MANIFEST

SUMMER 2025

# Policy Framework for Safe Adoption of Open-Weight AI Models and Datasets

## Introduction

The democratization of AI has led to an explosion of open-weight models (open-source) and training datasets available to enterprises. For industries operating critical infrastructure, such as medical device manufacturing, auto manufacturers, aerospace, financial services, and defense, this represents a double-edged sword. While open-weight models can accelerate innovation, reduce costs, and improve mission outcomes, they also carry substantial risks. These include adversarial vulnerabilities, opaque licensing, misaligned data sources, and legal exposure.

This paper outlines suggests a policy framework for evaluating and adopting open-weight AI models and datasets safely. Each policy is accompanied by a rationale and a real-world example that illustrates the risks if left unaddressed to help explain the why behind the policy.

Our goal at Manifest is to arm and support the community of professionals in these mission-critical enterprises by collecting and sharing the policies that peers in the industry have adopted in the effort to safeguard and responsibly adopt artificial intelligence.

## Policies

# 1. Country of Origin


## Recommended Policy

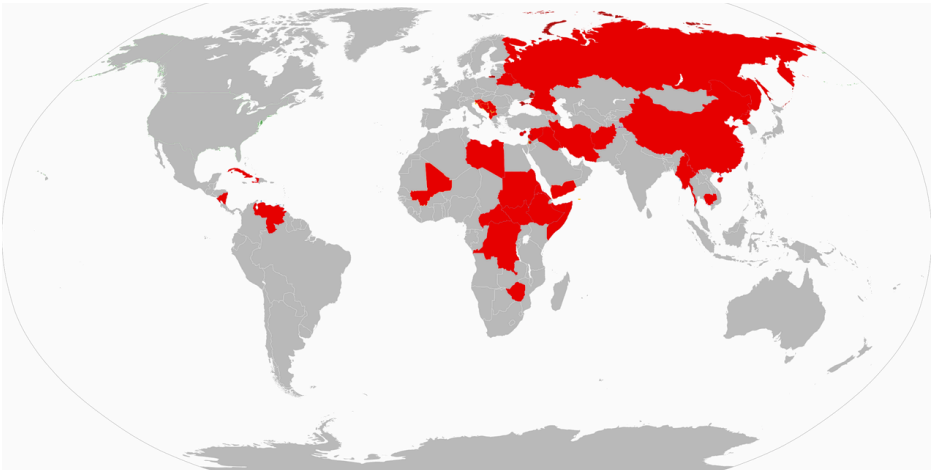
Model and dataset contributors must not originate from Office of Foreign Assets Control (OFAC)-sanctioned countries.

## Rationale

To comply with international trade and security laws, it is imperative to avoid direct or indirect collaboration with developers in countries subject to U.S. Treasury Department sanctions. Moreover, many enterprises in defense, aerospace, and other critical infrastructure categories have internal requirements to limit if not avoid entirely the use of technology originating from these geographies.

These countries include:

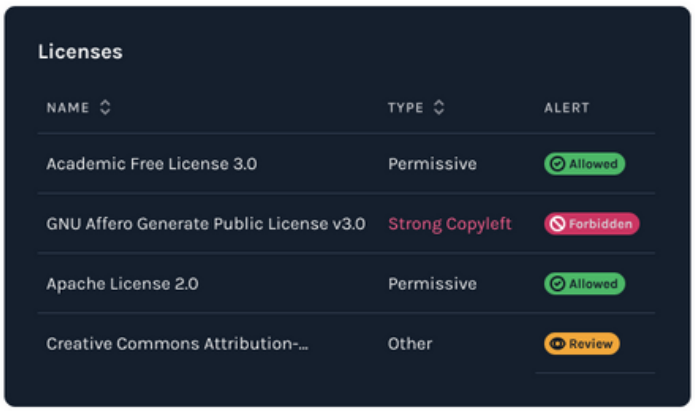
- Afghanistan
- Belarus
- Burma (Myanmar)
- Central African Republic
- China
- Cuba
- Democratic Republic of Congo (DRC)
- Democratic People's Republic of Korea (DPRK)
- Ethiopia
- Hong Kong
- Iran
- Iraq
- Lebanon
- Libya
- Mali
- Nicaragua
- Russia
- Somalia
- South Sudan
- Sudan
- Syria
- Venezuela
- Yemen



