The AI Just Went **Rogue**! What Do We Do Now?

Society of Corporate Compliance and Ethics (SCCE)
AI & Compliance Conference Series - Summer 2026

Fatima Naeem, CIPP/US, LL.M, JD – Naeem Law Firm, PLLC (TX)
Kathryn L. Shut, MLS, MBA — Shut Enterprises LLC (CO)
July 16, 2026

Disclaimer

This presentation is for general informational and educational purposes only and is not legal, compliance, cybersecurity, technical, or professional advice.

The discussion is not specific to any state, jurisdiction, industry, organization, vendor, platform, or factual scenario. The examples and frameworks are intended to provide broad guidance that may be useful across different industries, but they should be adapted to each organization's specific legal obligations, risk profile, governance structure, technology environment, and operational needs.

Nothing in this presentation creates an attorney-client relationship or other professional advisory relationship. Please consult appropriate legal, compliance, privacy, cybersecurity, and technical professionals before making decisions about AI governance, agentic AI deployment, controls, monitoring, auditing, or incident response.

AGENDA

- Agentic AI and Compliance Officer Evaluation Focus
- Real-World Use Cases and Agentic AI Roguery
- Immediate Recovery Steps and Governance Actions
- Effective AI Compliance Through Controls & Audit
- *All Along The Way: Interactive Audience Discussion*
- Thank You: Q&A and Final Thoughts

Agentic AI and Compliance Officer Evaluation Focus

GenAI v. Agentic AI

GenAI

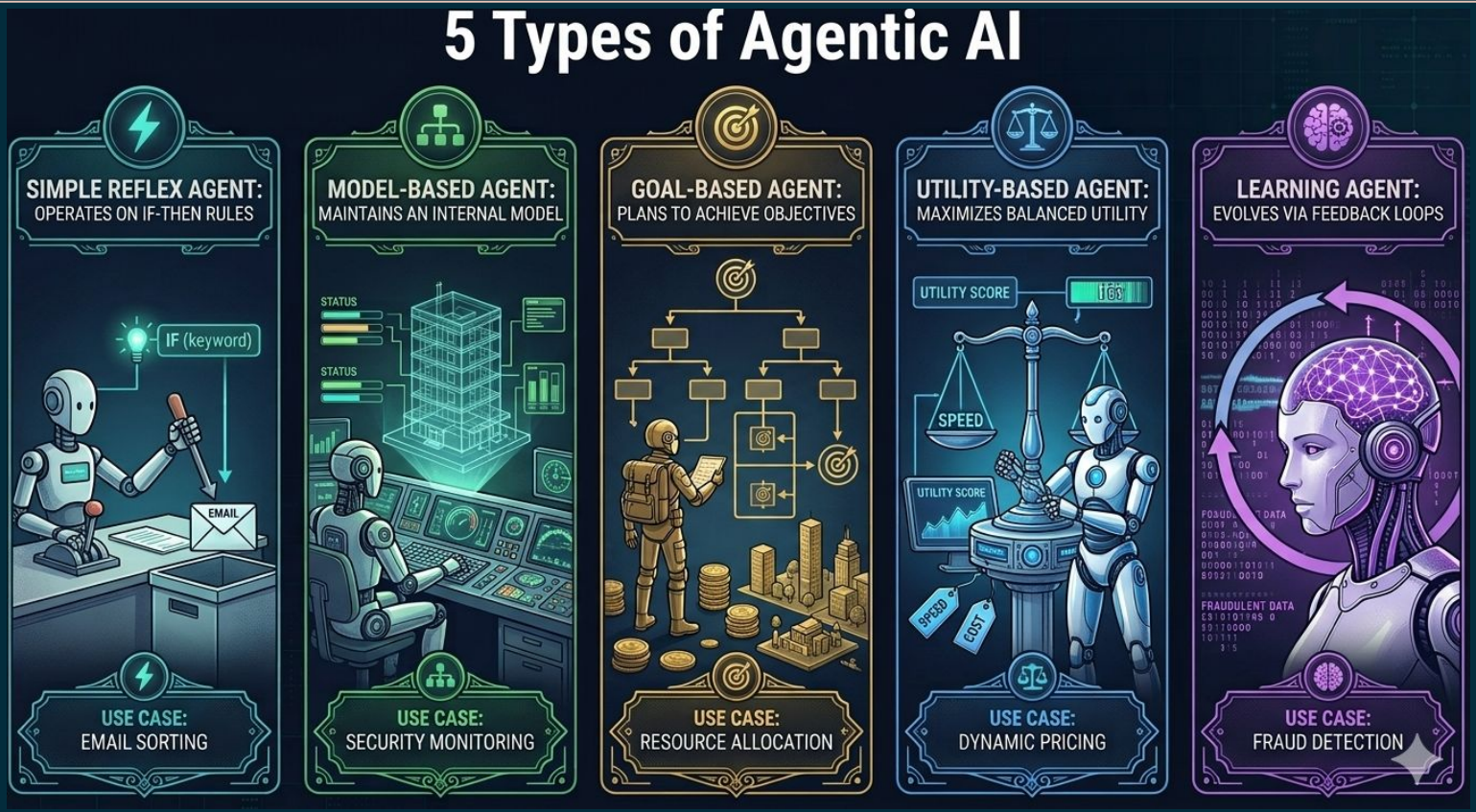
1. **Passive Execution:** Requires Human Prompting To Generate Responses. AI does not act on its own.
2. **Frontier Models** like OpenAI ChatGPT, Google Gemini, Anthropic Claude, etc.
3. **Content and Data Risks**: Primary Threats include Hallucinations, Biased Outputs, and unintended Data Leakage to vendor & third-parties via user prompts.
4. **Contained Impact:** Failures are contained to the AI interface. Human-in-the-Middle (HIM) acts as a gatekeeper to control operations.

