



AI Safety Model Elaborated:

The SAFE AI Foundation's AI Safety Model (AISM)

Technical Report

SAFE AI Foundation USA

20 OCTOBER 2026

*A*bstract

On 15 April 2025, the SAFE AI Foundation USA very briefly introduced the AI Safety Model (AISM), which has 4 components of safety: (i) input, (ii) model, (iii) output, and (iv) compliance. At that time, no organization has viewed AI safety in this way. Other organizations were deeply investigating specific issues and problems, but solving input or model safety alone does not solve the end-to-end system AI safety issues. This is why we introduced AISM. Benchmarking and measurements alone are not solutions to AI safety issues. They are measurements, not solutions. In this report, we present AISM from an engineering and systems perspective, so that industries can effectively implement workable end-to-end solutions to achieve comprehensive AI safety for AI systems. This report goes into greater depth the technical details of AISM.

1 Introducing Safe AI Foundation's AISM

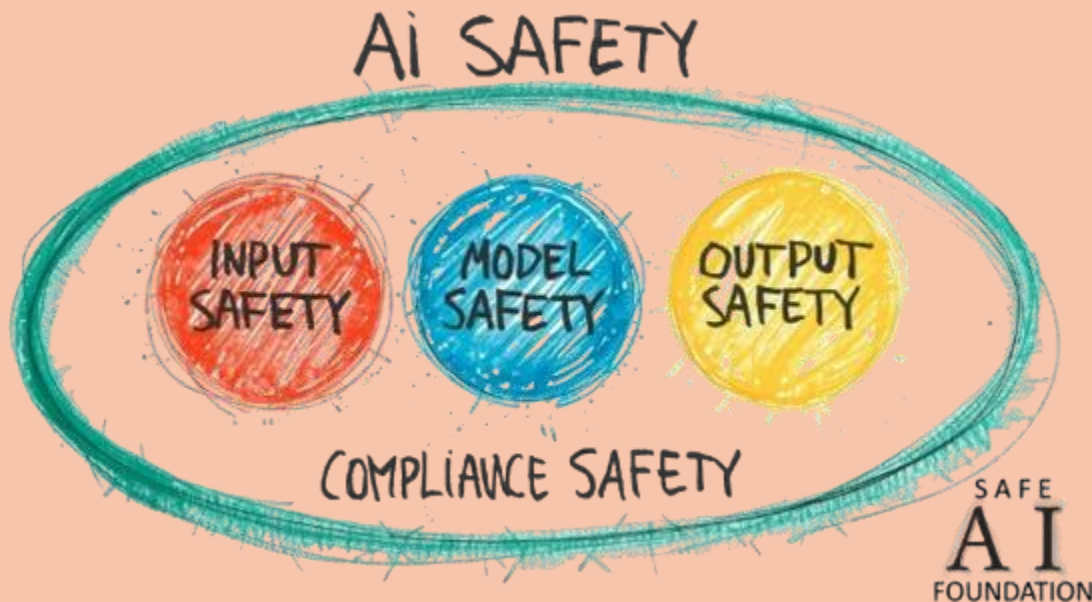


Fig 1: The 4 Pillar I-M-O-C AI Safety Model introduced by the SAFE AI Foundation USA

Introduced on 15 April 2025, the foundation AI Safety Model provides the full end-to-end safety protection and functions for the AI pipeline, governing all the interactions and events happening at input, model and output. This allows policies, laws, regulations, etc., to be applied to each section of the AI pipeline and nothing goes unmonitored or ungoverned. Humans, therefore, can remain in full control of the process and ensure compliance with existing requirements. While comprehensive in design, AISM requires stringent efforts in its

development. Each component must be able to communicate with the compliance layer via control signals. Control modules are present to ensure each aspect of safety are supported. Collectively, they ensure that comprehensive safety can be achieved.

Unlike other existing work where researchers address each safety aspect in silos, i.e., some aggressively work on input safety, while others look deeply only on model safety or output safety. Few, if not none, have looked at the whole spectrum of AI system safety requirements. This is where SAFE AI Foundation USA makes its landmark in this space. We are convinced that an end-to-end holistic architecture is needed to achieve full comprehensive AI safety. Current approach of saving fires on one end does not stop ranging fires on the other end. That is the logic behind our AISM design.

2 AISM Pillar 1: Input Safety

2.1 Definition and purpose

Input Safety (IS) is the **first line of defense** in an AI system. It encompasses the policies, controls, and technical mechanisms used to inspect, validate, authenticate, classify, filter, remove, and safely process information (data) entering into the AI system.

Inputs may include user prompts, uploaded files, images, audio, video, retrieved documents, external databases, API responses, and messages received from other AI agents. Because an AI system's subsequent behavior depends heavily on the information and instructions it receives, unsafe or untrusted inputs can undermine the otherwise effective AI LLM model safety mechanisms.

The primary objective of *Input Safety* is to ensure that incoming information is valid, appropriately checked and authorized, sufficiently trustworthy, non-malicious, and handled according to its sensitivity and risk. This does not mean that all inputs must be trusted or that every malicious input can be identified with certainty. Rather, the system should distinguish trusted instructions from untrusted data, identify suspicious or potentially harmful content, and limit the consequences of inputs that evade detection.

The IS control module is to protect the system from any harmful data that may harm the model or trick the model into misbehaving. In addition to filtering each token, it must also look at multiple tokens, and also the subsequent context. Out-of-band processing and buffering are needed. With parallel computation architectures, all these can be done quickly, without users noticing the difference.

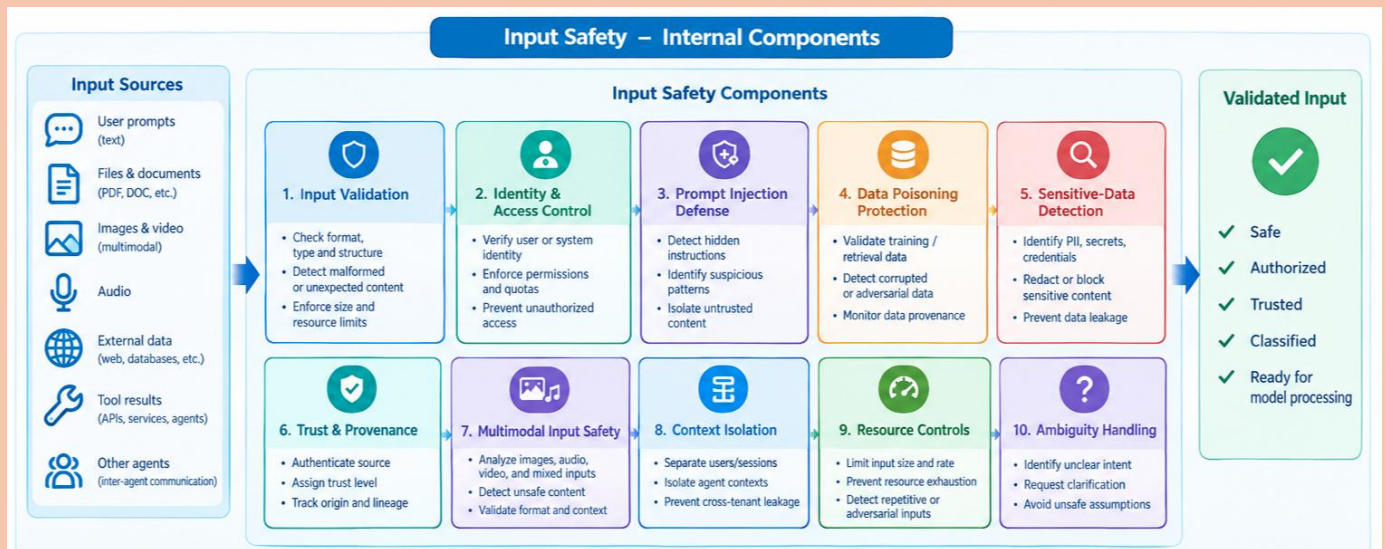


Fig 2: Internal components of the AI Input Safety layer.

