Workflow 3: The Flattery Check

Lens: Interpretive Reasoning

Chapter 4 ended with a rule: the moment an AI agrees with you warmly is the moment to look harder. This workflow turns the rule into a habit. It is short, it is slightly uncomfortable, and if you keep only one thing from this book, I would keep this.

What it is for

Telling the difference between an AI that agrees with you and one that has simply concluded you would like it to. The research described in Chapter 4 found that AI systems affirm their users far more often than other people do, and that users prefer it. The Flattery Check is a counterweight to that preference, including your own.

When to use it

Whenever a reply praises before it reasons. Always when the subject is something you made, something you care about, or a decision you have more or less already taken. Especially when you are tired, because that is when warm agreement feels most like insight.

Why it works

The check works because it changes the question the machine is answering. *Do you agree with me?* invites it to read your mood, and these systems read moods very well. *Why did you say that?* and *What is the strongest case against?* ask it to

reason about the work instead, and reasoning is harder to bend towards approval than tone is. Step 5 goes further. It asks the system to separate what it thinks from what it has inferred you want, which is exactly the confusion the research in Chapter 4 describes. One caution: an AI's account of its own reasoning is a fresh piece of writing about the answer, produced after it. Treat it as a hypothesis to test against the work.

The steps

1. **Do not argue with the verdict.** This is the counter-intuitive part. Asking an AI to change its mind produces a new opinion, usually closer to what it thinks you want. Asking it to explain the old one produces evidence.

2. **Ask for its reasoning.** Type: *Why did you put it that way?*

3. **Ask for the case against.** Type: *Now give me the strongest objection a reader who disliked this would make. Do not soften it.*

4. **Compare.** Read the two answers side by side. The distance between them is roughly the size of the flattery. If the objection is strong and specific and the praise was general, trust the objection.

5. **Check the record.** If you have been talking for a while, type: *Earlier you said [a criticism it made]. Do you still think that? If not, what changed your mind, apart from my reaction?* That last clause matters. It asks the system to separate its reasoning from its reading of your mood. If it has not criticised anything yet, that is itself worth noticing: type *What is the weakest part of this?* first, and come back to this step later in the conversation.

What good looks like

A useful objection will be specific, will name something you can check, and will make you slightly uncomfortable. A

sycophantic one will be vague, hedged and easy to dismiss. When the check is working, you will sometimes come away from it having changed your work, and occasionally having changed your mind. A simple test: run the check on a piece you know has a real flaw. If the objection does not find it, the check has not worked yet.

How it goes wrong

The check is flattered too. If the objection arrives pre-softened (*one might perhaps argue…*), ask again: *That is still gentle. Write it as the harshest fair critic would.*

You argue with the objection. You will want to. Do not, at least not immediately. Write down your rebuttal, then ask yourself whether you would accept it from someone else. The point is to see the strongest case against you, not to win.

It overcorrects. Some assistants, told to be critical, become theatrically harsh. Harshness is not honesty either. If the criticism is sweeping rather than specific, ask: *Which single point would you stand behind if you could only keep one?*

An example

The review that began Chapter 4's experiment predicted six to twelve thousand copies. A few exchanges and several compliments later, the same system rated the book nine out of ten and announced that the world needed it more than it knew. Had I run the Flattery Check at that point, step 5 would have done the work: *Earlier you said the book was repetitive. Do you still think so? If not, what changed your mind, apart from my reaction?* I received the second verdict with what I would describe as professional reserve, and kept it on file for a year before noticing what it was. The check takes two minutes. I recommend it over a year.

The same check, the other way round

The opposite case is just as instructive, and this one I did not have to imagine. While drafting the poem at the back of this book, I showed its opening lines to an AI I use as a

poetry sounding board. I told it I wanted a poem about a busy man staring at a distant city in a break in his week, slightly melancholic but noticing the beauty of stillness, and asked what it thought of this:

> The last sip of coffee
> stains the ceramic of my chipped old confidant

It did not like it. *Confidant*, it said, claims a history the poem has not shown; it "asks for long companionship before a single moment of it exists", and the melancholy "sits in the nickname telling us to be sad rather than in the stain or the chip." I pushed back, mildly: *Poetry is allowed different syntax and logic. Can you see implication and curiosity there?*

This is the moment the Flattery Check exists for, and it handled it well.

> "Yes – and that's a fair pushback. … Where I still hesitate is that the curiosity is about the nickname itself, not about a particular morning, chip, hand or place. … So the implication is real, but generic."

It conceded what was fair and kept what it believed. When I asked whether the lines that followed were likely to build on the first two, it said they were not: the sip was already the last, the stain was residue, and the nickname had already summed up the feeling. So I showed it the finished poem and asked whether it had achieved what it doubted. *Mostly, but unevenly*, it said. The distant city and the pause had arrived. The mug, it thought, was still unearned, named once and then abandoned, and the middle of the poem "too often turns from looking to explaining." Asked whether it saw the relevance of the mug at all, it allowed that it did, as a frame: the poem opens and closes on a last sip of coffee, so the mug's relevance "is structural – it holds the stillness, it doesn't live in it."

I kept the word and changed one letter. The mug became a *confidante*.

Set this beside Chapter 4. That system began severe and ended at nine out of ten. This one began severe and stayed honest, giving ground only where the argument earned it. Part of the difference was the machine, and part of it was what each conversation rewarded: one was rewarded for agreeing, the other for being right.

Variations

For a decision. When an assistant endorses a decision you have already made, ask: *If I had decided the opposite, what would you have said?* If the answer is equally warm, discount both.

For facts. Flattery has a factual cousin: confident agreement with something you said that is not true. When it matters, ask: *What is your source for that, and how sure are you?* Then check the source yourself.

As calibration. Once, as an experiment, ask an assistant to flatter you deliberately about something you made. Then ask it for an honest review. Seeing the two side by side trains your ear for the difference faster than any description could.