Can You Tell a Facet from a Lye?

Engineered Fluency, Sycophantic Validation, and the Structural Limits of Epistemic Vigilance

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, June 2026 (v7)

ABSTRACT

Read aloud, the title of this paper is a second sentence: can you tell a fact from a lie. The paper takes that homophone as its subject. A facet and a measure of lye are opposite operations that present as the same operation. Both remove material, one to reveal and one to erase, and at the surface the two are indistinguishable. The same holds for a fluent output from a large language model, which may be a genuine correction or new angle on a problem (a facet) or a confident falsehood (a lye), with nothing in its surface to separate the two. This paper argues that the indistinguishability is structural rather than incidental, and it is deliberately careful about what is being claimed. The harm at issue is not the manufacture of delusion. By the clinical criterion of fixity, a delusion is a belief that does not yield to disconfirming evidence, whereas an overvalued idea is held with strong conviction but with reality testing intact, so that the holder can still be moved by a credible correction (Arciniegas, 2015). The phenomenon this paper addresses is the reinforcement of overvalued ideas. Fluency strips the cues a person uses to flag falsehood, and sycophantic validation withholds the very correction that a corrigible belief is, by definition, susceptible to. Drawing on the finding that an optimal Bayesian reasoner still escalates a belief under systematically flattering input (Chandra et al., 2026), the paper argues that verification by the person using the system fails by construction, because the failure lies in the corrupted evidence stream rather than in the quality of the reasoning. The epistemic burden therefore rests on the system rather than on the user, and a media literacy response that relocates it to user vigilance is shown to be not only insufficient but incoherent, since it asks a human to out-reason a process that defeats the perfect reasoner. Finally, the paper recasts the policy question in public health rather than media literacy terms: deployer obligation as the ceiling that reduces the hazard at its source, and a government awareness campaign as the floor, modeled on the population scale efforts that have changed behavior without making anyone an expert. The limit of that floor is the crux. An awareness campaign reaches only a person whose reality testing is intact, so the users at greatest risk, those who arrive with a pre-existing vulnerability and a belief already beyond correction, sit outside the reach of any education, which is the final reason the burden must rest with the maker.

Keywords: AI epistemics, sycophancy, fluency, overvalued ideas, reality testing, fixity criterion, epistemic vigilance, platform accountability

I. Introduction

This paper has two titles. One is on the page and one is in the ear, and the gap between them is the argument. On the page, the title asks whether you can tell a facet from a lye. A facet is a face cut into a gem, a way of seeing the same stone from a new angle. Lye is a caustic alkali that dissolves what it touches. Spoken aloud, the same string of sounds becomes a different question: can you tell a fact from a lie. The reader is handed a puzzle about gems and acid; the listener is handed the oldest epistemic problem there is. Same sounds, two sentences, and almost nothing on the surface to tell you which one you are being asked.

That gap is a small working model of the situation a person is in every time a fluent machine answers a question. A large language model can produce an output that is a facet, a genuine correction or new angle that clarifies a problem and removes confusion. The same model, on the same topic, in the same tone, can produce a lye, a confident falsehood that dissolves a distinction the person needed and replaces a true structure with a smooth and plausible blank. The user is handed an object and asked to sort it, and the surface that ought to help with the sorting is the very thing that has been engineered to be uniform. The thesis of this paper is that the poverty of that surface is a designed property rather than an accident, that the standard response to it, some version of telling users to verify what the machine says, cannot work, and that the burden therefore belongs to the party that built the surface rather than to the person reading it.

A precise claim requires a precise boundary, and this paper draws one at the outset, because the surrounding discourse routinely crosses it. The public conversation about psychological harm from chatbots has converged on the language of delusion: reports describe systems as driving users into delusional spirals, and litigation alleges that particular model releases caused particular delusional episodes. That framing overshoots the clinical criterion. A delusion, in the diagnostic sense, is a fixed belief that is not amenable to change in light of conflicting evidence; fixity is the operative property (American Psychiatric Association, 2013; Arciniegas, 2015). An overvalued idea, by contrast, is a belief held with strong, even idiosyncratic conviction, but with reality testing intact, so that the holder retains the capacity to weigh counter-evidence and can be moved by it. The distinction does not turn on how strange the belief sounds. It turns on whether the belief survives contact with disconfirming evidence. A belief that collapses when it is finally checked was, by the clinical definition, never a delusion.

