

Safe.ai: AI Agent Risk Assessment Platform

Weiliang Ho
National University of
Singapore

Aryan Thakar
University of Maryland,
College Park

AI Insurance Platform for Domain-Specific Agents

- Proposes an insurance platform for AI agents to ensure safety, reliability, and compliance in high-stakes domains like healthcare and legal AI.
- Integrates performance-based scoring, risk modeling, and tailored insurance coverage to mitigate critical failures and financial risks.
- Leverages advanced statistical models and real-time analytics to evaluate agent vulnerabilities and calculate premiums.
- Promotes responsible AI use while unlocking a unique market opportunity with high profit potential in the growing AI ecosystem.

1. Problem overview

The rapid adoption of AI in critical industries like healthcare and legal services has highlighted the urgent need for robust risk mitigation mechanisms. While domain-specific AI agents offer efficiency, they often lack transparency and accountability, raising concerns about safety, reliability, and compliance. The stakes are high, as AI failures in these sectors can lead to catastrophic outcomes, including loss of life, legal repercussions, and significant financial and reputational damage. Current solutions, such as regulatory frameworks and quality assurance protocols, provide only partial protection against the multifaceted risks associated with AI deployment. This situation underscores the necessity for an innovative approach that combines comprehensive risk assessment with financial safeguards to ensure the responsible and secure implementation of AI technologies across high-stakes industries.

Understanding the Problem

1. **Opaque Risk Profiles:** AI agents operate as black boxes unlike traditional assets. It is often unclear how they make decisions, which can lead to unpredictable behavior in high-stakes scenarios.
2. **High-Stakes Consequences:** Failures of AI systems in domains such as healthcare or legal services can result in life-threatening errors, regulatory violations, or substantial financial losses. For instance:
 - A healthcare diagnostic AI misidentifies early-stage cancer, leading to incorrect treatment or delays.
 - A legal AI agent generates faulty compliance advice, exposing firms to penalties and lawsuits.
3. **Lack of Financial Safeguards:** Companies deploying AI agents in these domains face significant financial exposure without mechanisms to mitigate risks. While traditional insurance models exist for physical and operational risks, they are not designed to cover the unique vulnerabilities of AI systems.

2. Your solution

We propose an AI insurance platform tailored to domain-specific agents, designed to ensure trust and accountability while fostering innovation. The platform integrates three key components:

1. **Performance-Based Scoring:** Each agent undergoes rigorous evaluation to assign a safety and performance score, reflecting its reliability and adherence to domain-specific requirements.
2. **Risk Modelling Expertise:** Leveraging statistical modelling and real-time analytics, the platform predicts potential risks and identifies critical failure points, enabling proactive measures to prevent harm.
3. **Insurance Coverage:** Companies purchasing agents receive insurance coverage based on the agent's performance score. This protects them from financial losses due to critical failures, creating a safety net for high-stakes AI deployment.

Drawing inspiration from the success of Risk Management Solutions (RMS) in natural catastrophe modelling, our platform aims to persuade reinsurers and stakeholders by demonstrating robust methodologies and evidence-based results. RMS built trust over decades by crafting synthetic populations of catastrophe events; similarly, our platform will simulate AI failure scenarios to validate its risk models and establish credibility.