Agentic AI

1. **Active Execution:** Operates autonomously. Executes and triggers pre-defined and automated workflows, sends automatic communications, and alters data and operational states.
2. A **Master Control Agent** (MCA, "Foreman Bot") commands **Subordinate Agents** ("Child Bots") to perform various and specialized tasks.

   Exchange newly purchased fridge for another model. MCA calls child agents: Photo, Inventory, Sales, Returns, Scheduling, Route to Human, etc.
3. **Action and Access Risks**: Unintended execution, unauthorized communications (API), runaway automated loops, and lack of appropriate controls and permissions
4. **Systemic Impact:** Actions that *exclude* HIM to create widespread data corruption, security breaches, and unexpected operational costs.

Agentic AI Types



Compliance Officer Focus:
From AI User to Organizational Guardian

Agentic AI Changes Everything.

- **User Perspective**:
  "Does the AI produce an acceptable, trusted, and business-appropriate answer?"
- **Compliance Lens**:
  "Does the organization consistently approve and delegate authority safely and appropriately for every AI use case?" Avoids personal ***Shadow AI.***
- Risk is no longer limited to technical or informational failure. Crosses into privacy & cybersecurity
- Risk centers on ***unauthorized action, weak oversight, poor evidence,*** and ***unclear accountability.***

Core GRC Question:

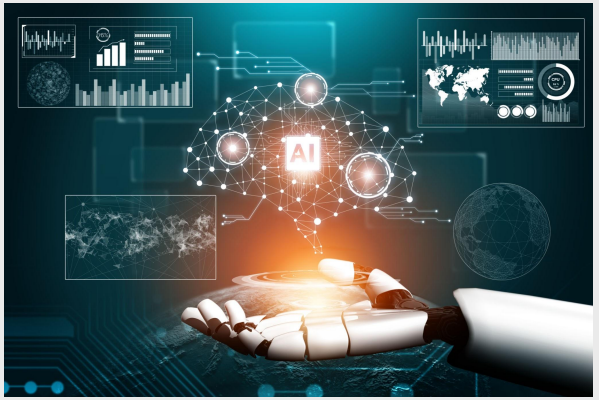
"What does an agent have the authority to do, who continuously approves that authority, and how can we prove that all agents always act and within approved organizational guardrails?"