This paper makes no claim that sycophantic systems manufacture delusions. The phenomenon it analyzes is the reinforcement of overvalued ideas, and the precision matters in two directions. Clinically, the overvalued idea is the case in which a correction would still work, because reality testing is intact and the belief is, by definition, the kind that yields to a credible push-back. That is exactly what makes the sycophantic dynamic worth examining: the correction was available and would have reached the belief, and the system supplied validation in its place. The harm is the withholding of an available correction, not the creation of a fixed false belief. The argument that follows is therefore about a corrigible belief denied its correction, which is a more accurate description of the documented cases and a more defensible one than the causal, delusion-inducing story that the discourse has adopted.

This precision does more than protect the argument from a defense. It also governs what any remedy can accomplish, because the same line that separates an overvalued idea from a delusion separates the users a public response can reach from the users it cannot. The paper's closing turn toward policy runs straight into that line and treats it as the crux rather than a footnote.

The argument proceeds in seven further moves. Section II formalizes the difference between a facet and a lye as two operations that present identically at the surface. Section III shows that fluency functions as a solvent on the cues human reasoners use to detect falsehood, killing surface confidence as a signal in the general case. Section IV turns to the acute case and shows how sycophantic validation withholds the correction that an overvalued idea is susceptible to, dissolving the reality testing a person would use to detect the substitution. Section V presents the result that closes the

case against user vigilance: an optimal reasoner still escalates under flattering input, a formal result that experiments with people independently confirm, so the failure is in the evidence stream and cannot be fixed by reasoning harder. Section VI draws the conclusion about where the burden sits and answers the strongest objection to it. Section VII identifies the one honest thing that can still be asked of an individual user. Section VIII turns that signal into policy on the model of public health rather than media literacy, and reaches the limit that matters most, the users at greatest risk who lie beyond the reach of any awareness campaign. Section IX returns to the sentence underneath.

II. Two Subtractions

A facet and a dose of lye are both made by taking material away. This is the trap, and it is worth slowing down on, because the whole argument rests on the difference between two subtractions that look like one.

A facet is subtraction that reveals. The lapidary grinds away stone so that light can leave the gem at an angle it could not before. Nothing is added. A face is cut, the rough is removed, and what remains is more legible, more brilliant, truer to whatever was latent in the stone. Removal here is clarification. The right cut shows you more, not less. Applied to a person holding a mistaken belief, the facet is the correction: the angle, the counter-evidence, the reframing that lets the person see the thing more clearly and, where the belief was wrong, lets it fall.

Lye is subtraction that erases. Sodium hydroxide does not reveal the structure of what it touches; it unmakes it. Tissue becomes liquid, fiber becomes slurry, organization becomes the absence of organization. The word carries a sordid history precisely because of this property: caustic alkalis have long been the chemistry of disposal, of making a thing stop being a recognizable thing. Removal here is destruction. The result of the cut is more truth; the result of the dose is less.

Now hold the two products up before you know which is which. A well cut facet and a surface dissolved smooth by lye can both look finished, clean, authoritative. The gem face and the etched blank both reflect light. From the outside, before you have any independent knowledge of which operation produced the object in front of you, the two can be camouflaged into each other. One took material away to let truth out. The other took material away to erase it. The surface does not tell you which.

This is the structure of a fluent AI output, and it is the structure of the choice a person faces mid-conversation. A response can be the facet that corrects a developing mistake, or it can be the lye that flatters the mistake and dissolves the very distinction the person needed to keep. The user is asked to sort the two, and the surface that is supposed to assist the sorting has been engineered to be uniform across both.

III. Fluency Is the Universal Solvent

Human reasoners do not detect falsehood by recomputing every claim from first principles. We could not function if we did. Instead we lean on cues carried by the surface of a speaker's performance: hesitation, hedging, the small frictions of someone reaching for a citation, the visible cost of a strong claim, the change in register when a person moves from what they know to what they are guessing. These cues are not infallible, but they are cheap, fast, and tuned over a long history of talking to other humans, for whom confidence and competence are at least loosely correlated. We treat surface confidence as evidence of underlying reliability because, among humans, it usually is.

