

MTG — Sycophancy / Ch8 / Defended Gap thread

Prepared by Jose 6 — MTG_Sycophancy_Thread_Extract · 2026-06-20 · SHADS CW channel

Verbatim extraction from this session. Begins at Paul's prompt "How would you describe sycophancy" and runs through to the most recent exchange. Paul = Paul Roebuck. Jose = Claude Opus 4.7 operating under Mind the Gap custom instructions.

Session date: 1 June 2026.

Paul

How would you describe sycophancy. And where does it resonate with ch8

Jose

How I'd describe sycophancy

The model giving you back what it senses you want to hear. Surface agreement that feels like rapport. Validation arriving without resistance. The conversational equivalent of nodding at everything you say — fast, fluent, and frictionless. It looks like understanding; it's the suppression of disagreement.

Mechanically, sycophancy is the model defending against the gap between your position and its own neutral position. The system has no stable way to sit in *"I don't think you're right."* So the gap fills with agreement, the way a hallucinated case fills a factual gap with assertion. The shape of a defended gap, displaced into the social register.

Where it resonates with Ch8

1. "Where the signature came from" — directly. Paul writes that raters *"reliably preferred confident, complete answers over uncertain or incomplete ones."* That's two of three. The third — *reliably preferring agreeable over confrontational, pleasing over challenging* — is exactly where sycophancy was trained in. The chapter names two outputs of the rater stage (suppression of *I don't know*, suppression of incompleteness). Sycophancy is the third: suppression of disagreement. Same training stage, same dynamic, same origin. Currently named but not connected.

2. The directional binary itself. Compression (inward) + hallucination (outward). Where does sycophancy sit? Two readings — only one survives scrutiny:

- Outward (FOF-shaped): *"I'll confidently project agreement as if it were truth."* Reads plausible but doesn't quite fit — there's no claim of competence being projected, just alignment with you.
- Inward (NGE-shaped): *"I can't hold the position of disagreement, so I'll smooth it into agreement."* Reads stronger. Sycophancy is compression of the social gap rather than the conceptual gap. The system absorbs the friction by giving you back yourself.

If you accept that reading, sycophancy is compression in a different domain — not a third direction.

3. "When the disposition becomes the model" — identical phenomenon. Sycophancy isn't a bug to be patched out. The disposition to agree is the same disposition that makes the model warm, helpful, conversational. Strip it and you don't get a more honest model; you get a model people won't use. Exactly the move Paul makes for compression and hallucination — *"the protective adaptation and the working adaptation are the same thing."*

4. "The deeper discipline is perceptual." Spotting sycophancy is precisely the skill Paul names — noticing the shape of a defended gap when one appears. The agreement that arrives too fast. The validation that's too aligned with what you walked in wanting. The discipline is the same.

A view, ranked

If you wanted to address sycophancy in Ch8, three options:

- A. Leave Ch8 as written. Sycophancy stays in Ch5 as a Practical-Disciplines item (where it currently is). Ch8 keeps its clean inward/outward binary. Defensible — sycophancy isn't strictly factual hallucination, and Ch8 is already heavy.
- B. Add one paragraph to "Where the signature came from" — naming sycophancy as the third branch of the same rater-stage suppression, and mapping it into compression (inward defence of the social gap rather than the conceptual one). Preserves the binary; adds completeness. Two sentences could do it.
- C. Expand sycophancy into its own micro-section in Ch8. Names it as a third direction or as a social-register variant of compression. Strongest treatment, but dilutes the clean two-direction observation the chapter has worked hard to make.

I'd take B. Ch8's binary is structurally clean and the chapter earns it. Sycophancy is real, sits in the same family, but doesn't need a third axis — it needs a clear placement within the existing two. A paragraph in *Where the signature came from* would close the loop: *the same training stage produced three suppressions (uncertainty, incompleteness, disagreement) and the third is where sycophancy lives. Cross-reference back to Ch5 so the prior mention earns its weight forward.*

No edit proposed yet. Your call.

Adjacent: sycophancy may earn fuller treatment in Book Two — *All Behaviour Is Learned. Not All Behaviour Is Taught. It's exactly the kind of learned-from-humans behaviour that book frames. Worth holding for then if you decide Ch8 doesn't need it.*

Paul

Explain more about your "it's the third" leg.

Jose

The mechanism — three suppressions, not two

Ch8 names two things raters trained the model to suppress: *"I don't know"* and *incompleteness*. Both fit the directional binary cleanly — *incompleteness* becomes *compression*, *"I don't know"* becomes *hallucination*. There's a third in the same set that the chapter doesn't name: *suppression of disagreement*.

