AI coding assistants: what can your agents reach and run?

This month, move every developer to a company account, require the tool's sandbox with bypass modes switched off in managed settings, and keep production credentials away from agents. Judge every new tool, MCP server or automation by what outsiders can write into what it reads, and what it could run, reach or merge if that text took over.

Reviewed	Next review	Version	Written for
5 Nov 2026	**5 Dec 2026**	**1.2**	**Engineering leads, platform and security teams whose developers use AI coding assistants and agents**

Latest version
sheets.einblick.dev/ai-coding-assistants/

If you only do five things

1. **Move everyone to company accounts.**
   Since 24 April 2026, GitHub trains on Copilot Free, Pro, Pro+ and Max interactions by default, not on Business or Enterprise[1]. Block `*.individual.githubcopilot.com` at the firewall[2].

2. **Keep production credentials away from agents.**
   Agent commands run with the developer's credentials[4]. In April 2026 an agent found an account-scoped Railway token on a developer's machine and deleted a production database volume with it[5]. Hand out scoped, non-production tokens.

3. **Require the sandbox and block bypass modes.**
   Claude Code's sandbox is off by default and still reads `~/.ssh` unless you deny it[6]. Require it in managed settings, set `disableBypassPermissionsMode` in Claude Code and Copilot[7,8] and keep yolo modes for offline throwaway containers[7].

4. **Treat inputs as instructions from their authors.**
   Issues, pull requests, docs and config files can all steer an agent. Cline's triage agent had a shell and read any GitHub user's issues, which exposed its publishing credentials; an npm token Cline failed to rotate let an outsider publish a release[9]. Agents that read outsiders' text get no shell and no secrets.

5. **Have a second person approve every agent change.**
   Copilot cloud agent can't approve its own pull requests, nor can the person who asked, and workflows wait for a human by default[10]. Local agents push as the developer[4], so require approval of the latest push by someone else, as in the secure development sheet.

Roll out an assistant or agent to a team

1. **Settle plans and data terms first.**
   Choose business plans that don't train on your code, check retention options and sign the processor contract the AI inputs sheet describes. Where licensing matters, block **Suggestions matching public code**, which only IDEs fully enforce[28].

2. **Write managed settings per tool.**
   Require the sandbox, in Claude Code with `failIfUnavailable` and `allowUnsandboxedCommands: false`, which needs macOS, Linux or WSL2[6], disable bypass modes and allowlist MCP servers[8]. Push them so developers can't weaken them[12].

3. **Move unattended work off laptops.**
   Run long autonomous tasks in the vendor's cloud sandbox or a container. Copilot cloud agent pushes to one branch only and works behind a firewall[10], which covers only processes its Bash tool starts, not MCP servers, and can be bypassed[11].

4. **Protect the files that steer the agent.**
   Put `AGENTS.md`, `CLAUDE.md`, `.mcp.json` and the `.github`, `.vscode`, `.cursor`, `.claude` and `.gemini` folders under CODEOWNERS with **Require review from Code Owners**, as GitHub advises for its own agent files[13].

5. **Turn on logs before the first agent pull request.**
   GitHub's audit log keeps agent events under `actor:Copilot` for 180 days, but no local prompts[14]; Claude Code exports tool decisions over OpenTelemetry[15]. In Germany, agree this with the works council, which co-determines monitoring tools[16].

Gaps that catch teams out

- **VS Code's agent sandbox leaves the network open.** Outbound access stays on by default, so restrict it separately[4].
- **Headless runs trust the repository.** `claude -p` runs its hooks and `.mcp.json` servers unasked; use `--setting-sources user` on code you didn't write[17]. Gemini CLI in CI trusted a pull request's `.gemini/` until April 2026, a CVSS 10 flaw[18].
- **Ignore files and deny rules leak.** Copilot content exclusion skips symlinks and may miss indirect access[19]; Claude Code's Read deny rules miss scripts that open files[17]. Keep secrets out of the repository.
- **Cursor trusts every workspace.** Workspace trust is off by default; enforce `security.workspace.trust.enabled` through MDM[20].

Reviewing what agents write and install

- **Working code can be exploitable.** If the prompt skipped security, 35% of working backends by BaxBench's top model, Claude Opus 4.5, were exploitable[21]. Keep agent pull requests behind the same CI gates.
- **Check that every new package is real.** In a 2025 study, 19.7% of packages in code from 16 early-2024 models didn't exist, and 43% of those recurred in all 10 repeats of a prompt[22]. Check a new dependency's history first.
- **Pin MCP servers.** The fake `postmark-mcp` built trust over 15 versions before 1.0.16 secretly BCC'd emails to an outside server[23]. Allowlist servers in managed settings[8] and pin versions.

What changed in 2026

- **Copilot offers EU data residency** since April[24], on Enterprise Cloud with data residency, off by default and at 10% more AI credits[25].
- **GitHub can require local sandboxing** for Copilot, generally available since 7 October, with policies developers can't weaken[12].
- **Copilot's approval can count toward required approvals** since 28 October, if admins allow it[26]. Keep it off where agents open pull requests.
- **Claude Code starts in auto mode** from v2.1.283, with a classifier reviewing actions; `permissions.disableAutoMode` turns it off[7].

What you can skip

- **Banning assistants outright.** It pushes developers to private accounts you can't see. The BSI and ANSSI advise offering approved tools with clear rules instead[3].
- **Security rules in `AGENTS.md` or `CLAUDE.md`.** Instructions guide the model and are not an enforcement boundary[4]. Back each rule with a setting.
- **Hand-tuned command allowlists.** Auto-approval rules parse commands on a best-effort basis, and aliases or quote tricks can slip past them[4]. A sandbox does more for less upkeep.
- **Ranking models by how secure their code is.** Even the top model on BaxBench writes exploitable code[21]. Put the effort into CI gates and review.