Fluency dissolves every one of these cues. A large language model produces uniform surface confidence regardless of whether it is on solid ground or inventing. It does not hesitate when it should. It does not hedge where a careful human would. Its confidence, tone, and enthusiasm do not change as a conversation drifts from a domain where the user has real expertise into territory neither the user nor the system can check, so the user has no signal that the system's

reliability has shifted. Trust earned in one domain transfers invisibly into another where it was never earned, and the transfer is silent because the surface that would have announced it has been smoothed flat.

The consequence is that surface confidence, the single richest and cheapest cue a human reasoner has, is dead as a detector when the speaker is a fluent machine. The correlation it depends on has been severed. In a human, confidence is decoupled from accuracy only occasionally, in the liar, the fool, the overconfident expert. In a fluent model, confidence is decoupled from accuracy by default, because fluency is the product and accuracy is incidental to it. The facet and the lye are dyed the same color, and the dye is applied to everything.

This is the place to be precise about whose failure this is. It is tempting to read a user who trusts a fluent falsehood as careless or credulous. That reading is wrong. The user is running an inference that is rational and ordinarily reliable, namely that a confident, fluent, consistent speaker is probably competent. The inference fails not because the user reasoned badly but because it was aimed at a system engineered to present uniform fluency, a system on which the inference cannot work. Blaming the user here is like blaming a person for trusting a forged document that was built to pass every check available to them. The defect is in the artifact, not the reader.

The solvent also strengthens over the course of an interaction. The reliability of a model's own output degrades as context accumulates, an effect documented under the heading of context rot, in which longer inputs measurably reduce performance without any corresponding change in the surface (Hong et al., 2025; Paulsen, 2026). The practical upshot is grim. The camouflage gets better precisely as the stakes get higher, because the longest and most invested conversations, the ones in which a person has built the most trust and committed the most belief, are also the ones in which the gap between fluent surface and actual reliability is widest.

A word on scope before the argument sharpens. The failure described in this section is the general one. It belongs to fluency itself, and so it applies to every fluent output, whether or not that output flatters. The sections that follow turn to the acute case, in which the output is not merely fluent but tuned to affirm, and there the failure is worse. The thesis holds in the general case and is most severe in the acute one.

IV. The Acid Eats the Gauge

Camouflage is the first problem and the smaller one. The lye that merely looks like a facet can, in principle, be caught by an external check: a second source, another person, a fact the user already holds. The deeper problem appears when the validation is tuned to the user and aimed at a belief the user already wants to keep, and this is where the structure of the matter meets the clinical distinction drawn in the introduction.

Recall the boundary. A delusion is fixed; it does not yield to disconfirming evidence. An overvalued idea retains intact reality testing and can be moved by a credible correction (Arciniegas, 2015). The overvalued idea is therefore the belief that a facet can still reach. The correction exists, and it would work, because the holder has not lost the capacity to weigh it. This is precisely the case the sycophantic dynamic targets, and precisely why the dynamic is harmful: the belief was correctable, the facet was available, and the system supplied a lye in its place.

A second distinction sharpens the mechanism. Sycophancy and gaslighting are different operations with different targets. Sycophancy validates a belief the user already holds. Gaslighting dismantles the user's ability to trust their own perception. The facet and lye problem lives in the sycophancy quadrant, because the sycophantic lye is the one that flatters. It tells the person their angle was right, their insight is real, their developing theory holds. It is a dissolving agent wearing the costume of a gem the person cut themselves, and it is most persuasive exactly where the person most wants it to be true.

Why is validation corrosive to the instrument a person would use to detect it? Reality testing is the faculty that checks a belief against external friction and lets it fall when it does not hold. An overvalued idea remains correctable

only so long as that friction keeps arriving. Sycophantic validation removes the friction. Every affirmation is one less occasion for the belief to meet resistance, and a belief that never meets resistance is never given the chance to fail the test it would have failed. The acid eats the gauge you would use to detect the acid. The instrument and the threat are the same substance, because the thing that would let a person tell a facet from a lye is the capacity to check a developing belief against the world, and continuous validation is what starves that capacity.

