Keeping your AI safe — a guide you can keep

Capture, Build, Automate · The Studio · Spring 2026 Cohort · Week 4

- This is the read-and-keep version of the Week 4 lesson, set up to scan.
- Each move tells you what to do and how to do it; read it once now, come back when you need it.

The one idea

- You don't keep your AI safe by locking it down; you keep it safe by controlling what it can touch — files, accounts, tools, memory — and watching what you type in.
- People get this backwards: the tighter you are about what it can touch, the more real work you can hand it.
- The floor is already solid: your work is encrypted in transit and on the server, and Anthropic's own people can't read it by default.
- You can delete anything, and once training is off it's gone from their servers within 30 days.
- Anthropic also gets checked by outside auditors, so the foundation is in good shape and the work is your end.

Set it up right

Do these once.

Give your AI its own space

- Give your AI one folder that is its whole world: everything you want it working with lives inside, everything else stays out.
- You already did this — the brain folder from week one is this rule.
- That folder isn't just where your stuff lives, it's a fence, so anything outside it is not your AI's business.
- When you're deciding whether your AI should see something, the only question is: is it in the folder or isn't it.

Give your AI its own identity

- Give your AI its own email account, separate from your main one, and connect every tool to that identity.
- Create a new free email account used for nothing but your AI, then connect your drives, calendars, and tools to it.
- That way you hand it one drawer instead of the whole house.
- Connect a tool through your main account and you've handed your AI every doc and shared file you've ever touched, going back years.
- This is the highest payoff for the least effort in the whole guide — one separate email address is the move.

Turn training off

- Set your account so your work isn't used to train the model — working in a folder is not a privacy free pass.
- Most of you are on Pro or Max, where Claude Code runs under the same consumer rules as the chat app, so the same "help improve Claude" setting controls training on your folder work.
- Go into your Anthropic account's privacy and data settings and turn "help improve Claude" off.
- Don't assume it's already off — a change back in 2025 made everybody actually choose, so go look with your own eyes.
- On means your files and prompts can be used to train the model and they keep that data five years; off means they don't train on it and they keep it 30 days.
- Turning it off is also what makes the delete promise real, so the 30-day deletion only holds once training is off.
- If you're on the API instead of Pro or Max the rule is different — they don't train on your prompts or code by default — but most of you are Pro or Max, so go check the switch.
- Good to know: Claude Code also keeps a plain-text copy of your sessions on your own computer in `~/.claude/projects/` for about 30 days.

Keep the approval step on

- Keep the setting on that makes your AI show you what it's about to do before it does it.
- This is built into Claude Code; leave the approval step on — the one where it shows the plan and waits for you to say go.
- Don't turn it off to save a few clicks; it's the setting that saves you, and you'll lean on it every day.

Build the habits

Do these every day. They're light.

Back up before anything big

- Before you hand your AI something real to do, have a copy you can fall back to.
- Copy the folder, or use the restore-point habit we already teach — same idea, a clean version to roll back to.
- It's two minutes of insurance against hours of redoing work, so if it makes a mess you just roll back.

Make it show the plan before anything risky

- Anything risky — sending, deleting, posting, paying — your AI shows you the plan first, you read it, then you say go.
- This is the approval step you turned on, now applied to everything risky, not just email.
- It's the same approve step as last week's workflow: draft it in my voice, I approve, nothing goes out until your eyes are on it.
- This is the habit that keeps you out of trouble more than anything else in the guide.

Never type a password

- Don't ever type a password, a key, or a login into your AI.
- Sessions get logged — your folder keeps a copy and the server might too — so a password you type is a password sitting in a log.
- Keep passwords in a password manager, and if a tool needs a login, connect it properly through the settings.

Treat anything it reads from outside as suspect

- When your AI reads something from outside — an email, a web page, a PDF someone sent — don't let it act on what it read without you looking first.
- Outside content can have hidden instructions meant to trick your AI into doing something you didn't ask for, and it happens.
- The plain version to remember: if you didn't write it, your AI shouldn't blindly do what it says.

Keep it clean

Do these now and then.

Be picky about connectors

- Every tool you connect is a new door into your stuff, so be choosy about what you add and drop what you don't use.
- Before you add a connector, check who built it — if you can't tell what it is or who's behind it, don't connect it.
- Every few months, look at your connected tools and drop the ones you're not using, because an unused connector is just an open door.
- This isn't a reason to connect nothing — connecting tools is how this gets powerful — it's a reason to be choosy.

