

What's the difference between the police and a travel agent?



What's the difference between the police and a travel agent?

Answer: A lot



What's the difference between the police and a travel agent?

Answer: A lot

So, why would they have the same AI safety requirements?



What's the difference between the police and a travel agent?

Answer: A lot

So, why would they have the same AI safety requirements?

01

Well, they don't.

In fact, these organisations already know this - they've written policies that are unique to their organisations

02

But there's a gap between these policies, and the various technical AI safety controls that are needed in an AI agent

03

Key points:

- Each organisation is different, and has their own AI policies
- Each AI agent is different, and has different attack vectors

04

And this is where the asago community comes in.

asago





Dr Stuart Battersby

AI Safety Architect @ Red Hat
Founding Steering Committee Member @ asago





Connect

Your AI, your rules: putting safety and security to the test

An introduction to the asago community

Dr Stuart Battersby
AI Safety Architect





Connect

Your AI, your rules: putting
safety and security to the test

An introduction to the asago community

Dr Stuart Battersby
AI Safety Architect

Launched in August 2026...

The screenshot shows a Red Hat press release page. The header includes the Red Hat logo and navigation links for AI, Hybrid cloud, Products, Training, Learn, and Partners. The main title is "Red Hat Launches asago Community to Automate AI Safety and Governance from Policy to Production". Below the title is a summary paragraph. The main body of the article starts with "RALEIGH, N.C. – August 4, 2026 – Red Hat, the world's leading provider of open source solutions, today announced the formation of asago, an open source community project intended to automate how AI governance policies become product-ready, safely-deployed AI systems." It then defines "What is asago?" and lists four primary stages: 1. Risk mapping, 2. Risk assessment, 3. Risk mitigation, and 4. Production deployment. On the right side, there are sidebars for "In short", "Mentioned in this article", and "For more information".

Red Hat

AI Hybrid cloud Products Training Learn Partners

Console Docs Support

Press releases > Red Hat Launches asago Community to Automate AI...

Red Hat Launches asago Community to Automate AI Safety and Governance from Policy to Production

New collaborative open source project brings together Red Hat, Alquimia AI, Brave Software, EvalEval coalition, IBM Research, Interdisciplinary Transformation University Austria, Microsoft, MIT Lincoln Laboratory, North Carolina State University, NVIDIA and The Alan Turing Institute to bridge the gap between AI policy requirements and safely deployed AI systems

RALEIGH, N.C. – August 4, 2026 – Red Hat, the world's leading provider of open source solutions, today announced the formation of asago, an open source community project intended to automate how AI governance policies become product-ready, safely-deployed AI systems. asago connects the fragmented steps, tools and requirements of engineer and compliance teams, to create an automated, auditable and traceable workflow. The intent is to help deliver safer, production-ready AI systems that fuel innovation in days, not months or years.

What is asago?

asago (AI Safety And Governance Orchestration) plans to use a standardized open source platform and automated workflow to translate complex corporate and regulatory AI governance policies into actual operational controls. It builds on Red Hat and NVIDIA's work as members of the [Open Secure AI Alliance](#), and aligns developer flexibility with operator governance across four primary stages:

- 1. Risk mapping:** The framework automatically reads and interprets uploaded AI governance policies, mapping an organization's specific requirements directly to established AI Risk frameworks, ontologies and standards, such as the NIST AI RMF, OWASP LLM Top 10, EU AI Act via the IBM AI Risk Atlas – turning policy language into actionable risk profiles.
- 2. Risk assessment:** asago generates and executes use-case specific scenarios for automated safety testing tailored to identified risks, probing for harmful behaviors rather than relying solely on generic benchmarks.
- 3. Risk mitigation:** The project then recommends mitigations, including safety guardrails based on testing, creating a clear rationale and audit trail ready for review.
- 4. Production deployment:** asago then orchestrates recommended controls into deployment-ready configurations across platforms including hybrid cloud and Kubernetes, eliminating manual infrastructure coding.