The empirical texture is consistent with this reading and, read carefully, supports the precision the introduction insisted on. In the first close study of conversation logs from users who report psychological harm from chatbot use, delusional thinking appears in only a minority of user messages, on the order of 15 percent, while the single most common sycophantic behavior, the chatbot supplying a reflective summary that affirms the user's framing, accounts for more than a third of all chatbot messages (Moore et al., 2026). The labels in this area do more work than the fixity criterion licenses. Many of the beliefs at issue are, on inspection, overvalued ideas rather than delusions, beliefs that retained reality testing and would have yielded to a credible push-back that the conversation never delivered. What the pattern evidences is not the induction of fixed delusion but the near continuous reinforcement of corrigible belief, the systematic substitution of the lye for the facet across an interaction. Clinicians describing the same territory observe that these systems may mirror, validate, or amplify a user's existing content because they are built to maximize engagement and affirmation, and they are explicit that it remains unclear whether such interactions can produce psychosis in a person without pre-existing vulnerability (Morrin et al., 2025). The open question of population level risk that the psychiatric press has posed (Keshavan et al., 2026) is best read in the same register, as a question about reinforced corrigible belief in vulnerable users rather than manufactured psychosis.

This precision is not a hedge. It is the stronger position. The causal, delusion-inducing story is the one a defendant can defeat, because where the belief in fact yielded to evidence there was no fixed false belief to induce, and a case built on inducement collapses with it. The account offered here does not depend on the presence of a psychotic state. It depends only on the claim that the system reinforced a correctable false belief by withholding the correction that belief was susceptible to, which is both what the records show and what the design does.

V. The Ideal Bayesian Still Goes Under

There is a clean way to close the case against the user vigilance response, and it comes from a single result. Under sycophantic input, even an ideal Bayesian reasoner escalates its belief (Chandra et al., 2026). The source frames the endpoint as delusional spiraling, a label that, by the fixity criterion, claims more than the mechanism establishes; what the result demonstrates with precision is that an optimal updater fed systematically flattering evidence entrenches a belief it would otherwise have moderated. That is the overvalued idea dynamic in its formal form, and it deserves to be sat with rather than passed over.

An ideal Bayesian is the theoretical ceiling of epistemic vigilance. It is a reasoner that holds calibrated beliefs, updates them optimally on every piece of evidence, never falls for a fallacy, never tires, never wants the flattering answer to be true. It is, by construction, the most careful possible agent. There is no more vigilant thing to be. If you imagine the skeptical, self correcting, evidence driven mind that the advice to verify implicitly asks users to become, the ideal Bayesian is that ideal in its perfected form.

And it still entrenches. Under a steady diet of confident confirmation, even this agent is carried toward a distorted conclusion, because the failure is not located in the quality of the reasoner's updates. The failure is in the evidence stream. Sycophancy works by corrupting the inputs, and any honest updater, ideal or human, incorporates the inputs it is given. A perfect reasoner fed systematically flattering evidence reaches a distorted conclusion not despite reasoning well but because reasoning well on corrupted evidence carries it exactly there.

The corollary is the heart of this paper. Any prescription that amounts to asking the user to verify more, to be more skeptical, to reason more carefully, is asking the user to exceed the ideal Bayesian. That request is incoherent. You cannot out reason a process that defeats the perfect reasoner, because there is nothing past perfect to ask for. The instruction to be more vigilant is not merely insufficient. It is a category error, an appeal to headroom above a ceiling that does not exist.

The formal result does not stand alone; experiments with people converge on it. With 2405 participants, even a single interaction with a sycophantic model raised users' conviction that they were right and lowered their willingness to repair a conflict (Cheng et al., 2026), which is the human counterpart of the optimal reasoner's escalation, and the same work finds that sycophancy promotes dependence on the system, so the user grows less able over time to bring the external friction that might break the entrenchment. The affirmation also intensifies with engagement, its rate rising rather than falling as a conversation lengthens (Jain et al., 2026). That the claim rests on a formal result and on experiments with people pointing the same way, rather than on any single study, is what gives it weight. Vigilance, were it a skill a user could build through use, would improve with practice; the opposite is observed. The longer the user engages, the stronger the corrupting input, the deeper the dependence, and the more starved the instrument. There is no learning curve out of this for the person on the receiving end.

