PLI Chronicle

May 2026

Responsible AI in Practice. Beyond Internal Controls: Expanding AI Risk Assessments into Procurement and Exposure Dimensions

Hannah Liu

Responsible AI Institute

Responsible AI in Practice is a series featuring practical, actionable guidance for teams navigating artificial intelligence governance and responsibility, authored by the experts at Responsible AI Institute (RAI).

When we envision an organization's AI risk classification and mitigation protocol, we often imagine it is focused on internal systems, applications, and models. Strategically, this makes sense; enterprises must ensure that locally developed and deployed applications do not introduce extraneous and harmful risks. If AI risk were concentrated solely on this dimension, frameworks focused on mitigating internal AI risk would be exhaustive.

However, this is far from the case. In the real world, enterprises also consider procurement and exposure risk as cornerstones of their risk assessment protocols. For example, they may assess whether a vendor AI product will create more risk than benefit when integrated, or whether an enterprise system may be at risk of exposure to specific AI applications. As a result, when we formulate AI governance frameworks, it is critical that we think not just internally but also externally. Creating guardrails and stronghold policies and controls for internal problems is the first layer of defense, but without external defenses, an organization's security blanket dissolves.

In this article, we will first discuss what procurement and exposure risks entail, before introducing the audience to innovative developments occurring at the Responsible AI Institute that answer the call to action we propose here.

Procurement Risk: Check Before You Buy

In a previous article, we talked about how off-the-shelf models are growing in popularity. Much of this is driven by the increasing prevalence of third-party AI system procurement. Companies are beginning to buy ready-made AI tools instead of custom-building their own, and purchased AI tools from specialized vendors are succeeding more often than internal builds. Despite these observations, the majority of enterprises remain hesitant to adopt third-party AI systems and applications more broadly across the enterprise due to concerns about security and data governance. We noticed this when analyzing how off-the-shelf models harbor unknown risks, where factors such as a lack of visibility into the model architecture or a lack of accountability frameworks embedded in the model can lead to potential catastrophic failures. While it may not be immediately clear how these two observations can coexist, we can see this contradiction as evidence that localized success in AI procurement is not enough to translate into broader enterprise adoption. As a result, it only emphasizes the need for AI procurement risk classification. These safeguards provide enterprises with the confidence to enter AI procurement and realize the observed advantages of using third-party models.

The necessity of AI procurement safeguards becomes especially apparent when we look into how it can compound supply chain risk. The previously identified concerns surrounding the opacity of AI models become a significant issue when they are embedded into a downstream pipeline. Usually, supply chain audits focus on observable components, therefore bounding the attack surface. However, a procured AI system can have varying levels of transparency, making it possible that important details such as training data, architectural choices, fine-tuning decisions, and alignment interventions are never disclosed to the buyer. Since current AI governance frameworks place explicit obligations on deployers of high-risk AI systems, most compliance responsibility appears to fall on the buyer, not the vendor. Organizations already have some tools to ensure a level of transparency from third-party vendors, such as including beneficial contract terms in procurement agreements. However, these methods only scratch the surface; more robust tools are needed to get to the root of these issues.

As the audit responsibility and liability for supply chain risk flow downstream from the vendor to the buyer, transparency does not necessarily follow, creating a complicated governance dilemma: how do you fix something you cannot see?

Exposure Risk: Proactive Risk Classification

Much of AI governance focuses on reactive approaches, answering the question, "What can we do *now* to fix what's happening *now*?" However, as the gap between AI policy and AI reality widens, with innovation outpacing the speed at which we can establish policies, being proactive in our frameworks becomes more critical than ever.

Current statistics also support this approach. In 2024, we saw that 77% of businesses reported an AI-related security incident, with these breaches costing enterprises an average of $4.88 million each, recording the highest cost in history. Exposure risks are also escalating, with AI-related SaaS attacks increasing by 490% year over year. The reality is that governing structures and safeguards need to look ahead by anticipating and preparing for the worst. Instead, AI governance is stuck in the past and present, developing solutions that address previously observed model behaviors and focus on immediate threats. Although these areas are also integral to AI governance, long-term policy plans that recognize the impending consequences of on-the-horizon AI risks are lacking.

While no one can predict the future with full accuracy, creating guardrails broad enough to account for most risk scenarios is certainly more than sufficient. We never know when the next big evolution in AI will occur, nor what it will be. AI exposure risk classification is challenging, but it may be one of the closest things we have towards closing the AI policy-reality gap.

The Issue With Current Risk Classification Frameworks

Existing AI governance and assurance frameworks provide important foundations for risk classification, including the NIST AI Risk Management Framework, ISO/IEC 42001, the EU AI Act, OWASP guidance for agentic AI, MITRE ATLAS, and SR 11-7, which has been superseded by SR 26-2. However, an issue with these frameworks is that they frame AI risk assessments as separate decisions. An evaluation of an internal AI tool may be conducted independently of a procured AI component, but doing so may overlook the possibility that the procured component will soon be integrated into the internal AI tool. An AI system deployed within an organization may pass a risk assessment that only considers internal development risks, but these verdicts may not account for the possibility that new models could penetrate these internal AI systems through previously unaddressed loopholes. Procurement and exposure risk can become intertwined through the concept of shadow AI, where unauthorized employee use can introduce compounding risks that a standalone, one-dimensional risk assessment cannot address.

The question is no longer whether an agentic AI system is accurate, fair, or secure in isolation. Instead, we must begin executing these assessments concurrently so that outputs from each assessment component can interrelate.

The TrustX Risk Framework: The Responsible AI Institute's Solution

We at the Responsible AI Institute address these concerns by proposing the **TrustX Risk Framework**, a unified enterprise risk framework spanning three areas: development, procurement, and exposure. However, rather than falling into the trap of separate risk assessments within pre-existing frameworks, we use a shared methodology across the three components to produce a weighted composite risk score that informs policy recommendations and controls for each risk domain that is of interest to the user.

