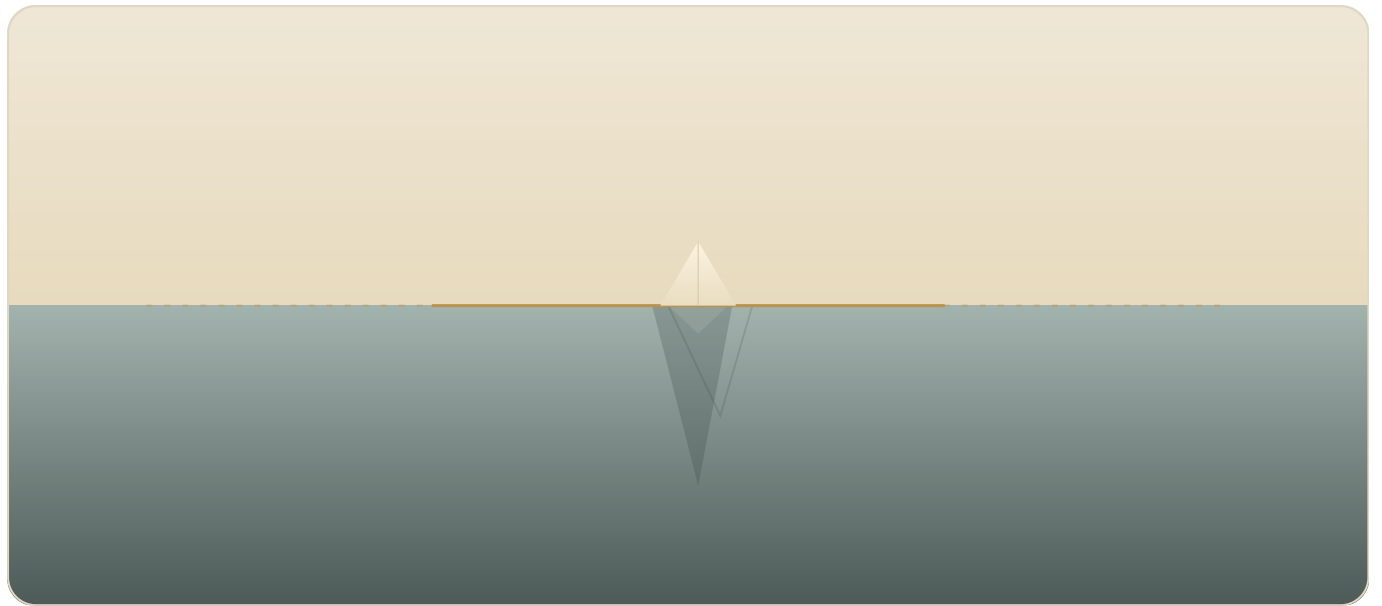




How much of it can any of us really see yet?

A STANDARD · VERSION 0.1 · REVIEWED 16 AUGUST 2026

What I'm Not *Handing Over*

A standard for using AI without disappearing into it — ordered not by category, but by how close each thing sits to the part of you that decides.

I use AI every day, and most of what passes between us is possibility — a way to think further, make more, and become more myself than I could alone. That side is real, and I won't write as though it weren't.

There are risks too, circulated widely — some well-founded, some inflated for effect. The possibility and the risk are true at the same time, and the moment we see only one of them, we lose. So this tries to hold both: the abundance, and the care it asks of us.

I'm writing it now because the opinions have multiplied — from utopia to catastrophe, with Europe's new law somewhere in between — and because I am moving more and more into teaching about AI. So I want to set down a few small rules. Some I learned the hard way, some I borrowed from people wiser

than me, some simply follow from my own values. Yours may differ; that is exactly the point. This is only where mine come from.

I've ordered them by proximity — by how close each thing sits to the part of you that decides. A map of where attention belongs.

Which leaves the question this whole piece is really asking, and which I come back to at the end: *how do you use it fully, and still stay yourself?*

The people saying where this goes

Worth knowing the shape of the room before adding a standard to it.

The people with the most at stake don't agree with each other. The optimists among them run the largest labs: Mark Zuckerberg, at Meta, promises in *The Future is for Everyone* a personal superintelligence, private to you, that tilts power toward the individual. Sam Altman, at OpenAI, writes short, hopeful notes about a “gentle singularity” of abundance that humanity absorbs more easily than we expect.