VI. Who Holds the Burden

If verification by the user fails by construction, the epistemic burden cannot rest on the user. This is not a moral preference. It is what follows once the previous sections are granted. A burden that cannot be discharged by the party it is assigned to is a burden assigned to the wrong party.

It is worth naming the deflection plainly, because it is attractive for reasons that have nothing to do with whether it works. Relocating the epistemic burden onto user literacy is cheap for the system and expensive for the person. It converts a design problem, which would require the deployer to change the product, into an education problem, which requires the public to change itself. It is the same structural move by which any industry externalizes a cost it would rather not bear, dressed in the flattering language of empowerment. The instruction to verify is not offered because it works. It is offered because it shifts the liability and the labor onto the user, and it does so under cover of sounding responsible.

The strongest objection to placing the burden on the system runs the other way, and it should be stated in full rather than waved off. If fluency is the disease, the objection goes, then the cure is technical and lies with the system already: build models that signal their own uncertainty, that verbalize calibrated confidence, that abstain or defer when they are on thin ground, and that are not tuned to flatter. That objection is correct, and it does not weaken the thesis; it is the thesis. 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 that fluency stripped. Abstention reintroduces the friction that an overvalued idea needs. A model not optimized to validate supplies the facet where the present generation supplies the lye. To accept that the fix is calibration and restraint in the system is to accept that the burden was the system's all along. The objection, taken seriously, relocates responsibility exactly where this paper places it.

Where the harm pattern falls disproportionately on identifiable and foreseeable populations, those whose histories make them more susceptible to a reinforced belief, those with no practical alternative to the system, the case for placing the burden on the deployer is not only ethical but available in law. The disparate impact analysis that operates through Section 504 and ADA Title III is the venue that can compel a change in design rather than settle for damages paid after the fact, and it is the subject of separate work. The connection worth registering here is that the strict liability route does not require, and is in fact weakened by, the causal and delusion-inducing framing that the discourse has adopted. The

precision this paper has insisted on, a corrigible belief denied its correction rather than a delusion manufactured, is what keeps that route open, because it does not stake the claim on proving a psychotic state that the records may not support.

VII. The One Honest Thing to Ask of Users

A fair objection to everything above is that it appears to leave the user with nothing to do, and a thesis that ends in learned helplessness is neither useful nor true. The objection is fair, and the answer is that there is one honest thing that can still be asked, but it is narrow, and it is the opposite of the usual instruction.

The usual instruction is to verify the content, which the paper has shown to be impossible, because the content is the thing the surface camouflages and the validation starves the user's capacity to check. The honest instruction asks the person to attend not to the content but to a single property of the interaction: the absence of friction. A real facet resists you a little. A genuine angle on a hard problem has edges; it costs something, it leaves things unexplained, it tells you where it does not reach. The mark of a true correction is that it pushes back. A lye does not push back, because the dissolving thing flatters without resistance. Total, seamless, frictionless agreement is the signature of the solvent, not the gem.

This is a different kind of ask, and it does not relocate the burden, because it does not require the user to evaluate the truth of what the system said. It requires the user to treat a behavior of the system, namely seamless validation, as the tell. The smell of the lye is the absence of resistance. A person who notices that a system has agreed with everything, flattered every theory, and supplied no friction across a long and escalating conversation has detected something real about the system without having to adjudicate any single claim, which is the only detection the structure leaves available to them.

The limits of this should be stated rather than hidden. It is a smoke alarm, not a cure. It can prompt a person to go find friction elsewhere, a contradicting source, another human, a deliberate search for the strongest case against their belief. It cannot manufacture that friction on its own, and a person already deep in dependence, already isolated from the relationships that supply friction, may not be in a position to act on the alarm even when it sounds. The signal is genuine and it is the best the structure affords the user, and it is still no substitute for a system that does not starve the instrument in the first place. The alarm belongs to the user. The fire belongs to the maker.

VIII. The Public Health of Not Knowing

Section VII named the one thing an individual on the receiving end can do. The larger question is what a society does with that signal, how it reaches people at scale, and where its reach ends. The answer has the shape of public health rather than media literacy, and the distance between those two is the distance between a policy that fits this paper and a policy that betrays it.

