methods for sharing that information.**
4. **Set out the types of information to be shared by different actors along the value chain.**

Because different companies have different roles in creating and providing AI systems, the type of information to be shared will vary based on a company's role and the availability of information shared in particular value chains.

[1] General purpose AI (GPAI) models can be used for a wide variety of tasks and may be integrated into a variety of downstream AI systems. As a result, there are a range of different entities along the GPAI value chain. For this reason, BSA's best practices focus on the GPAI value chain and identify the information that different entities should provide to others.

SHARING INFORMATION ALONG THE GENERAL PURPOSE AI VALUE CHAIN



Developer of GPAI Model

Developers of GPAI models should share information about the model with downstream providers.

Downstream Providers

Downstream providers may modify a GPAI model as they integrate it into an AI system. They should share information about the underlying model and their changes to that model with other actors farther downstream the AI value chain.

Deployment

Companies often deploy AI systems that use GPAI models, and have been modified by downstream providers to fit their use case.

1 ENTITIES SHARING INFORMATION SHOULD INCLUDE[2]

- Providers that develop a GPAI model for use by other businesses
- Downstream providers that integrate GPAI models developed by others into an AI system for use by their business customers
- Companies that make available GPAI models to their business customers, without integrating the GPAI model into an AI system[3]

2 AUDIENCE FOR INFORMATION

- Relevant actors situated farther downstream in the GPAI value chain

3 POTENTIAL METHODS FOR SHARING INFORMATION

- Model or system cards
- Public disclosures with standard template
- Standardized technical documentation
- Incorporating links to publicly available information, when available

4 INFORMATION TO BE SHARED

As a general matter, information sharing obligations will depend on what is reasonable and appropriate based on the risks and capabilities of a GPAI model or AI system, the different roles of different actors in the AI value chain, and the working relationship between the various actors, including business customers using AI-related services.

---

[2] A company may be involved in one or more of these activities.

[3] These companies would only provide information about the underlying GPAI model and describe the lack of changes to the model, as the information about a standalone AI system would not apply.

BEST PRACTICES FOR SHARING INFORMATION ALONG THE AI VALUE CHAIN

PROVIDERS OF A GPAI MODEL should share the following information with downstream providers:

- GPAI model name
- Developer name
- Date of release
- Date(s) of revision, if any
- Modality/format (e.g., text, video, image)
- Intended use
- Known limitations
- Model characteristics
- Related software and hardware
- Information reasonably necessary to integrate the GPAI model
- An overview of the training data, including the type and provenance of data and curation methodologies
- Information on data used for testing or validation of the GPAI model, if applicable
- Acceptable use policy, including any prohibited uses
- Contact information to report concerns/issues with the GPAI model

DOWNSTREAM PROVIDERS that integrate a GPAI model into an AI system for use by business customers should share some or all of the following information with relevant actors situated further downstream in the AI value chain:

1. **Information about the underlying GPAI model(s), including:**
   - GPAI model name(s)
   - GPAI developer name(s)
   - Available documentation from GPAI model provider (as set out above)
2. **Information about relevant changes made to the GPAI model by the downstream provider, namely:**
   - Statement summarizing relevant changes by the downstream provider. These may include, as applicable:
     - Changes to the intended use of the GPAI model
     - Changes that alter the model weights (e.g., retraining model, fine tuning model, response learning from human feedback (RLHF)) and a summary of data used to retrain or finetune the model, if applicable
     - Changes that customize the GPAI model without altering model weights (i.e., retrieval augmented generation (RAG), prompt optimization, use of supporting models)
     - Changes to privacy or security controls of GPAI model
     - Information about changes to performance or known limitations if the downstream provider has put the model to use after the relevant changes
     - Estimated cadence of changes
   - Statement of no relevant changes by the downstream provider.[4] These may include:
     - Statement that the provider did not make changes that alter the model weights (e.g., translating the model into a different language without retraining it)
     - Statement that the provider did not customize the GPAI model

[4] These statements may be particularly relevant for companies that provide business customers the ability to access a GPAI model but do not change the model itself.

3. Information about the AI system into which the GPAI model is integrated:

- AI system name
- Downstream provider name
- Date of release
- Date of revision(s), if any, or cadence of system updates
- Modality/format (e.g., text, video, image)
- Intended use
- Known limitations
- AI system characteristics
- Related software and hardware
- Information reasonably necessary to integrate the AI system
- Information regarding testing or evaluation of the AI system, if applicable
- Acceptable use policy, including any prohibited uses
- Contact information to report concerns/issues with the AI system

The level of detail for this information may appropriately vary based on the risks presented by foreseeable uses of the AI system.

In sharing this information, companies should respect confidentiality obligations and share information in line with protections for trade secrets, intellectual property, or other business confidential information.

ADDITIONAL INFORMATION THAT MAY BE APPROPRIATE IN SPECIFIC SCENARIOS.

Further information may be shared by either the GPAI model provider or the downstream provider, as appropriate and applicable, including:

- Instructions for use for the GPAI model or AI system, as applicable
- Information on data sets, including any datasheets for data sets
- Employee training guidance
- Certifications, if applicable
- Related research

HOW SHOULD INFORMATION BE SHARED?

Information may be shared between different entities in the GPAI value chain in different ways, depending on the AI model and AI system at issue. For example:

- Information may be made publicly available, such as publishing information on the date a model is released or significantly revised. In these scenarios, publishing the information in a manner that downstream providers may link to or otherwise pass to other actors is helpful.
- Information that is not public may also be shared between different actors in the course of a business relationship.
- Providers may also develop standardized ways of notifying others in the value chain when they update such information.

