

Agentic AI Guardrails Checklist

For ChatGPT, Microsoft Copilot, Copilot agents, and any AI agent that can take actions. Adapt each item to your team, check it off, and revisit it every quarter.

ORGANIZATION _____ COMPLETED BY _____ DATE _____

01 What agents can access

- ☐ List every AI tool and agent in use and what it connects to (email, files, CRM, calendar).
- ☐ Grant least privilege: only the folders, mailboxes, and apps the task truly needs.
- ☐ Use read-only access wherever it works.
- ☐ Keep sensitive client data (health, financial, IDs, case notes) out of reach.
- ☐ Remove access when an agent, project, or staff role ends.

02 What agents can do

- ☐ Label each agent: read only, draft only, or can take action.
- ☐ Require a human to approve before an agent sends email or messages, makes payments, deletes records, or shares anything outside your organization.
- ☐ Turn off actions an agent never needs.
- ☐ Start new agents in draft-only mode until you trust them.

03 Who owns it

- ☐ Name one owner for each agent: a real person, not just "IT."
- ☐ The owner knows what it does, what it can reach, and who uses it.
- ☐ New agents and new permissions get a quick approval: who asked, why, what data, who said yes.
- ☐ Keep a simple register: agent · owner · access · approval date.

04 Logging & review

- ☐ Turn on activity or audit logs in your admin console (e.g., Microsoft 365).
- ☐ Do a monthly review: who used it, what it touched, anything surprising.
- ☐ Spot-check a few outputs each month for accuracy, tone, and privacy.
- ☐ Write down what you found and fixed in the register.

05 Staff rules

- ☐ Check AI outputs before they go to a client, funder, or the public.
- ☐ Never paste confidential or client data into tools that aren't approved.
- ☐ Use work accounts only, not personal AI accounts.
- ☐ Report mistakes or odd behavior right away. No blame, just fix it.

06 Stop switch

- ☐ Write down how to pause or disable each agent, and who is allowed to.
- ☐ Know how to disconnect its access (apps, mailboxes, API keys).
- ☐ Incident steps: 1) Pause it. 2) Save screenshots and logs. 3) Tell the owner and leadership. 4) Fix it and notify anyone affected. 5) Review before turning it back on.

Tip: Start with your highest-risk agent, the one that can send, pay, delete, or share, and work through all six sections for it first.