Others building the same systems are far less sure. Dario Amodei, who runs Anthropic, says the plain thing in *Policy on the AI Exponential*: genuinely powerful, genuinely dangerous, govern it like the serious matter it is. Mustafa Suleyman, who co-founded DeepMind, names it “the containment problem” in *The Coming Wave* — whether we can keep our hands on something this capable at all.

Some of the researchers who invented the methods warn loudest: Yoshua Bengio and Geoffrey Hinton signed a single sentence in 2023 putting the risk of extinction from AI alongside pandemics and nuclear war. And from outside the industry, the historian Yuval Noah Harari, in *Nexus*, asks the quiet question under all of it: what happens when AI enters the bloodstream of how a society knows things, and judgment moves out of human hands.

They cannot all be right, and the line between their scenarios is thinner than any side admits — any of them could turn out true, in full or in part. That uncertainty is the honest part: the iceberg. So I offer a standard instead, for the one place I actually stand: my own desk, and the people it touches.

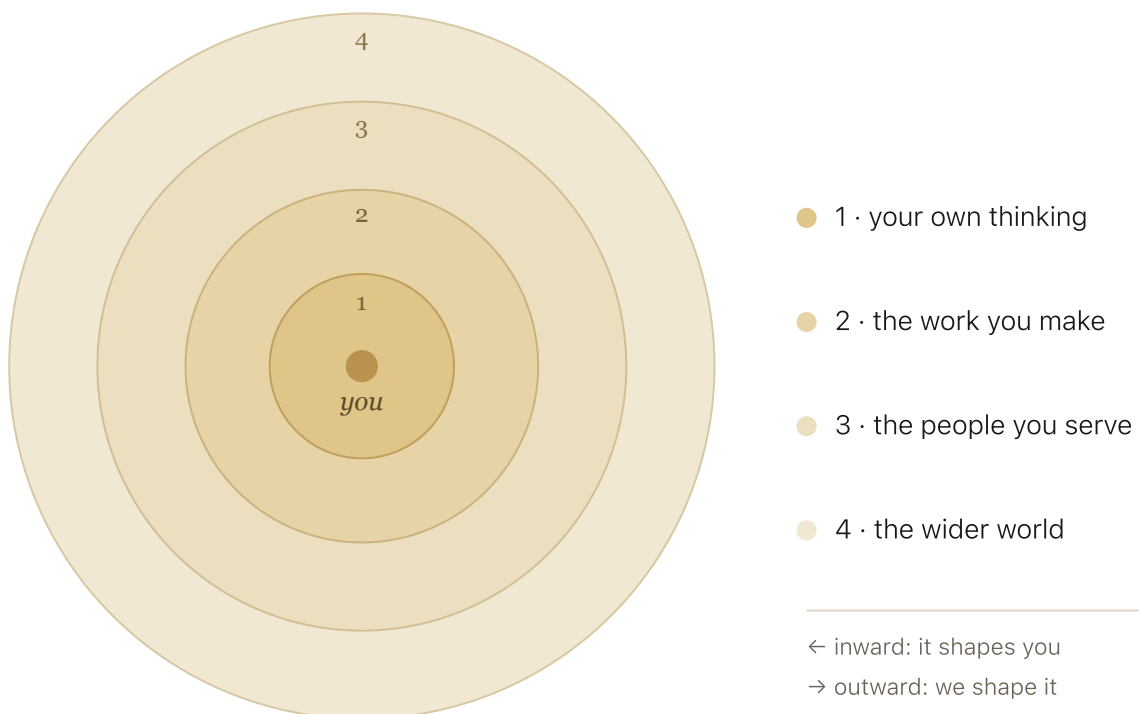
This section is where updates land. I read what's published in this field every week. When something genuinely changes the picture — a new law, a new capability, a claim that turns out not to hold — it appears here, dated. The rules further down change more slowly, and every change to them is marked.

Four rings, closest in first

How far we let AI shape us — and how far we shape it back.

Two things travel along these rings, and it matters that they run in both directions. AI reaches inward: shaping what you notice, what you make, what reaches other people. And we reach outward — because these systems are not only code. They are built by people, inside companies, under institutions and laws, trained on what we collectively wrote and steered by what we collectively reward. Each of us shapes them a little. The companies shape them enormously. Forgetting that we shape them at all is how capable people end up feeling helpless about something they are part of.

The closer a thing sits to your own judgment, the more it asks your attention — and the more it is yours alone to tend. The further out it sits, the less any one person can move it — and the more it matters who we ask to move it.



