Responsible AI Roadmap

Risk & Quick Mitigations (High-Impact Essentials)

Navigate AI. Advance what matters.

📋 ***Human-led interventions*** | 🖥️ ***AI prompt tools***

v2026.02.13
latest version at
alignIQ.ai

🎲 Step 1 Sandbox → Basics to experiment safely

🛡️ *Identify Safe vs. Unsafe Data*

- ❑ Use org-approved tools & connectors
- ❑ **Turn off training** in public chats
- ❑ **Classify data Red / Amber / Green**

Pro tip: *Know retention/logging defaults*

Data traffic light defaults (org policy overrides)
Red = secrets/regulated → never paste.
Amber = internal/drafts → mask & minimize.
Green = public→ safe. When unsure, treat as Amber

- ❑ Never paste Red data

🔎 *Expect & Detect AI Hallucinations*

- ❑ Follow the 3-source rule
- ❑ Verify numbers & dates

Pro tip: ***Run the "could I defend this?" test:*** *rewrite it, flag what you can't defend, and verify those parts*

- ❑ Ask for citations, link & quote
- ❑ Request a brief counter- argument
- ❑ Experiment regularly to map the edge of AI capabilities

🤝 *Maintain Stakeholder Trust*

- ❑ **Disclose** AI editing/summarizing/drafting help
- ❑ Don't forward raw output
- ❑ Keep **human sign-off** and name the owner
- ❑ High-stakes: two-person review
- ❑ Log prompts (date + purpose) for stakeholder work

Pro tip: *Keep a disclosure template bank, and match details to stake and sensitivity*

🏗️ Step 2 Workflow → Daily use with approved AI tools

🛡️ *Protect Data & Tools at Work*

- ❑ Store and delete data per IT policy
- ❑ Only approved AI tools + extensions
- ❑ Treat new extensions/skills as untrusted (audit/allowlist)

Pro tip: *Pilot in a sandbox; inspect traffic if possible*

🔎 *Prioritize & Mitigate Bias*

- ❑ **Triage** risks by likelihood and severity
- ❑ Review outputs across multiple audience perspectives
- ❑ Document criteria

- ❑ Run persona-swap tests
- ❑ Add **inclusive constraints**

Pro tip: *Ask for plain language, alt-text/captions, or accessibility options before finalizing*

🌿 *Keep Healthy AI Boundaries*

- ❑ **Your POV first, then AI**
- ❑ AI windows; breaks stay breaks
- ❑ Single-task & batch check-ins
- ❑ Key work: quick peer sync

Pro tip: *Pause: What's the goal? What's non-negotiable? Is this my job?*

- ❑ Ask it to first ask you 3 questions
- ❑ Ask for 1 risk + 1 verify step

🏰 Step 3 Strategy → Advanced/long-term alignment

🛡️ *Access, Compliance, & Agent Controls*

- ❑ ✓ permission, data region, retention
- ❑ Least privilege by default
- ❑ Allowlisted extensions (owner reviewed)
- ❑ Require agents confirm before action
- ❑ Report incidents to AI lead/IT

Pro tip: *Review access quarterly; retire what doesn't work*

- ❑ If unsure, treat it as **RED** and stop

🤝 *Align AI with Mission & Values*

- ❑ Keep do/don't-automate list
- ❑ Mission in custom instructions
- ❑ Start projects with a **values preamble** + audience in mind

Pro tip: *Occasionally log when choosing not to delegate to AI and why*

- ❑ **Write mission** into the prompt's require-ments

🌿 *Use AI Sustainably*

- ❑ Choose right-sized models for the task
- ❑ Ask vendors for energy, water, and carbon disclosures

Pro tip: ***Track environmental impact at the org-level, not per use***

- ❑ Prefer text over media
- ❑ Batch requests