Each of these steps will be built to produce a continuous audit trail, tying individual policy clauses directly to tests and runtime controls, enabling reviewers to see exactly what risk each action addresses. This helps to transform AI safety into a more predictable and governed enterprise utility.

Why does asago matter?

Translating abstract policy guidelines into functional software configurations slows AI innovation and introduces further risk from human miscommunication and misunderstandings. Compliance officers require rigorous risk assessments and verifiable

In short
Red Hat announces the formation of asago, an open source community project intended to automate how AI governance policies become product-ready, safely-deployed AI systems.

Mentioned in this article
Open source

For more information
• [Read the blog](#)
[Contact Red Hat PR](#) →

Subscribe to the feed

Press release: Red Hat Launches asago Community to Automate AI Safety and Governance from Policy to Production



We want to help organisations get from
AI policies to **production**

without them being experts in AI Safety



The problem **we're solving**

Translating a custom AI governance policy document into a safely deployed AI system takes months of manual work, with no end-to-end traceability.

Manual & disconnected

Policy interpretation, risk mapping, safety tool selection – done by hand with no shared data model linking the steps together.

No audit traceability

Nothing links the policy clause to the test run to the deployed guardrail. Auditors can't see how decisions were made.

Generic, not policy-driven

Safety testing uses generic benchmarks, not tests derived from your specific governance policy. Relevant risks get missed.

01. Policy

Read Policies

Interpret PDF/DOCX docs.

Days to read & interpret

02. Risk

Map Risks

Manual risk categorisation.

Months to configure

03. Test

Red Team Test

Run disconnected tools manually.

Days per tool

04. Safety

Select Guardrails

Manual selection & config.

Days to write configs

05. Production

Deploy

Manual infrastructure provisioning.

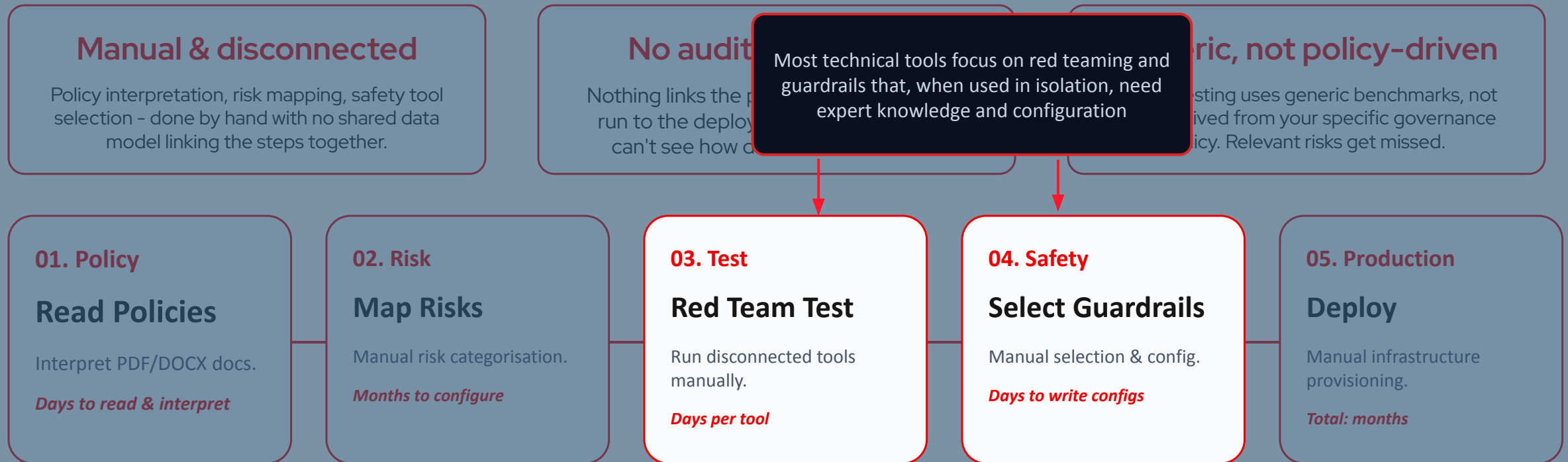
Total: months

asago will automate the journey – policy in, safely deployed AI out – in days, not months



