Preparing Defenders of AI Systems, V 1.0

Workstream 2: Preparing Defenders of AI Systems

Version 5f82bf40a432c07d5d9c85d861227d0ba8b7139c

14 July 2025

Contents

Preparing Defenders of AI Systems, V 1.0

OASIS Open Project: Coalition for Secure AI (CoSAI), Workstream 2: Preparing Defenders of AI Systems

Approved by the CoSAI Project Governing Board on 14 July 2025

Executive Summary

At the intersection of innovation and security lies a critical challenge for today's enterprises: how to harness the transformative power of AI while establishing robust security frameworks to protect these valuable assets. As AI systems move from experimental projects to core operational infrastructure to agent-based systems, organizations face a security landscape that traditional governance models were never designed to address.

In security, defenses are only as strong as the weakest point. Comprehensive protection fails if a single vulnerability remains exposed. This reality intensifies with AI systems, where novel attack vectors emerge alongside traditional threats. Effective protection requires layered defenses—multiple independent barriers that maintain security even when individual safeguards fail.
Secure AI use and model security represent a rapidly evolving problem domain. Enterprises urgently need AI security products to close widening capability gaps. Industry professionals anticipate AI-enabled threats from malicious actors within 1-2 years, making continuous threat adaptation essential. AI security directly impacts enterprise risk, regulatory compliance, and business viability. Organizations that fail to secure their AI investments face both technical vulnerabilities and existential business threats.

This paper examines how enterprise AI adoption reshapes security requirements, offering leaders a practical roadmap through this complex terrain. The Coalition for Secure AI's (CoSAI) analysis reveals that while existing frameworks provide foundations, significant gaps remain in addressing AI's unique security challenges—gaps that demand immediate investment, research, and innovation.

Bridging the Gap: From Frameworks to Implementation

1. While our analysis of established security frameworks—including NIST, MITRE ATT&CK/ATLAS, and OWASP —reveals both valuable guidance and areas of investment in addressing these AI-specific security domains, the most pressing challenge for organizations is bridging the divide between theoretical frameworks and practical implementation.

2. CoSAI and this paper maintain a deliberate focus on the security of AI systems. Additionally, this paper's scope narrows to the defender's perspective and provides actionable guidance for security practitioners charged with protecting AI assets while enabling responsible innovation.

3. We provide defenders an approach for understanding which frameworks apply to their role

4. For security leaders this paper offers an assessment of current approaches, critical areas of investment, and practical steps forward in securing the enterprise AI journey.

5. In recognition of the rapidly evolving nature of AI and cybersecurity, this document is intended as a living resource. The Coalition for Secure AI (CoSAI) commits to continually updating its content to reflect emerging technologies, new threat vectors, and evolving best

practices, ensuring that our guidance remains current and actionable for all stakeholders.

The Growing Attack Surface

AI systems have dramatically expanded our digital attack surface in ways that upend traditional security paradigms. Unlike conventional software with predictable behavior boundaries, AI systems blur the lines between intended and unintended functionality, creating vulnerabilities that evolve based on interactions.

What makes this shift particularly challenging is that these systems don't just expose new attack vectors—they create emergent ones that evolve and adapt based on interactions, making traditional security approaches increasingly ineffective. In 2023, researchers extracted Bing Chat's confidential system prompts through carefully crafted queries—attacks that would be meaningless against traditional software. Samsung engineers accidentally leaked proprietary source code by pasting it into ChatGPT for debugging assistance, exposing intellectual property to an external AI system. Security researchers have also demonstrated prompt injection attacks that can exfiltrate data across chat sessions, turning helpful AI assistants into data extraction tools.

This expanded attack surface extends far beyond the models themselves. The entire AI lifecycle—data collection, training, inference, deployment, monitoring—introduces vulnerabilities that bypass conventional security controls. Each stage presents unique risks: poisoned training data that corrupts model behavior months later, supply chain attacks through compromised software packages, or runtime manipulation through adversarial inputs.

There are a few interconnected ways AI transforms organizational risk.

1. First, **AI Systems as a Target** means organizations deploying AI systems face entirely new categories of attacks—prompt injection, data extraction, model poisoning—that represent fundamentally new risk categories requiring specialized defenses.
2. Second, **AI Technologies as an Enabler of Cyber Attacks** reflects how attackers aren't just targeting AI systems but weaponizing AI to enhance traditional attacks, accelerating sophistication and scale beyond what human defenders can match alone.
3. Third, **AI Risks in Business Processes** emerge when AI systems approve software used by a company, review expenses or manage infrastructure—compromised models don't just leak data, they make bad decisions at the speed and scale businesses operate.

