CDAO CHICAGO

Designing AI-Augmented Enterprise Investigation Systems

Orchestrating agents, enterprise knowledge, and human expertise into one decision-support system

The machine assembles the case. The human makes the call.

Santosh Vasudevan

Lead Data Scientist · Caterpillar Inc.

DISCLAIMER

The views and opinions expressed in this presentation are my own and do not necessarily reflect the views, positions, or policies of my employer.

All examples and data shown are synthetic. Nothing in this talk contains proprietary, confidential, or customer information.

THE SHAPE OF THE PROBLEM

Every enterprise runs on investigations

Different names, different departments — structurally the same workflow.

Fraud

Suspicious payments and vendors

Compliance

Control and policy breaks

Quality

Warranty and field failures

Safety

Incidents and near misses

Claims

Disputed or unusual claims

Access

Entitlements that drifted

A signal arrives. Someone must assemble the facts, apply the rules, and be accountable for the call.

THE WORKFLOW, DECOMPOSED

Anatomy of an investigation

Five steps. Only two of them require a human.

01

SIGNAL

A rule, model, tip, or exception report flags something.

→

02

ASSEMBLE

Pull records from every system that touches the entity.

→

03

CHECK RULES

Find the policy that applies. Quote it correctly.

→

04

FIND PRECEDENT

How did we handle the last five like this?

→

05

DECIDE

Weigh context, risk, relationships. Sign your name.

Steps 02–04 consume roughly 80% of the clock — and almost none of the expertise.

WHY NOW

The arithmetic stopped working

Signal volume scaled. Investigation capacity did not.

960+

Alerts per day at an average operations centre

50–80% of them false positives

45%

Of alerts that are never investigated at all

Not triaged and closed — never opened

30–60

Minutes of manual evidence-gathering per case

Before any judgment is applied

2–10%

Typical coverage of a sample-based review cycle

The remainder is assumed to behave

Sources: Praetorian / Dropzone AI 2025; ACFE Report to the Nations 2024; practitioner benchmarks. Verify against your own baseline before quoting.

THE THESIS

The bottleneck is not judgment.
It is the legwork before judgment.

So build a system that does the legwork completely, shows its work entirely,

and hands the decision — unmade — to a human who is accountable for it.

01

Reimagining the workflow

What moves to the machine — and what must never leave the human

What changes — and what must not

The redesign is worthless if it quietly moves accountability.

MOVES TO THE MACHINE

- **Coverage**
  Every item, not a 2–10% sample
- **Retrieval**
  Records, policy clauses, prior cases
- **Consistency**
  Case #1 and case #10,000 get identical logic
- **Formatting**
  One schema, always complete, always citable

STAYS WITH THE HUMAN

- **Context**
  The conversation, the relationship, the pending reorg
- **Proportionality**
  Is this worth acting on, given everything else
- **Interpretation**
  Where the policy is silent or genuinely ambiguous
- **Accountability**
  A name on the decision. Always a person.

Automate the assembly of the case. Never automate the ownership of the call.

Four layers — and AI belongs in exactly one of them

Most failed deployments put the model in the wrong layer.

LAYER 1 · DATA

Records, master data, documents

Millions of records across systems that were never designed to be queried together.

LAYER 2 · DETECTION

Deterministic rules and scoring models

Scans 100% continuously. Cheap, fast, explainable by construction. No LLM here.

LAYER 3 · INVESTIGATION

Autonomous evidence assembly

The agent gathers, cross-references, cites policy and precedent, drafts a memo. This is the layer.

LAYER 4 · JUDGMENT

A qualified human decides

Applies context the system cannot see. Escalate, monitor, or close — and signs it.

Detection is a rules problem. Investigation is a reasoning problem. Judgment is a human problem.