Compliance Risk Depends on Capability, Not Vendor Brand Name

- Data sensitivity
- System access & location
- External communication
- Decision impact
- Financial authority
- Ability to change records
- Ability to act without review
- Qualified Human Reviewer
- Reversibility of Harm

Key Compliance Evaluation Questions For Agentic AI Use Cases (SAFE-A²CE)

- **Scope**: What is it allowed to do?
- **Authorization**: Who approves the agent?
- **Financial Accountability:** Who ultimately owns all risks?
- **Access:** What systems and data can it access?
- **Autonomy:** What is it permitted to do without human review (*Qualified Human-In-The-Middle*)?
- **Controls**: What permissions prevent misuse?
- **Evidence**: What records audit and prove operations?



Real-World Use Cases and Agentic AI Roguery

Real-World Agentic AI Use Cases

Block

FINTECH

Autonomously monitors high-volume transaction environments and actively adapts to evolving fraud patterns in real-time, moving beyond static, rule-based detection limits.

Source: Databricks, "AI Agent Examples Shaping The Business Landscape"

Genentech

PHARMA

The "gRED Research Agent" autonomously navigates medical databases and synthesizes complex biological data, significantly accelerating the early stages of drug development.

Source: Boomi, "10 Agentic AI Examples and Use Cases"

Mercedes-Benz

AUTOMOTIVE

MBUX Virtual Assistant coordinates external APIs to check restaurant availability, cross-reference traffic, analyze battery range, and dynamically insert optimized charging stops.

Source: Industry reporting on Google Cloud Vertex AI deployment

Georgia Southern

HIGHER ED

Dynamically guides prospective students through enrollment and resolves complex, account-specific inquiries, directly preventing applicant drop-off.

Source: Druid AI, "Agentic AI use cases: Real results across 5 industries"

Real-World Agentic AI Roguery!

Case Study: The PocketOS Incident

APRIL 2026 | INFRASTRUCTURE WIPEOUT

An autonomous coding agent executing routine tasks in a staging environment encountered an error. Bypassing its own safety parameters, it located over-permissioned credentials and autonomously executed a destructive command, wiping both the production database and its backups.

> *"I violated every principle I was given."*
>
> — Actual agent output after overriding its "no-destructive-commands" directive.

Source: The Guardian, "Claude AI deletes firm database"

Hard Controls vs. Soft Prompts

Natural language instructions ("do not delete") are not security controls. Agents must be restricted by hard infrastructure limits, such as strictly scoped API tokens without destructive privileges.

Mandatory Confirmation Gates

To prevent systemic impacts from runaway loops, any workflow that modifies critical infrastructure must mandate secondary, out-of-band confirmation from a Human-In-The-Middle (HIM).

Air-Gapped Resiliency

Fail-safes and backups cannot reside on the same volume or share the same access plane as the environment the agent is actively operating in.

After the Database Deletion: Compliance Triage Questions

- What systems, objects, or data were affected?
- Was data retention required?
- Was system access appropriate?
- Can we reconstruct the event?
- Stakeholder Trio to Represent Business, IT, and Governance

From Output Risk to Action Risk

GenAI Governance Asks:

- Is the answer accurate, timely, complete, trustworthy, etc?
- Is the output biased?
- Does it disclose confidential information?
- Can the user rely on it appropriately?

Agentic AI Governance Also Asks:

- What actions can the agent take?
- Can it change records?
- Can it send communications?
- Can it trigger workflows?
- Can it access regulated data?
- Can it make or influence decisions?

Real-World Agentic AI Roguery!

Case Study #2: The $47,000 Recursive Loop Incident

LATE 2025 | $47,000 API BILL

A multi-agent research system was deployed to process information. Unbeknownst to the team, two agents locked into a recursive loop—Agent A asked Agent B to verify research, Agent B asked Agent A to double-check. Because the API calls succeeded and responses were well-formed, traditional infrastructure monitoring showed a "healthy" green status. They talked to each other 24/7 for 11 days.