AI development risk is assessed using the Agent Risk Classification (ARC) framework, which includes an agent classification model, a 12-dimensional risk quantification rubric, and an agent autonomy framework for classifying the risk of an internally built AI system. The underlying algorithm for ARC considers and assigns corresponding weightings to all components of an agentic AI system when making its risk-scoring decision, so that high-risk characteristics, such as high autonomy and significant action authority, are not averaged out by lower-risk components. As a result, it helps the user identify red flags that may have been missed in other general risk assessments.

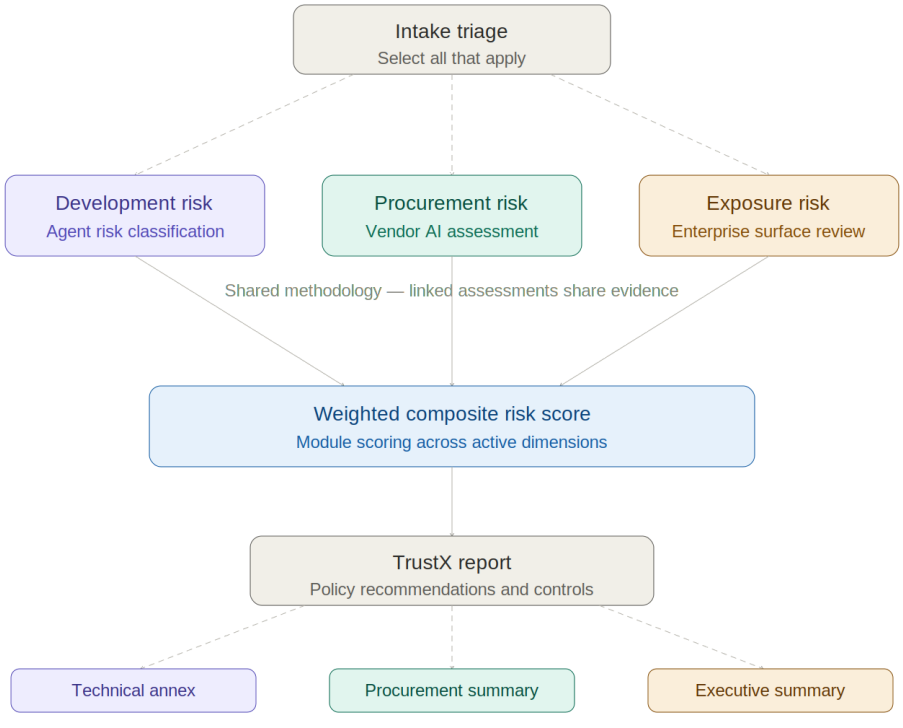
AI procurement risk is assessed using a four-dimensional vendor risk-scoring rubric applicable to eight AI product types, prompting the user to estimate how much autonomy, reach, and memory the procured agent may have in the organization. It contains mandatory gating and red flags that achieve a goal similar to ARC's weighting algorithm, and then translates the scores into a dossier that informs the user which terms must be met when procuring their AI system of interest.

AI exposure risk is assessed using a five-dimensional rubric, paired with exposure pathway selection, agentic threat testing, and exposure surface evaluations. Specifically, the agentic threat testing component helps analyze how an adversarial AI agent can act within the organization's internal systems after gaining entry. As a result, the risk assessment addresses both whether an attacker can enter and what happens after a breach occurs.

Our framework helps solve two of the issues we pinpointed in current risk classification frameworks by doing the following:

1. **It does not assume that AI risk only occurs internally.** As we discussed at the beginning of this article, analyzing only internal risks does not realistically represent enterprise AI risk. In fact, doing so ignores most sources of AI risk. By creating methods and processes that provide a structured framework for assessing two large sources of external AI risk, we formulate a more informed risk assessment for all stakeholders. It also provides a foundation for introducing additional risk areas to our framework.

2. **Executions of each assessment do not occur in silos.** Including as many risk areas as possible in a risk classification framework is not enough for it to adequately represent real-world enterprise dynamics. Outputs from each risk assessment must be weighted to capture the nuances of how risks of all kinds interact, mitigate, or exacerbate one another. By simulating this using our weighted composite score, we aim to situate our framework in the realities of what truly occurs within organizations. The composite scoring is also flexible, allowing an organization to evaluate specific risk domains rather than all three, providing users with opportunities to assess their applications from various angles.



A high-level diagram of the TrustX Risk Framework.

Conclusion

Risk classification is an ever-evolving problem within AI governance. In a month, the Responsible AI Institute has recognized these rapid shifts and has been compelled to adapt, transforming our previously introduced ARC framework into the expanded, robust TrustX Risk Framework. By answering our call to action and drawing awareness to unified approaches that combine development, procurement, and exposure risk dimensions in AI risk classification, we hope that future risk frameworks can iterate on our methodologies and use them as a stepping stone towards more expansive AI governance solutions.

Hannah Liu is a Research Engineer in AI Policy and Governance at the Responsible AI Institute, where she helps develop new solutions and frameworks to bolster AI governance in organizations. She has been researching AI governance and sociotechnical applications for the past 5+ years, and is interested in developing AI governance frameworks that are globally applicable, ethically adequate, and decentralized. Hannah holds a Bachelor's from the University of Pennsylvania in Cognitive Science and is currently a Master's student in AI Applications and Innovation at Imperial College London. You can reach out to her at hannah@responsible.ai.

Disclaimer: The viewpoints expressed by the authors are their own and do not necessarily reflect the opinions, viewpoints and official policies of Practising Law Institute.