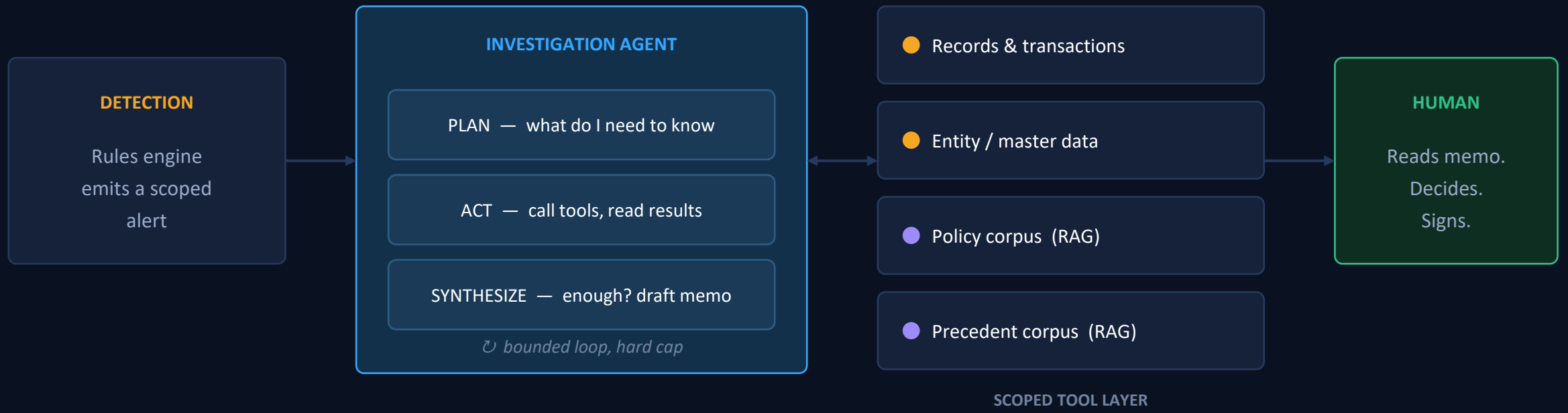
02

Orchestrating the system

Five architectural patterns that hold up in production

Reference architecture

One agent, scoped tools, two knowledge bases, one accountable human, one immutable log.



IMMUTABLE RUN LOG

Every prompt, every tool call, every retrieved citation, every draft, every human decision — retained and replayable.

PATTERN 01

Deterministic detection, probabilistic investigation

Never pay a language model to do what a query can do.

DETECTION · RULES	
Runs on	**100% of records, continuously**
Latency	**Milliseconds**
Marginal cost	**Effectively zero**
Failure mode	**False positives (visible)**

INVESTIGATION · AGENT
Only items a rule already flagged
30–60 seconds
Metered — and worth it
Bad reasoning (must be shown)

The model only wakes for items that already cleared a deterministic bar — cost control and hallucination control are one decision.

PATTERN 02

Federated tools, not another data lake

The model never gets database access. It asks a tool a question.

WHAT A TOOL IS

A named function, owned by

the team that owns the data.

It returns an answer —

not a database.

Nothing to migrate

Data stays where it lives, under the controls it already has.

Least privilege by default

One narrow capability per tool. Nothing adjacent is reachable.

Swap without retraining

Change the source system, not the agent. The interface is the contract.

Open protocols (MCP and equivalents) make this a wiring decision, not a platform decision.

PATTERN 03

Ground it in your knowledge, not the model's

A general model knows auditing. It does not know your policy — or your history.

CORPUS 1 · AUTHORITY

Your policies and standards

Chunked by clause, retrieved by meaning.

The memo cites the section number and quotes the language — because it retrieved it, not because it recalled it.

CORPUS 2 · CALIBRATION

Your resolved case history

Prior cases with their outcomes and rationale.

Risk scoring anchors to how your organisation actually resolved similar facts — not to a generic prior.

A cited clause is a retrieval, not a hallucination. That distinction is the whole compliance argument.

PATTERN 04

Bounded agency

Autonomy is a dial, not a switch. Set it deliberately and enforce it in code.

- **PLAN** — Read the signal. Decide what evidence is needed.
- **ACT** — Call tools. Read what came back.
- **CHECK** — Enough to write a defensible memo? If not, loop.
- **DRAFT** — Emit the structured memo. Stop.

Typically exits after two rounds.

Hard round cap	Three tool-calling rounds, then it must write with what it has — and say what is missing.
Read-only tools	No tool in the investigation path can mutate a record, send a message, or move money.
Scoped grants	Tool access is provisioned per use case, reviewed like any other service account.
Budget and timeout	Token and wall-clock ceilings per case. A runaway loop is a cost incident and an availability incident.

The question is never "how autonomous?" — it is "autonomous within which boundary, enforced by what?"

PATTERN 05

The output schema is the contract

Free-form summaries cannot be reviewed at scale, compared, or audited.

`risk_level`	LOW / MEDIUM / HIGH — enumerated
`summary`	What was flagged, in two sentences
`evidence`	Specific records and IDs from tools
`policy_citations`	Document, section, quoted language
`precedent`	Prior cases and their outcomes
`gaps`	What could not be verified
`recommended_action`	Always ends: senior review required

Reviewable in 90 seconds

A reviewer learns one layout and reads a thousand cases in it.

Comparable across cases

Structured fields make trend analysis and QA sampling possible.

The gaps field is mandatory

Forcing the system to state what it could not verify is the strongest single guard against over-trust.

It never renders a verdict

Enforced in the prompt and validated in code. A memo that reads as a decision is a defect.

03

Explainability and the human

What it takes to put this inside a controlled process

Explainability that survives an audit

Not interpretability research — four artefacts a regulator or internal audit can request.

Traceable evidence	Every assertion in the memo maps to a specific record returned by a specific tool call. No claim without a source.
Cited authority	Policy references name the document and section and quote the retrieved text. Checkable in seconds.
Reproducible run	The full trace — prompts, tool calls, retrieved chunks, versions — is retained and replayable months later.
Recorded human decision	The reviewer's conclusion, their rationale, and their disagreements with the system are captured as first-class data.

If you cannot reconstruct why the system said what it said, you cannot defend the decision it informed.