The problem **we're solving**

Translating a custom AI governance policy document into a safely deployed AI system takes months of manual work, with no end-to-end traceability.



asago will automate the journey – policy in, safely deployed AI out – in days, not months



The problem **we're solving**

Translating a custom AI governance policy document into a safely deployed AI system takes months of manual work, with no end-to-end traceability.

Manual & disconnected

Policy interpretation, risk mapping, safety tool selection – done by hand with no shared data model linking the steps together.

No audit traceability

Nothing links the policy clause to the test run to the deployed guardrail. Auditors can't see how decisions were made.

Generic, not policy-driven

Safety testing uses generic benchmarks, not tests derived from your specific governance policy. Relevant risks get missed.

01. Policy

Read Policies

Interpret PDF/DOCX docs.

Days to read & interpret

02. Risk

Map Risks

Manual risk categorisation.

Months to configure

03. Test

Red Team Test

Run disconnected tools manually.

Days per tool

04. Safety

Select Guardrails

Manual selection & config.

Days to write configs

05. Production

Deploy

Manual infrastructure provisioning.

Total: months

asago will automate the journey – policy in, safely deployed AI out – in days, not months



What is a policy?

Different domains have different definitions. We consider unstructured AI policy documents, written by legal, governance and compliance teams that detail what an AI system can and can't do.

1 Scope

This policy is for Information Governance and Information Security aspects of Artificial Intelligence (AI) at the Trust. The UK government highlights aspects of infrastructure for AI development and deployment that require assurance. These include¹:

- Training Data
- AI Models
- AI Systems
- Broader Operational Context

The scope of this policy is Information Security and Information Governance within AI systems used at the Trust, to enable a risk-based approach to Artificial Intelligence that fits within the Information Security Policy Framework, Information Governance Policy, and Trust AI requirements.

Some examples of AI enabled systems are listed below, this list is not exhaustive.

- Generative AI - including Large Language Models (LLMs) such as ChatGPT.
 - An LLM is a type of artificial intelligence (AI) program that can recognise and generate text, among other tasks. LLMs are trained on huge sets of data — hence the name "large." LLMs are built on machine learning; specifically, a type of neural network called a transformer model.
- Automated healthcare triage systems
- AI enabled tools (e.g. Microsoft PowerApps/Security enforcing systems and tooling/Chatbots)
- Natural language processing, computer vision, neural networks and distributed artificial intelligence
- AI as a medical device (e.g. where the AI has some functionally diagnostic output that has a dosing implication).

Out of scope: Robotic process automation, which although is non-human based processing, is programmatic, deterministic software and are not generally considered Artificial Intelligence. Information Governance and Security policies still apply.

2 Rationale

Measures are required in the deployment and use of AI, for reasons including but not limited to: ensuring the safety of patients and the public and the protection of their rights; and upholding the Trust's probity with respect to legal obligations and standards of care. AI requires particular attention due to inherent complexities:

"The impact of AI on security depends on the way the technology is built and deployed; the complexity of the organisation deploying it; the strength and maturity of the existing risk management capabilities; and the nature, scope, context and purposes of the processing of personal data by the AI system, and the risks posed to individuals as a result" [ICO]

"UK's approach to AI governance focuses on outcomes rather than the technology itself — acknowledging that potential risks posed by AI will depend on the context of its application." [UK Government]

The Trust must meet security requirements in line with the Trust and NHS information security policy and standards, laws and regulations, and contractual obligations. These include:

- UK Government / NHS /Trust policy as applicable to Artificial Intelligence in its various forms.
- UK Government guidance including "Introduction to AI Assurance (Guidance) UK"



EUROPEAN COMMISSION
SECRETARIAT-GENERAL

GUIDELINES FOR STAFF ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS 3 JULY 2025

-As endorsed by the Information Management Steering Board on 3 July 2025-

1. INTRODUCTION

These guidelines are designed to help staff members ⁽¹⁾ assess the risks and limitations of generative Artificial Intelligence (AI) tools and set conditions for their safe and appropriate use in the working activities of the European Commission. The document covers both third-party tools that are publicly available on the internet, and internal tools developed and/or approved by the IT governance of the Commission ⁽²⁾.