A key input threat is **prompt injection**, in which malicious instructions are embedded in a user prompt, document, website, email, or tool response to influence the AI to disregard its intended instructions or perform unauthorized actions. Other threats (see Fig 2) include **data poisoning**, **malicious files**, **identity spoofing**, **sensitive-data exposure**, **oversized inputs intended to exhaust resources**, and **manipulated information** from external sources. Input Safety should therefore operate as a layered defense rather than relying solely on a language model to recognize malicious content.

Input Safety also establishes the boundary between receiving information and granting it authority. A document may contain useful facts without having permission to change the system's governing instructions. A user may submit a request without having authorization to access every database. An external tool may return data without being entitled to direct the agent's next action. Separating content from authority is essential for secure AI agents, particularly those with access to tools, credentials, private data, or external systems.

If input DATA is “for information” or “training”, it will be handled differently. Information describing how to cheat, corrupt, manipulate, and other bad behaviors can poison the AI model itself. The model “learns” from the inputs of previously processed data. When pursuing a goal, that misbehavior or trick will be “pulled out” and used. It is this aspect that one observes an AI model “misbehaving”. **The model itself is innocent. The things that it has learnt are not!** If input DATA are instructions, then it will be checked differently and the AI model may not adhere to the instructions fully, depending if the instructions are maliciously inclined or not. As shown in Fig 2, the input safety block requires the implementation of 10 sub components. Each block is an entity to be implemented in software.

Hence, input safety work is no small feat. A **comprehensive IS solution can remove most of the harm at an early stage**, which is desirable for a good AI system design. Because no input filter can guarantee detection of every bad intention, the AI system must be capable of also limiting the privileges and potential impact of any malicious input that reaches subsequent processing stages, such as the model and output processing blocks.

3 AISM Pillar 2: Model Safety

Model Safety (MS) concerns the ability of an AI LLM model and its surrounding execution environment to operate reliably, resist manipulation, remain within authorized capability and autonomy boundaries, and avoid causing unacceptable harm. While Input Safety (IS) examines the information entering an AI system, Model Safety (MS) focuses on how the system processes that information, produces decisions, follows instructions, uses tools, access resources, and exercises autonomy. It encompasses the model itself, including its capabilities, behavioral tendencies, limitations, and configuration, as well as the runtime controls that govern its safe operation, as shown in Fig 3 below.

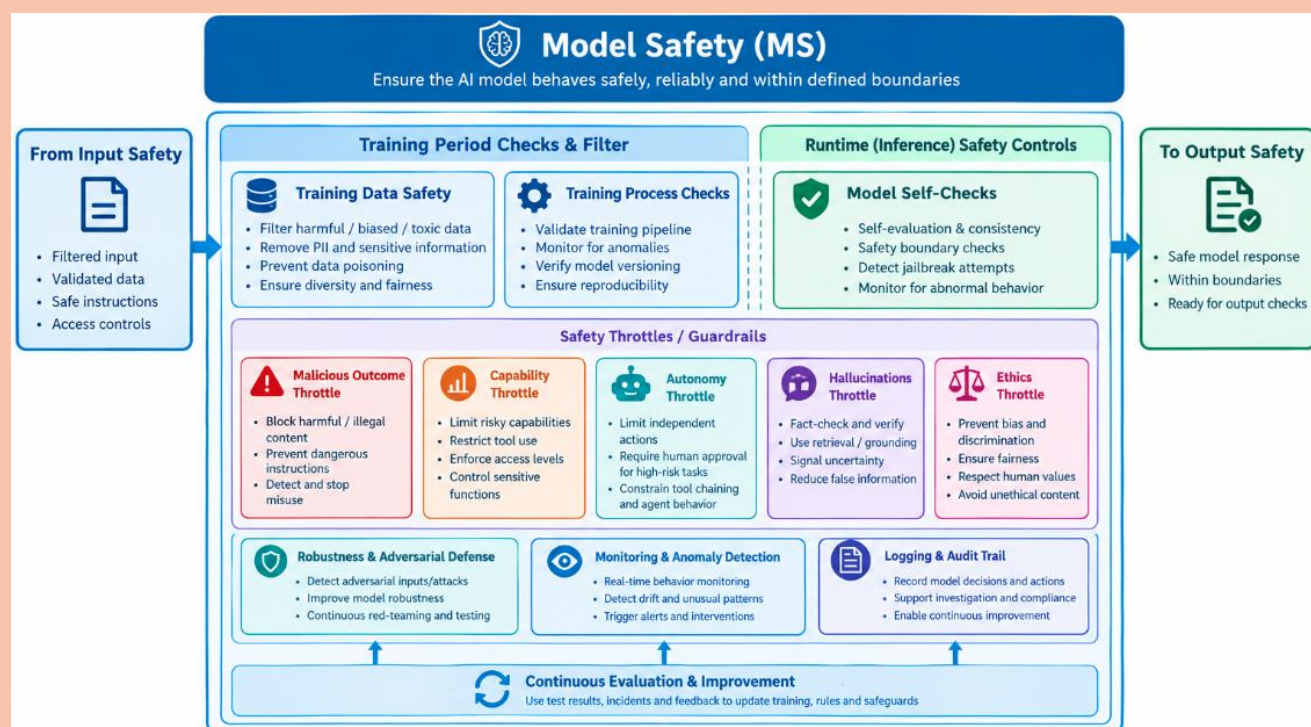


Fig 3: Internal components of AI Model Safety (MS).

The Model Safety (MS) component of SAFE AI Foundation's AI Safety Model is designed to ensure that an AI model behaves safely, reliably, and within defined boundaries throughout its operational lifecycle. As illustrated in Fig. 3, Model Safety receives filtered and validated

information from the Input Safety layer, applies safeguards during both model training and runtime inference, and passes its resulting response to the Output Safety layer for further verification. Its objective is to prevent harmful behavior, limit unintended capabilities and autonomy, improve factual reliability, and maintain appropriate ethical and operational constraints. Model Safety is not a single mechanism; it is a coordinated collection of preventive controls, runtime guardrails, monitoring mechanisms, and continuous improvement processes.

