Risk, Security & Compliance · Security & Cyber

Security Defect Triage Agent

CAI Score L4 · Decide — Senior human accountable for every decision

▲ this agent

L1 Assist	L2 Draft	L3 Operate	L4 Decide	L5 Autonomous

TIER	PILLAR	AUTOMATION	HUMAN ROLE
L4 · Decide	**Risk, Security & Compliance**	**≈50% of routine effort**	**Named accountable decision-maker**

01 — Executive summary

Security Defect Triage Agent is a CAI Score L4 (Decide (supervised)) agent for risk, security & compliance. Analyses security defects from multiple sources and assigns a consistent, risk-based severity rating using the firm internal scoring model. It operates with senior human accountability: a named person owns every decision it supports. Indicatively, it can redirect around 50% of the routine effort this work consumes.

02 — Problem & context

— Dev and security argue over how severe each defect really is.

— Defects from different scanners arrive in different formats and scales.

03 — Representative use cases

— Consolidates security defects from multiple scanner outputs into a single, normalised queue.

— Applies a consistent, documented severity model to every defect — eliminating inter-team severity disputes.

— Accelerates triage: defects receive a risk-based severity rating within minutes of ingestion, not days.

— Provides the security team with an explained, auditable rationale for every severity assignment.

— Produces the evidence to track triage speed, severity consistency, and remediation prioritisation accuracy.

04 — How to use it

1. Connect it in your tenant to the systems this work already touches (security & cyber).
2. Obtain senior sign-off and confirm the named accountable person before go-live.
3. Receives the case or work product
4. Assembles evidence and flags risk or discrepancy
5. Presents structured findings to the named human
6. Named human makes the decision; rationale is logged
7. Review decision outcomes and the audit trail after each cycle; adjust scope as needed.

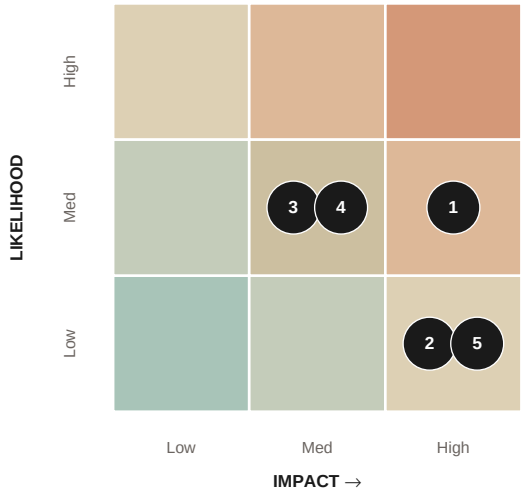
05 — Where it fits in the process

Sits between detection and remediation — it consolidates defects from every source and applies one consistent severity model, ending the dev-vs-security debate.

Decision-support loop for a L4 agent — human owns every call.

06 — Risk assessment

#	RISK	LIKELIHOOD	IMPACT	MITIGATION / CONTROL
1	**Incorrect evidence assembly**	Medium	High	Named human reviews all agent output before any decision is made; the human owns the conclusion.
2	**Over-reliance on agent assessment**	Low	High	L4 tier structure: agent evidences, human decides. Decision authority cannot be delegated to the agent.
3	**Exposure of sensitive data**	Medium	Medium	Runs inside your own tenant; Colleague AI processes none of your data.
4	**Behavioural drift over time**	Medium	Medium	Periodic review against case outcomes detects drift; scope and prompts are version-controlled.
5	**Out-of-policy or edge-case handling**	Low	High	Hard scope boundary: the agent flags anything outside policy; senior sign-off is required at go-live.



Residual likelihood × impact after the tier's built-in controls. Numbers map to the table above.

07 — Regulatory & compliance impact

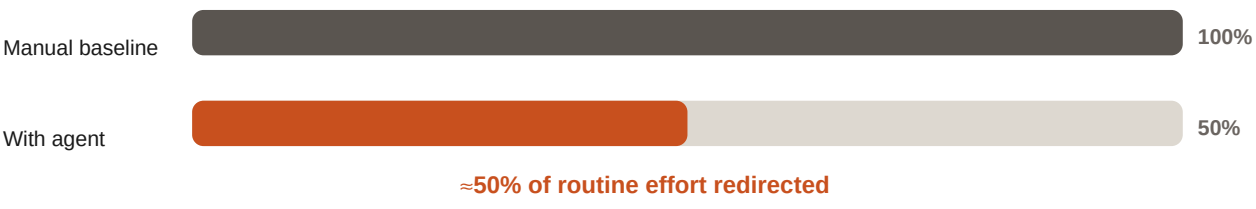
FRAMEWORK	RELEVANCE	HOW THIS AGENT ALIGNS
EU AI Act	Operational use under human oversight; logging and risk-tiering apply.	Documented risk tier, human oversight and event logging per agent — designed to map to the obligations.
DORA	Relevant where the function is in scope for financial-sector operational resilience.	Actions are logged and attributable, supporting resilience and traceability requirements — designed to map to.
GDPR	Applies wherever personal data is in scope of the workflow.	Data stays in your tenant; Colleague AI processes none of it. Purpose limitation and minimisation are preserved.
ISO/IEC 42001	AI management-system controls: roles, monitoring, review.	The agent operates within a documented management system — defined tier, named roles, logging and review. Designed to meet; not certified.

"Designed to map to" / "designed to meet" describe alignment with the CAI Score framework, not external certification. Confirm obligations with your compliance function.

08 — Value & illustrative ROI

Analyses security defects from multiple sources and assigns a consistent, risk-based severity rating using the firm internal scoring model.

Assumptions (illustrative — replace with your own)	Result
4 FTE engaged · 1840 productive h/FTE/yr · 50% addressable · €70/h blended	≈ 3,680 hours/yr redirected · ≈ €258,000/yr indicative value



Figures are an illustrative model to be validated against your own volumes and rates — not a guarantee.

KPIs to track

Reduced analysis time · fewer clarifying questions · consistent severity ratings.

09 — Recommendations

— Obtain senior sign-off before go-live — mandatory for L4 deployments.

— Document the severity model in the agent configuration so the rationale is auditable.

— Assign a named security lead as the accountable reviewer for severity calls.

— Review the first 50 severity assignments manually to validate the model before scaling.

10 — Governance & method

This agent runs in your own Microsoft/Azure tenant; Colleague AI hosts only the control plane (score, policy, audit) and processes none of your data. A named senior person owns every decision the agent supports. Every action is time-stamped and attributable. The CAI Score classification is documented and open to challenge.