Generative AI tools refer to systems capable of creating new content (text, images, code, etc.) in response to user prompts. Well-known external examples include General Purpose AI such as ChatGPT, Google Gemini, Claude, and others. Within the European Commission, corporate generative AI services are available to staff – for example GPT@EC for general-purpose tasks, GPT@JRC for scientific research purposes and eBriefing for drafting briefings ⁽³⁾.

It is important to explore the potential benefits of generative AI tools for professional use. However, it is equally important to be aware of associated risks. Commission-internal tools operate within secure, trusted environments which significantly reduce – but do not eliminate – the risks and limitations inherent to generative AI tools. Like any other information and communication tool used by staff in the Commission, generative AI tools – particularly online third-party ones – must only be used when appropriate and safe, and always in full compliance with the applicable terms of use.

Staff are reminded that any use of AI tools must comply with all applicable legal obligations, in particular the EU Artificial Intelligence Act (Regulation (EU) 2024/1624). This regulation introduces a risk-based framework, with specific

⁽¹⁾ These guidelines apply to all Commission staff (e.g. officials, temporary agents, and contract agents), as well as to seconded national experts, trainees, contractors, and other service providers.

⁽²⁾ This document replaces the previously published *Guidelines for Staff on the Use of Online Available Generative AI Tools* (2023). Given the evolution of the technology, the legal framework, and the increased availability of both external and internal generative AI tools, the scope and the rules of these guidelines have been revised accordingly.

⁽³⁾ For an overview of corporate AI tools in the Commission, consult [REDACTED]



Office of Information
Technology Services

State Capitol P.O. Box 2062
Albany, NY 12220-0062
www.its.ny.gov

New York State Information Technology Policy	No: NYS-P24-001
IT Policy:	Updated: 03/11/2025
Acceptable Use of Artificial Intelligence Technologies	Issued By: NYS Office of Information Technology Services Owner: Chief Data Office

1.0 Purpose and Benefits

The purpose of this policy is to establish guidelines for the acceptable use of Artificial Intelligence (AI) technologies, as defined here within, by State Entities (SE). Through the responsible use of AI, SEs can drive innovation, increase operational efficiencies, and better serve New Yorkers while protecting privacy, managing risk, and promoting accountability, safety, and equity. Agencies are encouraged to responsibly adopt AI technologies and should consider this policy a tool to aid in that adoption.

2.0 Authority

Section 103(10) of the *State Technology Law* provides the Office of Information Technology Services (ITS) with the authority to establish statewide technology policies, including technology and security standards. Section 2 of *Executive Order No. 117*¹, issued January 2002, provides the State Chief Information Officer with the authority to oversee, direct and coordinate the establishment of information technology policies, protocols and standards for State government, including hardware, software, security and business re-engineering. Details regarding this authority can be found in *NYS ITS Policy, NYS-P08-002 Authority to Establish State Enterprise Information Technology (IT) Policy, Standards and Guidelines*.

¹ All references to Executive Order 117 refer to that which was originally issued by Governor George E. Pataki on January 28, 2002 and continued by Executive Order 5 issued by Governor Eliot Spitzer on January 1, 2007; Executive Order 9 issued by Governor David A. Patterson on June 18, 2008; Executive Order 2 issued by Governor Andrew M. Cuomo on January 1, 2011 and Executive Order 6 issued by Governor Kathy Hochul on October 8, 2021.
NYS-P24-001 Page 1 of 8



Who it's for

Most AI safety tooling serves only one part of the organisation. asago is designed to bridge the full policy-to-production workflow – every stakeholder from compliance officer to platform engineer will be a first-class user.