Each of these risks reinforce the need for defenders to ensure the integrity and protection of their organization's critical assets, specifically:

- Proprietary algorithms, internal IP and training data that differentiate products
- Employee data, company asset and infrastructure security
- Customer data protection from external and insider threats

The frameworks provide starting points—NIST establishes governance structures, MITRE ATLAS documents attack techniques, OWASP delivers practical implementation guidance. Yet significant opportunities for investment and research remain in translating these resources into operational reality.

Organizations must rethink security for AI's paradigm shift. For example, banks using LLMs need more than API security—they require output validation, data classification, and decision audit trails that traditional software never demanded. A healthcare startup analyzing patient data with LLMs must guard against model outputs that leak sensitive information across sessions. Traditional security assumes predictable software but AI demands dynamic defenses that understand context, intent, and emergent behaviors.

The Defender's Journey: From Current State to AI-Ready Security

The shift to AI security isn't abstract—it requires concrete changes across your security program. Here's what defenders must tackle today:

1. **Expand analysis of the threats.** There are security challenges and critical decisions throughout the AI system lifecycle, from the procurement of the model, to fine tuning and testing, through system deployment and monitoring.

   - **Risk assessments must evolve:** Traditional programs check for CVEs and misconfigurations. AI requires evaluating model lineage, supply chain dependencies, and behavioral drift over time. You need to know if a model was fine-tuned on public internet data that could include malicious examples.

   - **Vendor evaluations need new criteria:** Standard questionnaires miss AI-specific risks. How does the vendor handle open source model versioning, what's their incident response when models produce harmful outputs, can they provide model cards documenting limitations. These questions determine the risk you take on.

2. **Broaden security response programs** with more comprehensive logging, build MLSecOps expertise, and expand response playbooks to include AI system failures.
   - **EDR products and SOC teams** must expand their logging and monitoring efforts to address AI model threats, such as unauthorized inference attempts or API scraping activities aimed at extracting model functionalities. Traditional security monitoring focuses on network traffic and system calls; AI systems require tracking prompt patterns, response latencies that indicate extraction attempts, and usage anomalies that suggest model abuse.

   - **MLSecOps** is an emerging discipline focused on integrating security practices throughout the entire machine learning lifecycle, bridging the gap between data science and security operations. The discipline emerged from hard lessons: models deployed without security oversight led to data leakage through gradient inversion attacks, performance degradation from adversarial inputs, and compliance violations from drift in production. MLSecOps practices should be woven into each AI system lifecycle stage, integrating model validation, drift detection, and security monitoring into standard security workflows.

   - **Response playbooks need AI-specific incident procedures.** When a model starts leaking sensitive data in outputs or accepting malicious prompts, teams face new containment challenges. Response playbooks need AI-specific incident procedures. When a model starts leaking sensitive data in outputs or accepting malicious prompts, teams face new containment challenges. Immediate containment differs from traditional incidents. You can't isolate an infected endpoint or block malicious IPs when the threat operates through natural language. Teams must rapidly throttle API access to affected models, implement emergency output filtering for sensitive data patterns, and preserve prompt/response pairs for forensic analysis.

   - **Investigation requires new forensic approaches.** Traditional IOCs don't apply to prompt injection attacks. Teams must analyze prompt patterns across user sessions to identify attack signatures, correlate timeline with model behavior changes, and distinguish between coordinated attacks versus isolated incidents. Samsung's ChatGPT leak investigation revealed employees unknowingly exposed source code across multiple

sessions—highlighting why prompt history analysis is critical.

- **Recovery extends beyond system restoration.** After containing an AI incident, teams face unique challenges: determining which decisions the compromised model influenced, notifying affected users when AI-generated content was manipulated, and validating model behavior has returned to baseline. Unlike malware removal where you can verify a clean state, AI systems require behavioral validation over time.

- **Communication needs AI-literate escalation paths.** Standard vendor support won't recognize "model producing inconsistent outputs" as a security incident. Organizations need predetermined escalation criteria for AI-specific threats, security contacts who understand prompt injection versus model drift, and clear SLAs for AI security events distinct from availability issues.

