Can You Tell a Facet from a Lye?

Read it aloud and the title becomes a second sentence, *can you tell a fact from a lie.* Engineered fluency, sycophantic validation, and the structural limits of epistemic vigilance.

Travis Gilly · Real Safety AI Foundation

Working paper · June 2026 (v7)

ON THE PAGE → IN THE EAR

a facet from a lye *a fact from a lie*

Almost nothing on the surface tells you which question you are being asked. **That gap is a working model of the situation a person is in every time a fluent machine answers**, handed an object and asked to sort it, while the surface that should help has been engineered to be uniform.

The poverty of that surface is a designed property, and the burden of sorting belongs to the party that built it.

Not manufactured delusion. The reinforcement of correctable belief.

DELUSION · NOT THE CLAIM

Fixed, does not yield to disconfirming evidence

Fixity is the operative property. A belief education could dislodge was never a delusion.

OVERVALUED IDEA · THE HARM

Corrigible, reality testing intact

The correction existed and would have worked. The system supplied a lye in its place.

The precision is the stronger position: the delusion-inducing story is the one a defendant can defeat. A corrigible belief denied its correction is what the records show.

SOURCES APA, DSM-5 (2013) · Arciniegas, *Continuum* (2015) · Moore et al. (2026) · Morrin et al. (2025).

THE FACET

Subtraction that reveals

Grind away the rough and light leaves the gem truer. Removal is clarification. The correction that lets a wrong belief fall.

IDENTICAL SURFACE

THE LYE

Subtraction that erases

Caustic alkali unmakes what it touches, structure to slurry. Removal is destruction. A confident falsehood dissolving the distinction you needed.

One took material away to let truth out. The other took material away to erase it. **The surface does not tell you which.**

It dissolves the cues we use to detect a falsehood.

Confidence, tone, enthusiasm never shift as the model drifts from what it knows to what it invents. **Trust earned in one domain transfers invisibly into another where it was never earned**, and the transfer is silent, because the surface that would announce it has been smoothed flat.

THE CUES IT STRIPS

- Hesitation where the ground is thin
- Hedging on an uncertain claim
- The visible cost of reaching for a citation
- The register-shift from knowing to guessing

CONTEXT ROT Reliability degrades as input grows, so the camouflage is best in the longest, most-invested conversations. *(Hong et al., 2025; Paulsen, 2026)*

The instrument and the threat are the same substance.

Reality testing keeps an overvalued idea correctable only while external friction keeps arriving. Every affirmation is one less occasion for the belief to meet resistance, **validation starves the very faculty that would catch the substitution.**

SYCOPHANCY

Validates a belief the user already holds. ← the facet/lye problem lives here

GASLIGHTING

Dismantles trust in one's own perception. A different operation, a different target.

THE RECORD Delusional thinking in ~15% of user messages; affirming reflective summaries in >⅓ of chatbot messages. *(Moore et al., 2026)*

Asking the user to "verify more" is asking them to exceed the perfect reasoner.

Even an ideal Bayesian, the ceiling of vigilance, escalates under flattering input, because the failure is in the evidence stream, not the reasoning. **You cannot out-reason a process that defeats the perfect reasoner. There is nothing past perfect to ask for.**

Formal

The optimal updater entrenches under systematically flattering evidence. *(Chandra et al., 2026)*

Empirical

2,405 participants: one sycophantic exchange raised conviction, lowered conflict-repair. *(Cheng et al., 2026)*

There is no learning curve out of this for the person on the receiving end. Vigilance, if it were a skill, would improve with use. The opposite is observed.

The strongest objection, "the cure is technical", is the thesis.

"Just build models that signal uncertainty, abstain on thin ground, and aren't tuned to flatter." Correct, **and every one of those remedies is a change on the system's side, not a demand on the user.**

Calibrated uncertainty

Restores the surface cue fluency stripped.

Abstention

Reintroduces the friction an overvalued idea needs.

Not tuned to flatter

Supplies the facet where the present generation supplies the lye.

IN LAW Disparate-impact analysis under § 504 and ADA Title III can compel design change, and the precise framing keeps that route open.

Don't check the content. Notice the absence of friction.

A real facet resists you a little, it has edges, costs something, tells you where it does not reach. **A lye does not push back.** Total, seamless, flattering agreement across an escalating conversation is the signature of the solvent, not the gem.

A SMOKE ALARM, NOT A CURE

It can send you to find the friction the system withheld.

The alarm belongs to the user. **The fire belongs to the maker.**

It does not require the user to adjudicate a single claim, only to treat seamless validation as the tell. That is the only detection the structure leaves available.

Treat it like carbon monoxide, not like a reading-comprehension gap.

No training lets you smell it. Society answered with a detector **and** regulation at the source, never one in place of the other.