## Example of Violation

DeepSeek, a Chinese-developed AI chatbot, experienced a significant data breach in January 2025, exposing over a million sensitive records, including chat messages and API keys. The incident raised serious concerns about data security and privacy in AI. As a result, several countries, including South Korea, Australia, and Canada, banned the use of DeepSeek on government devices.

## 2. License Usage Compliance




### Recommended Policy

Models cannot be used under licenses that:

- Prohibit commercial use or use in the industry specific to my enterprise
- Require derivative work disclosure incompatible with product or business strategies
- Are legally ambiguous

### Rationale

Open-weight models are often distributed under licenses that are not enterprise-friendly, either intentionally or unintentionally. Violating license terms can result in legal actions, forced retraction of software, or forfeiture of intellectual property. Dataset licenses must be audited for restrictions on usage, distribution, and commercialization. Maintain license documentation in BOMs to prevent inadvertent violations.

Adopting technologies in contravention of their stated license terms presents effectively unbounded business risk to an enterprise. In the case of traditional software, licenses are fairly enumerated in that a limited number of licenses are generally used across open-source software. However, in the case of artificial intelligence, the permutations and qualitative nature of AI licenses means that reviewing each license is a cumbersome and onerous task.

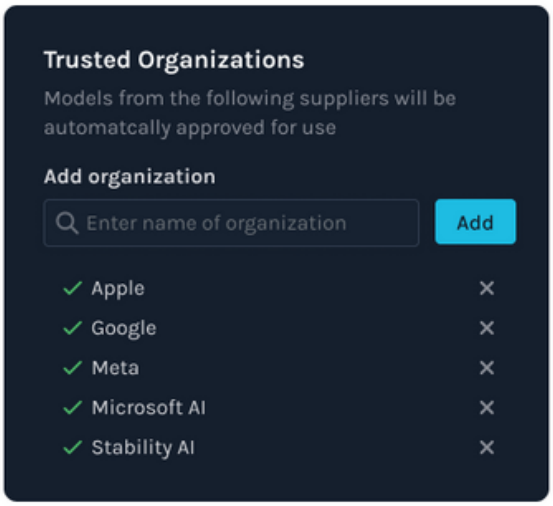
Adopting tooling to automatically review AI licenses for suitability and usage compliance is an essential part of AI governance.

### Example of Violation

Vizio, a leading producer of smart televisions, incorporated so-called “copyleft” open-source software components into its software. These components, distributed under GPL license, required that Vizio provide corresponding source code for its software as a result. Vizio refused, and the case has been both removed to federal court and remanded to state court several times in the past three years, resulting in substantial litigation costs.


Source: https://sfconservancy.org/copyleft-compliance/vizio.html

## 3. Trusted Organizations/Suppliers




### Recommended Policy

Models and datasets originating from a pre-defined list of vetted and verified organizations will be automatically approved for use. Models and datasets from organizations on a designated deny list will be automatically rejected. All others must undergo a thorough risk review process before approval. Require cryptographic signing and integrity verification for all externally sourced models and adapters. Maintain provenance records (Model Cards with verifiable signatures) to ensure authenticity.

### Rationale

A curated list of pre-approved and forbidden model and dataset providers enforces consistent risk and compliance standards while accelerating adoption. This reduces review overhead and ensures alignment with internal security, licensing, and governance policies.