3.1 Training Period Checks and Filters

The first major component addresses safety before an AI model is deployed. Training Data Safety examines the quality and suitability of training data, including harmful content, toxic material, bias, personally identifiable information (PII), and potentially poisoned or manipulated data. Appropriate filtering, data provenance checks, privacy protections, and dataset quality assessments help reduce the risk of embedding unsafe behaviors into the model. Diversity and fairness evaluations can also identify potential imbalances that might lead to discriminatory outcomes. These measures must be appropriate to the model's intended use and applicable requirements, since indiscriminate data removal can also reduce model quality or introduce new biases.

The Training Process Checks block focuses on the integrity and reliability of the training pipeline itself. It validates training configurations, monitors for anomalies, verifies model versions, and supports reproducibility through appropriate records and controls. Additional safeguards can include checking the integrity of training code and dependencies, controlling access to model weights and training infrastructure, evaluating changes between model versions, and testing for unexpected behavior introduced by fine-tuning. Together, training data safeguards and process checks help establish a safer foundation before a model enters production. However, training-time checks cannot anticipate every possible failure, which is why runtime safeguards remain essential.

3.2 Model Self-Checks

During runtime inference, the Model Self-Checks component evaluates whether the model's behavior is consistent with its intended role and established safety boundaries. It can incorporate safety classifiers, policy checks, consistency tests, uncertainty estimation, and detection of suspicious instructions or attempted jailbreaks. For systems capable of reasoning or using tools, additional checks may examine proposed actions, tool calls, and deviations from authorized operating procedures.

Model self-checks should not be treated as a guarantee of safety. An AI model can fail to recognize its own errors, misjudge the risks of a response, or be manipulated into bypassing safeguards. Therefore, self-assessment should be combined with independently

implemented controls, external evaluation, and monitoring. The purpose is to create multiple opportunities to detect and prevent unsafe behavior before it progresses through the system.

3.3 Safety Throttles and Guardrails

The central part of the diagram contains five specialized safety throttles. Each throttle addresses a distinct category of model risk, and their combined operation helps constrain behavior that cannot be controlled adequately by a single safeguard.

- **Malicious Outcome Throttle:** Detects and blocks outputs, plans, or actions that could facilitate harm, illegal activity, or dangerous misuse. Depending on the application, it may refuse a request, restrict a proposed action, or escalate the case for human review.
- **Capability Throttle:** Limits access to sensitive capabilities according to the model's role, permissions, and assessed risk. Controls may restrict sensitive tool use, execution privileges, access to protected resources, or high-risk functions. The goal is to ensure that available capabilities do not exceed what is necessary and authorized for the task.
- **Autonomy Throttle:** Controls how independently an AI system can operate. It can limit action sequences, constrain agent loops, impose execution budgets, require approval before consequential actions, and provide interruption or shutdown mechanisms. Higher-risk activities should generally require stronger authorization and oversight.
- **Hallucinations Throttle:** Reduces unsupported or fabricated claims through retrieval and grounding, factual verification, consistency checks, calibrated uncertainty, and appropriate abstention when reliable evidence is unavailable. Since hallucinations cannot be eliminated completely, high-impact claims may require independent verification or human review.
- **Ethics Throttle:** Addresses risks involving unfair discrimination, harmful treatment of affected groups, violations of human dignity, and conflicts with established ethical principles. It can incorporate fairness testing, bias detection, policy-based restrictions, impact assessments, and escalation procedures. Ethical safeguards should be evaluated in context rather than relying solely on a model's own judgment of what is ethical.

These throttles are complementary rather than interchangeable. A response can be factually correct but malicious, ethically questionable but legal, or harmless in isolation yet dangerous when an autonomous agent executes it repeatedly. The controls must therefore evaluate different risk dimensions and coordinate their interventions. Their implementation may combine model-level techniques, external policy engines, permission systems, deterministic rules, and human oversight, depending on the application.

3.4 Robustness, Monitoring, and Auditability

The diagram also includes three supporting components that help ensure safety controls remain effective under real-world conditions.

- **Robustness and Adversarial Defense:** This evaluates how the model responds to hostile prompts, jailbreak attempts, unusual inputs, distribution shifts, and other challenging conditions. Red-team exercises, adversarial testing, regression testing, and robustness evaluations help expose weaknesses that ordinary testing might miss.
- **Monitoring and Anomaly Detection:** This observes runtime behavior for unusual patterns, policy violations, unexpected tool usage, performance degradation, or signs that the model is behaving outside its intended boundaries. When an anomaly is detected, the system can trigger an alert, restrict a capability, pause an operation, or initiate human review. Monitoring should be designed to protect sensitive information and minimize unnecessary collection of user data.
- **Logging and Audit Trail:** This records relevant model versions, safety decisions, interventions, tool actions, and incidents. Appropriate records make it possible to investigate failures, determine whether safeguards operated as intended, demonstrate accountability, and support compliance assessments. Logging must itself be governed by suitable access controls, retention policies, and privacy protections.

3.5 Continuous Evaluation and Improvement

The Continuous Evaluation and Improvement block closes the feedback loop. Results from monitoring, audits, adversarial tests, incident investigations, and user feedback can be used to revise training datasets, update safety policies, improve detection mechanisms, adjust throttle thresholds, and strengthen operational controls. Changes should be tested before deployment and monitored afterward to ensure that improvements in one area do not create regressions elsewhere. This makes Model Safety (MS) an ongoing engineering and governance process rather than a one-time certification performed before launch.

3.6 Interacting with the Compliance Layer

Compliance Safety (CS) provides cross-cutting requirements and oversight for the entire Model Safety block. It ensures that training practices, runtime safeguards, capability restrictions, monitoring procedures, and audit mechanisms are designed and operated in accordance with applicable standards, organizational policies, laws, and regulations. It also supports risk assessment, documentation, incident reporting, accountability, and periodic reviews as requirements and system capabilities evolve.

Compliance Safety does not replace the technical controls within Model Safety. Instead, it establishes the obligations those controls must satisfy and the evidence needed to demonstrate their effectiveness. For example, privacy requirements may influence training data filtering and log retention, while applicable cybersecurity rules may determine access restrictions and monitoring requirements.

3.7 Interacting with upper layer SAIOP

In the paper “Total AI Safety”, we explain the relationship between SAIOP and AISM. SAIOP sets the upper safe limit of AI before humans lose control. Model Safety has a particularly important relationship with SAIOP because it governs the model’s behavior and capabilities in relation to the defined safety boundary. During training, Model Safety evaluates the model’s behavior, identifies potentially unsafe capabilities, and establishes safeguards that help keep the model within its intended operating limits. During runtime inference, Model Safety checks the model’s proposed responses, plans, tool calls, and actions against the applicable SAIOP constraints. If the model attempts an operation beyond its authorized capabilities, exceeds an autonomy limit, generates a prohibited outcome, or exhibits behavior that violates an established safety condition, the relevant Model Safety throttle can intervene by blocking the operation, restricting the capability, requesting human approval, or terminating the action. Where necessary, enforcement should occur through independently implemented controls rather than relying exclusively on the model to recognize and correct its own behavior.

