HOUSE STRATEGIES GROUP

Responsible AI: A Practitioner's Primer

The shared foundation every House Strategies Group team member works from — the language, frameworks, and judgment behind our Responsible-AI practice.

House Strategies Group · Responsible-AI Practice · Developed by HSG leadership · Required reading for all team members

1 · The one idea to anchor on

You are not managing AI. **You are managing the governance of AI** — the rules, checks, evidence, and people that make an organization's AI use accountable, fair, transparent, and trustworthy over time. The models are the team's job; the *system that keeps the models honest* is yours.

Responsible AI exists because AI fails in ways ordinary software doesn't: it can be **biased** (treat groups unequally), **opaque** (no one can explain why it decided something), **brittle** (it silently degrades as the world changes), and increasingly **autonomous** (it acts without a human in the loop). Three forces converge to demand a response:

- **Policy** — federal guidance now expects AI governance (NIST, OMB, GAO).
- **Risk** — a bad model decision in an oversight body can harm people and the institution's credibility.
- **Trust** — an Inspector General that uses AI must be able to defend *how* it uses it, to Congress and the public.

> **THE LENS THAT ANCHORS EVERYTHING**
>
> Every Responsible-AI deliverable is an answer to one question an auditor will eventually ask: *"Show me how you knew this AI was working, fair, and under control."* If a deliverable doesn't help answer that, it's decoration.

2 · The canon, decoded

Four bodies of guidance matter. You don't need to memorize them — you need to know what each is *for* and how they fit together.

GAO-21-519SP — the backbone (read this one properly)

The Government Accountability Office's **AI Accountability Framework.** It is the lens the Inspector-General community uses to *audit* AI, which is exactly why we hold the programs we build to it. Four principles, each answering a plain question:

Principle	The question it answers
Governance	Who is accountable for this AI? Are there policies, roles, and oversight?
Data	Is the data appropriate, representative, well-documented, and properly handled?
Performance	Does it actually work as intended — accurately, reliably, and fairly?
Monitoring	Does it keep working over time, and who is watching for drift and incidents?

NIST AI RMF — the operating model

The National Institute of Standards and Technology's **AI Risk Management Framework** turns the principles into a lifecycle with four functions: `Govern` (the culture/policies that run across everything) → `Map` (understand each AI's context and impact) → `Measure` (test it — accuracy, bias, explainability) → `Manage` (treat the risks, monitor, respond). Its companion **Playbook** is the practical "what to actually do" per function. Think: GAO tells you *what good looks like*; NIST tells you *how to operate it*.

OMB M-25-21 & M-25-22 — the federal rulebook (the floor)

White House Office of Management and Budget memos (April 2025). **M-25-21** tells federal agencies to appoint a **Chief AI Officer**, stand up an **AI Governance Board**, maintain a **use-case inventory**, and apply **minimum risk practices to "high-impact" AI. M-25-22** governs how they *buy* AI (lifecycle monitoring, no vendor lock-in).

> **AN IMPORTANT NUANCE**
>
> These OMB memos **do not legally bind an Inspector General** — IGs are independent, and the Postal Service sits outside the FAR entirely. So we treat M-25-21 as a *floor to meet-or-exceed* and build the framework around the IG's independence, rather than bolting on generic compliance.

NIST Generative AI Profile + the Traffic Light Protocol

The **GenAI Profile (NIST AI 600-1)** is a checklist of risks unique to generative AI — the headline one is **confabulation** (the model states false things confidently). The **Traffic Light Protocol (TLP)** is a simple data-sharing classification (RED/AMBER/GREEN/CLEAR) we extend into an "Enhanced TLP" for the AI lifecycle.

3 · Ten concepts you must be able to discuss

For each: what it is, why it matters here, and the red flag that should make you stop the team.

1. Use-case intake & AI inventory

WHAT IT IS	A standard "front door" every proposed AI use passes through, and a living register of every AI system in use.
WHY IT MATTERS	You can't govern what you can't see. The inventory is the foundation everything else hangs on.
RED FLAG	A model is in production that never went through intake, or isn't in the inventory.

2. Risk tiering

WHAT IT IS	Sorting each AI into **High-Impact / Limited / Minimal** based on consequences — so scrutiny matches stakes.
WHY IT MATTERS	High-impact systems (e.g., a fraud score that triggers an investigation) get the heavy controls; a doc classifier doesn't.
RED FLAG	A consequential, rights-affecting model is being treated as "low risk" to skip review.

3. Bias & fairness

WHAT IT IS	Whether a model treats groups unequally. The classic test is the **4/5ths rule**: if one group's selection rate is below 80% of the top group's, that's a disparate-impact flag.
WHY IT MATTERS	An oversight body that enforces fairness cannot field a biased model. Mitigation happens pre-, in-, and post-processing.
RED FLAG	Geographic or proxy variables (ZIP, route) that quietly stand in for race or income.

4. Explainability (XAI)

WHAT IT IS	Making a model's decision understandable. **SHAP** and **LIME** show which inputs drove a given prediction.
WHY IT MATTERS	No one should act on a score they can't explain — especially in law enforcement. "Show your work" is non-negotiable.
RED FLAG	A high-stakes recommendation with no explanation attached, or "the model just said so."

5. Drift & monitoring

WHAT IT IS	Models silently decay as the world shifts. **PSI** (Population Stability Index) measures how far inputs have moved; cross a threshold and you retrain.
WHY IT MATTERS	"It worked at launch" is not a defense. Governance is continuous, not a one-time sign-off.
RED FLAG	A production model with no monitoring, or no defined action when it drifts.