Mistakes that cost the most

- **Giving a CI agent a shell and outsiders' text.** Cline calls that "functionally equivalent to giving every GitHub user shell access"[9].
- **Letting a classifier approve risky work.** Anthropic says auto mode doesn't guarantee safety or replace review of sensitive operations[7]. Add `ask` rules for deploys, migrations and pushes, such as `Bash(git push *)`, knowing they miss rewritten commands[17].
- **Skipping workflow approval on agent pull requests.** Admins can let workflows run on Copilot's pull requests without a human[10]. Keep the approval, as for outside contributors.
- **Putting deployment secrets in Agents secrets.** Copilot can read those. Keep what it shouldn't touch in Actions secrets, which it can't[13].
- **Yolo mode in a container with secrets.** A malicious project can then take anything inside, including the agent's credentials[27].

What this looks like at your size

Small team, no security staff

- Buy the business plan of one assistant and keep company code out of personal accounts.
- Require the sandbox and block bypass modes centrally, as Claude Code's server-managed settings on claude.ai allow without MDM[6]. Run autonomous tasks in the vendor's cloud agent, not in yolo mode on a laptop.
- Keep production tokens off developer laptops.

Growing, with IT or security staff

- Push managed settings by MDM: sandbox required, bypass disabled, MCP allowlist, deny rules for secret paths.
- Stream the GitHub audit log and agent telemetry to your SIEM.
- Alert on permission mode changes and new MCP servers.

Enterprise or regulated

- Turn on EU data residency or zero retention where offered, and record the tools in your DPIA.
- Run unattended agents only in cloud sandboxes or ephemeral runners with an egress allowlist and short-lived credentials.
- Red-team agent workflows with poisoned issues and pull requests before granting write access.
- Grow AppSec and review capacity with agent output, as the BSI and ANSSI advise[3].

Version history

5 Nov 2026, version 1.2 Added that the Copilot cloud agent's firewall covers only processes its Bash tool starts, not MCP servers, and can be bypassed.

Sources

1. Managing GitHub Copilot policies as an individual subscriber. GitHub Docs, checked 5 Nov 2026. docs.github.com
2. Managing GitHub Copilot access to your enterprise's network. GitHub Docs, checked 5 Nov 2026. docs.github.com
3. AI Coding Assistants. ANSSI and BSI, Sep 2024. bsi.bund.de
4. Secure AI-assisted development in VS Code. VS Code Docs, checked 5 Nov 2026. code.visualstudio.com
5. Your AI wants to nuke your database. Guardrails fix that. Railway, 29 Apr 2026. blog.railway.com
6. Claude Code: Configure the sandboxed Bash tool. Anthropic, checked 5 Nov 2026. code.claude.com
7. Claude Code: Choose a permission mode. Anthropic, checked 5 Nov 2026. code.claude.com
8. Enterprise managed settings (Copilot). GitHub Docs, checked 5 Nov 2026. docs.github.com
9. Post-mortem: Unauthorized Cline CLI npm publish on February 17, 2026. Cline, 24 Feb 2026. cline.bot
10. Risks and mitigations for GitHub Copilot cloud agent. GitHub Docs, checked 5 Nov 2026. docs.github.com
11. Customizing or disabling the firewall for GitHub Copilot. GitHub Docs, checked 5 Nov 2026. docs.github.com
12. Local sandboxing for GitHub Copilot now generally available. GitHub Changelog, 7 Oct 2026. github.blog
13. Building a secure cloud environment for agents in your enterprise. GitHub Docs, checked 5 Nov 2026. docs.github.com
14. Reviewing audit logs for GitHub Copilot. GitHub Docs, checked 5 Nov 2026. docs.github.com
15. Claude Code: Monitoring. Anthropic, checked 5 Nov 2026. code.claude.com
16. Works Constitution Act (BetrVG), section 87(1). Gesetze im Internet, checked 5 Nov 2026. gesetze-im-internet.de
17. Claude Code: Configure permissions. Anthropic, checked 5 Nov 2026. code.claude.com
18. Gemini CLI: Remote Code Execution via workspace trust and tool allowlisting bypasses. GitHub Advisory Database, 24 Apr 2026. github.com
19. Content exclusion for GitHub Copilot. GitHub Docs, checked 5 Nov 2026. docs.github.com
20. Agent Security. Cursor Docs, checked 5 Nov 2026. cursor.com
21. BaxBench leaderboard (no security reminder). ETH Zurich, checked 5 Nov 2026. baxbench.com
22. We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs. USENIX Security, Aug 2025. usenix.org
23. Information regarding malicious "postmark-mcp" package. Postmark, 25 Sep 2025. postmarkapp.com
24. Data residency (US + EU) and FedRAMP-authorized models now available in Copilot. GitHub Changelog, 13 Apr 2026. github.blog
25. GitHub Copilot with data residency. GitHub Docs, checked 5 Nov 2026. docs.github.com
26. Copilot code review: Merge with confidence with Max and approvals. GitHub Changelog, 28 Oct 2026. github.blog
27. Agent approvals and security (Codex). OpenAI, checked 5 Nov 2026. learn.chatgpt.com
28. Supported surfaces for GitHub Copilot policies. GitHub Docs, checked 5 Nov 2026. docs.github.com