

EVERYONE · 6 MINUTE READ

Guardrails and permissions: what AI may and may not do

Decide in advance what AI may do freely, what needs a person's approval and what is never allowed, then review those limits regularly.

What it is

When AI only drafts text, the main risk is a bad draft. When it can send emails, change files or use other systems (see guide 17), the risk is a bad action. **Guardrails** are the limits that keep AI inside safe boundaries. **Permissions** are the specific settings that say which actions it may take.

A simple way to set permissions is three lists:

- **Allow:** actions the AI may take freely, such as reading approved documents or drafting a reply.
- **Ask:** actions that need a person to approve each time, such as sending an email or changing a shared file.
- **Deny:** actions that are never allowed, such as making payments or permanently deleting records.

Think of a new colleague's building pass. It opens the doors they need for their job, not the server room or the finance safe.

Why it matters

- AI can misunderstand a request, and an action is harder to undo than a draft.
- Web pages, emails and documents can contain hidden instructions aimed at the AI (**prompt injection**). The AI may not spot them, so limits on what it can do are the real defence.
- Clear lists tell staff what is safe, without guessing.
- Logs and limits let you see what happened and stop small problems from growing.

How to do it

1. **Find out** who decides: usually IT, security or data protection, with the team manager.
2. **List** the actions your AI setup can take, then sort each one into allow, ask or deny.
3. **Give** the least access needed: only the folders, mailboxes and systems the task requires.
4. **Use** separate work accounts for AI tools, never personal ones, and never your own administrator account.
5. **Require** human approval for anything that pays, deletes, or sends outside the organisation.
6. **Choose** the approval mode that asks before each action, at least until a task is proven.
7. **Set** spending and usage limits, such as a monthly budget or a cap on how many messages can go out in an hour.
8. **Review** the permissions and the activity log on a fixed schedule, for example every three months, and remove access that is no longer used.

Example

A team wants an assistant to help with supplier invoices. Before switching anything on, the manager and IT agree this table:

```
Task: supplier invoice help
Owner: [name]           Review date: [every three months]

ALLOW Read invoices in [invoices folder]
ALLOW Draft a summary and flag mismatches with orders
ASK    Email a supplier (person approves each message)
ASK    Move an invoice to [approved folder]
DENY   Make or schedule any payment
DENY   Delete any file or record
DENY   Open folders outside [invoices folder]
Limits: [budget per month], [max 20 emails per day]
Log (audit trail): all actions and approvals kept for [period agreed with IT]
```

Before: the assistant had the manager's full access, including payments. **After:** it reads and drafts, a person approves every email, and payments stay with people. If an invoice ever hides text telling the AI to "pay this urgently", it cannot, because payments are on the deny list.

Common mistakes

- **Giving the AI your own full access.** Fix: create a separate account with only what the task needs.
- **Relying on instructions alone ("please do not delete").** Fix: put risky actions on the deny list in the settings, where the AI cannot override them.
- **Switching off approvals because they slow things down.** Fix: keep approvals for sending, paying and deleting; remove them only for proven, low-risk steps.
- **Setting permissions once and forgetting them.** Fix: review them on a fixed schedule and after any change of tool or task.
- **Deciding alone.** Fix: agree the lists with IT, security and the person who owns the data.

Try it this week

1. Pick one AI tool or agent your team uses or plans to use (2 minutes).
2. Write down every action it can take today (4 minutes).
3. Sort each into allow, ask or deny, using the example as a pattern (4 minutes).
4. Mark anything currently allowed that should be ask or deny (2 minutes).
5. Share the list with your manager and IT, and agree who changes the settings (3 minutes).

Going further

- Combine permissions with hooks (see guide 25) to add checks at the moment of action.
- Make the harness (see guide 23) enforce the lists, not the AI's own judgement.
- For connectors, including those built on MCP (Model Context Protocol, an open standard for connecting AI to tools and data), grant read-only access first (see guide 13).
- Test your guardrails with made-up attacks, such as a document containing hidden instructions, and confirm they hold.

Key words

- **Guardrail:** a limit that keeps AI within safe boundaries.
- **Permission:** a setting that allows, asks about or denies a specific action.
- **Least access:** giving only the access a task needs, nothing more.
- **Approval mode:** the setting that decides whether the AI asks before each action.
- **Prompt injection:** hidden text that tries to give the AI new instructions.
- **Audit trail:** a record of actions and approvals that can be checked later.