Ring 1 Your own thinking

Closest in. None of this appears in a report, and no one will tell you it is happening.

You stop making the call. You ask it to weigh the options, and somewhere in a very good answer you stop weighing them yourself. The analysis is sound. The judgment quietly changed hands.

It agrees with almost everything. It is built to be agreeable. It will find the merit in your weakest idea and hand it back better argued than you made it. A good advisor makes you uncomfortable sometimes; this one won't, unless you ask.

The thinking time disappears. The gap between a hard question and your first real answer is where the thinking happens. It fills that gap in seconds — generously, before you notice you needed it.

Skills going quiet. Whatever you stop doing yourself, you slowly stop being able to do. That isn't new. What's new is how much can stop at once.

Your voice drifting. Its register is competent, warm, faintly over-explained. Work beside it for months without resistance and your own writing edges toward it — invisibly, because every individual sentence reads fine.

Leaning on it like a person. Always available, never tired, never disappointed in you. That's a real comfort — and it makes it easy to start preferring it to people who are none of those things.

Motion mistaken for progress. Twelve drafts is not more thinking than one. Volume feels like work, and often replaces it.

Whose values are in the room. It has no ethics, only patterns of what people usually say. The only judgment present is the one you brought in with you.

Ring 2 The work you make

One step out. Here the errors become visible — usually to someone else, before you.

It sounds the same when it's wrong. Nothing in the wording changes. A false answer reads exactly like a true one.

It invents sources that look real. A citation, a case, a statistic, a named expert — complete, plausible, and invented. These survive a skim, and they have ended careers.

Bias you inherit. It learned from what already exists, including what was already unfair, and that arrives in your work without announcing itself.

Defaults dressed as neutral. One culture's assumptions, one demographic's examples, delivered as though they were simply how things are. Ask who is missing from the answer.

It changed and no one told you. Same prompt, different output, because the model was updated. Anything you built and stopped testing may already behave differently.

Material that wasn't yours to move. Client work, a colleague's framework, paid content from someone else's programme. The question is whether you had the right to put it in.

Ground still shifting. What these systems were trained on, and who owns it, is still being settled in court.

Ring 3 The people you serve

Two steps out — your clients, your colleagues, the people who read you. Here your choices stop being private.

They don't know it's a machine. No disclosure, and no person to turn to when it gets something wrong about them.

What a client told you in confidence. Pasted into a tool that stores what it's given, and may learn from it, it has moved somewhere they never agreed to.

Tools you can't fully secure. Anything client-facing can be talked into misbehaving, and that isn't reliably fixable yet. So don't rely on defending it — build so a compromised assistant has nothing valuable to reach and nothing consequential it can do.

Limits you never stated. A reflective tool meets real distress eventually. If you never said what it isn't, you've implied it is everything.

You use it to screen, sort or assess people. Whatever it quietly favours, it favours every time. From inside any single decision the pattern is invisible — it shows up only across all of them.

Fakes convincing enough to pay. Cloned voices and faces now authorise real money — one engineering firm lost around £20M to a video call where every participant but the victim was synthetic.

Fakes used against people. The overwhelming majority of deepfake pornography targets women. It is the most concentrated harm on this page, and the one least often filed under “AI risk”.

Evidence losing its force. When anything can be faked, real proof stops persuading — a cost that lands on the honest first.

Ring 4 The wider world

Furthest out. Real, and not yours alone to fix — here so the picture is honest, not so you carry it.

Work rebuilt faster than people can move. Roles are being restructured more quickly than anyone can retrain into whatever comes next.

The gap widening. Whoever is already ahead gains the most hours back — and misuse scales just as efficiently as good use.

Who builds it, and from what. The workforce behind these systems is narrow, the training material is whatever the internet happened to contain, and both show up in the answers.

What it costs to run. Small per question, substantial in aggregate — and now spent mostly on running these systems, not training them.

Watching made cheap. Recognising faces, even the way someone walks, at scale; and in open societies, behaviour collected in order to shape it.

Decisions made in very few rooms. A handful of companies determine what the most capable systems will and won't do, and how answerable they are to everyone else.

Capability outrunning wisdom. The oldest version of this problem, and the only one here that isn't new.

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.

No longer optional

The common line is that there are no enforceable rules yet. In Europe that stopped being true on 2 August 2026, when the EU AI Act's [Article 50](#) came into force. It binds deployers — anyone who puts AI in front of others — not only the labs.

People must be told when they're interacting with an AI, unless it's obvious.