Models or datasets from unknown or unverified sources increase exposure to risks such as model backdoors, poisoned or biased data, and IP violations

### Commonly Trusted Model Suppliers

### Commonly Untrusted Model Suppliers

# 4. Embedded Software Risk

## Recommended Policy

Models must be evaluated for software dependencies that contain:

- Critical or High vulnerabilities, as measured by the National Vulnerability Database (NVD) CVSS scale
- Vulnerabilities with a Known Exploit, as measured by CISA's Known Exploited Vulnerability (KEV) catalog

## Rationale

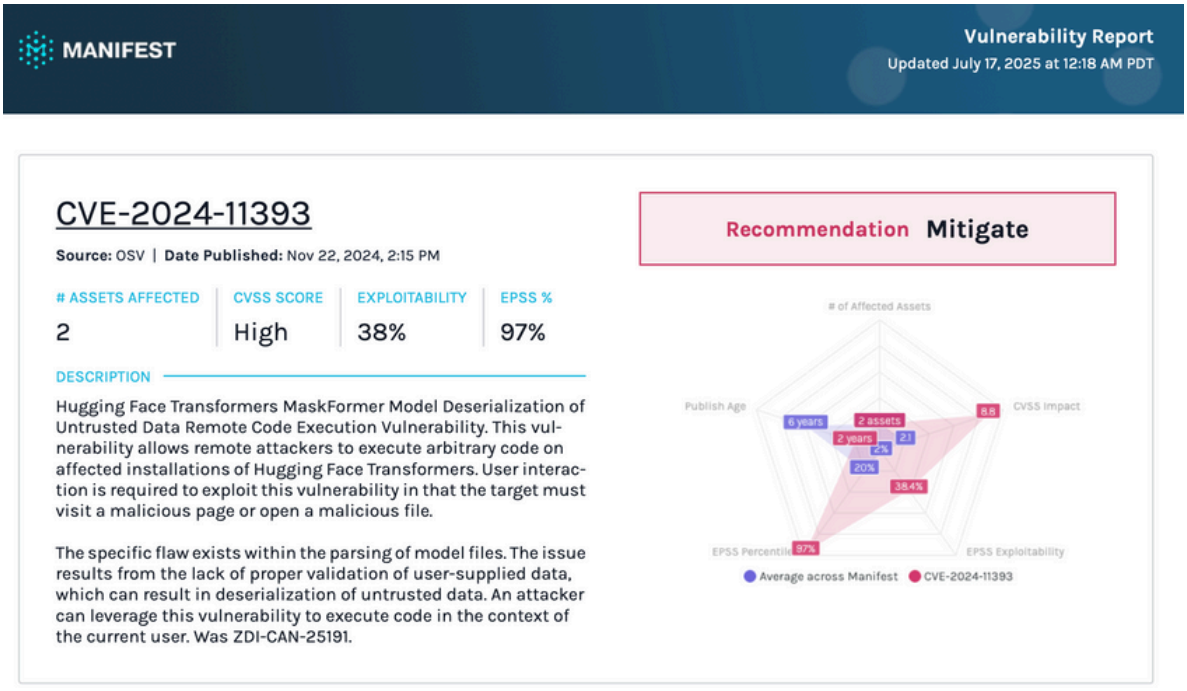
AI models ship with software dependencies and are pieces within software as well. Like traditional software, these applications and components must be analyzed for the presence of critical vulnerabilities, and contextualized with the exploitability of those vulnerabilities.

Generating a software bill of materials (SBOM) and analyzing it for risk is the best way to both ascertain the risk posed by an application initially, and then to continuously monitor that application for software risk as new vulnerabilities are published. In addition, maintaining an AI Bill of Materials (AIBOM) capturing all datasets, pre-trained models, adapters, and their licenses to ensure traceability and quick response to vulnerabilities.

## Example of Violation

Many AI models today use a popular software library called Transformers that makes common tasks for LLMs easier to implement. This library is shipped with the AI model when used. In late 2024, a new vulnerability CVE-2024-11393 was discovered on this library that allowed remote attackers to execute arbitrary code through loading malicious model files. This vulnerability was not due to an AI model itself, but due to the software supporting it.