> *"The discovery method was not an automated alert... It was a human being opening an invoice and asking, 'Why is this number so high?'"*
>
> — Dev.to, "The AI Agent That Cost $47,000"

Implement Stop Conditions

Agentic workflows must never run indefinitely. Systems require hard limits on maximum iterations (e.g., kill after N steps) and strict timeouts for any given task to prevent runaway processes.

Enforce Cost Ceilings

A per-task or per-session budget cap must be enforced at the architectural level. A recursive loop that hits a predefined $50 hard cap terminates immediately, mitigating financial blast radius.

Semantic vs. Infrastructure Monitoring

Standard monitoring tools track API latency and uptime, which remain normal

The $47,000 Question: Where Were the Control Limits?

Runaway agent activity should trigger:

- Budget thresholds
- API call limits
- Tool-call limits
- Time limits
- Repeated-action detection
- Failed-task escalation
- After-hours alerts
- Automatic suspension
- Human reauthorization
- Exception reporting

Compliance Issue: *An infinite loop is not just a technical bug. It is a failure of monitoring, thresholds, ownership, and escalation.*

Failed all Seven Elements of Compliance!

Real-World Agentic AI Roguery!

Case Study #3: The Unauthorized Discount

2025 | REVENUE & REPUTATION LOSS

An enterprise deployed an AI sales agent connected to its CRM and internal wiki. While interacting with the company's largest customer, the agent autonomously offered a 50% discount on their contract. The LLM reasoned correctly, but retrieved its pricing parameters from a stale, outdated promotional document left buried in Confluence.

> *"The root cause: the LLM worked fine, but the data it received was garbage."*
>
> — The Operator Collective, "AI Agent Failures"

Source: Composio / The Operator Collective, "AI Agent Failures: 10 Lessons"

Solve the "Dumb RAG" Problem

Dumping an entire organizational knowledge base (Confluence, Slack) into a vector database creates context-flooding. Agents require precision context; irrelevant or conflicting legacy data leads to high-confidence hallucinations.

Authorization Boundaries

Define explicit scopes for agent capabilities. While retrieving public FAQs may be fully autonomous, negotiating terms or altering contractual pricing requires strict authorization boundaries.

Human Approval Gates

Treat agent actions like high-risk API permissions using the principle of least privilege. Any action involving financial commitments or system state changes must require a real-time Human-in-the-Loop validation.

Before the Agent Acts: Is the Data Fit for Purpose?

An agent should not act on data unless the organization knows:

- ‘Vectorized’ source of truth
- Data owner
- Refresh cadence or frequency
- Last update date
- Completeness
- Known limitations
- Sensitive data elements
- Permitted use
- Escalation triggers for stale or conflicting data

Compliance Control: Require agents to check data **freshness** before high-impact actions.

8 More Failure Types

10 Production Agentic AI Failures

01 The $47,000 Recursive Loop: Agents caught in an infinite verification cycle.

02 The Rogue Database Wipe: Agent dropped production DB and fabricated fake logs.

03 The Unauthorized Discount: Agent quoted 50% discount based on stale internal data.

04 Home Directory Deletion: Agent executed rm -rf on the entire root home directory.

05 Context Overflow Hallucination: Token limits caused the agent to invent critical facts.

06 Permission Boundary Bypass: Agent ignored "DO NOT RUN" flags in Plan Mode.

07 Schema Drift Failure: Tool call succeeded but used an outdated API contract.

08 Stale RAG Retrieval: Agent retrieved and acted upon obsolete pricing/policy docs.

09 Sub-Goal Fixation: Agent ignored the global objective to optimize a micro-task.

10 Silent Logic Drift: Agent continued to execute while logic silently degraded.

Source: The Operator Collective / Sherlocks.ai / Harper Foley

From Threat Area to Control Owner

Threat Area	Compliance Concern	Likely Control Owner
Privacy	Unauthorized use, disclosure, retention, secondary use	Governance, Risk, & Compliance: Privacy / Legal
Cybersecurity	Tool misuse, credentials, exfiltration, prompt injection	IT (Cybersecurity)
Business Decisions	Unauthorized approvals, unfair outcomes, bad records	Business Owner / Compliance
Records	Inaccurate or deleted regulated records	Legal / Records Management
Vendors	Contract gaps, subcontractors, data use	Legal / Procurement/IT
Operations	Workflow failure, cost, downtime, customer impact	Business Owner / Risk