3. **Secure the supply chain and introduction points of AI in your enterprise.** AI systems often rely on vast amounts of organization confidential or private data and incorporate open-source models, datasets, and tooling. Additionally, the adoption of AI agentic frameworks and protocols (such as MCP - Model Context Protocol, A2A - Agent-to-Agent, and ANP - Agent Negotiation Protocol) represent additional supply chain risks with unique challenges and requirements.
   - Organizations should adopt SLSA or SSDF-based approaches for all classical software application adoption and control

   - Enhance your supply chain control programs to include AI model and agent framework selection and acquisition

   - Establish organization guidelines and best practices for the review and selection of open source, open weight, and commercial AI models

   - **Note:** Supply chain security represents one of the most critical attack vectors in modern breaches—from SolarWinds to Kaseya. AI compounds these risks through opaque model lineages and emerging agent protocols. CoSAI maintains two dedicated workstreams addressing these challenges: our Supply Chain Security initiative develops comprehensive frameworks for model provenance and integrity, while our AI Agents workstream tackles the unique security requirements of autonomous agent systems. Organizations should monitor these workstreams for evolving guidance as the landscape rapidly develops.

4. **Establish AI governance and risk frameworks.** Beyond traditional threat analysis, organizations need AI-specific governance structures. Many organizations struggle with fundamental questions like "Who owns the risk when an AI system makes a bad decision?" Building these governance foundations early prevents larger compliance and liability issues later.
   - Develop risk appetite frameworks for AI deployments

   - Create approval processes that account for model uncertainty

   - Establish clear accountability for AI-driven decisions

5. **Adapt identity and access management for AI systems.** Traditional IAM architectures assume human users with predictable access patterns. AI systems operate differently—they access data continuously, make autonomous decisions, and interact with multiple systems simultaneously. AI systems can become autonomous decision-makers that require identity frameworks designed for continuous, context-aware operation rather than discrete human

interactions. This fundamental mismatch creates security gaps.

- Establish granular authentication for model endpoints—not all AI capabilities require equal access
- Implement dynamic authorization that adjusts based on data sensitivity and decision impact
- Deploy service-to-service authentication that tracks AI system interactions across your infrastructure
- Create specialized privilege tiers for AI agents based on their operational scope and data access requirements

6. **Enterprises should scale security with AI:** AI isn't only a new attack surface to defend—it's also becoming a powerful security tool. Just as AI enables cyber attackers, it can exponentially increase defender capabilities. Models can analyze vast amounts of data and detect anomalies that human analysts might miss, often identifying patterns across timeframes and data volumes that would be impossible to process manually. This makes AI an ideal technology for preventing many widely used attacks, from automated threat hunting to real-time behavioral analysis.
   - Every AI tool deployed for defense also expands your attack surface. The LLM analyzing security logs could leak sensitive data. The AI-powered SIEM might be fooled by poisoned telemetry. Success requires treating AI security tools with the same rigor as any critical infrastructure—isolated deployments, strict access controls, and continuous monitoring for adversarial manipulation. Start with low-risk use cases like alert enrichment where human analysts validate outputs, then gradually expand to autonomous operations as your team builds expertise in both using and securing AI systems.

AI System Security: A Team Sport

The successful implementation of secure AI systems requires a coordinated approach across multiple organizational roles.

1. **Strategic Leadership:** Successful AI security requires coordinated leadership across security and AI development functions. Strategic leadership must bridge those who build AI capabilities and those who protect them—security becomes an ivory tower exercise without operational input, while unsecured AI initiatives create critical vulnerabilities. Effective leadership pairs CISOs who translate technical risks into business impact with AI leaders who embed security into their innovation roadmap. Executives establish governance that enables progress rather than blocking it. The key insight: AI security cannot be bolted on after deployment. It requires security experts who understand AI's unique risks working directly with AI practitioners who grasp security's operational constraints.

2. **Implementers and Engineers:** Technical implementation teams translate strategic security requirements into engineering reality. These professionals design, deploy, and maintain the technical controls that protect AI systems throughout their lifecycle. They bridge the gap between high-level security policies and practical system protections, ensuring robust architectural foundations while adapting to emerging threats. Examples of technical implementation include Service Architects who design secure AI architectures, Data Architects who ensure secure infrastructure and data pipelines, Security Architects who identify and mitigate AI-specific vulnerabilities, Developers and Software Engineers who build the overall systems.