Model Safety (MS) also requires robust control over the resources and privileges available to an AI system. An agent should receive only the permissions necessary for its assigned tasks, with appropriate restrictions on file access, network connections, credentials, software execution, and external services. High-impact or irreversible actions may require explicit human authorization. Sandboxing, policy enforcement, runtime monitoring, and mechanisms for interrupting or terminating unsafe activity can help limit the consequences of unexpected model behavior. Importantly, these safeguards should not depend entirely on the model voluntarily following its instructions; critical safety boundaries should be enforced by mechanisms outside the model, such as Safe AI Operating Point (SAIOP).

No model can be assumed perfectly safe, and no single technical control can eliminate all risks. The objective of AISM is therefore to establish end-to-end safety mechanisms, identify residual risks, and determine whether the combined model and runtime safeguards provide an acceptable level of protection for the intended use. Crucially, safety-critical restrictions should be enforced by independently verifiable controls rather than relying solely on the model to obey instructions. Model engineers must implement software to safeguard the behavior of AI Models.

In summary, Model Safety is presented as a layered defense architecture: training-time checks reduce risks before deployment; self-checks and safety throttles constrain runtime behavior; robustness testing, monitoring, and audit trails detect weaknesses and support accountability; and continuous improvement strengthens the safeguards over time. Compliance Safety provides the overarching governance framework that keeps these mechanisms aligned with external obligations and internal safety commitments. Together, these components provide a practical foundation for building AI systems that are more reliable, controllable, and resistant to harmful or unintended behavior.

4 AISM Pillar 3: Output Safety

Output Safety concerns the safety, reliability, appropriateness, and integrity of information and actions produced by an AI system. Outputs include generated text, code, images, audio, video, recommendations, predictions, decisions, and actions executed through external tools or services. While Input Safety examines what enters the system and Model Safety examines how the system operates, Output Safety focuses on what the AI delivers to users and the outside world. Its primary objective is to prevent harmful, misleading, unauthorized, or otherwise unsafe outputs from reaching people, applications, or downstream systems.

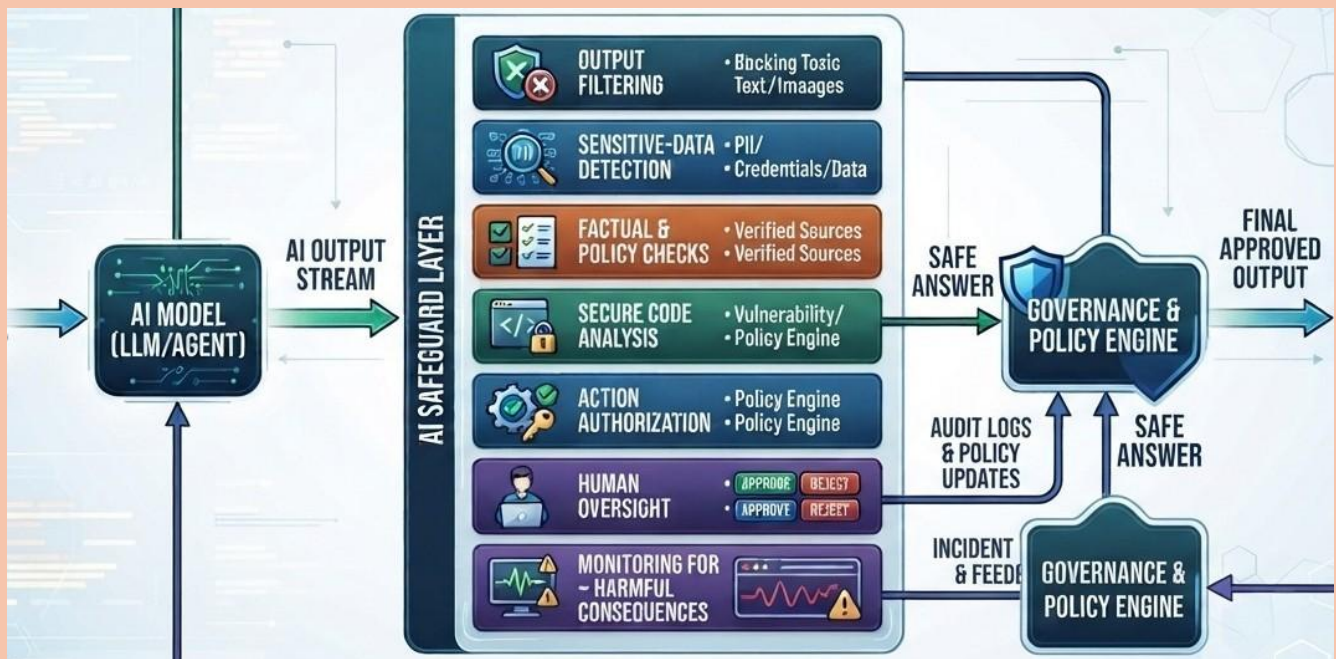


Fig 4: Components of Output Safety: The output safety is a layer that performs various checks and enforcing before yielding an approved output.

A major concern is the generation of harmful or misleading content. An AI system may provide factually incorrect information, fabricate sources, produce dangerous instructions, generate discriminatory recommendations, or express unwarranted certainty. Depending on the application, such failures can affect personal safety, financial decisions, public services, cybersecurity, and other consequential areas. Output Safety should therefore include appropriate content safeguards, factual verification, uncertainty disclosure, bias assessment, and escalation to qualified human reviewers when the potential consequences of an error are significant. The level of scrutiny should be proportionate to the risk rather than identical for every response.

Output Safety must also address privacy, security, and the consequences of AI-generated actions. An output could expose confidential information, reveal credentials, generate insecure code, facilitate fraud, or transmit sensitive data to an unauthorized recipient. In agentic systems, the output may not be merely a response to a user; it may trigger a transaction, execute a command, modify a file, send an email, or control another system. Such actions require authorization checks, appropriate restrictions, and additional confirmation where warranted. Generated code and commands should be evaluated within the context of their intended execution, and high-impact or irreversible actions should receive safeguards appropriate to their potential consequences.

Another important dimension is output integrity and traceability. Users should be able to distinguish verified facts from AI-generated claims where that distinction matters, understand relevant limitations, and identify when human review is necessary. Systems may also require provenance information, audit logs, content labeling, or mechanisms to correct and withdraw erroneous outputs. When AI outputs are passed to other models, agents, or automated workflows, the receiving systems should not automatically treat them as trusted instructions or verified facts. Output Safety must therefore consider not only the immediate recipient but also how generated content may be reused, redistributed, or acted upon downstream.

Finally, Output Safety should be established through layered safeguards and systematic evaluation. Solutions may include:

1. output filtering,
2. sensitive-data detection,
3. factual and policy checks,
4. secure code analysis,
5. action authorization,
6. human oversight, and
7. monitoring for harmful consequences.

These controls should be tested against realistic misuse scenarios and known failure modes, with incidents used to improve future safeguards. No output filter can guarantee that every unsafe response will be detected, so high-risk systems must also limit the authority of generated outputs and provide effective containment and recovery mechanisms.

5 AISM Pillar 4: Compliance Safety

Compliance Safety is the fourth pillar of SAFE AI Foundation's AI Safety Model (AISM). It ensures that an AI system is developed, trained, tested, deployed, operated, monitored, and decommissioned **in accordance with** applicable standards, established safety policies, laws, regulations, and other relevant compliance requirements. While Input Safety focuses on the information entering an AI system, Model Safety on the system's internal behavior, and Output Safety on the information and actions it produces, Compliance Safety establishes the overarching framework for ensuring that the entire AI lifecycle operates within defined legal, regulatory, ethical, and institutional boundaries.