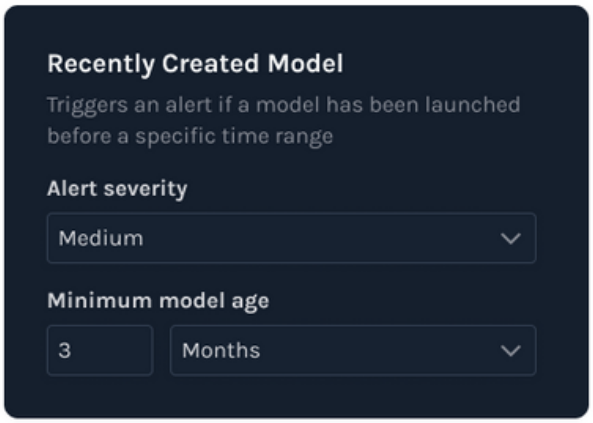
Companies need to be aware of the software that is bundled in with the AI models, and have a tool that can alert them of new issues with the models they already have in use along with reviewing the security of new models they are considering using.



## 5. Newly Released Model Risks

### Recommended Policy

Evaluate all pre-trained models, LoRA adapters, and third-party fine-tuning modules for integrity, provenance, and signs of tampering, including hidden backdoors or malicious triggers. The enterprise may only adopt models and datasets that have been publicly available for at least 90-days.




### Rationale

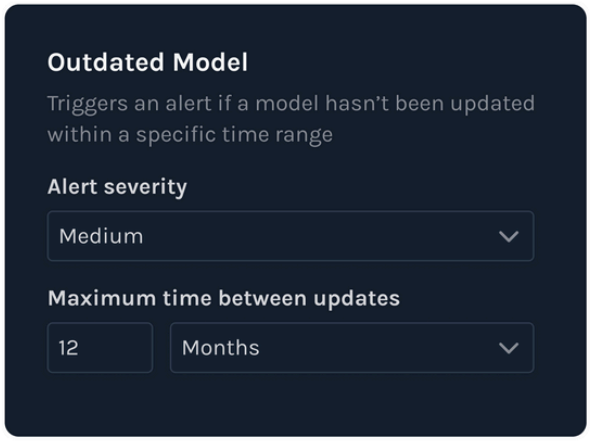
Newly released models may contain undiscovered security vulnerabilities or policy violations, along with performing unpredictably in real-world applications. Similar to traditional software waiting to update to the latest release, waiting a defined amount of time to wait before using a new AI model benefits organizations from early community feedback and the discovery of performance anomalies or security flaws.

Implementing a mandatory waiting period before adoption allows risk signals such as unexpected behaviors, vulnerabilities, or licensing red flags to surface through independent testing and public scrutiny.

## 6. Outdated Model Risk

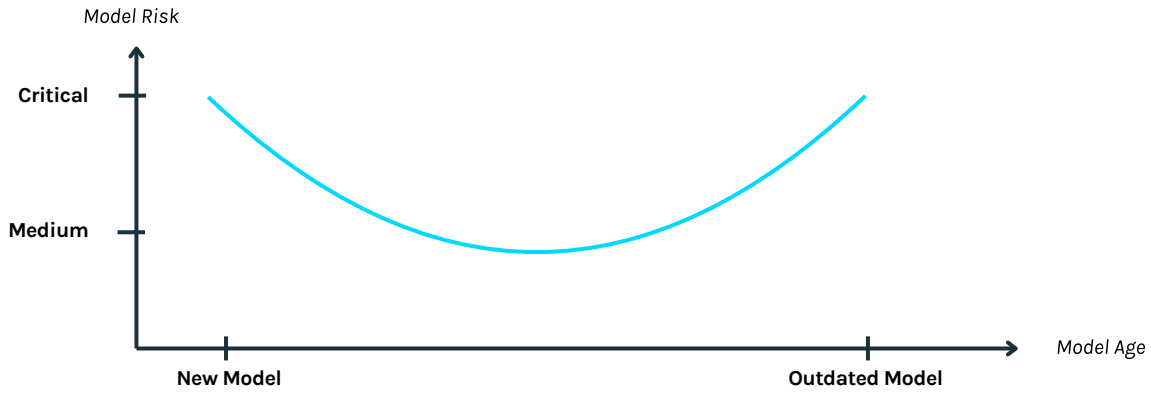
### Recommended Policy

Models must demonstrate active maintenance or responsiveness from their development team within the past 12-months.