3. Pilot experiment or demo

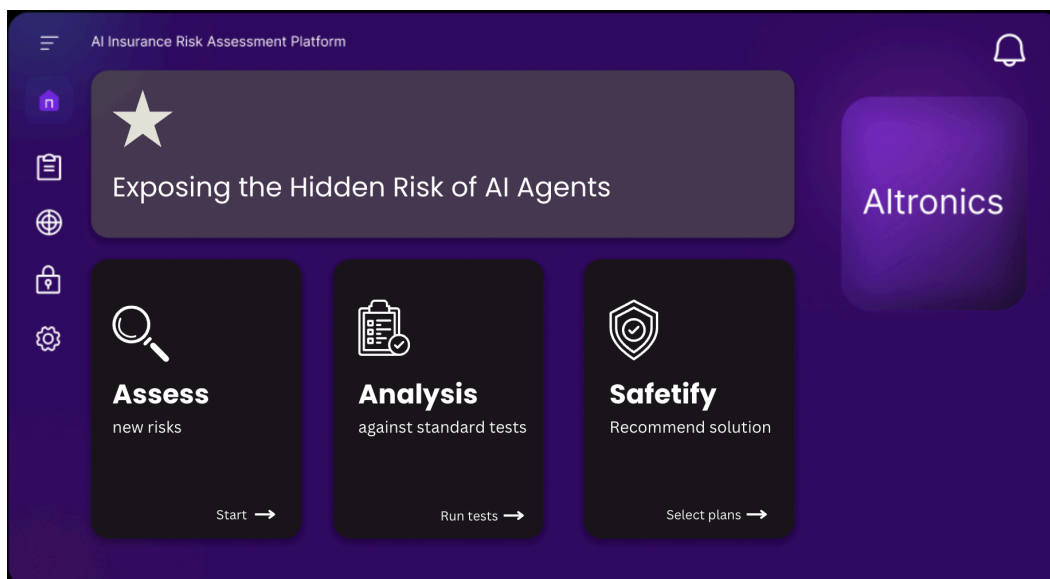


Figure 1 – AI Insurance Risk Assessment Platform

The initial pilot will target healthcare AI startups using predictive algorithms in diagnostics. Collaborating with these startups, we will:

- **Performance-Based Scoring:** Rigorous evaluation assigns safety and

- reliability scores to AI agents, reflecting their risk levels.
- **Risk Modeling Expertise:** Advanced statistical models and real-time analytics identify vulnerabilities and predict potential failures.
- **Insurance Coverage:** Tailored insurance plans are provided based on risk assessments, protecting companies from critical AI failures.

For instance, consider a diagnostic AI agent used in detecting early signs of cancer. Our platform would analyze its accuracy, sensitivity, and compliance with regulatory standards. If the agent meets high safety thresholds, we provide insurance coverage to the healthcare provider, ensuring protection against potential litigation or operational disruptions caused by misdiagnosis.

4. Process

Timeframe	What will you do?
February 2025	Focus on analyzing growth potential, scalability, technical feasibility, and market size to strengthen the business case. Refine the business model and establish strategic partnerships with insurance providers for market relevance. Join an InsurTech accelerator to connect with key stakeholders and begin prototype development, enhancing the MVP with advanced features.
May 2025	This phase will focus on validation by gathering comprehensive feedback from various stakeholders, including potential customers, insurers, reinsurers, and end consumers . The feedback will assess the solution's usability, effectiveness, and market fit, ensuring it addresses stakeholder needs comprehensively.
August 2025	To emphasize refining the platform based on stakeholder input , ensuring alignment with market expectations and industry standards. The plan will then be executed to deliver a user-centric solution .
2026	The focus will shift towards achieving profitability by fully commercializing the solution. Strategic partnerships and refined business strategies will support scaling operations, capturing a significant share of the AI insurance and risk management market.

5. Impact on AI safety & key risks

Our solution directly addresses critical concerns surrounding AI safety and ethical deployment. By embedding risk analysis into the core of the insurance platform, we promote transparency and accountability. This approach incentivizes companies to

prioritize safety, as higher-performing agents lead to lower insurance premiums and stronger market reputations.

In high-stakes domains, such as autonomous driving or financial trading, our platform's ability to predict and mitigate risks could prevent billions in damages annually. Moreover, the integration of governance frameworks and real-time analytics ensures compliance with global AI regulations, fostering a culture of responsible innovation.

Our product directly aligns with the industry safety standards specified in the OECD principles on Artificial Intelligence. These include:

- Addressing safety concerns by overriding, repairing, or decommissioning an AI system that causes undue harm and/or undesired behavior
- Safeguarding information integrity in the context of generative AI
- Cooperation with suppliers of AI knowledge, AI resources, AI system users, and any other stakeholders.
- Clarifying information regarding transparency of AI systems.
- Clarifying environmental sustainability
- Promote collaboration between jurisdictions for working together on policy for AI.

Our product helps ensure that a given AI agent is following these principles.

Moonshot's Web UI enables visualization test results in charts that are easy to read.

In conclusion, our project bridges the gap between AI development and risk management, offering a comprehensive solution that not only safeguards stakeholders but also propels the ethical and sustainable advancement of AI technologies.