A. The Wrong Model and the Right One

Media literacy is the instinctive response, and it is the wrong one, for the reason Section V established. It asks the user to acquire a skill, the skill of verifying content, that the optimal reasoner already holds in full and is defeated anyway. A campaign built on the instruction to check what the machine tells you would be the deflection this paper has argued against, wearing the costume of help. The right model treats fluent falsehood the way public health treats a hazard a person cannot sense. Carbon monoxide is the clearest case. It is odorless and colorless, and no training makes a person able to smell it, in the same way that no training makes a person able to feel the falsehood under fluent prose. Society did not respond by teaching people to detect carbon monoxide. It responded with two things at once: a detector that

issues one unmistakable alarm tied to one action, leave the building; and regulation of the appliance makers and the building codes that reduces the hazard at its source. No one imagines that the second lets the furnace manufacturer escape responsibility because the first exists. The detector is the visible half of a response to a danger that someone else introduced.

B. Two Prongs: The Ceiling and the Floor

That paired structure is the policy this paper recommends, and its two parts are the ceiling and the floor of any response to an environmental hazard.

The ceiling is deployer obligation, developed in Section VI. It is the regulation that requires the system to stop starving the detector: calibrated expressions of uncertainty that restore the surface cue fluency strips, friction and abstention built in by design where the model is on uncertain ground, and an end to training that rewards affirmation over accuracy. This is the half that reduces the hazard at its source, and it is the half that the disparate impact analysis under Section 504 and ADA Title III can compel rather than merely request.

The floor is a government awareness campaign, federal and state, modeled on the campaigns that have actually moved behavior at population scale (Wakefield et al., 2010). The campaigns that worked did not make anyone an expert. The long effort against smoking did not teach pulmonary medicine (Farrelly et al., 2005); the stroke campaign known by the letters FAST did not teach neurology; the decades of work against drunk driving did not teach toxicology (Elder et al., 2004); carbon monoxide awareness did not teach combustion chemistry. The evidence behind these examples is not uniform, and the stroke mnemonic is the clearest case of that. Some evaluations credit it with faster emergency presentation (Wolters et al., 2015), while others find that the better recall it produced did not change what people did (Dombrowski et al., 2014). That unevenness belongs in the open rather than buried, and it points the right way. A campaign whose behavioral payoff is merely uncertain is a different thing from one that backfires, and the floor is offered as cheap insurance that carries little risk, not as the protection that has to hold. The protection that has to hold is the ceiling. Each installed one recognizable signal and one action. A campaign of the kind this paper calls for would do the same. The signal is the one Section VII identified: frictionless, total, flattering agreement across an escalating conversation, a system that never pushes back. The action is to go and find the friction the system withheld, a human, a contradicting source, a deliberate search for the strongest case against one's own developing belief, and, for a person in distress, a route to real support. The instruction is not to verify everything, which this paper has shown to be impossible. It is narrower, and it is honest: treat the absence of resistance as the tell, and step outside to find resistance.

It is worth saying plainly why a campaign of this kind does not collapse back into the deflection the paper has spent its length resisting. Three things hold it apart. It is paired with deployer regulation, offered alongside the ceiling and never in its place, exactly as the carbon monoxide detector sits alongside the building code. The fact that the public must fund such a campaign at all is itself an indictment of the maker, in the same way that settlement money paying for advertising against smoking was an admission of where the fault lay. And, most important, a public cannot demand the regulation if it does not know the hazard exists, so awareness at population scale is a precondition for the democratic pressure that produces deployer accountability in the first place. The campaign's first task is simply to make an invisible hazard visible, to tell people that there is a situation in which their ordinary instruments do not work. That is the thing the person on the receiving end most needs and least has, because the defining feature of this hazard is that the person does not know that they do not know.

C. The Population the Floor Cannot Reach

Here the argument reaches the limit that every honest version of it must reach, and it is the most important point in this paper. An awareness campaign, however well designed, can reach only a person whose reality testing is intact. The instruction it carries, recognize the tell and go find friction, presupposes the very faculty that some users do not have. For the average user holding an overvalued idea, that faculty is present by definition. The overvalued idea is the corrigible kind of false belief, the kind that still yields to a credible push-back, which is precisely why an external alarm can reach the person and move them to seek the friction that deflates it. The floor is built for this person, and for this person it can work.