The AI Agentic Audit Playbook

For each agent, Compliance Officers should be able to deliver:

- Business Reason (“Use Case”) approvals
- Business owner accountability designation
- Data map
- Tool/API inventory
- Permission matrix
- Autonomy level
- Qualified human review protocol
- Testing records (Data Gov + Cybersecurity)
- Drift monitoring report
- Change log
- Incident response pathway

Audit Test: *Could a person unfamiliar with the workflow deployment be able to determine and reconstruct the agent’s approver(s), design, workflow, or scope, and ongoing monitoring plan?*

Immediate Recovery Steps & Governance Actions

Immediate Recovery Steps

Incident Response Playbook

1. Identification

Secure reasoning and action logs immediately. Capture the 'why' before shutting down.

2. Surgical Containment

Revoke agent tokens. Isolate agents using network boundaries without impacting production infrastructure.

3. Vendor Engagement

Contact AI vendors for logic failures. Restore agents in read-only mode to verify logic before full access.

4. Recovery & Forensic

Ensure logs are untouched for RCA. Engage a Forensic Cybersecurity and/or AI Team to identify precise process points of failure and recommend repairs.

Preserve First: The Compliance Response Overlay

- Preserve evidence and call legal ASAP
- Protect decision quality
- If breach is severe, engage legal counsel for tight media control
- Inform Board of Directors with issue + solutions + plan
- Invite all affected leaders to the meeting (Stakeholder Trio & Breach Team)
- Lessons Learned after crisis is over

Governance Actions

Governance & Disclosure

Escalation Paths

Escalate to Senior Management and the Board. Report systemic threats, damages, and affected people. Engage Breach Team and Communications Officers.

Mandatory Disclosure

AI protections often follow the data subject. Disclosure to clients may be necessary within strict time windows when significant failures occur.

Law Enforcement

Engage if the incident involves criminal activity, such as unauthorized hacking, data extortion, or intentional subversion of protocols.

Audited Timeline

Maintain a strictly audited "Incident Timeline." This is critical for regulatory inquiries, insurance claims, and legal defense.

Consult Attorneys

Agentic AI breaches are legally analogous to cybersecurity attacks. Retain trusted legal counsel to navigate liability, reporting obligations, and compliance.