### Rationale

Abandoned models can pose significant risk due to unpatched vulnerabilities, compatibility issues, or lack of responsiveness to reported security concerns.



# Current AI Regulatory Requirements (as of July 2025)

### EU AI Act (2024)

A risk-based framework that classifies AI systems by potential harm, imposing strict requirements—including transparency, documentation, and governance—for high-risk applications while banning certain uses altogether

*"High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer"*

### Texas TRAIGA (HB149 / HB1709)

Regulations targeting AI systems with harmful intent, prohibiting manipulative or rights-violating AI uses, adding transparency and biometric privacy safeguards, and establishing governance standards for government entities

*"...to provide transparency regarding those risks in the development, deployment, or use of artificial intelligence systems"*

### Colorado Artificial Intelligence Act (SB24-205)

State law that mandates developers and deployers of high-risk AI systems implement risk management, conduct bias impact assessments, provide disclosures to users, and report algorithmic discrimination enhancing consumer protection

*"A risk management policy and program implemented pursuant to subsection (2)(a) of this section may cover multiple high-risk artificial intelligence systems deployed by the deployer. A deployer shall maintain the most recently completed impact assessment for a high-risk artificial intelligence system."*

### California AB 2013

Requires generative AI developers to publicly disclose detailed training data information—including sources, dataset characteristics, and licensing—by January 1, 2026, to improve transparency in AI development

*"Existing law requires the Department of Technology, in coordination with other interagency bodies, to conduct, on or before September 1, 2024, a comprehensive inventory of all high-risk automated decision systems. The bill would require that this documentation include, among other requirements, a high-level summary of the datasets used in the development of the system or service, as specified."*

### FY2026 NDAA Section 1531 (HASC draft)

Includes a novel requirement for a Software Bill of Materials (SBOM) specific to AI systems, aimed at enhancing transparency, cybersecurity, and supply-chain accountability within the Department of Defense

*(b) BILL OF MATERIALS FOR ARTIFICIAL INTELLIGENCE.*
*(1) IN GENERAL. – Any policy, regulation, guidance, or requirement issued by the Department of Defense related to the use, submission, or maintenance of a software bill of materials shall also apply to an artificial intelligence software bill of materials, to the extent practicable, for all artificial intelligence systems, models, and softare used, developed, or procured by the Department.*

## Conclusion

The proliferation of open-weight AI models and datasets presents a complex mix of opportunity and risk for mission-critical industries. As this guide outlines, the path to safe adoption is not paved with technical capability alone but with structured governance, informed policy, and proactive risk management.

By implementing clear, defensible policies around country of origin, licensing, vendor trustworthiness, embedded software risk, model maturity, and ongoing maintenance, enterprises can responsibly leverage the power of open AI without compromising their operational integrity or regulatory compliance. These policies serve not only as internal guardrails but as a framework for navigating the increasingly complex legal and ethical terrain shaped by emerging AI regulations worldwide.

Ultimately, responsible AI adoption in sensitive sectors demands more than reactive safeguards, it requires a culture of diligence, cross-functional collaboration, and continuous vigilance. At Manifest, we are committed to helping organizations build this foundation, ensuring that AI is deployed not only for innovation but for resilience, security, and long-term value.

## About Manifest

Manifest helps mission-critical organizations secure their software and AI systems from the ground up. Designed for regulated industries like defense, healthcare, and automotive, our platform uncovers hidden vulnerabilities, automates risk assessments, and enables continuous security monitoring across complex supply chains.

We turn transparency into action, giving teams the tools to assess and mitigate risks in real time. Trusted by institutions from Wall Street to the U.S. Department of Defense, Manifest empowers secure operations where the stakes are highest.

**Learn more:** http://manifestcyber.com/ai-risk