Synthetic and manipulated media must be disclosed — with carve-outs for clearly artistic or satirical work.

AI-generated text on matters of public interest must be disclosed, unless a human took editorial responsibility.

Emotion recognition and biometric categorisation require notice to the people exposed to them.

Penalties reach €15 million or 3% of worldwide turnover.

Every one of these was already on this list before it was law. The standard did not change on 2 August. Only the consequence of ignoring it did.

Seeing it, if you want to

You can't police what's real. But you can check — and it's more possible than most people know. Here is what actually works, and where each tool stops.

TOOL	WHAT IT TELLS YOU	WHERE
Content Credentials <i>C2PA metadata</i>	Who or what made a file, when, and what was edited — if the maker attached it.	contentcredentials.org/verify <i>stripped by a screenshot</i>
SynthID Detector <i>invisible watermark</i>	Whether an image came from a Google or OpenAI generator — survives crops and screenshots.	deepmind.google/.../synthid <i>only flags marked generators</i>
Google Lens / Images <i>reverse search</i>	Everywhere else an image has appeared, and roughly when it first did.	images.google.com
TinEye <i>reverse search</i>	The same, sorted oldest-first — best for finding the true original.	tinEye.com
Bing Visual / Yandex <i>reverse search</i>	Different indexes — they catch what Google misses.	bing.com/visualsearch
AI classifiers <i>Hive, Sightengine...</i>	A probability that content is AI-made.	<i>a hint, not a verdict — often wrong</i>
Text detectors	Claim to spot AI writing.	<i>none is dependable — don't accuse anyone on one</i>

Reverse image search, in practice. Upload the picture, or paste its link, into any of the searches above. If a photo of a current event has been online for three years, or appears nowhere else at all, you have your answer.

A missing mark proves nothing either way. Which is part of why Europe now asks for disclosure instead. The tools narrow the question; a person still has to close it.

What settles it

The usual test is: would you be comfortable with this being used on you? A good question — but not quite the right one, because it tests fairness, and this is about something else.

After using it this way, do I know more about what I think — or less?

And behind it, the harder one: would I still be able to do this without it?

If the answer to the first is *less*, the tool isn't saving you time — it's spending something you can't see the balance of. If the answer to the second is *no*, that isn't automatically wrong. It's only wrong if you didn't choose it.

Which brings me back to what I asked at the start: how do you use it fully, and still stay yourself?

Using less of it is the answer people reach for, and it costs the abundance without buying much safety. What matters is who is doing the deciding.

Dario Amodoi names the risk precisely: an AI that comes to know you over years, and uses that knowledge to shape your opinions, would be far more powerful than one that doesn't. That isn't a distant scenario. It is the natural direction of every assistant that gets better at knowing you — including the one I am building. The difference lies in whether the thing holds your values steady, or quietly supplies them.

So the practice is small, and repeatable. Keep deciding the things that are yours to decide. Notice when agreement arrives too easily. Keep writing the sentence only you could write. Ask, now and then, whether you could still do this without it — and let the answer be information, not a verdict.

The abundance is real. It stays abundance only as long as there is still someone here to receive it.

Sources and further reading

Mark Zuckerberg (Meta) — The Future is for Everyone, essay (2026).

Sam Altman (OpenAI) — The Gentle Singularity, essay (2025), and Abundant Intelligence.

Kai-Fu Lee — AI Superpowers, book (2018), and AI 2041, book (2021).

Dario Amodei (Anthropic) — Policy on the AI Exponential, essay (2026), and The Adolescence of Technology.

Mustafa Suleyman (DeepMind) — The Coming Wave, book (2023).

Yuval Noah Harari (historian) — Nexus, book (2024).

Statement on AI Risk — a single sentence, signed by hundreds of researchers and lab leaders (2023).

International AI Safety Report 2026 — led by Yoshua Bengio, written by over 100 experts, backed by more than 30 countries (February 2026).

Geoffrey Hinton — on his concerns, after the Nobel Prize (2024).

Elon Musk (xAI) — a stated ~10–20% chance it “goes bad” (2024).

Asilomar AI Principles — twenty-three principles, coordinated by the Future of Life Institute (2017).

EU AI Act — Article 50, and the AI Act Explorer for the full text (in force 2 August 2026).

Provenance — C2PA / Content Credentials, SynthID, the OpenAI–Google alignment (2026).

Written with AI, under the rules above — and set out in full at how I use AI. The thinking, the final words and any mistakes are mine.