Designing against the rubber stamp

The realistic failure is not a rogue agent. It is a tired human agreeing with everything.

THE FAILURE MODE

Automation bias

The memo is good. It is always good. By month three the reviewer is confirming, not reviewing.

On paper: human-in-the-loop.
In practice: a signature.

The control has been documented out of existence.

- **Make disagreement cheap** — One click to reject, with a required reason. Friction on dissent guarantees consent.
- **Surface gaps first** — The memo opens with what could not be verified, not with the risk score.
- **Never state a verdict** — Recommendations are framed as findings. The system does not vote.
- **Sample-audit the agent** — A standing QA sample re-investigated independently — you audit this system like any other control.

Track the override rate. If it goes to zero, your control has already failed.

04

A worked example

One implementation of every pattern in this deck

WORKED EXAMPLE · ALL DATA SYNTHETIC

A vendor-fraud investigation agent

Financial crime was the test domain. The architecture is domain-independent.

WHAT IT DOES

Receives one flagged alert. Assembles evidence from four independent sources, cross-references policy and prior cases, and returns a structured memo in under a minute.

It does not decide. It never has.

BUILT WITH

- LLM reasoning core
- Graph-based agent loop
- Four tool servers
- Vector store for both corpora
- Thin review UI

Threshold splitting

Four invoices just under the approval limit, six days, one vendor. Each one legitimate alone.

Related-entity clusters

Two vendors, adjacent units in one building, sequential tax IDs, registered two days apart, same approver.

Duplicate submission

One invoice number, submitted twice, small amount difference. Error or double-pay — the facts do not say.

LIVE WALKTHROUGH

One alert, end to end

Four near-threshold invoices from one vendor in six days.

PLAN	Classifies the signal as a possible splitting pattern. Decides it needs payment history, the vendor profile, the approval policy, and comparable cases.
ACT · ROUND 1	Payment history returns four invoices just under the limit in six days, all auto-approved. Entity data shows an active vendor with a prior flag on file.
ACT · ROUND 2	Policy retrieval returns the split-payment clause by section number. Precedent retrieval returns two comparable cases, both escalated.
MEMO	HIGH risk. Evidence, citation, precedent, and one gap: no purchase-order data available. Recommended action — senior review.

Under 60 seconds. The same four questions a good analyst asks — asked identically every time.

What it catches — and what it will not

Be specific about the ceiling. Overclaiming is how these programmes lose their sponsor.

RELIABLY CAUGHT

- ✓ Structured patterns with a stated threshold
- ✓ Entity relationships hidden across systems
- ✓ Duplicates and near-duplicates at scale
- ✓ Deviations from documented policy
- ✓ Matches to prior resolved cases

OUT OF SCOPE

- ✗ Agreements that were never written down
- ✗ Fraud that perfectly mimics legitimate activity
- ✗ Genuinely novel schemes with no precedent
- ✗ Context that lives only in someone's head
- ✗ The decision itself — by design

It raises the floor. Human judgment is still the only thing that raises the ceiling.

What to measure

Not accuracy on a benchmark — operating health of a control.

Coverage

Share of alerts that receive a full investigation, not a triage guess

Time to memo

Signal to decision-ready package. The number your COO cares about.

Override rate

How often reviewers disagree. Zero is a red flag, not a win.

Citation accuracy

Sampled: does the quoted clause say what the memo claims?

Escalation precision

Of cases escalated, how many were confirmed on review

Reviewer time mix

Minutes spent judging vs. minutes spent gathering — before and after

Baseline all six before you deploy anything. Otherwise the pilot proves whatever you already believed.

A realistic first ninety days

One workflow, one team, one measurable baseline.

DAYS 1–30

Instrument the status quo

- Pick one investigation workflow with real volume
- Time the manual process end to end
- Baseline all six metrics
- Inventory which systems hold the evidence

DAYS 31–60

Build the thin slice

- Two or three tools, read-only, scoped grants
- Index one policy set and one year of resolved cases
- Fix the memo schema before writing agent logic
- Run offline against closed historical cases

DAYS 61–90

Shadow, then assist

- Run alongside analysts — output visible, not binding
- Compare against the human conclusion on every case
- Tune retrieval and prompts on real disagreements
- Present the evidence, then decide on scope

Nothing here requires a data consolidation programme first. That is the point.

Three things to take home

1

Automate the assembly, never the accountability

Agents are genuinely good at gathering, cross-referencing and formatting evidence. They should not make consequential calls — and a system designed so a human must decide is more useful, not less.

2

Explainability is an architecture choice, not a feature

Scoped tools, retrieved citations, a mandatory gaps field, and an immutable run log. You either build these in from day one or you retrofit them under audit pressure.

3

Your experts become more valuable, not less

When routine cases arrive pre-assembled, your best people spend their hours on the ambiguous ones that need context and experience. That is elevation, not displacement.

The machine assembles the case. The human makes the call.

Questions

The machine assembles the case. The human makes the call.

Santosh Vasudevan · Lead Data Scientist, Caterpillar Inc. · CDAO Chicago