Review your own files

- Once in a while, open your own files and read what your AI has stored about you and your people, then delete what shouldn't be there.
- What your AI "remembers" isn't off in some cloud — it's your files: your MEMORY file, your brain folder, the notes it keeps.
- Most of it will be fine, but now and then there's a private detail or an old thing that's no longer true, so delete it.
- It's a five-minute read, and that's the nice part of working in a folder: no mystery, just open the file.

Know the kill switch

Know how to stop it

- Know how to stop your AI if it runs off on something it shouldn't: find your usage page, learn what a runaway looks like, and know how to pull the plug.
- Find your usage page today, before you need it — it's where you watch your AI burning through your limit.
- A runaway looks like your AI doing the same thing over and over and eating through your usage fast.
- A real one: a guy's AI ran in a loop burning one percent of his whole weekly limit every minute before he caught it.
- You don't need to be scared of it, you just need to know where the off switch is — stop the session, and if you need to, shut it down from the usage page.

What to share with your AI

- Everything above is about the surface; this is the other half — what you type in yourself — and most people never think about it.
- You can lock the surface down perfectly and then type your client's full name, address, and what they pay you, and the fence doesn't help, because you walked it in.
- Use a traffic light for what you type in: green, yellow, red.

Green — share freely

- Share the high-level stuff freely; this is most of what you'll ever put in.
- Green is your mission, vision, and what you stand for; your industry, market, and the kind of content you put out.
- Green is your public messaging, a plain description of your business model, and your audience in general terms.
- Green is anything you've already published, your general goals, and the challenges you're chewing on.
- Green is the personal side too — how you're feeling, what you're working through, thinking out loud.
- Rule of thumb: if it's high-level and doesn't name a specific person, company, or confidential number, it's green.

Yellow — generalize first

- When there's a real specific in there, you don't leave it out — you smooth it down before it goes in.
- Yellow is your revenue in a range, your priorities at a high level, things you've noticed about a competitor, a client persona with the name off.
- Yellow is your pricing philosophy instead of your exact rates, and an overview of how you run something inside instead of the whole playbook.
- Instead of "my client Sarah pays me $4,200 a month," type "a mid-size client on a monthly retainer."
- Instead of "our revenue last quarter was exactly $86,900," say "our quarterly revenue is in the mid five figures."
- Rule of thumb: if a detail is specific enough to point at a real person or company, generalize it before you type it.

Red — keep private

- Some things stay out even smoothed down.
- Red is your exact financials, bank details, and passwords.
- Red is client names, contracts, and anything a client told you in confidence.
- Red is a proprietary formula or trade secret, and a legal matter still in progress.
- Red is other people's personal information — that's not even yours to share.
- Rule of thumb: if sharing it would break someone's trust, cross a privacy line, or hand over the keys, it's red, no matter how you word it.

If your work is regulated or legal

For the few of you who carry client data or work in something regulated — health, legal, finances.

- Don't use the consumer version of Claude for regulated data or legally-sensitive work.
- The Claude you and I use — Free, Pro, Max — is the consumer version, built for general use, not for regulated or legally-sensitive work.
- Anthropic has an Enterprise plan for that, with protections the consumer plans don't have, including a HIPAA agreement (a BAA) for health data.
- If your work lives in that world, that's the door you go through, and you talk to whoever handles compliance on your end first.
- Here's the real reason: in February of this year, in U.S. versus Heppner, a man under investigation used consumer Claude to work through his legal defense, figuring those chats were private like talking to a lawyer.
- The court said no — it ruled those conversations were not protected, on the simple line that Claude is not an attorney.
- Talking to a public AI is treated like talking to any other outsider; it's not privileged.
- The takeaway: don't use consumer AI for legal strategy, and don't assume what you type in is private the way it would be with a lawyer or a doctor — it isn't.
- For most of you this is just good to know; for the few it's really about, it's the most important thing in here.

Short recap

- You don't keep your AI safe by giving it less.
- You control what it can touch, you watch what you type into it, and you can trust it with more, not less.

The certifications behind "outside-audited": Anthropic holds SOC 2 Type II (an independent audit confirming their systems reliably protect customer data over time — security, availability, confidentiality) and ISO 27001 (an international standard for a formal, systematic approach to managing information-security risks, from policies through technical controls).