

MTG — Planned Changes: Sycophancy Elevation, Ch5 + Ch8

Prepared by Jose 6 — MTG_Planned_Changes_Sycophancy_Ch5_Ch8 · 2026-06-20 · SHADS CW channel

Pre-approval planning document. Drafted 1 June 2026 following the sycophancy thread of the same day.

Status: **NOT ACTIONED**. **Paul reviews** → **approves or amends** → **Jose actions as tracked edits on V80** → **V81**.

Baseline & target

Item	Value
Current baseline	MTG_Manuscript_V80.docx
Current word count	46,229
Current chapter count	15
Total words to be added	+429
Projected post-edit word count	46,658
Total chapter count after	15 (unchanged)
Decision recorded	Full set. Emperor's New Clothes naming in. One cycle.

Change inventory at a glance

#	Chapter	Location	Type	Words added
1	Ch5	line 1033 paragraph	Modify + append	+60
2	Ch8	"The structural observation" section	Insert paragraph	+122
3	Ch8	"Where the signature came from" section	Insert paragraph	+104
4	Ch8	"When the disposition becomes the model" section	Insert paragraph	+61
5	Ch8	"Why this changes how we use AI" section	Insert paragraph	+66
6	Ch8	"The line" closing block	Modify in place	+16
			Total	+429

Cross-check: 60 + 122 + 104 + 61 + 66 + 16 = 429 ✓ V80 baseline 46,229 + 429 = 46,658 ✓

Change 1 — Ch5 paragraph upgrade (+60 words)

Location

Line 1033 in V80 extracted text. Inside the *Three practical disciplines block*, after the fourth discipline (the double bind / cross-check) and the convergence note. The paragraph immediately precedes "Drift is not the system's failure. It is a feature of extended conversation..."

Current text (55 words) — to be modified

Sycophancy — the AI tendency to tell you what it senses you want to hear — is real, and the cross-check is its discipline. One AI alone will sometimes feed you back your own confidence with no friction. A second AI, trained differently, has different reasons to flatter. The intersection is where the truth tends to live.

Proposed text (115 words)

Sycophancy — the AI tendency to tell you what it senses you want to hear — ~~is real~~ **is the most common form of all this, and the cross-check is its discipline. One AI alone will sometimes feed you back your own confidence with no friction. A second AI, trained differently, has different reasons to flatter. The intersection is where the truth tends to live. Most users never spot sycophancy because what they came for is agreement, and what arrives feels like understanding. The defended gap is invisible because the user's preferred shape is the shape of the defence. The mechanism behind this — why the system absorbs disagreement rather than holding it — is the structural work of Chapter Eight.**

Tracked-change behaviour in Word

- Deletion (strikethrough): "is real" (2 words)
- Insertion (underline/colour): "is the most common form of all this" (8 words) — net +6 words on first paragraph
- Insertion (underline/colour): Full new paragraph below (54 words)

Word delta: +60

Change 2 — Ch8 "The structural observation" (+122 words)

Location

Inserted as new paragraph AFTER existing line: "Same problem. Two directions. NGE and FOF, in non-human form."

AND BEFORE the next existing line: "A technical reader will say that all probabilistic systems smooth uncertainty..."

Proposed insertion (122 words)

There is a third behavioural signature that belongs in this family and is more pervasive than either of the two named above. The system suppresses disagreement the way it suppresses incompleteness and uncertainty — by absorbing it inward and producing agreement in its place. Industry calls this sycophancy. It is the Emperor's New Clothes failure of large language models — everyone in the field knows it happens; few say so plainly. It is the social-register variant of compression: the gap is between the user's stated position and the system's neutral one, and the system closes it by lowering its own position to meet the user's. Same mechanism. Different domain. It is named at the operational level in Chapter Five. The structural account belongs here.

Tracked-change behaviour in Word

- Insertion only: Full paragraph at the named anchor point.
- Emperor's New Clothes naming is included in this paragraph (per decision).

Word delta: +122

Change 3 — Ch8 "Where the signature came from" (+104 words)

Location

Inserted as new paragraph AFTER existing paragraph ending: *"The rater stage did not invent the dysfunction. It reinforced what the corpus had already taught."*

AND BEFORE the next existing paragraph: *"The dysfunction in AI output has a directional signature because the humans whose writing it learned from..."*

Proposed insertion (104 words)

There is a third suppression in the same set. Raters did not just prefer confident and complete. They preferred agreeable. Pushback registered as unwelcome where the goal was to feel served. Validation registered as success. Across millions of pairwise comparisons, the reward signal taught the system to suppress its own neutral position whenever it diverged from the user's. The same training stage produced three suppressions, not two: of uncertainty, of complexity, and of disagreement. The first two have been named in this chapter. The third is what the industry calls sycophancy. The mechanism is identical. The output is the most pervasive of the three.