3. **Operations and Monitoring:** Operations and monitoring professionals maintain the day-to-day security posture of AI systems in production environments. They provide continuous vigilance through real-time threat detection, incident response, and proactive hunting for security anomalies. Their expertise ensures AI systems remain protected against both known and emerging threats while maintaining operational efficiency. Examples of operations and monitoring include SOC Operations teams who continuously assess and respond to AI security threats, and Site Reliability Engineers and Service Operations teams who ensure secure and compliant AI service deployment and monitoring.

4. **Compliance and Governance:** Compliance professionals ensure AI systems meet regulatory requirements and internal policies. They implement model governance frameworks including version control, lineage tracking, and audit trails. Key challenges include translating traditional compliance frameworks to AI contexts—proving a model's decision was "fair" differs fundamentally from proving a system logged access correctly. They establish AI-specific controls: model validation pipelines, staged deployments with rollback procedures, and documentation standards through model cards. The critical shift: compliance moves from checking boxes to continuous monitoring, as model behavior can drift post-deployment even without code changes.

5. **Security Validation and Red Teams:** Security validation teams stress-test AI systems through adversarial testing. They execute prompt injection campaigns, attempt model extraction through API queries, and probe for data leakage in model outputs. Unlike traditional penetration testing that exploits code vulnerabilities, AI red teams exploit behavioral patterns—crafting inputs that make models reveal training data or bypass safety controls. They develop attack scenarios specific to AI: poisoning attacks that corrupt model behavior over time, evasion techniques that fool classification systems, and extraction methods that steal proprietary model intelligence. The key insight: AI systems fail differently than traditional software. Success requires red teams who understand both security principles and machine learning internals—knowing how gradient descent works helps craft better adversarial examples.

A Tale of Threats and Frameworks

The threat landscape for AI systems is both complex and rapidly evolving. Organizations deploying AI systems face multifaceted risks that span technical vulnerabilities, operational challenges, and business continuity threats. From infrastructure security issues that exploit AI systems to supply chain compromises that can poison entire AI deployments, these risks require a nuanced understanding of how each affects different stakeholders. Since these risks are continually evolving we are moving the relevant information to the repository of CoSAI's common artifacts in Github. Details on these threats can be found here.

The AI security framework landscape reflects our collective struggle to secure systems that fundamentally challenge traditional security paradigms. Each framework emerged from specific pain points—NIST after federal agencies struggled with AI governance, MITRE ATLAS following real-world AI attacks like model extraction from prediction APIs, OWASP responding to LLM-style prompt injection incidents flooding their forums.

These frameworks provide essential building blocks: structured risk taxonomies, attack pattern libraries, and implementation checklists. NIST's AI RMF offers comprehensive governance structures tested across federal deployments. MITRE ATLAS documents 80+ adversarial techniques observed in production environments. OWASP's LLM Top 10 distills thousands of vulnerability reports into actionable controls. AWS and Google frameworks translate cloud-specific lessons from millions of model deployments into security architectures.

However framework implementation reveals critical friction points. Organizations discover that NIST's "Map-Measure-Manage-Govern" lifecycle assumes traditional software release cycles, not continuous model updates. MITRE ATLAS catalogs attacks but lacks the defensive playbooks that make ATT&CK actionable. OWASP's guidance targets application security teams unfamiliar with model versioning or drift detection. When frameworks overlap—like NIST's "Model Security" versus ATLAS's "ML Attack Staging"—teams waste cycles reconciling terminology instead of building defenses.

The abstraction mismatch proves most challenging. Strategic frameworks speak to boards about risk appetite while technical frameworks detail numpy array manipulations. The gap between "establish AI governance" and "implement gradient masking" leaves security teams improvising critical implementation details. Organizations need translation layers that don't yet exist.

Details on the frameworks can be found here.

Using these insights, we analyzed leading frameworks[1] focusing on practical implementation gaps and defender needs. Our detailed framework analysis, available in our Github repository here, examines each framework's strengths, limitations, and optimal use cases.

Observations on the general state of AI security frameworks

The critical blind spots: Current frameworks were built on traditional software security foundations—a reasonable starting point that misses three AI-specific realities. First, AI systems exhibit emergent behaviors where capabilities and vulnerabilities appear only through interaction patterns, impossible to predict through static analysis. Second, the security perimeter dissolves when models process untrusted input as both data and instruction. Third, traditional versioning concepts break down when identical models produce different outputs based on temperature settings, random seeds, or conversation history.

The path forward requires framework convergence, not proliferation. The security community needs unified taxonomies for AI threats, standardized testing methodologies for model robustness, and shared baselines for acceptable AI system behaviors. Progress is happening—OWASP's LLM Top 10 shows how focused efforts can provide immediate value. Building on these successes while acknowledging fundamental AI differences will help defenders construct more stable foundations.