Annex 1
TEMPLATE FOR USE BY: GPAI MODEL PROVIDER

The template below can be completed by the provider of a GPAI model to share information along the AI value chain, consistent with BSA's best practices. This template may also be further developed by companies to document their information-sharing practices.

Information About GPAI Model	
GPAI model name	
Developer name	
Date of release	
Date(s) of revision, if any	
Modality/format (e.g., text, video, image)	
Intended use	
Known limitations	
Model characteristics	
Related software and hardware	
Information reasonably necessary to integrate the GPAI model	
Overview of training data, including the type and provenance of data and curation methodologies	
Information on data used for testing or validation of the GPAI model, if applicable	
Acceptable use policy, including any prohibited uses	
Contact information to report concerns/ issues with the GPAI model	
Additional Information—May Be Appropriate in Specific Scenarios	
Instructions for use for GPAI model	
Information on data sets, including any datasheets for data sets relevant to GPAI model	
Employee training guidance	
Certifications, if applicable	
Related research	

Annex 2
TEMPLATE FOR USE BY: **DOWNSTREAM PROVIDERS**

The template below can be completed by a downstream provider integrating a GPAI model into an AI system to share information along the AI value chain, consistent with BSA's best practices. This template may also be further developed by companies to document their information-sharing practices.

Information About Underlying GPAI Model	
Name of GPAI model(s) integrated into AI System	
Developer(s) of GPAI model(s) integrated into AI system	
Further information on GPAI model, as provided by the GPAI model provider	
Information About Relevant Changes to the GPAI Model by Downstream Provider	
Statement of relevant changes by the downstream provider to the GPAI model. These may include: » Changes to the intended use of the GPAI model; » Changes that alter the model weights (e.g., retraining model, fine tuning model, response learning from human feedback (RLHF)) and a summary of any data used to retrain or finetune the model, if applicable; » Changes that customize the GPAI model without altering model weights (e.g., retrieval augmented generation (RAG), prompt optimization); » Changes to privacy or security controls of GPI model; » Information about changes to performance or known limitations if the downstream provider has put the model to use after the relevant changes; and » Estimated cadence of changes.	
Statement of no relevant changes by the downstream provider to the GPAI model. These may include: » Statement that the provider did not make changes that alter the model weights (e.g., translating model into other languages without retraining it), or » Statement that the provider did not customize the GPAI model	

Information About AI System Into Which GPAI Is Integrated	
AI system name	
Downstream provider name	
Date of release	
Date of revision(s), if any, or cadence of system updates	
Modality/format (e.g., text, video, image)	
Intended use	
Known limitations	
AI system characteristics	
Related software and hardware	
Information reasonably necessary to integrate the AI system	
Information regarding testing or evaluation of the AI system, if applicable	
Acceptable use policy, including any prohibited uses	
Contact information to report concerns/issues with the AI system	
Additional Information—May Be Appropriate in Specific Scenarios	
Instructions for use for AI system	
Information on data sets, including any datasheets for data sets relevant to AI system	
Employee training guidance	
Certifications, if applicable	
Related research	

Annex 3
COMPARISON: BSA BEST PRACTICES AND EU AI ACT OBLIGATIONS FOR GPAI PROVIDERS

This table compares BSA's recommended best practices for GPAI model providers with the obligations for such providers to share information under the EU AI Act. Article 53(1)(b) of the AI Act requires providers of GPAI models to give downstream providers that integrate the model into their AI system certain documentation, set out in Annex XII.[5]

BSA BEST PRACTICES Information to Be Shared by GPAI provider	EU AI ACT Annex XII Obligations for GPAI Providers to Share Information With Downstream Providers
GPAI model name	
Developer name	
Date of release	1.c: The dates of release and methods of distribution
Date(s) of revision, if any	1.c: The dates of release and methods of distribution
Modality/format (e.g., text, video, image)	1.g: The modality (e.g., text, image) and formats of inputs and outputs
Intended use	1.a: The tasks that the model is intended to perform and the type and nature of systems into which it can be integrated
Known limitations	1.h: The license for the model
Model characteristics	1.f: The architecture and number of parameters 2. A description of the elements of the model and the process for its development, including: (b) the modality (e.g., text, image, etc.) and format of the inputs and outputs and their maximum size (e.g., context window length, etc.)
Related software and hardware	1.d: How the model interacts, or can be used to interact, with hardware or software that is not part of the model itself, where applicable 1.e: The versions of relevant software related to the use of the GPAI model, where applicable
Information reasonably necessary to integrate the GPAI model	2. A description of the elements of the model and the process for its development, including: (a) technical means (e.g., instructions for use, infrastructure, tools) required for the GPAI model to be integrated into AI systems
Overview of training data, including the type and provenance of data and curation methodologies	2. A description of the elements of the model and the process for its development, including: (c) information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies
Information on data used for testing or validation of the GPAI model, if applicable	2. A description of the elements of the model and the process for its development, including: (c) information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies
Acceptable use policy, including any prohibited uses	1.b: The acceptable use policies applicable
Contact information to report concerns/ issues with the GPAI model	

[5] Outside of the EU, California is considering draft regulations that would apply more broadly than GPAI. The draft regulations would require companies that make or train automated decision-systems to provide information to other downstream actors, including a "plain language explanation of any requirements or limitations that the business identified as relevant to the permitted use." California's draft regulations are expected to undergo public comment in late 2024.