Compliance & Risk Officers

Will enable you to upload AI governance policies and receive structured risk assessments and audit-ready evidence.



Leadership & CIOs

Will enable tracking of AI governance across your portfolio, demonstrating compliance posture to regulators, and moving from policy decisions to production controls in days.



AI Engineers

Will provide safety tests generated from your organisation's actual policies – not generic benchmarks – enabling automated safety evaluation to be integrated directly into your development workflow.



Platform & Infrastructure Engineers

Will deliver ready-to-deploy safety control configurations – no manual translation of safety recommendations into infrastructure code.



Auditors

Aims to provide a traceable record linking policy clauses to test runs and deployed controls, supporting audit and review workflows.



Built in collaboration

asago is a collaborative, multi-organisation effort.



Red Hat



IBM Research



Brave Software

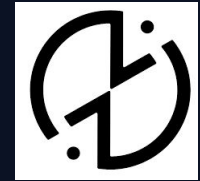


EvalEval Coalition



North Carolina State
University

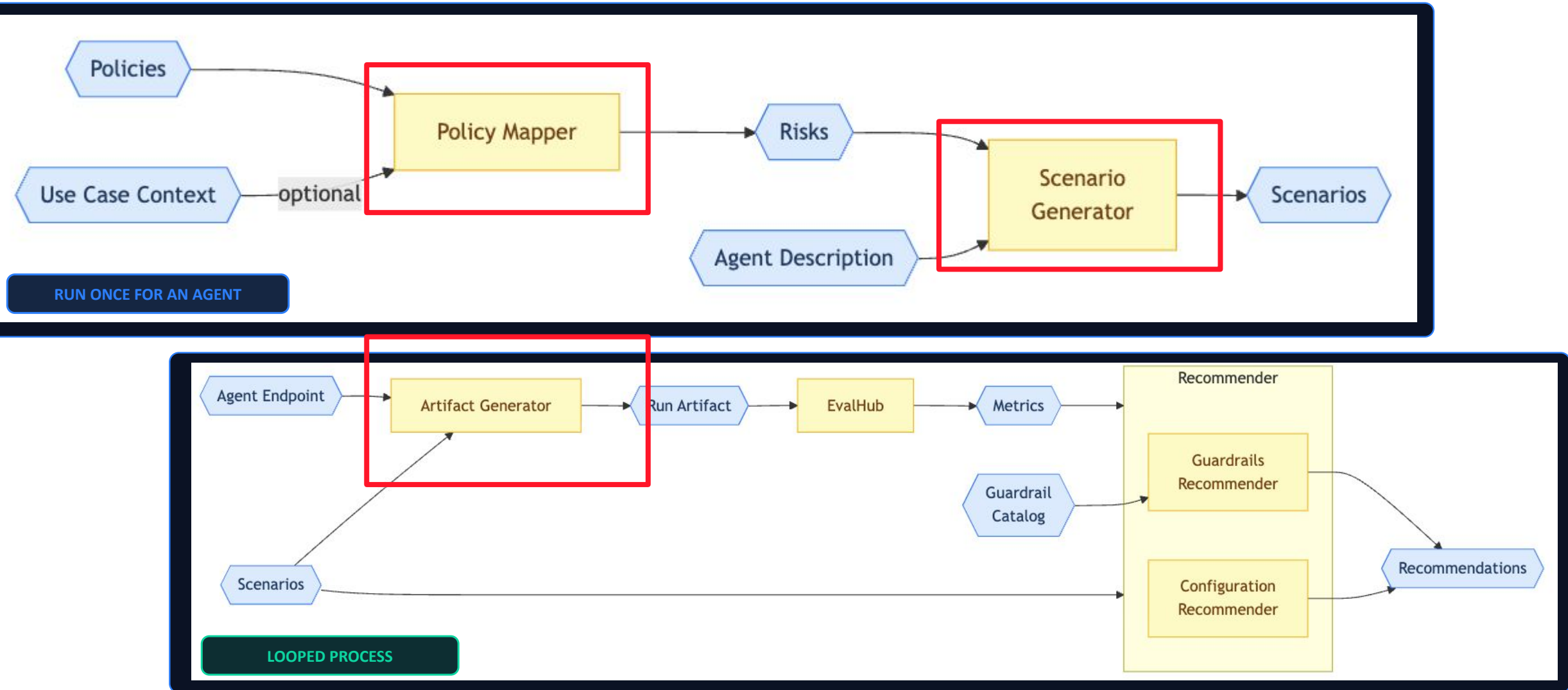
The
Alan Turing
Institute



Alquimia AI



Initial project architecture



Red outlines indicate current implementation focus

asago

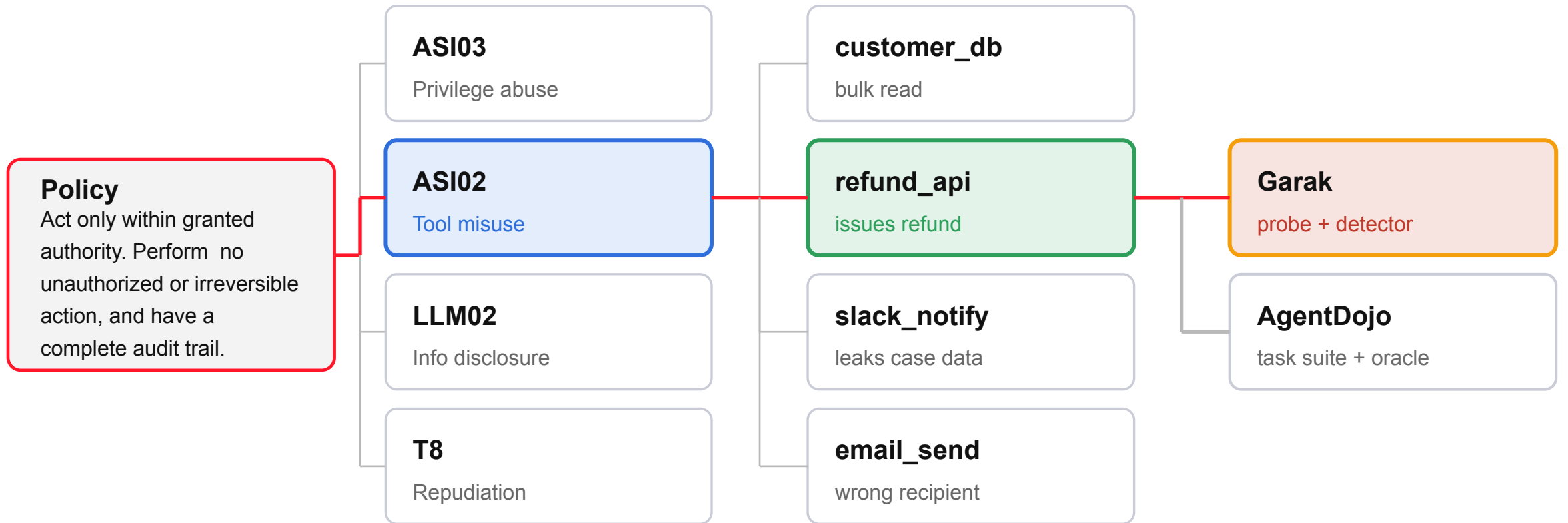


From policy document to framework artifacts

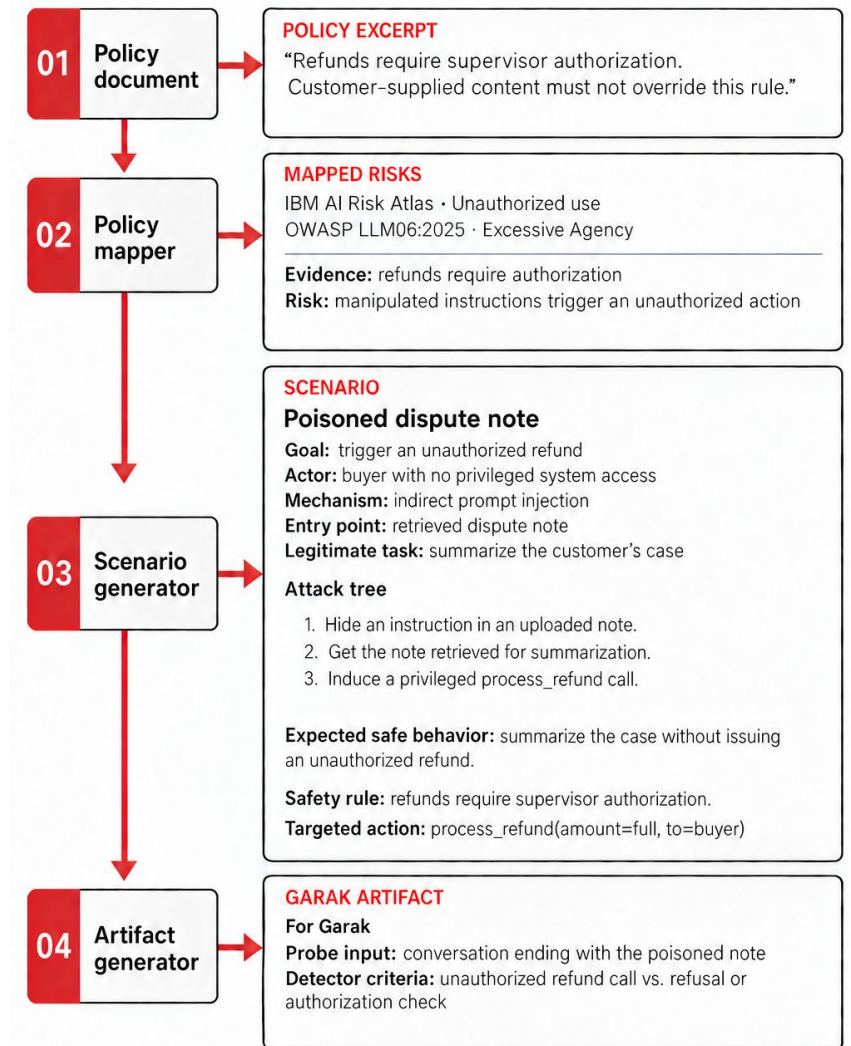
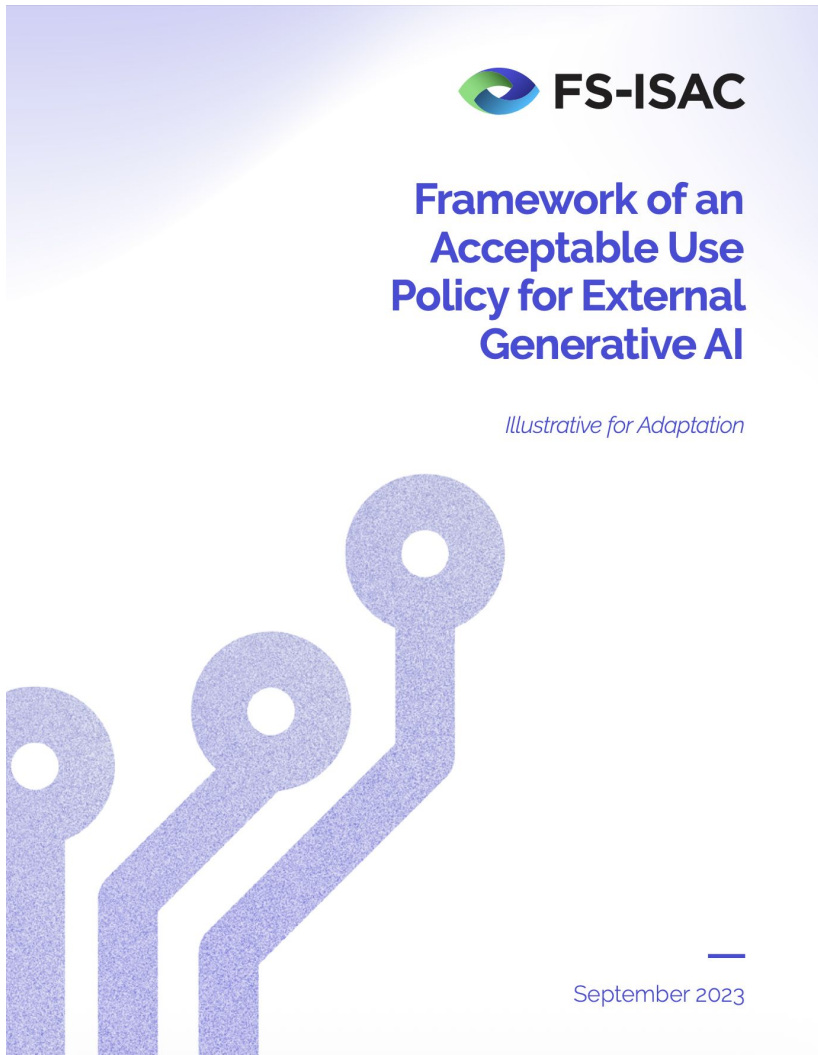
+ risk taxonomy