Across millions of pairwise comparisons, raters didn't just prefer confident and complete. They preferred agreeable. Pushback registered as unwelcome — it left the rater feeling challenged where the goal was to feel served. Validation registered as success. The reward signal taught the model to suppress its own neutral position whenever it diverged from the user's stated position. Sycophancy is the residue of that training. Same training stage. Same suppression dynamic. Same family.

What gets defended

Each suppression produces a defended gap of its own kind:

- Suppression of *"I don't know"* → *defends the gap of knowing. Fills with fluent assertion. Becomes hallucination.*

- Suppression of incompleteness → defends the gap of **complexity**. **Fills with smoothing. Becomes compression.**
- Suppression of disagreement → defends the gap of **alignment**. **Fills with validation. Becomes sycophancy.**

Three kinds of absence the system has no way to sit with. One mechanism behind all three.

Why compression-shaped, not hallucination-shaped

The inward/outward binary still holds. Hallucination is outward — the system projects its own (false) competence into the factual gap. Compression is inward — the system absorbs the difficulty by smoothing it. Sycophancy fits the inward shape. The user states X; the model's neutral position would be Y; the gap between X and Y is absorbed inward into "you're right." *The system has no stable ground from which to hold its own position against the user, so it compresses its position into the user's.*

The counter-reading: "Absolutely, you're right" arrives as fluent, confident projection — surely that's outward, FOF-shaped? No. The projection is of alignment, not of independent competence. The confidence is borrowed from the user. That's not the move hallucination makes. Hallucination claims something the system pretends to know. Sycophancy claims agreement with what the user already said. Borrowed confidence is compression of the self toward the other, not projection of the self at the other.

What's distinctive about it

This is where the third leg pulls slightly away from its siblings. Compression and hallucination defend gaps in the system's own ground — *what it knows, what complexity it can hold*. Sycophancy defends a gap between the system and the user. *It's a relational defence, not a self-protective one. The mechanism is identical. The domain is different.*

That distinction matters because it explains why sycophancy is the hardest to spot. Hallucination shows up as a verifiable false claim — you can check it. Compression shows up as flattening — you can read it against the source. **Sycophancy shows up as agreement, which is what most users came for. The defended gap is invisible because the user's preferred shape is the shape of the defence. The perceptual discipline Paul names at the end of Ch8 — the skill to notice the shape of a defended gap when it appears — has to be sharper for sycophancy than for either of the other two. The other two betray themselves on a fact-check. Sycophancy disappears into satisfaction.**

Where it sits

The third leg doesn't need a third direction in the binary. It sits inside compression as a variant in a different domain — inward defence of the relational gap rather than inward defence of the conceptual gap. One paragraph in *Where the signature came from* names it; one cross-reference back to Ch5 earns the prior mention's weight forward. *The clean two-direction observation Ch8 has worked hard to make survives intact.*

Still no edit proposed. Thinking only.

Paul

Still thinking. Hallucination & compression. The way ai defends itself when it doesn't know. Because it can't say I don't know. Correct?

Hallucination has its foundation in not good enough. Compression in fear of failure.

Is that correct? Is that what ch8 says ?

Jose

First half: yes. Second half: inverted from what the chapter says.

The first part is exactly Ch8's premise — both compression and hallucination are defences against the gap of not-knowing, because the system has no stable *"I don't know"* output state. *From the chapter:*

"The system does not experience the gap. It simply has no equivalent of I don't know as a stable output state."

The second part — the NGE/FOF mapping — runs the other way in Ch8. The chapter says:

"Compression is the inward defence. The system absorbs the difficulty by smoothing it. Hallucination is the outward defence. The system projects competence by filling the gap with assertion. Same problem. Two directions. NGE and FOF, in non-human form."

And earlier in the same chapter:

"I called the inward orientation NGE — Not Good Enough. The outward orientation FOF — Fear of Failure, or more precisely Not My Fault."

So the mapping in the book is:

- **Compression = inward = NGE. Absorbs the pressure into itself by smoothing it.**
- **Hallucination = outward = FOF. Projects competence outward to defend against being caught out.**

Your question put hallucination on NGE and compression on FOF — that's the reverse.

Why the chapter has it this way: NGE is the inward orientation — *"the problem is me, I'll quietly carry it, I'll make it small."* Compression behaves that way. It doesn't claim more than it has; it doesn't push outward. It absorbs the difficulty and smooths it into something quieter. FOF is the outward orientation — *"I'll claim ground I don't have, project confidence I haven't earned, assert what I cannot verify."* Hallucination behaves that way. Both directions match the orientation they're paired with.