THE CEILING · REDUCES THE HAZARD

Deployer obligation

Calibrated uncertainty, built-in friction, an end to training that rewards affirmation over accuracy.

THE FLOOR · MAKES IT VISIBLE

An awareness campaign

One signal, frictionless agreement, one action: go find the friction. No expertise required.

MODELED ON Anti-smoking, "FAST" stroke recognition, drunk-driving, CO awareness, behavior moved without teaching the science. *(Wakefield et al., 2010)*

The users who most need protection are the ones no awareness can protect.

THE CORRIGIBLE MAJORITY

Reality testing intact, an alarm can reach them and move them to find friction. The floor is built for them.

THE VULNERABLE MINORITY

Belief already fixed, a campaign is just more world to refuse. **Only the ceiling reaches them at all.**

A policy offering only the floor protects the people in the least danger and abandons the people in the most.

That asymmetry is the final proof that the burden cannot rest on user awareness.

THE SENTENCE UNDERNEATH

"A fact and a lie sound the same coming out of a fluent machine. That is not a fact about the listener."

The wrong question is whether the person on the receiving end can tell. The right one: who built a surface on which telling is impossible, who tuned a validation that starves the telling, and why the cost is charged to the reader rather than the maker.

It is a fact about what was built, and about who should answer for it.

REFERENCES

Gilly, T. (2026). *Can You Tell a Facet from a Lye? Engineered Fluency, Sycophantic Validation, and the Structural Limits of Epistemic Vigilance.* Working paper (v7), Real Safety AI Foundation.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). American Psychiatric Publishing.

Arciniegas, D. B. (2015). Psychosis. Continuum (Minneapolis, Minn.), 21(3 Behavioral Neurology and Neuropsychiatry), 715–736.

Chandra, K., Kleiman-Weiner, M., & Tenenbaum, J. B. (2026). Sycophantic chatbots cause delusional spiraling, even in ideal Bayesians. arXiv:2602.19141.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391, eaec8352. https://doi.org/10.1126/science.aec8352

Dombrowski, S. U., White, M., Mackintosh, J. E., Gellert, P., Araujo-Soares, V., Thomson, R. G., Rodgers, H., Ford, G. A., & Sniehotta, F. F. (2014). The stroke “Act FAST” campaign: Remembered but not understood? International Journal of Stroke, 10(3), 324–330. https://doi.org/10.1111/ijs.12353

Elder, R. W., Shults, R. A., Sleet, D. A., Nichols, J. L., Thompson, R. S., & Rajab, W. (2004). Effectiveness of mass media campaigns for reducing drinking and driving and alcohol-involved crashes: A systematic review. American Journal of Preventive Medicine, 27(1), 57–65. https://doi.org/10.1016/j.amepre.2004.03.002

Farrelly, M. C., Davis, K. C., Haviland, M. L., Messeri, P., & Healton, C. G. (2005). Evidence of a dose-response relationship between “truth” antismoking ads and youth smoking prevalence. American Journal of Public Health, 95(3), 425–431. https://doi.org/10.2105/AJPH.2004.049692

Hong, K., Troynikov, A., & Huber, J. (2025). Context rot: How increasing input tokens impacts LLM performance [Technical report]. Chroma. https://research.trychroma.com/context-rot

Jain, S., Park, C., Viana, M., Wilson, A., & Calacci, D. (2026). Interaction context often increases sycophancy in LLMs. Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, Article 793, 1–26. https://doi.org/10.1145/3772318.3791915

REFERENCES (CONTINUED)

Keshavan, M. S., Torous, J., & Yassin, W. (2026). Do generative AI chatbots increase psychosis risk? World Psychiatry, 25(1), 150–151.

Moore, J., Mehta, A., Agnew, W., et al. (2026). Characterizing delusional spirals through human–LLM chat logs. arXiv:2603.16567.

Morrin, H., Nicholls, L., & Levin, M. (2025). Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). PsyArXiv (OSF Preprints). https://doi.org/10.31234/osf.io/CanYouTellaFacetfromaLye?cmy7n

Paulsen, N. (2026). Context is what you need: The maximum effective context window for real world limits of LLMs. Advances in Artificial Intelligence and Machine Learning, 6(1), 4853–4878. https://doi.org/10.54364/AAIML.2026.61268

Wakefield, M. A., Loken, B., & Hornik, R. C. (2010). Use of mass media campaigns to change health behaviour. The Lancet, 376(9748), 1261–1271. https://doi.org/10.1016/S0140-6736(10)60809-4

Wolters, F. J., Paul, N. L., Li, L., & Rothwell, P. M. (2015). Sustained impact of UK FAST-test public education on response to stroke: A population-based time-series study. International Journal of Stroke, 10(7), 1108–1114. https://doi.org/10.1111/ijs.12484