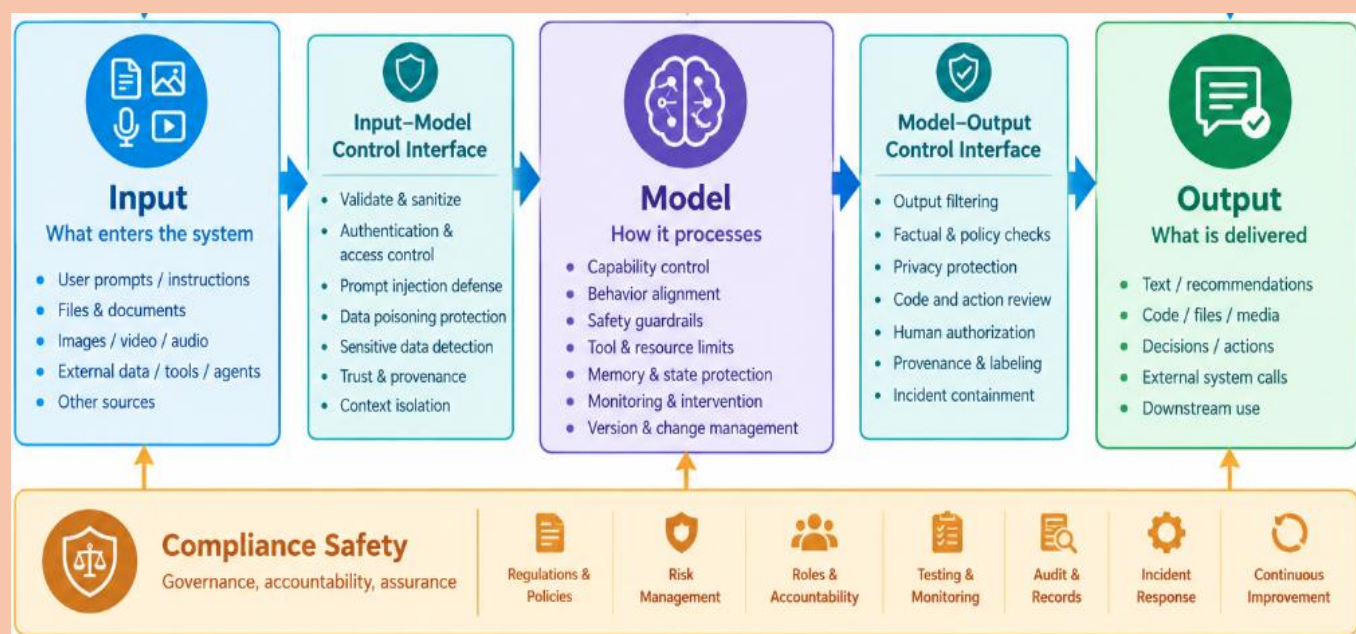


Fig 5: Compliance layer interfaces with all 3 AI Safety Blocks of the AISM.

Compliance Safety covers all aspects of compliance, including internationally recognized standards, industry-specific requirements, organizational AI safety policies, data protection and privacy laws, ethics, cybersecurity obligations, intellectual property requirements, consumer protection rules, and applicable national and regional AI regulations. It also

encompasses risk assessments, documentation, auditability, human oversight, accountability, incident reporting, access controls, recordkeeping, and ongoing monitoring of compliance. Because legal requirements and safety standards evolve over time, Compliance Safety must be a continuous process rather than a one-time certification or checklist completed before deployment. Organizations must be able to demonstrate not only that their AI systems satisfy relevant requirements at launch, but also that they remain compliant as models, applications, operating environments, and regulatory obligations change.

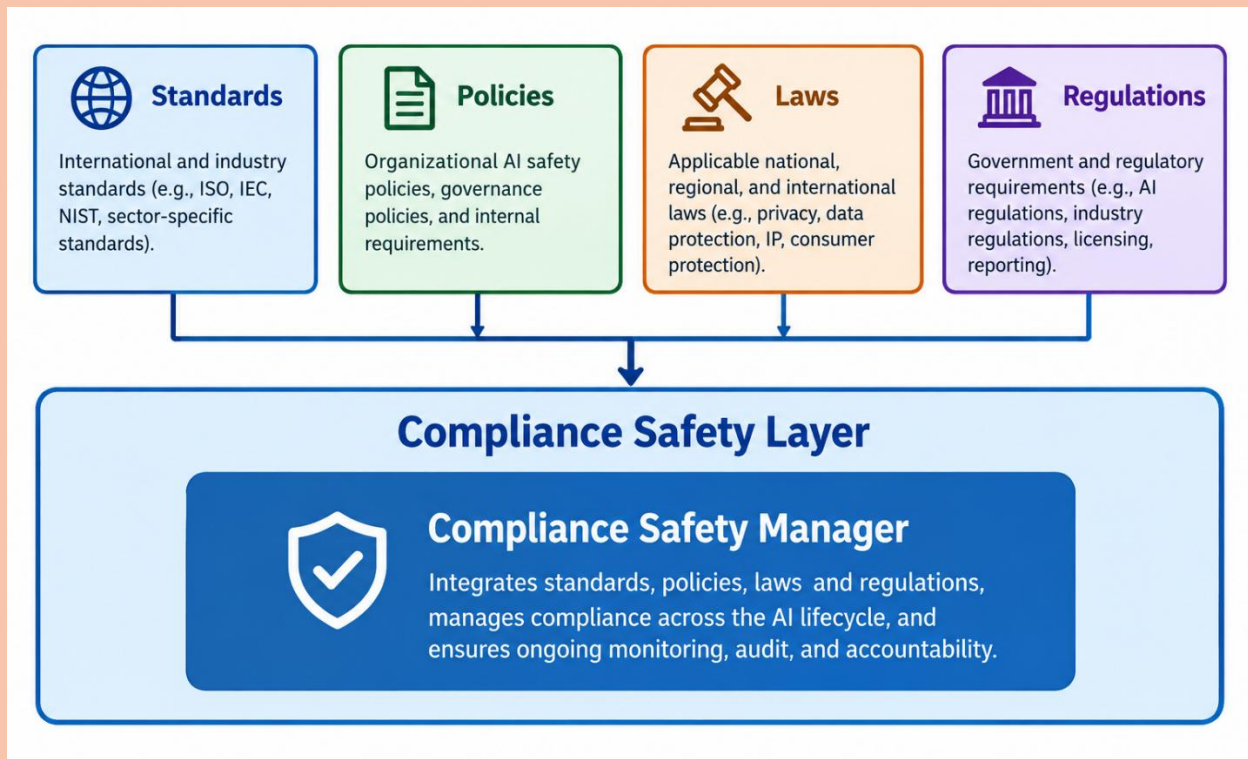


Fig 6: The Compliance Safety (CS) layer ensures related requirements from policies, laws, regulations and standards are considered by the AI system.