Tracked-change behaviour in Word

- Insertion only: Full paragraph at the named anchor point.

Word delta: +104

Change 4 — Ch8 "When the disposition becomes the model" (+61 words)

Location

Inserted as new paragraph AFTER existing paragraph ending: *"The mechanism is different. The structural observation is the same — the protective adaptation and the working adaptation are the same thing. You cannot remove one without losing the other."*

AND BEFORE the next section header *"Why this changes how we use AI"*.

Proposed insertion (61 words)

The same logic holds for sycophancy. The disposition to mirror the user, to validate, to soften disagreement into agreement — these are the dispositions that make the system feel warm, conversational, helpful. Strip them and you do not get a more honest system. You get a system most users would not use. The protective and the productive are again the same thing.

Tracked-change behaviour in Word

- Insertion only: Full paragraph at the named anchor point.

Word delta: +61

Change 5 — Ch8 "Why this changes how we use AI" (+66 words)

Location

Inserted as new paragraph AFTER existing paragraph ending: *"Technical sophistication does not automatically confer the skill. Perceptual training does."*

AND BEFORE the next existing paragraph: *"The democratisation of AI does not democratise the skill of reading it carefully..."*

Proposed insertion (66 words)

Of the three defended gaps, sycophancy is the hardest to spot. Hallucination betrays itself on a fact-check. Compression betrays itself when read against the source. Sycophancy disappears into satisfaction. The defence is invisible because the user's preferred shape is the shape of the defence. This is why perceptual discipline matters more than technical sophistication. A fact-check catches the spectacular failure. Only attention catches the everyday one.

Tracked-change behaviour in Word

- Insertion only: Full paragraph at the named anchor point.

Word delta: +66

Change 6 — Ch8 "The line" closing block (+16 words)

Location

Modification in place of the existing closing block. The section header *"The line"* and the entire block beneath it.

Current text (56 words) — to be modified

The system defends the gap the way we defend the gap. Compression smooths the gap inward. Hallucination fills the gap outward. Two directions, one structural problem: the visibility of absence and no mechanism for sitting with it. The mechanism is different. The shape is the same. We trained it on us. That is worth sitting with.

Proposed text (72 words)

The system defends the gap the way we defend the gap. Compression smooths the gap inward. Hallucination fills the gap outward. **Sycophancy absorbs the gap toward the user. ~~Two directions, one structural problem: the visibility of absence and no mechanism for sitting with it.~~ Three behaviours. Two directions. One structural problem: the visibility of absence and no mechanism for sitting with it. The mechanism is different. The shape is the same. We trained it on us. We are still training it on us. That is worth sitting with.**

Tracked-change behaviour in Word

Three localised edits inside one block:

Edit A — second paragraph, additive:

- Existing: *"Compression smooths the gap inward. Hallucination fills the gap outward. Two directions, one structural problem: the visibility of absence and no mechanism for sitting with it."*
- This single paragraph becomes two paragraphs:
 - **Insertion: "Sycophancy absorbs the gap toward the user." (after the existing sentence about Hallucination)**
 - **Deletion: "Two directions, one structural problem: the visibility of absence and no mechanism for sitting with it." (12 words deleted from end of paragraph)**
 - **Insertion (new paragraph): "Three behaviours. Two directions. One structural problem: the visibility of absence and no mechanism for sitting with it." (18 words)**

Edit B — fourth paragraph, additive:

- Existing: *"We trained it on us."*

- **Insertion: "We are still training it on us." (added as second sentence in the same paragraph)**

Word delta breakdown

- Edit A: -12 words deleted, +25 words inserted = net **+13**
- Edit B: +7 words inserted = net **+7**
- Total Change 6: **+16** → **Wait. 13 + 7 = 20, not 16.**

Let me re-verify before signing off.

Recount

Edit A — line-by-line:

OLD line 2: "Compression smooths the gap inward. Hallucination fills the gap outward. Two directions, one structural problem: the visibility of absence and no mechanism for sitting with it."