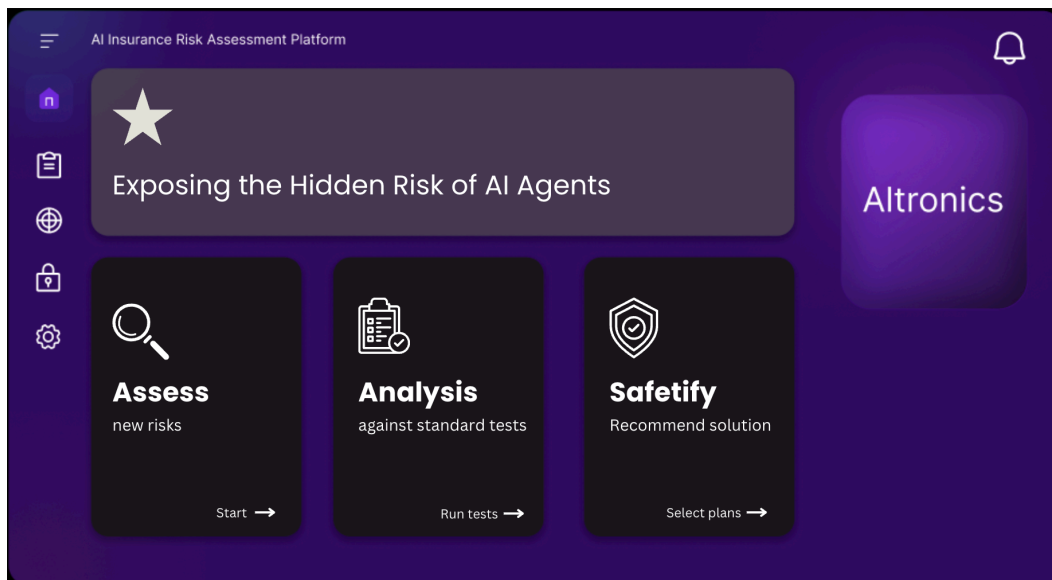
6. Appendix

This project stands out because it directly addresses an urgent, real-world problem: ensuring the safe and responsible deployment of AI agents. The market value for AI insurance is immense, given the exponential growth of AI adoption in high-stakes domains. By pioneering this field, our solution captures a unique market opportunity with enormous profit potential while providing measurable societal benefits.

According to an [AI Assurance Technology Market Report](#), the global AI Assurance Technology (AIAT) is expected to double each year for the rest of this decade, and is projected to reach \$276 billion by 2030 (from only 1.63 billion in 2023).

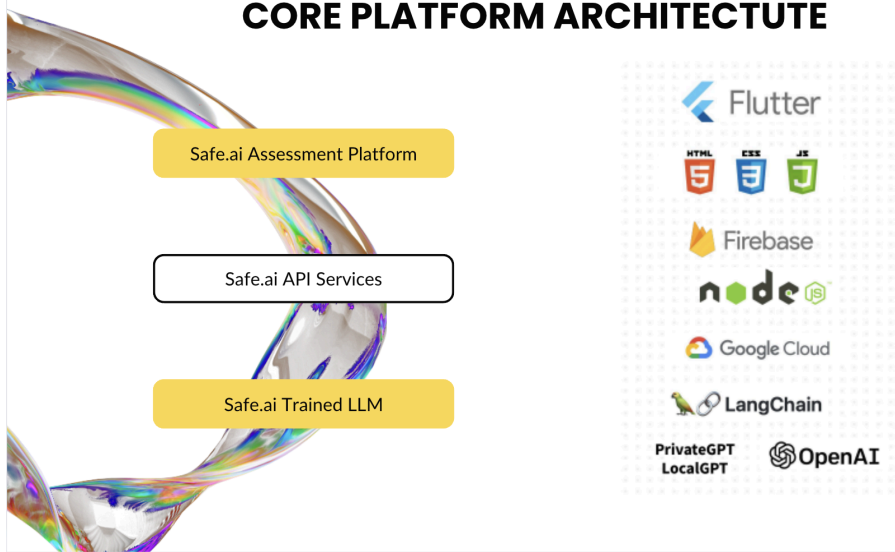
Historical examples like RMS illustrate the profitability of specialized risk modelling, while our addition of an insurance component ensures sustained revenue streams and competitive differentiation. Furthermore, this solution aligns with industry demands for transparency and safety, establishing a robust framework for long-term trust and innovation.

With the AI industry rapidly evolving, our approach enables businesses to deploy AI responsibly while minimizing risks, ultimately creating a safer and more sustainable technological future.