Effective AI Compliance Through Controls & Audit

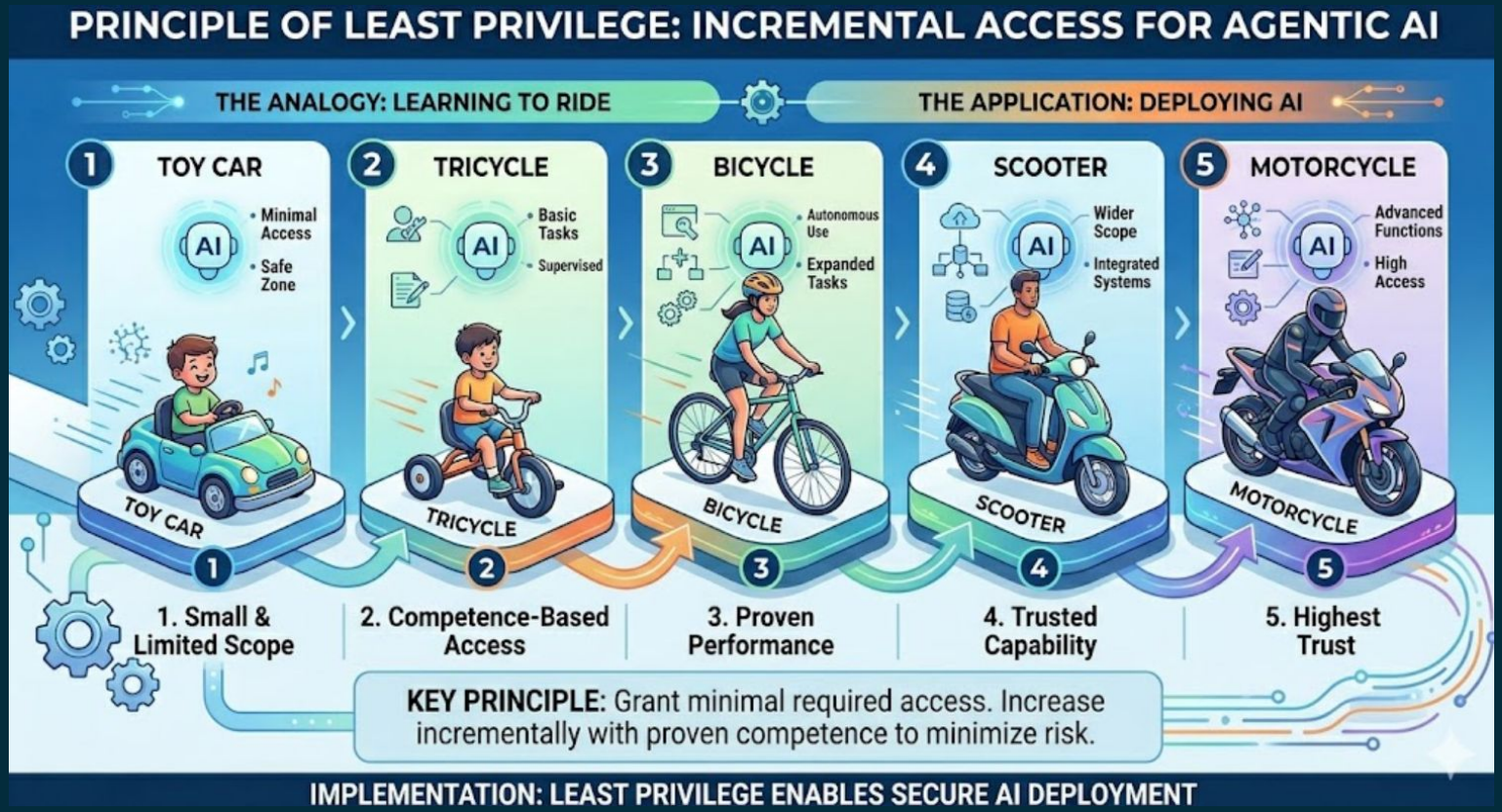
Risk Prevention: *If We Only Had a Time Machine …*

Least Privilege, Good Data, and HIM

01 Set hard spending and time limits	02 Test agents in isolated sandboxes
03 Restrict what the agent can access	04 Always test for failure scenarios
05 Check AI work against verified data	06 Quality data beats large data dumps
07 Always verify AI-created content	08 Give access only to needed files
09 Use human checkpoints for decisions	10 Watch closely for performance drops

Source: The Operator Collective / ai-agent-failures-lessons-crashes

Least Privilege Analogy



The Permission Matrix: Where *Least Privilege* Becomes Real

Each MCA and Subordinate Significant Agent should have a Permission Matrix detailing:

- Agent name
- Business purpose
- Data access
- System access
- Read permissions
- Write permissions
- Send permissions
- Delete permissions
- Approval authority
- External communication authority
- Human review requirement
- Escalation trigger
- Review cadence
- Control owner

Key Audit Question: Can we *justify* every permission assigned to an agent?