6. Human oversight

WHAT IT IS	**In-the-loop** (human approves each action) · **on-the-loop** (human supervises, can intervene) · **in-command** (human sets bounds, AI runs within them).
WHY IT MATTERS	The oversight model must match the stakes. Consequential decisions need a human in the loop, logged.
RED FLAG	An enforcement action taken on an AI output with no documented human review.

7. Evidence logs, model cards & datasheets

WHAT IT IS	The paper trail. **Model cards** document a model (purpose, data, limits); **datasheets** document a dataset; **evidence logs** capture decisions and reviews.
WHY IT MATTERS	This documentation discipline *is* half of AI governance — it's what makes the program audit-ready.
RED FLAG	"We'll document it later." Undocumented = ungoverned.

8. Agentic AI risk

WHAT IT IS	AI that takes actions autonomously (multi-step workflows, tool use). New failure modes: **prompt injection**, **scope creep**, runaway actions.
WHY IT MATTERS	Agencies are deploying agentic capability now — governing it (action allowlists, immutable logs, approval gates, red-teaming) is fast becoming essential.
RED FLAG	An agent with permission to act on systems and no allowlist or human approval gate.

9. Data classification (Enhanced TLP)

WHAT IT IS	Labeling data/outputs by sensitivity (RED/AMBER/GREEN/CLEAR) plus AI overlays: provenance, PII/CUI layering, role-based access, time-based reclassification.
WHY IT MATTERS	An IG handles law-enforcement-sensitive data; mishandling it in an AI pipeline is a serious incident.
RED FLAG	Sensitive case data flowing into a tool not cleared for that tier.

10. Performance KPIs & ROI

WHAT IT IS	The metrics that prove the AI program is working and worth it — accuracy, override rates, time saved, recoveries attributed.
WHY IT MATTERS	Leadership funds what it can measure. KPIs turn "trust us" into a dashboard.
RED FLAG	A program that can't state, in numbers, whether its AI is helping.

4 · Why an Inspector General is different

This is where Responsible-AI work for an oversight body diverges from generic AI governance. An Inspector General is not a normal agency:

- **Independence.** An IG audits its agency; it must be able to audit *itself* to the same bar. Governance can't be borrowed boilerplate — it has to be defensible under the IG's own scrutiny.
- **Law-enforcement-sensitive data.** Investigations, whistleblower identities, active cases. The framework must protect this above ordinary CUI — the Enhanced TLP exists for exactly this.
- **Audit-your-own-AI.** The most powerful framing: the OIG should govern its AI to GAO-21-519SP — the same framework it would use to audit anyone else's AI. Practice what you oversee.
- **Community leadership.** To our knowledge, the IG community (CIGIE) hasn't published a Responsible-AI framework — so an IG that builds one well can lead the community, not follow.

> **THE PRINCIPLE IN ONE LINE**
>
> Build the oversight body a Responsible-AI program it could audit itself — tuned to its independence and its law-enforcement data — not a corporate template with a logo swapped in.

5 · Operating discipline

Questions worth asking every week

- What's in the inventory that wasn't last week — and did it go through intake?
- Any model drift or fairness flag triggered? What was the action?
- Which high-impact systems still lack explainability or a human-review gate?
- What evidence did we capture this week that an auditor could pull?
- Are we documenting as we go, or building a documentation debt?

Failure modes to watch

- **"Maintain" treated as one-time.** A living inventory and risk registry are ongoing operations, not a build-once deliverable.
- **Replacing instead of integrating.** Governance should travel with the client's existing analytics stack, not impose a closed system.
- **Boilerplate drift.** The moment a deliverable reads like a generic policy, it stops being defensible under real scrutiny.
- **Documentation debt.** Undocumented governance is exactly what an auditor flags first.

6 · Cheat sheet

Glossary — the terms you'll hear

AI RMF — NIST's AI Risk Management Framework (Govern/Map/Measure/Manage).

GAO-21-519SP — GAO's AI Accountability Framework; the IG audit lens.

M-25-21 / M-25-22 — OMB memos on federal AI use / acquisition.

CAIO — Chief AI Officer; the accountable AI owner.

High-impact AI — AI with significant effect on rights, safety, or resources.

Bias / disparate impact — unequal outcomes across groups.

4/5ths rule — fairness test: a group below 80% of the top selection rate is flagged.

SHAP / LIME — methods that explain why a model made a prediction.

Drift — a model degrading as real-world data shifts.

PSI — Population Stability Index; a drift metric.

Confabulation — a GenAI model stating false things confidently.

HITL / HOTL — human in-the-loop / on-the-loop oversight.

Model card / datasheet — standard docs for a model / a dataset.

Agentic AI — AI that autonomously takes multi-step actions.

Prompt injection — malicious input that hijacks an AI's instructions.

TLP — Traffic Light Protocol; data-sharing sensitivity labels.

PII / CUI — personal / controlled-unclassified information.

Red-teaming — adversarial testing to find an AI's failure modes.

Use-case inventory — the register of every AI system in use.

Evidence log — auditable record of AI decisions & reviews.

> **THE WHOLE PRIMER IN ONE SENTENCE**
>
> Govern the AI to the same standard an oversight body would use to audit it — accountable, fair, explainable, monitored, documented — and you can approach any Responsible-AI engagement with confidence.