
THE STANDARD

What I hold to

The rules I actually work by, kept daily.

These are what I do, and what I refuse to do, when something this capable is sitting open beside me. They come from three places: things I learned by getting them wrong, things I borrowed from people who think about this more rigorously than I do, and things that simply follow from what I value. The backbone is the [Asilomar AI Principles](#) — twenty-three principles coordinated by the Future of Life Institute at the Beneficial AI conference in 2017, written by researchers for laboratories, and translated here for one person working with other people.

Yours would look different, and should. What matters is that they exist *before* you need them. In the moment you need one, you will be busy, and the easy thing will be very close to hand.

A principle you can't picture yourself keeping is a sentence, not a standard. So each of these is written to be kept.

ON JUDGMENT

- 1 Bring the question, not just the task.** Say what you're actually trying to work out before you open the window. If you can't put it in one sentence, it will help you produce something impressive instead of finding out.
- 2 It drafts. You decide.** Use it for analysis, options, first versions. What you believe, what you'll stand behind, what you won't do — those never leave your hands, even when its answer is better than yours.
- 3 Treat easy agreement as a signal to slow down.** It is built to be agreeable, so its approval tells you nothing. Ask for the strongest argument against you, written by someone who wants you to be wrong, and read that one twice.

4 **Keep the pause.** Form your own rough answer before you ask for its polished one — even sixty seconds will do. Otherwise its framing becomes your starting point, and you never find out what you would have thought.

5 **Say it in your own words first.** Raw and unedited, straight from your head, then let it sharpen. Once you've read its version, you can't get yours back.

ON THE WORK

1 **Open every source before you cite it.** Facts, figures, names, quotes. It will produce a citation that is complete, plausible and fictional, and that mistake survives a skim all the way into print. If you haven't clicked it, don't quote it.

2 **Only put in what's yours to put in.** Know where the material came from and what the tool keeps. Client and confidential work doesn't go into anything that stores inputs or trains on them by default.

3 **Ask who's missing from the answer.** It speaks in one culture's defaults and presents them as neutral. Read the output once more for who it wasn't written for — then ask someone who would actually know.

4 **Treat imported prompts and agents like a stranger's attachment.** A bought template carries instructions you haven't read. Read all of it before you run it, and buy only from people you'd name out loud.

5 **Re-test what you stopped watching.** Models change without notice. Anything you built, launched and left alone is running on assumptions that may have quietly expired — check it against the case you first tested.

ON THE PEOPLE YOU SERVE

1 **Disclose it, and say what it isn't.** Tell people plainly when they're dealing with AI, state its limits in the same breath, and leave a way to reach a human.

- 2 **Build so a compromised assistant can't hurt anyone.** Anything customer-facing can be talked into misbehaving, and that isn't reliably fixable — so give it nothing confidential to reveal and no consequential action to take. A person approves whatever matters.

- 3 **Know how to switch it off, and write the failure plan before launch.** Decide now what you'll do the first time it is wrong in public. You won't think clearly once it happens.

- 4 **Say what it isn't before distress arrives.** Anything reflective eventually meets something real. One calm line written in advance — what this is not, and where else to go — is worth more than anything improvised in the moment.

- 5 **If it displaces someone, the person comes first.** Not the efficiency case. Honesty, notice, time, and proper advice — the law differs sharply by country, and this is the decision, not a footnote to it.

ON POWER

- 1 **Prefer narrow, fenced, and understood.** A small tool you built deliberately beats a capable one you don't understand. Look into the company before its agent goes near your work or your clients. Nothing here is safe; some things are well-governed.

- 2 **Reach for it on purpose.** Because this question needs it — not because it's already open. Habit is how the other seventeen quietly stop applying.

- 3 **Keep this revisable.** Date it, version it, change it in public when it's wrong. A standard that never changes has stopped paying attention.