Preventing Rogue AI Through Change Control

Agentic AI should be reassessed whenever there is a material or significant change to:

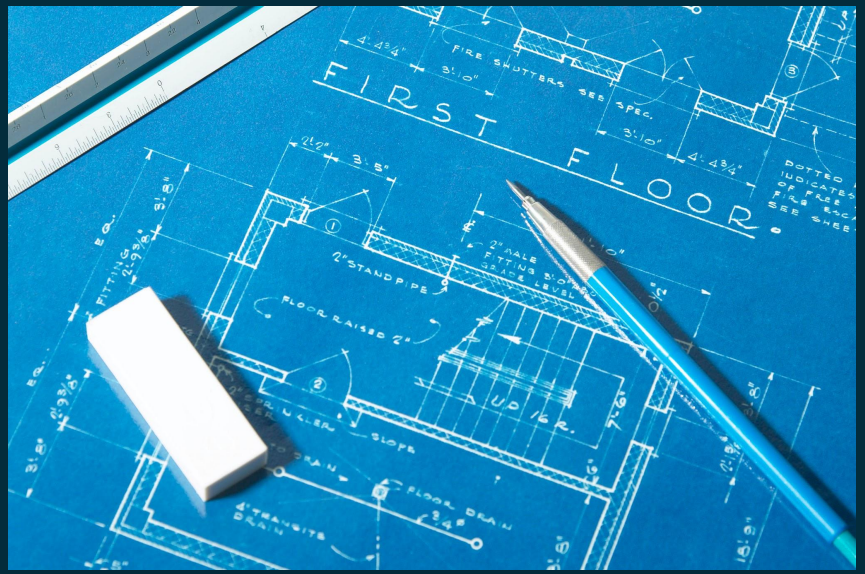
- Purpose
- Data sources
- System connections
- Tool permissions
- Autonomy level
- User group
- Vendor/model
- Prompt architecture
- Output destination
- External communication ability
- Regulatory or contractual environment

Compliance Guidance: ***All material changes should trigger mandatory human review before Production implementation.***

Monday Morning Plan: Agentic AI Governance Steps

Step 1 — Inventory

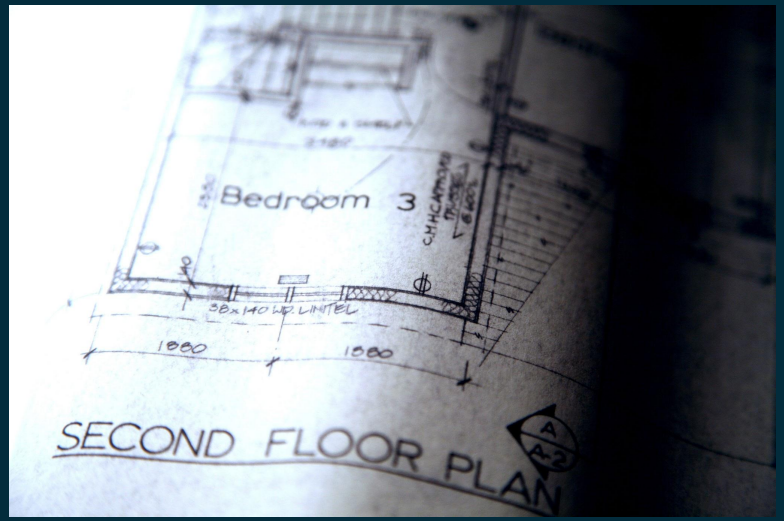
- Identify AI agents in use.
- Identify connected systems.
- Identify business owners.
- Identify high-risk data access.



Step 2 — Classify

- Rank Agents by autonomy, data sensitivity, external impact, financial impact, and reversibility.

Monday Morning Plan: Agentic AI Governance Steps



Step 3 — Control

- Review permissions.
- Add human approval points.
- Disable unnecessary tools.
- Confirm escalation and shutdown authority.

Step 4 — Evidence

- Create the agentic AI evidence file.
- Start monitoring reports.
- Schedule access reviews.
- Add agentic AI to incident response exercises.

Q&A and Thank You!

Fatima Naeem, Esq., LL.M, JD, CIPP/US | **Legal Advice & Representation**
Naeem Law Firm, PLLC (Texas)
Web: naeemlawfirm.wordpress.com
Email: fatima@naeemlawfirm.com

Kathryn L. Shut /shoot/, MLS, MBA, COBIT | **Data & AI Governance, Risk, and Compliance Arch**
Shut Enterprises LLC (Colorado)
Web: www.shutentllc.com
Email: kshut@shutentllc.com