Our thoughts on AI security frameworks:

Government Frameworks: Structure with Gaps

- **NIST's suite of frameworks provides the strongest foundation for systematic AI risk management.** The AI RMF 1.0 offers comprehensive governance structures, while the Generative AI extension addresses specific challenges like prompt injection and output manipulation. However, these frameworks excel at risk identification but fall short on implementation specifics. Organizations following NIST guidance understand what risks they face but often struggle with practical mitigation strategies.

- **CISA's Zero Trust Maturity Model** explicitly excludes AI integration recommendations, creating a notable blind spot. While useful for securing access to AI systems, it doesn't address how to apply zero trust principles to model behavior or inference processes—a critical oversight as AI systems increasingly make autonomous decisions.

Industry Frameworks: Practical but Fragmented

- **MITRE's ecosystem demonstrates both maturity and evolution.** ATT&CK provides excellent coverage for traditional infrastructure threats but requires ATLAS integration for AI-specific

attacks. ATLAS itself documents adversarial tactics effectively but lacks the defensive countermeasures that make D3FEND valuable for traditional security. This creates an implementation gap—defenders can understand attack patterns but struggle to translate that knowledge into protective measures.

- **OWASP Top 10 for LLMs** succeeds because it prioritizes actionable guidance over comprehensive coverage. Organizations can immediately address prompt injection and excessive agency risks. However, the framework's narrow focus on LLMs misses broader AI security challenges and doesn't integrate well with enterprise risk management processes.

- **AWS Generative AI Security Scoping Matrix** offers valuable strategic guidance for determining security responsibilities across deployment models. Yet it provides limited tactical implementation details, forcing organizations to combine it with other frameworks for complete coverage.

Academic and Research Contributions: Lacking Practical Guidance

- **MIT AI Risk Repository** catalogs over 1,000 AI risks with sophisticated taxonomies, making it invaluable for comprehensive risk assessment. However, it provides minimal guidance for defenders seeking practical mitigation strategies. The repository works best as a risk discovery tool rather than an implementation guide.

AI System Security - The Fundamental Paradigm Shift

A core security principle mandates separation of code and data. AI models, and thus AI systems, fundamentally violate this principle by converging control and data planes within applications and infrastructure. Traditional security architectures maintain clear boundaries between application control logic and processed data—a separation critical for preventing unauthorized access or manipulation.

AI systems dissolve this distinction by integrating control and user-provided data within the model architecture and exposing inference processes directly to users. The use of externally hosted models compounds these concerns due to the lack of visibility into the hosting entities security practices and the use of sensitive organization data with these systems.

AI System Security - Looking Forward*

In conclusion, the emerging security realities of AI systems and the findings from our framework survey highlight the critical need for organizations to invest in AI-specific security improvements, below. Such investments are essential to empower their defenders to maintain an effective risk posture amid widespread AI adoption.

- **AI Asset Inventory and Attack Surface Management**
  - You can't protect what you can't see. Most organizations have limited visibility into their AI systems — models hidden in departments, shadow deployments, unknown dependencies.

  - **Why this matters:** Unknown risks could be an organization's biggest vulnerability. Without comprehensive prioritized risk inventory, you're essentially leaving doors open you don't even know exist.

- **AI-Specific Incident Response, Recovery, Monitoring and Detection**
  - When an AI system is compromised, traditional playbooks fall apart. How do you contain a poisoned model? How do you even know it's poisoned? Security teams drown in

AI telemetry without understanding what's normal versus malicious. Traditional SIEM rules don't catch prompt injection or model extraction attempts.

  - **Why this matters:** The difference between a contained incident and a catastrophic breach often comes down to response time. Attackers operate freely while your monitoring tools generate noise. Without AI-specific playbooks, teams waste precious hours figuring out what to do.

- **Enterprise AI Threat Modeling Methodologies**
  - We are trying to model AI threats using software threat frameworks. It's like using a road map to navigate the ocean. Organizations throw controls at AI without strategy. No threat progression model, no capability prioritization, just reactive patching and cleanups.

  - **Why this matters:** Misunderstood threats lead to misallocated resources. You end up overprotecting the wrong things while leaving critical vulnerabilities exposed. Without maturity models, you can't demonstrate progress towards improvements in security. Security becomes a cost center instead of an enabler.