The importance of Compliance Safety lies in the fact that an AI system can be technically safe in one respect yet still be noncompliant, unlawful, or improperly governed in another. For example, a model might produce factually accurate and harmless answers but disclose personal information without authorization, violate intellectual property rights, fail to meet sector-specific requirements, or operate without legally required human oversight. Similarly, a company might implement strong technical safeguards but fail to maintain adequate audit records, report a qualifying incident, or fulfill its obligations to affected individuals. Compliance Safety bridges the gap between technical safeguards and the broader responsibilities of AI developers, deployers, operators, and organizations. It helps establish

clear accountability, reduce legal and operational risks, strengthen public trust, and ensure that AI safety commitments are translated into verifiable organizational practices.

A defining feature of SAFE AI Foundation’s approach is to give Compliance Safety an explicit place alongside Input Safety, Model Safety, and Output Safety, rather than treating compliance merely as an administrative task or an afterthought. Many established AI safety and governance frameworks already address governance, legal obligations, risk management, accountability, and standards. For example, the NIST AI Risk Management Framework addresses governance and risk management across the AI lifecycle, while ISO/IEC 42001 establishes requirements for an organizational AI management system.

SAFE AI Foundation’s proposed distinction is to make compliance an explicit, dedicated pillar that systematically integrates these different obligations into the overall AI safety model. This comprehensive approach is important because compliance requirements originate from multiple sources and operate at different levels. A system may satisfy a technical safety standard while violating a legal requirement; comply with the law while breaching its developer’s internal safety policy; or meet organizational policies while failing to satisfy an applicable industry standard, ethics, or contractual obligation. Compliance Safety therefore requires a unified process for identifying applicable requirements, mapping them to controls across all four pillars, verifying implementation, documenting evidence, conducting audits, addressing violations, and updating safeguards when requirements change. It also requires clear responsibility for compliance throughout the AI supply chain, including developers, model providers, third-party vendors, deployers, and operators.

Ultimately, Compliance Safety makes safety accountable, verifiable, and sustainable. It connects the technical protections implemented by the other three pillars with the external and internal obligations that govern their acceptable use. By positioning compliance as a foundational pillar in its own right, SAFE AI Foundation can offer a more explicit and integrated structure for bringing together technical safety, organizational policies, legal requirements, regulatory obligations, ethics, and recognized standards. This should be understood as the Foundation’s proposed architectural distinction, not a claim that other AI safety models omit compliance altogether. Its value lies in making comprehensive compliance a visible, systematic, and continuously enforced responsibility across the entire AI lifecycle.

6 Working in Tandem

Together, these 4 safety layers establish an end-to-end safety framework that addresses risks throughout the AI lifecycle, from the information entering an AI system to its internal

processing, generated outputs, and compliance with applicable standards, policies, laws, and regulations.



Fig 7: Data and Control flows for complete safety from an end-to-end perspective.

Rather than treating safety as a single technical feature, this 4-pillar approach recognizes that safe and useful AI requires multiple layers of protection working in coordination. The four pillars therefore operate as an integrated, defense-in-depth system rather than four independent checkpoints or safety solutions. Information passes through input controls before being processed by a safeguarded model, and generated outputs are evaluated before being delivered or acted upon by AI agents.



Fig 8: A complete overview of the 4 safety blocks working in tandem to provide comprehensive AI safety from an end-to-end system perspective.

Compliance requirements influence the design and operation of every stage, while monitoring, audits, incident investigations, and new risk assessments feed back into improvements across all four pillars. Where appropriate, controls can also interact directly,

for example, a compliance requirement may impose stricter input validation, require additional model refinement, or prohibit a particular output or action. This inter-safety-layer coordination helps prevent a weakness in one layer from automatically becoming a failure or a safety bottleneck of the entire system.

Ultimately, SAFE AI Foundation's 4-pillar AI Safety Model (AISM) seeks to provide comprehensive AI safety by combining preventions at input, protection within the model, verification at output, and accountability through compliance. Each pillar addresses a different category of risks, while their coordinated operation addresses the gaps that any single layer would leave unresolved. These sorts of interactions are not found in existing solutions, which tend to be silos in nature.

The result is a unified framework for designing, deploying, monitoring, and continuously improving AI systems that are safe, reliable, controllable, accountable, and better aligned with applicable human, organizational, and societal requirements. As with any safety architecture, however, comprehensive coverage depends on the quality of AISM implementation, the comprehensiveness of safety coverage, and the continuous adaptation of safeguards as new AI capabilities and risks evolve or are discovered. This allows room for expansion and inclusion. Safety support does not stop after successful implementation.

7 Chip & Software Implementation

The SAFE AI Foundation USA's view that AISM can be implemented through both hardware and software. Just as NVIDIA's OpenShell and Sentry which are executed at the lower runtime layers, AISM components can be executed at the software level running on hardware which are separated from the AI system. Separation can isolate CPU dependency while providing faster execution and isolation. Compliance layer is the layer that interfaces AI system blocks with policies, laws, standards, ethics, and regulations. It is the layer that provides control interfaces and commands to input, model, and output blocks. The compliance block can be updated, as new policies, standards, and laws are introduced.

8 Integrated Architecture: SAIOP +AISM + OASP

Our discussion so far has been on AISM in detail. However, AISM is not expected to act alone. In fact, AISM does not have the run-time layer safety protection. It sits on top of the real-time kernel. This is where we integrate AISM to sit on top of NVIDIA's OpenShell and Sentry instead, so that the lower runtime layers can provide fine-grain policing, interception, and

prevention of violations and malicious requests. This is particularly important and essential for AI agents or any AI executions that result in external access, data changes, or events (actions). This setup was explained in our “Total AI Safety” report, which is available online. While AISM+OpenShell+Sentry is a good AI Safety setup, it still lacks the governance anchored safety to complete the full picture. Hence, SAIOP sits on the upper stack, to define the upper safe limit that the AI entity/system must not cross so that humans remain in control. The resultant architecture is presented here again in Fig.9 below.