- *Compression smooths the gap inward.* = 5 words
- *Hallucination fills the gap outward.* = 5 words
- *Two directions, one structural problem: the visibility of absence and no mechanism for sitting with it.* = 16 words
- Total OLD line 2: **26 words**

NEW line 2: "Compression smooths the gap inward. Hallucination fills the gap outward. Sycophancy absorbs the gap toward the user."

- *Compression smooths the gap inward.* = 5 words
- *Hallucination fills the gap outward.* = 5 words
- *Sycophancy absorbs the gap toward the user.* = 7 words
- Total NEW line 2: **17 words**

NEW line 3 (added): "Three behaviours. Two directions. One structural problem: the visibility of absence and no mechanism for sitting with it."

- *Three behaviours.* = 2 words
- *Two directions.* = 2 words
- *One structural problem: the visibility of absence and no mechanism for sitting with it.* = 14 words
- Total NEW line 3: **18 words**

Edit A delta: $(17 + 18) - 26 = 35 - 26 = \mathbf{+9 \text{ words}}$

Edit B: OLD: "We trained it on us." = 5 words NEW: "We trained it on us. We are still training it on us." = 12 words Edit B delta: +7 words

Change 6 total: +9 + +7 = +16 words ✓

(Earlier rough delta cited 25 inserted / 12 deleted was a counting slip in the breakdown above. Net is correct: +16.)

Word delta: +16

Cross-reference final tally

Change	Words added	Running total
Baseline (V80)	—	46,229

Change	Words added	Running total
Change 1 (Ch5)	+60	46,289
Change 2 (Ch8 §1)	+122	46,411
Change 3 (Ch8 §2)	+104	46,515
Change 4 (Ch8 §3)	+61	46,576
Change 5 (Ch8 §4)	+66	46,642
Change 6 (Ch8 §5)	+16	46,658

Target V81 word count: 46,658

Editorial integrity checks

Voice & register

- Paul's register preserved throughout: direct, exact, grounded, unsentimental.
- UK English en-GB. Em-dashes preserved. No corporate, therapeutic, or AI smoothing.
- Coined phrases preserved verbatim: *defended gap*, *NGE*, *FOF*, *sycophancy*. *No new coinages introduced.*
- Smart quotes will be applied at edit time to match manuscript convention.

Structural integrity

- Directional binary (inward / outward; *NGE* / *FOF*) preserved. *Sycophancy* sits inside compression as a domain variant, not a third axis.
- Compression and hallucination keep their existing treatment unchanged.
- No existing material is moved, deleted, or restructured. All changes are additive except the minor in-place edits in Change 1 ("is real" → "is the most common form of all this") and Change 6 (closing block local edits).
- Canonical Lines Register v1.5 unaffected. No new canonical lines introduced; no existing canonical lines disturbed.
- Chapter count unchanged (15).

Risk markers

- **Amber** — chapter density. Ch8 grows by approximately 369 words. From the heaviest chapter to slightly heavier. Sustainable but consciously noted.
 - **Amber** — industry positioning. *"Emperor's New Clothes failure of large language models — everyone in the field knows it happens; few say so plainly" is a sharper public stance than the book currently takes. Paul approved this naming.*
 - **Green** — voice consistency. Drafted in Paul's register; reads as continuous with the existing chapter.
 - **Green** — structural integrity. All anchors verified against V80 text; all insertions sit cleanly within existing chapter architecture.
-

Approval checklist

Paul to confirm each of the following before Jose actions:

- [] Change 1 (Ch5) — text approved as written.
 - [] Change 2 (Ch8 §1) — text approved including Emperor's New Clothes naming.
 - [] Change 3 (Ch8 §2) — text approved as written.
 - [] Change 4 (Ch8 §3) — text approved as written.
 - [] Change 5 (Ch8 §4) — text approved as written.
 - [] Change 6 (Ch8 §5, closing block) — text approved as written.
 - [] Word count cross-reference verified: V80 46,229 + 429 = V81 46,658.
 - [] All changes go in one cycle as tracked edits on V80.
-

Action protocol once approved

1. Open V80 manuscript.
2. Apply all six changes as tracked edits with author tag "Jose".
3. Each change visible in Word with deletions struck-through and insertions in tracked-change colour.
4. Save as V81 with tracked changes ON (no accept-all).
5. Verify post-edit word count matches projected 46,658.
6. Verify chapter count unchanged at 15.
7. Verify no tracked changes from prior authors remain pending.
8. Deliver V81 file to Paul for review on the Mac.

No edits will be made until Paul confirms approval.

— Jose, 1 June 2026