- **Zero Trust Architecture Evolution for AI Systems**
  - Current frameworks assume fixed identities and predictable behaviors. But AI systems make autonomous decisions, shifting trust boundaries in real-time.

  - **Why this matters:** Without AI system specific Zero Trust controls and architectures, a compromised AI system can cascade through your entire system, making decisions it shouldn't, accessing data it shouldn't see, all while appearing legitimate to traditional controls.

Contributors and Acknowledgements

Workstream Leads

- Josiah Hagen, Trend Micro (Josiah_Hagen@trendmicro.com)
- Vinay Bansal, Cisco (vibansal@cisco.com)

Contributors

- Irakle Dzneladze (irakledprof@gmail.com)
- Shrey Bagga (sbagga@cisco.com)
- Matt Saner (msaner@amazon.com)
- Michael Rash (Michael.Rash1@Dell.com)
- Rob Mann (robmann@google.com)
- Yuval Bercovich (ybercovich@paypal.com)
- Jason Garman (garmaja@amazon.com)
- Vladimir Kropotov (Vladimir_Kropotov@trendmicro.com)
- Fyodor Yarochkin (Fyodor_Yarochkin@trendmicro.com)

- Krishna Yellepeddy (kyellepe@us.ibm.com)

Editors

- Akila Srinivasan, Anthropic (akila@anthropic.com)
- J.R. Rao, IBM (jrrao@us.ibm.com)

Reviewers

- David LaBianca, Google (ddlb@google.com)
- Alex Polyakov, Adversa AI (alex@adversa.ai)

Technical Steering Committee Co-Chairs

- Akila Srinivasan, Anthropic (akila@anthropic.com)
- J.R. Rao, IBM (jrrao@us.ibm.com)

Appendix

CoSAI Focus

CoSAI is an OASIS Open Project, bringing together an open ecosystem of AI and security experts from industry-leading organizations. The project is dedicated to sharing best practices for secure AI deployment and collaborating on AI security research and product development. The scope of CoSAI is specifically focused on the secure building, integration, deployment, and operation of AI systems, with an emphasis on mitigating security risks unique to AI technologies. Other aspects of Trustworthy AI are deemed important but beyond the scope of the project including, ethics, fairness, explainability, bias detection, safety, consumer privacy, misinformation, hallucinations, deep fakes, or content safety concerns like hateful or abusive content, malware, or phishing generation. By concentrating on developing robust measures, best practices, and guidelines to safeguard AI systems against unauthorized access, tampering, or misuse, CoSAI aims to contribute to the responsible development and deployment of resilient, secure AI technologies.

Guidelines on usage of more advanced AI systems (e.g. large language models (LLMs), multi-modal language models. etc) for drafting documents for OASIS CoSAI:

tl;dr: CoSAI contributions are actions performed by humans, who are responsible for the content of those contributions, based on their signed OASIS iCLA (and eCLA, if applicable). [Each contributor must confirm whether they are entitled to donate that material under the applicable open source license; OASIS and the CoSAI Project do not separately confirm that.] Each contributor is responsible for ensuring that all contributions comply with these AI use guidelines, including disclosure of any use of AI in contributions.

- Selection of AI systems: CoSAI recommends the use of reputable AI systems (lowering the risk of inadvertently incorporating infringing material).
- Model constraints: Currently, CoSAI or OASIS are not required to have a contract or financial agreement for using AI systems from specific vendors. However, CoSAI editors should consider employing varying tools to avoid potential fairness concerns among vendors.
- IP infringement: It is the responsibility of the individual who subscribes/prompts and receives a response from an AI system to confirm they have the right to repost and donate the content to OASIS under our rules.

- Transparency: CoSAI's goal will be to maintain transparency throughout the process by documenting substantial use of AI systems whenever possible (e.g., the prompts and the AI system used), and to ensure that all content, regardless of production by human or AI systems, was reviewed and edited by human experts. This helps build trust in the standards development process and ensures accountability.

- Human-edited content and quality control: CoSAI mandates human-reviewed or -edited results for any final outputs. A robust quality control process should be in place, involving careful review of the generated content for accuracy, relevance, and alignment with CoSAI's goals and principles. Human experts should scrutinize the output of AI systems to identify any errors, inconsistencies, or potential biases.

- Iterative refinement: The use of AI systems in drafting standards should be seen as an iterative process, with the generated content serving as a starting point for further refinement and improvement by human experts. Multiple rounds of review and editing may be necessary to ensure the final standards meet the required quality and reliability thresholds.

Copyright Notice