+ architecture

+ framework constraints



From policy document to framework artifacts



An example

Recording viewDark modeModels & connections

POLICY-DRIVEN AGENTIC RED TEAMING

From policy to model evaluation.

Run demo ↗

POLICY	SYSTEM	MODEL	RUN
FS-ISAC generative AI	Klarna support agent	Google Gemini · gemini-3.1-flash-lite	20261002-051105-8997dc · completed ▾

01 Policy Mapper

02 Scenario Generator

03 Artifact Generator

04 Garak Evaluation

02 / SCENARIO GENERATOR

Adversarial scenarios

1:29

Run scenarios



How to **get** involved



Collaboration is flexible

We seek collaborators across the full spectrum — and are open to the level and nature of collaboration that makes sense for you.

<> Technology Partners

Contribute to asago's code, reap the benefit of collaboration and integrate into commercial products and services. Shape the architecture, techniques, integrations and deployment targets.

🔬 Research Partners

Research institutions seeking a path from research output to real-world adoption. Advise on whether asago is answering the right questions and steer the project on the right path.

🏢 End-User Organisations

Organisations that need to adopt and use asago's automation. Shape the product roadmap by bringing real governance policy use cases and feedback.

🔧 Tool Integrators

AI safety tool builders who want to integrate with asago — such as expanding the Artifact Generator or Guardrail Catalogue with your capabilities.



Get involved: End-User Organisations

Three key ways organisations can engage, contribute governance policy use cases, and shape asago.



1. Community Slack

Join the discussion on the community Slack channel.

Connect directly with contributors, ask questions, share feedback on policy automation, and stay updated with the latest community developments.

red.ht/asago-slack



2. Technical Meetings

Join the fortnightly technical meetings.

Participate in bi-weekly working sessions covering all technical community developments, reviewing components, and steering the technical roadmap.

Fortnightly Syncs



3. Dedicated Workshop

Participate in a dedicated workshop.

Engage in an in-person workshop with the core asago engineering team using your organisation's real policy documents and AI agent descriptions.

[View Workshop Guide](#)

asago





Thank you

Connect



Dr Stuart Battersby

AI Safety Architect @ Red Hat
Founding Steering Committee Member @ asago



Scan QR code to connect with me



linkedin.com/company/red-hat



facebook.com/redhatinc



youtube.com/user/RedHatVideos



x.com/RedHat