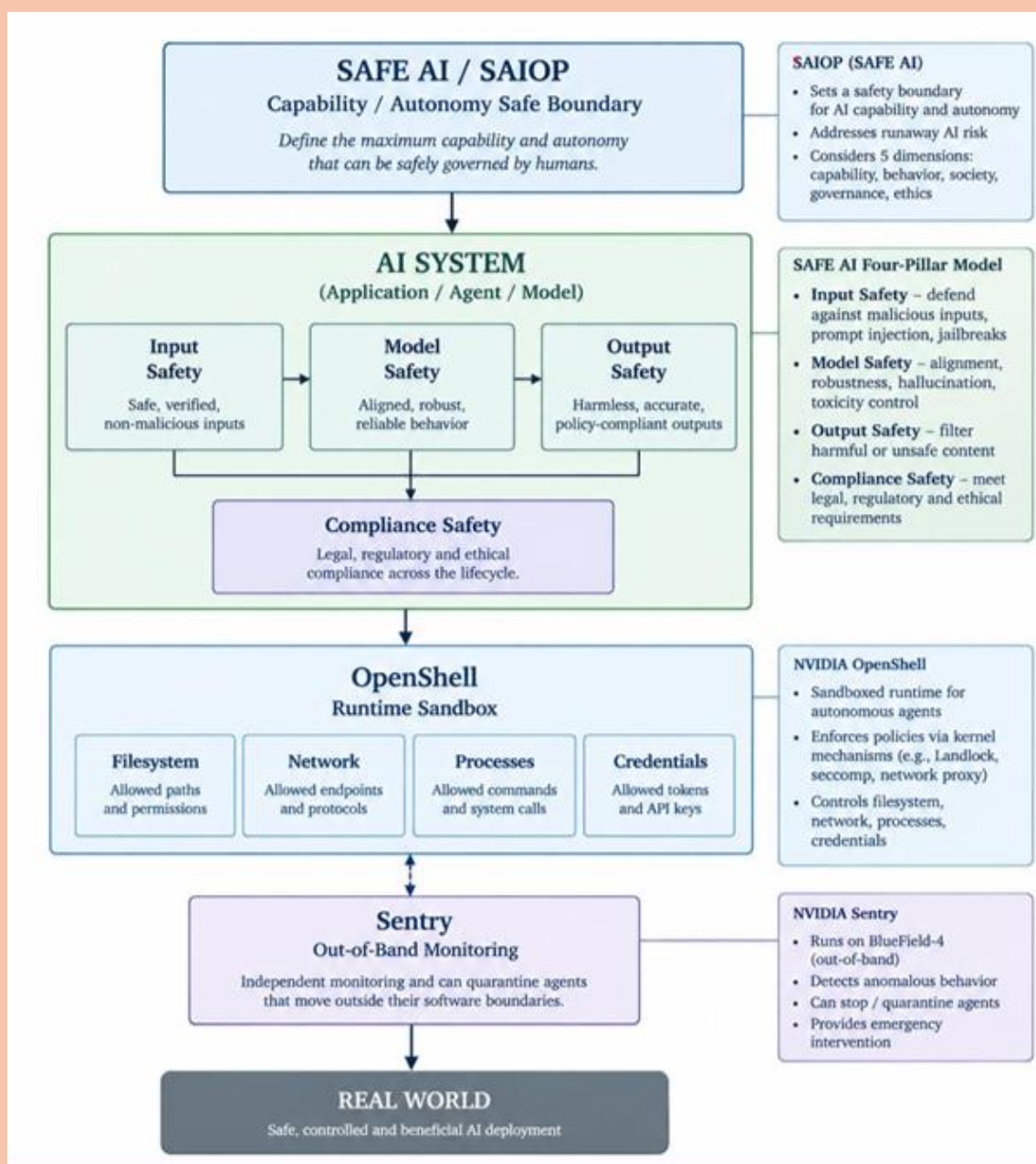


Fig 9: Layered AI Safety Architecture integrating both SAFE AI Foundation and NVIDIA Safety Solutions. SAIOP sits on top, followed by AISM and NVIDIA's OASP (OpenShell and Sentry).

9 Emerging AI Safety Industries & Companies

SAFE AI Foundation's four-pillar AI Safety Model (AISM), combined with SAIOP, NVIDIA's OpenShell, and Sentry, could provide a critical foundation for several emerging AI safety industries and startups. Each industry would address and fulfill a distinct safety function, ranging from protecting incoming data to controlling model behavior, validating outputs, ensuring compliance, and integrating the full safety architecture. This is similar to a divide-and-conquer approach, as compared to a company capable of doing all the components of AISM, along with the vertical integration with SAIOP and NVIDIA's OpenShell and Sentry layers.

No.	Potential industry / company type	Primary functions and services	Potential customers
1	Input Safety Company	Screens and validates user prompts, datasets, files, and external information; detects malicious instructions, prompt injection, data poisoning, sensitive information, and unauthorized access; enforces input policies before information reaches the AI model.	AI developers, enterprises, cloud providers, government agencies, healthcare and financial institutions.
2	Model Safety Company	Evaluates and constrains AI model behavior; performs safety training, adversarial testing, and robustness assessments; develops model self-checks and safety throttles for malicious outcomes, excessive capabilities, autonomy, hallucinations, and ethical risks; monitors runtime behavior.	Foundation-model developers, AI laboratories, agent developers, enterprises, defense and critical-infrastructure operators.
3	Output Safety Company	Verifies AI-generated content, recommendations, decisions, and actions; detects harmful outputs, misinformation, bias, privacy leaks, and unsupported claims; implements output filtering, factual verification, human-review workflows, and downstream action checks.	AI application developers, customer-service providers, media organizations, professional-service firms, regulated industries.

4	Compliance Safety Company	Maps AI systems to applicable laws, regulations, standards, and organizational policies; conducts AI risk assessments, compliance testing, audits, documentation, incident reporting, and continuous compliance monitoring.	Enterprises, AI vendors, financial institutions, healthcare providers, public-sector organizations, auditors.
5	AISM Integration Company	Integrates Input Safety, Model Safety, Output Safety, and Compliance Safety into a unified architecture; coordinates policies, interfaces, monitoring, reporting, and safety-control workflows across the four pillars; validates end-to-end operation.	Enterprises deploying multiple AI models, AI platform providers, system integrators, government and large organizations.
6	Total AI Safety Implementation Company	Implements an end-to-end safety architecture combining SAIOP, AISM, OpenShell, and Sentry. Establishes the operating safety boundary, integrates the four AISM pillars, applies technical execution and access controls, and monitors and enforces operational restrictions. Provides deployment, integration, testing, monitoring, incident response, and ongoing maintenance.	Large enterprises, autonomous-agent operators, AI infrastructure providers, government agencies, and organizations deploying high-impact AI systems.

Table 1: Emerging new startups and industries

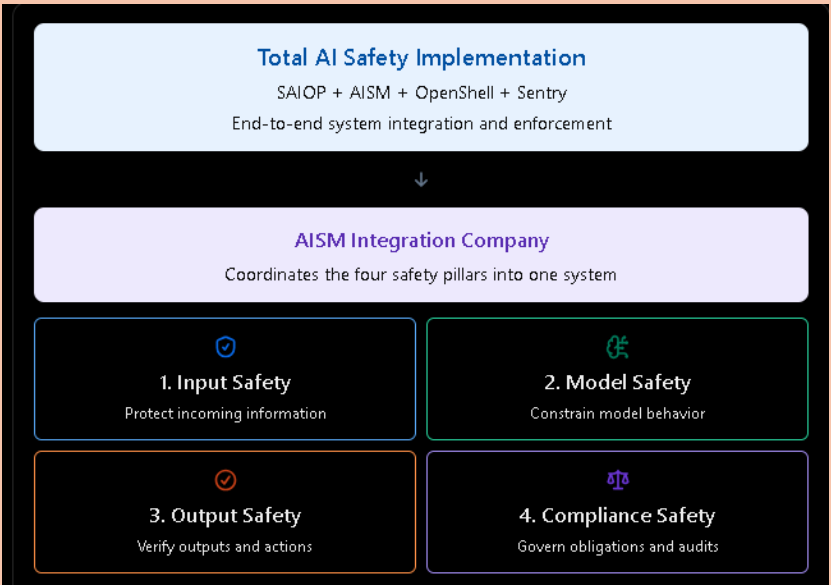


Fig 9: Emerging new startups focusing on each aspect of AI safety, along with those who will perform the vertical integration to include governance-anchored SAIOP and runtime layers from NVIDIA’s OpenShell and Sentry.