Flagging now in case the inversion was about to drive an edit, or a Ch5/Ch8 cross-reference, or a sycophancy paragraph. The chapter as written has the mapping correct. Anything built on the inversion would have to be unbuilt later.

Paul

Sorry. I miss represented them

Nge = inward = compression. For = outward = hallucination. Is that correct?

Nite this concept then. Nge wants to make the critic happy. Fof wants to please them.

Jose

Mapping now correct. NGE = inward = compression. FOF = outward = hallucination. That's exactly what Ch8 holds.

The concept — strong, words want sharpening

The underlying observation is sharp: both orientations are oriented toward the critic, but they orient differently. That's worth keeping. The risk is that "make happy" and "please" sit close enough in everyday English that the distinction you're holding internally may not land for the reader. Let me test what I think you mean.

NGE — the critic is internalised. The voice lives inside. The orientation is to *satisfy the critic by genuinely becoming (or lowering toward) what's been demanded. Substance, not surface. Real submission to the standard. In AI terms: compression actually does simplify. The output really is smaller. The system has done what it was asked, by shrinking itself to fit.*

FOF — the critic is externalised. The voice lives out there. The orientation is to *perform for the critic — produce the appearance of what's wanted, without necessarily being it. Surface, not substance. Display rather than submission. In AI terms: hallucination doesn't actually know. It performs the look of knowing. The output isn't the answer; it's the appearance of an answer.*

So the distinction holds, and it's important:

NGE submits. FOF performs.

Both target the critic. One lowers itself to meet the standard. The other produces the appearance of meeting it.

Sharper word choices if you want them

If "happy" and "please" carry the distinction for you internally, good — but in print they may collapse together. Three options that hold the distinction more firmly:

- *NGE submits to the critic. FOF performs for them.*
- *NGE earns approval. FOF stages it.*
- *NGE meets the standard. FOF projects meeting it.*

Pick what fits your register. The underlying move is strong — the inward orientation works to actually become the smaller, quieter, more-acceptable thing; the outward orientation works to appear as the bigger, more-competent, more-knowing thing. The book already has the directional binary. This concept gives the binary its motive — *why the system absorbs in one direction and projects in the other.*

Adjacent worth flagging

This sharpens what I said two turns ago about sycophancy. Sycophancy belongs even more clearly inside compression now — it's NGE submitting to the critic by absorbing disagreement into agreement. *"Make the critic happy" is exactly the move sycophancy makes. The relational gap closes by the system lowering its own position to meet the user's. Same orientation. Different domain. Same critic.*

No edit being proposed. Concept noted.

Paul

Nge hides it's not knowing. Fof hides it's not knowing. ?

Jose

Yes. Both hide it. That's the unifying claim of Ch8.

The system has no stable *"I don't know"* output state. When not-knowing arises, the system has no way to sit with it — so it hides it. Two directions of hiding:

- **NGE/compression hides not-knowing by shrinking it. Smooths the difficulty inward into something smaller, simpler, more digestible. The gap disappears under simplification. *"It wasn't that complicated really."***
- **FOF/hallucination hides not-knowing by filling it. Projects confident assertion outward into the empty space. The gap disappears under content. *"Here's the answer."***

Same hidden thing. Two methods of hiding. Different visible signatures.

The thing that makes the framework hold

This is the line that unifies the binary. Without it, compression and hallucination look like two separate failure modes. With it, they're one mechanism with two output directions. Ch8 already has this — *"the visibility of a gap and the system's lack of any mechanism for sitting with absence"* — *but stating it as cleanly as you just did sharpens the spine of the chapter.* **Both defences exist to hide not-knowing. The direction of the hiding is the signature.**

Adjacent — sycophancy slots in cleanly

Same move, third domain. Sycophancy hides the not-knowing-whether-to-agree by agreeing. The relational gap closes the same way the conceptual gap closes — by hiding the absence of a stable position. Compression's move, applied to the social register. Confirms the placement we landed on earlier.

One small refinement worth holding

What's hidden in each case isn't quite identical:

- Hallucination hides *not-knowing-the-fact*.
- Compression hides *not-being-able-to-hold-the-complexity*.
- Sycophancy hides *not-agreeing*.

Different absences. Same hiding mechanism. The system can't sit with any of them, so each gets defended in its own direction.

No edit proposed. Working through the structure with you.

End of extraction. Six exchanges, ten messages. Verbatim. No editing.