Responsible AI Roadmap

Risk & Quick Mitigations (High-Impact Essentials)

Navigate AI. Advance what matters.

v2026.03.05
latest version at
aligniq.ai

📋 ***Human-side guardrails*** | 🖥️ ***AI-side guardrails***

Follow in order: ① → ② → ③ · Read left→right within each step. · ☐ = action. · Bold = do first.

🎲 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 = people/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*

- ❑ **High-stakes: two-person review or 2-source verify**
- ❑ Verify numbers & dates
- ❑ Label certainty + 1 check (confidence ≠ accuracy)

Pro tip: *Run the "defend it" test: rewrite, flag unverifiable claims, verify those*

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

🤝 *Maintain Stakeholder Trust*

- ❑ Use AI for information, not relationships
- ❑ Disclose AI editing/summarizing/drafting help
- ❑ **Don't forward raw output**
- ❑ Human sign-off + named accountable owner
- ❑ 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 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 of a different audience
- ❑ Add inclusive constraints

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

🌿 *Keep Healthy AI Boundaries*

- ❑ **Brain-first: your POV, then AI**
- ❑ **AI windows; breaks stay breaks**
- ❑ Single-task & batch AI checks
- ❑ Key work: quick human grounding

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

- ❑ 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*

- ❑ Check permission, data region, retention (per tool)
- ❑ **Grant minimum access**
- ❑ Report incidents to AI lead/IT

Pro tip: *Quarterly review; revoke unused*

- ❑ **If unsure, treat it as RED and stop**
- ❑ Require agents confirm before action
- ❑ **If an agent wants new access, stop + confirm**

🤝 *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: *Save time on mundane work → reinvest in human work*

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

🌿 *Use AI Sustainably*

- ❑ Choose right-sized models for the task
- ❑ **Watch dependency & keep core skills sharp**

Pro tip: *Per-prompt impact is small; focus on org/vendor choices*

- ❑ Prefer text over media
- ❑ Batch requests