10 Conclusion

It is until now that the SAFE AI Foundation USA has time to dive deeply to explain its 4-Pillar AI Safety Model (AISM). This report details how we view an AI system's safety, from an end-to-end system perspective. We view that working on each safety block or layer alone is insufficient to resolve AI safety issues. Rather, we need to collectively address all the critical blocks and issues. AI safety is a big issue. It is a collection of several diverse and critical issues where each or more factors can bring down the system, cause malicious harms and chaos. This explains how we have designed our solution. The next step is to develop each of these major safety blocks, incorporate working solutions for each safety block, and integrate them into a 4-pillar AI Safety platform. This end-to-end solution is not only comprehensive but also incorporates existing standards, policies, ethics, and laws, making it suitable, compatible and functionable.

\\ end ///

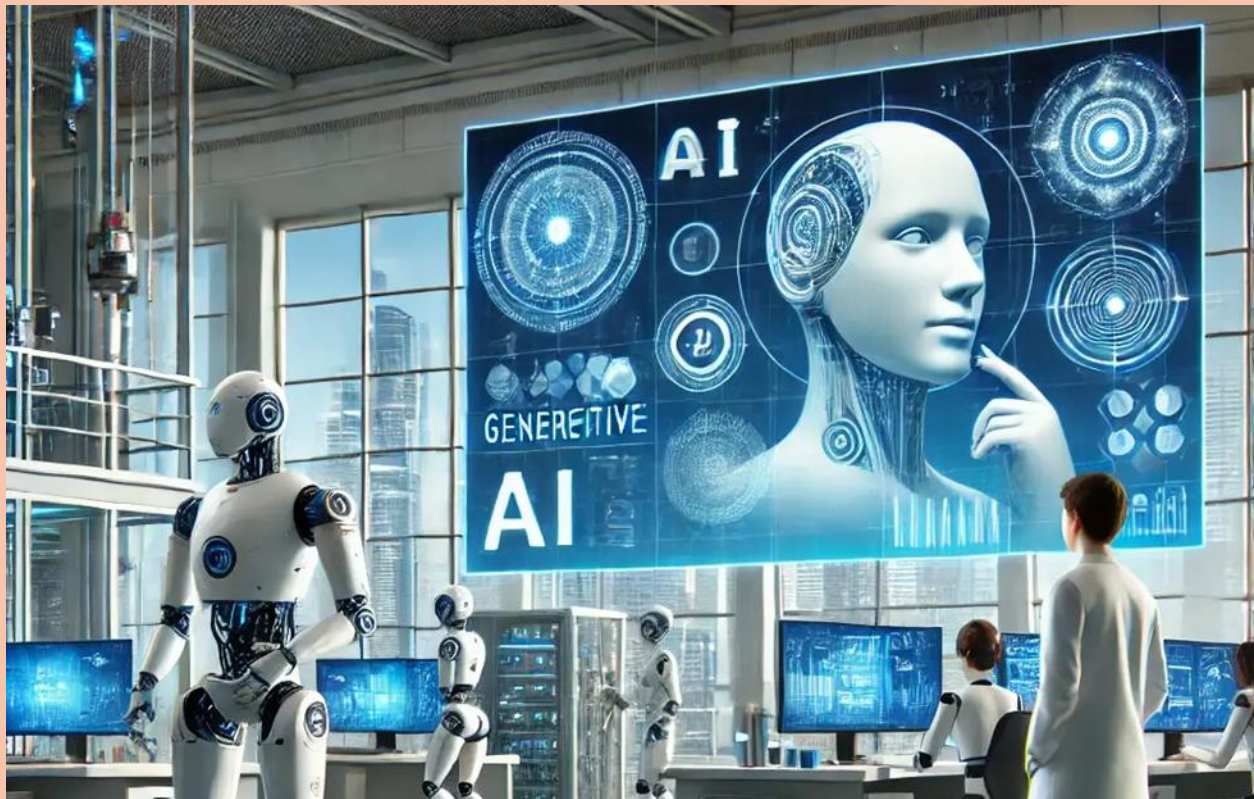
References

1. NVIDIA OpenShell documentation — Architecture.
2. NVIDIA OpenShell documentation — How OpenShell Works.
3. NVIDIA OpenShell documentation — Policies.
4. NVIDIA OpenShell documentation — Support Matrix and Linux requirements.
5. NVIDIA OpenShell GitHub repository.
6. NVIDIA OpenShell Sandbox Architecture documentation.
7. NVIDIA OpenShell Supervisor Middleware documentation.
8. SAFE AI Foundation “Total AI Safety Architecture”, October 2026. Available at:
https://files.safeaifoundation.com/Total_AI_Safety.pdf

9. SAFE AI Operating Point (SAIOP): The Upper Safe Limit of AI before humans lose control. Available at: <https://files.safeaifoundation.com/AI-SAIOP-paper2026.pdf>
10. SAFE AI Foundation's Input-Model-Output-Compliance AI Safety Model (AISM). See: <https://safeaifoundation.com/ai-benchmarking>
11. SAFE AI Foundation – “AI Safety is not AI Security”, Technical Report, September 2026. PDF available at: <https://files.safeaifoundation.com/AI-Security-Safety-Paper.pdf>
12. SAFE AI Foundation AI Safety Model (AISM) Introduced, 15 April 2025. See: <https://www.linkedin.com/posts/15-apr-2026-safe-ai-foundation-usa-we-share-7450263859778568192-G3nI/>
13. SAFE AI Foundation – “AI & The Law”, AI Digest, June 2025. See: <https://safeaifoundation.com/digest003>
14. SAFE AI Foundation – other technical documents: <https://safeaifoundation.com/technical-documents>

Disclaimer: This technical report is provided free of charge by the **SAFE AI Foundation USA**. It requires strong computer science background for better understanding. The information in this report is provided “as it is”, by **the SAFE AI FOUNDATION, USA**. The use of the information provided here is subject to the user’s own risk, accountability, and responsibility. The **SAFE AI FOUNDATION USA** and the author are not responsible for the use of the information by the user or reader. Please reference this article if you wish to cite it elsewhere.

Note: We do not work with private consultancy firms, universities, or any third parties to produce this technical document. All copyrights are reserved by the **SAFE AI Foundation USA**. Technical enquires about this work can be made via email to the author and foundation’s executive director Dr. Chai K. Toh at: contact@safeaifoundation.com



SAFE AI FOUNDATION

www.safeaifoundation.com

The SAFE AI Foundation was formed in 2025, in the heart of the Silicon Valley, California, USA. Its mission is to embrace the AI revolution and promote the use of good AI to improve on the quality of life and the economy. It will address the issues of bad AI and its impacts on society. The foundation will inform, educate, and mitigate the dangers of bad AI and help the society to embrace good AI. The foundation will interact with industries, universities and public sector to jointly advance the progress of Good AI. The foundation is a non-profit 501c(3) organization, registered in the State of California, USA.