

The Agent Operating Agreement

Write the rules down before you hand over the keys. This is for any AI that does work for you: a personal assistant, a team agent, a tool your people already use. Most of these rules get learned the hard way, by something going wrong once. Filling this in first is the cheaper way to learn them.

The job

What it's for, in one line. If you need two lines, you're describing two agents.

How much it can do, job by job

One row per task, and note what's at stake: who sees it, whether money moves, whether it can be undone. The higher the stakes, the lower the rung. Start lower than feels right; page two tells you when it's earned a move up.

THE TASK, AND WHAT'S AT STAKE	NOT HANDED OFF I DO IT MYSELF	RUNG 1 READS ONLY	RUNG 2 DRAFTS, I APPROVE	RUNG 3 ACTS ON ITS OWN
	☐	☐	☐	☐
	☐	☐	☐	☐
	☐	☐	☐	☐
	☐	☐	☐	☐
	☐	☐	☐	☐

It never does these

Not without me saying yes, every single time. Tick what applies and add your own. A draft is always fine. Hitting send isn't.

- ☐ Send anything
 ☐ Buy anything
 ☐ Publish anything
 ☐ Contact anyone
 ☐ Delete anything
 ☐ _____

How it reports

Every run gets a report, even when there's nothing to say. A silent run looks the same as one that never finished. Where, and when?

What counts as proof

"Done" isn't proof. What does it have to show you before it gets to say it finished? The changed rows, a link, a before and after?

Who has the final call, and what happens when I say no

Name the person. When they pass on something, it stays passed. The agent writes it down; it doesn't argue it again.

The Correction Log

The agreement changes every time it gets something wrong. **Log it**, so a correction turns into a rule instead of a repeat. It works both ways: when the agent catches you and it's right, log that too.

DATE	WHAT IT GOT WRONG	WHAT I TOLD IT	STANDING RULE NOW?	HELD THE NEXT TIME?
			Y / N	Y / N
			Y / N	Y / N
			Y / N	Y / N
			Y / N	Y / N
			Y / N	Y / N
			Y / N	Y / N
			Y / N	Y / N
			Y / N	Y / N

BEFORE YOU MOVE A JOB UP A RUNG

Five questions. All yes, or it stays where it is.

Ask these one job at a time, not for the agent as a whole. It can be ready for one thing and nowhere near ready for another.

01	Has it reported every run for at least a week, including the empty ones?	Y / N
02	Can you see what it changed without taking its word for it? If the only evidence is "updated and correct," the answer is no.	Y / N
03	Has it taken a correction on this job and not repeated the mistake?	Y / N
04	Would you notice within a day if it got this wrong?	Y / N
05	If it did get it wrong, could you undo it? Does a backup exist before it changes anything?	Y / N

Any no is fine. It just tells you what to fix before you give it more room. Write the fix in the log above, then ask again next week.

It was so sure it was done that I started doubting my own screen.

I STOPPED BUILDING AI TOOLS. NOW I MANAGE ONE.