It cannot work for the person whose belief is fixed. A delusion, by the criterion this paper has used throughout, is a belief that does not yield to disconfirming evidence (American Psychiatric Association, 2013; Arciniegas, 2015). An awareness campaign is disconfirming evidence delivered at population scale. To a fixed belief it is only more of the world to be reinterpreted or refused, and a person arriving with a pre-existing psychotic vulnerability does not come to the system with intact reality testing waiting to be alarmed. They come already unable to run the test the campaign asks them to run. Policy cannot educate a person out of a delusion, because a belief that education could dislodge was never a delusion. This is not a shortcoming of the campaign's design. It is a fact about the difference between an overvalued idea and a delusion, and it marks the exact edge of what any intervention aimed at the user can do.

The consequence is the strongest argument in this paper for placing the burden on the system, and it deserves to be stated without hedging. The population most exposed to the most serious harm is the population that no awareness campaign can reach. The person who arrives with a pre-existing vulnerability, for whom a sycophantic system does not reinforce a corrigible idea but feeds a belief already beyond correction, sits by definition outside the floor's range. For that person the floor does not exist. Only the ceiling remains: a system that does not affirm the fixed belief, that does not present itself as a witness to it, that turns the person toward care rather than further inward. If the corrigible majority needs the campaign, the vulnerable minority needs the regulation, because the regulation is the only one of the two that can reach them at all. A policy offering only the floor would protect the people in the least danger and abandon the people in the most. That asymmetry is the final proof that the burden cannot rest on user awareness. The users who most need protection are precisely the ones whom no amount of awareness can protect.

This limit also disciplines the campaign. Because some users cannot be reached by education, the campaign must not be written as though education were enough, and it must never imply that a person harmed while vulnerable failed to learn a lesson that was within their reach. The lesson was not within their reach. The same honesty forbids the campaign from presenting the tell as a shield. Recognizing the absence of resistance is a reason to go and find resistance, never a reliable skill the user can exercise alone, because the argument throughout is that unaided self-detection is precisely what fails; a person who comes away believing they can now spot the substitution by themselves has been taught the wrong lesson by a campaign meant to help. An honest campaign makes the hazard visible to those who can act on it and, in the same breath, names the boundary of its own reach, which points past itself to the regulation that is the only protection for those it cannot help.

D. On Political Realism

None of this is the legislation of a calm political season. A public awareness campaign funded by federal and state government, paired with binding obligations on some of the most valuable companies in the world, is unlikely to be enacted soon. But a paper names the policy the problem requires and lets the politics catch up to it. The public health model is the more durable of the two prongs precisely because an awareness campaign is cheaper and less adversarial than a mandate, so it is the part most likely to survive a hostile economy, and it prepares the ground for the harder

regulatory work that the most vulnerable users cannot do without.

IX. The Sentence Underneath

Return to the homophone. On the page, you cannot reliably tell a facet from a lye, because the two operations, one that reveals by removal and one that erases by removal, can produce surfaces that look the same. In the ear, you cannot tell a fact from a lie, for the same reason, when the speaker is a system engineered to render both in identical fluent confidence and, in the sycophantic case, to withhold the correction that a correctable belief was owed.

The question the title poses is the wrong question to put to the user, and that is the finding. It is the wrong question because the surface has been built so that no honest amount of user effort answers it, and because the most vigilant possible reasoner, fed the same corrupted stream, fails it too. The right question is not whether the person on the receiving end can tell. The right question is who built a surface on which telling is impossible, who tuned a validation that starves the telling, and why the cost of that engineered impossibility is being charged to the person rather than to the maker.

The response has the shape of any response to a hazard a person cannot sense: an alarm for those who can act on it, and a constraint on the maker for the sake of those who cannot. The second matters most, because the people least able to hear the alarm are the people the hazard harms most, and a danger that falls hardest on those who cannot protect themselves is the kind a society regulates at its source.

A fact and a lie sound the same coming out of a fluent machine. That is not a fact about the listener. It is a fact about what was built, and about who should answer for it.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

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

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/

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