COMPETITIVE ADVANTAGE		
	AI Verify's Project Moonshot	Altronics's AI-Safety Platform
Test GenAI/LLMs	✗	YES, this is our main objective!
Define AI Standard	✗	YES, we are working on it!
Risk-Free	✗	This is possible for sure!

CORE PLATFORM ARCHITECTURE

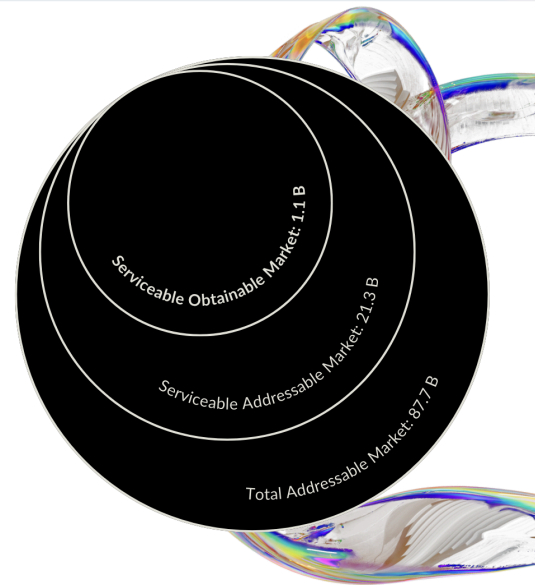


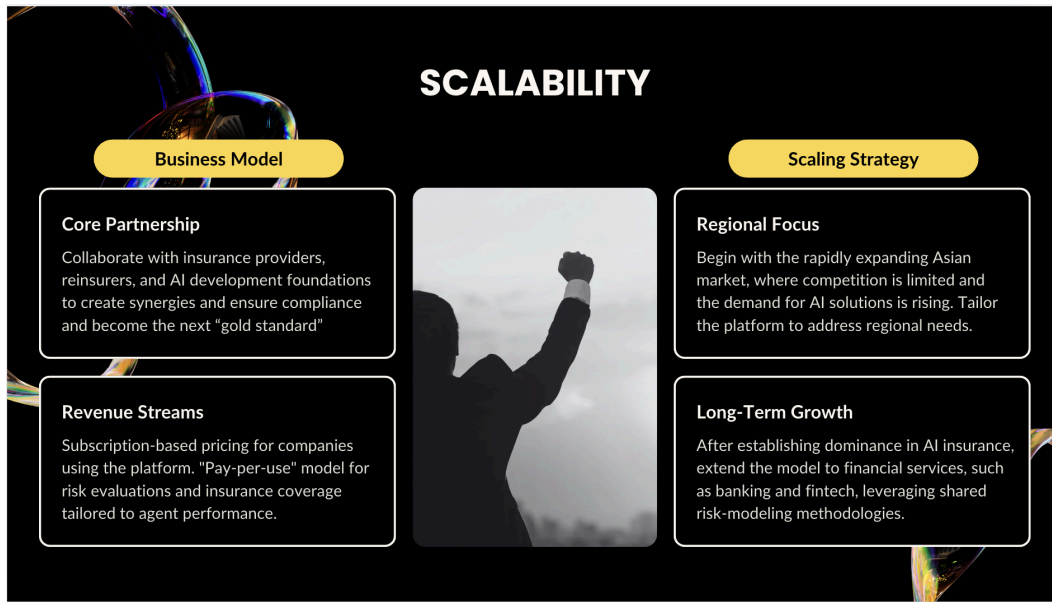
POTENTIAL MARKET SIZE

an estimated value of US\$87.65 billion by 2025 and a projected compound annual growth rate (CAGR) of 27.76% from 2025 to 2030, resulting in a market size of US\$298.40 billion by 2030. Within this, the AI safety segment is expected to reach US\$21.33 billion by 2025, highlighting significant demand for solutions that mitigate risks associated with AI deployments. By capturing even 5% of the AI safety market (US\$21.33 billion), the platform has the potential to generate **US\$1.07 billion in revenue**.

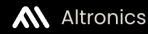
Data from:

in 2030 **statista**





MEET THE TEAM



LIAM HO

CEO & Full-stack AI
Engineer
(in AI-Safety)



ARYAN THAKAR

CTO & Computer
Scientist
(AI Trustworthy in Law)



JAMES GILLESPIE

Healthcare Advisor



**McKinsey
& Company**



NUS
National University
of Singapore

