Mine Guard AI

MineGuard Solutions Pty Ltd – AI Risk & Compliance Statement (NIST AI RMF 1.0)

Product: Incident AI
Version: 1.4 (April 2025)

Purpose and Context

Incident AI is a generative AI-powered investigation tool built to assist mining organizations in analyzing incident data, producing consistent event narratives, identifying contributing factors, and supporting organizational learning. It supports ICAM-aligned methodologies and is used as part of post-incident analysis workflows.

This document outlines how Incident AI adheres to the **NIST AI Risk Management Framework (AI RMF 1.0)** by addressing AI-specific risks and aligning with principles for trustworthy, responsible AI.

Govern

NIST Governance Focus: AI policies, accountability, lifecycle management, supply chain, and risk culture

- **Governance Structure**: Incident AI operates under a defined governance model. All AI system updates, data processing protocols, and use-case deployments are reviewed by technical director and safety domain experts.
- **Legal Compliance**: Our AI services are used under Australian safety law and comply with privacy legislation including the Australian Privacy Act and GDPR.
- **Supply Chain Risks**: All third-party AI models (e.g., OpenAI GPT-4, Claude 3.5 by Anthropic, Azure OCR) are independently assessed and selected based on their published safety, privacy, and data retention guarantees.
- **Lifecycle Management**: Version control, sunset procedures, and decommissioning processes are in place for all modules, ensuring outdated models are not exposed to users.
- **Fair Use Policy**: AI usage is capped under contract to a percentage of subscription value to control compute costs and manage ethical use of powerful generative systems.

Map

NIST Mapping Focus: Contextual understanding, intended use, limitations, stakeholders

- **Use Case Framing**: Incident AI is designed exclusively for post-incident analysis, safety investigations, and learning. It is not used for decision automation or HR evaluation.

Mine Guard AI

- **Stakeholder Engagement**: Feedback loops are maintained with frontline investigators, Site Senior Executive's, and safety managers to understand how the tool affects workflows, trust, and culture.
- **Limitations Disclosure**: Users are informed that generative models may produce hallucinations or blank outputs and are prompted to verify results. The tool also includes re-run options and quality control overlays.
- **Training Transparency**: The models used do not train on user data (OpenAI and Anthropic APIs operate under no-training, enterprise-grade agreements).
- **Risk Mapping**: Tools are tested for unintended consequences such as bias, over-simplified causality, or premature blame. Design prioritizes context-first prompts. Multiple tools exists to cross reference input files with bias.

Measure

NIST Measurement Focus: Testing, accuracy, robustness, explainability, bias management

- **Model Accuracy & Reproducibility**: All outputs (e.g., 5-Why analysis, PEEPO tables) are generated from the same inputs deterministically to support traceability.
- **Explainability**: Each major insight includes a "source attribution" showing which parts of the uploaded evidence informed the result.
- **Error Handling**: Known issues like blank outputs on first run are documented with suggested user actions and are being addressed with retry logic and dual-model support.
- **Bias Mitigation**: Language models are prompted with framing that avoids assumptions or blame. Users are encouraged to review system and contextual factors.
- **Independent Review**: Human-led review and feedback features are embedded to cross-check AI conclusions before final reporting.

Manage

NIST Management Focus: Risk prioritization, incident response, improvement cycles

- **Risk Triage**: The tool does not automate critical decisions. All AI-derived outputs are designed as drafts requiring human verification.
- **Incident Learning Loops**: Incidents analyzed through Incident AI are tagged and archived for learning pattern analysis across time.
- **Continuous Improvement**: Regular model evaluations, user feedback, and prompt tuning cycles are maintained to improve relevance, safety, and trustworthiness.
- **Trial Disclosure**: Users of free trials must sign a Service Agreement acknowledging known limitations and confirming data suitability.

Mine Guard AI

- **Deactivation Controls**: MineGuard can immediately disable modules or endpoints found to be producing unsafe or non-compliant content.

Trustworthiness Characteristics (NIST AI RMF 3.x Mapping)

Trust Characteristic	How Incident AI Addresses It
Valid & Reliable	Outputs are reproducible, sources cited, and feedback loop embedded
Safe	Human-in-the-loop model, designed for assistive—not automated—use cases
Secure & Resilient	Hosted on secure Render platform with ephemeral file handling and encrypted APIs
Accountable & Transparent	Prompt logs, user actions, and output audit trails are retained and reviewable
Explainable & Interpretable	Summaries come with source references and logic previews for AI-generated outputs
Privacy-Enhanced	Data encrypted in transit and not stored. No model training on customer data.
Fair & Bias-Managed	Language and framing explicitly designed to avoid premature judgment or bias

---

Reputational Risk Mitigation

Incident AI is designed to protect both individual and organizational reputations through careful prompt engineering, language control, and human-in-the-loop oversight. All outputs are reviewed for tone and neutrality, and the system emphasizes learning over blame. By standardizing investigations with consistent, system-focused analysis and encouraging reflective, respectful questioning, Incident AI helps prevent reputational harm that can arise from poor investigations, reactive conclusions, or emotionally charged reports. Any AI-assisted content presented for organizational reporting includes clear attribution as AI-generated drafts, further reinforcing accountability and transparency.

Summary

MineGuard Solutions is committed to responsible AI use in mining safety applications. Incident AI is designed with transparency, safety, and human-centric decision support in mind, and adheres closely to the NIST AI RMF 1.0 framework. As the system evolves, we will continue to strengthen our governance and measurement strategies in line with NIST guidance and international safety expectations.