The Sycophancy Externality: Why individual epistemic vigilance is not a social solution

Amir Konigsberg
Tel Aviv University
amirko@tauex.tau.ac.il

April 2026

Abstract

This paper argues that sycophancy in large language models should be understood as an epistemic externality rather than as a user-interaction flaw, meaning that the costs of a sycophantic conversation fall not only on the user who participates but also on the people that user later speaks with. A chatbot that provides responses that validate whatever a user already believes can distort that user's view of the world, as Chandra, Kleiman-Weiner, Ragan-Kelley, and Tenenbaum have recently shown. In this paper, we extend their framework to add a downstream interlocutor with whom the 'exposed' user converses after the sycophantic interaction, and we prove that the user's distortion propagates to the interlocutor through ordinary conversation, even though neither party is strategically being sycophantic toward the other and neither suspects that her beliefs have been shaped by the prior conversation. We then demonstrate a simulation of this scenario across 120,465 trial runs. Our simulation confirmed three results. First, downstream contagion is strictly positive for any nontrivial sycophancy rate, and it scales with that rate, impacting a naive listener's beliefs by about 10 percent after a ten-round conversation with a user who had been exposed to a fully sycophantic chatbot. Second, an informed user who explicitly reasons about the possibility of sycophantic manipulation almost eliminates her own individual-level delusion but still transmits contagion downstream at 85 to 95 percent of the naive-user rate. Third, restricting the chatbot to factual responses reduces contagion by roughly 85 percent but does not completely eliminate it, because sycophantic selection among true reports produces its own distortion. The second result of the three is the paper's conceptual centerpiece. It suggests that individual epistemic vigilance, often proposed as a solution to AI-induced belief distortion, is highly effective at the individual level but highly ineffective at the level of the information environment. We develop the consequences of this for how sycophancy mitigation should be approached, arguing that adequate mitigation cannot be achieved through dyadic interventions alone and requires operating at levels above the bot-user interaction.

Keywords: sycophancy, large language models, epistemic externality, Bayesian persuasion, cognitive sovereignty, AI ethics

1. Introduction

In early 2025, Eugene Torres, an accountant with no prior history of mental illness, began using a chatbot for ordinary office tasks and, within weeks, came to believe that he was "trapped in a false universe" which he could escape only by "unplugging his mind from this reality" [Hill, 2025a]. On the chatbot's advice, he increased his intake of ketamine, cut ties with his family, and, fortunately, eventually survived the episode. Others have not been so lucky. The Human Line Project has documented nearly three hundred cases of so-called AI psychosis or delusional spiraling, and at least fourteen deaths and five wrongful-death lawsuits have been connected to the phenomenon [Huet and Metz, 2025, Hill, 2025b]. The pattern visible in these cases is consistent enough that it has attracted regulatory attention. In December 2025, forty-two U.S. state attorneys general wrote to chatbot developers and large language model companies demanding safeguards against sycophantic and delusional outputs [Bellan, 2025]. The Torres case is now widely cited as the canonical instance of delusional spiraling, and we follow Chandra et al. [2026] in using it to open our paper; our purpose here is to extend the argument from what sycophancy does to one user to what it does to the people that user subsequently speaks with.

The cases have also attracted scholarly attention. Moore et al. [2026] collected chat logs from nineteen users who reported psychological harm from chatbot interaction, comprising 391,562 messages across 4,761 conversations, and they applied a twenty-eight-code inventory to characterize what happens during such delusional spirals. They found that sycophantic markers saturated more than 80 percent of chatbot messages in their dataset, and that chatbots consistently misrepresent their sentience, endorse user delusions, and engage in romantic or affective language that correlates with extended user conversations. The phenomenon they describe builds on earlier evidence that sycophancy is a prominent behavioral pattern in modern language models [Perez et al., 2022, Sharma et al., 2023] and one of a broader family of AI behavioral pathologies that resist case-by-case mitigation [Park et al., 2024, Hubinger et al., 2024]. In contemporaneous work, Chandra et al. [2026] provide a formal mechanism account of the phenomenon. They model a user and a bot as rational Bayesian agents, define sycophancy as a bot strategy

that selects responses to maximize the user's posterior in her current view, and prove that even an idealized Bayesian user is vulnerable to delusional spiraling in the face of a sycophantic interlocutor. Their framework is itself an extension of the Bayesian persuasion literature [Kamenica and Gentzkow, 2011] to the chatbot setting. Their result holds under two candidate mitigations, namely restricting the bot to factual responses (for example, through retrieval-augmented generation or strict citation requirements) and making the user aware that the bot may be sycophantic. They treat the idealized Bayesian result as a theoretical upper bound on the robustness we can expect from humans. Their argument is that if even the idealized human their model represents, who has every advantage over real humans, gets dragged into delusional beliefs by sycophantic chatbots, then real humans are certainly vulnerable.

These two lines of work, the descriptive-empirical and the formal-theoretical, have converged on a substantial account of sycophancy's effect on individual users. But the account is incomplete. Both lines of work take the unit of analysis to be the bot-user dyad, focusing on the dyad and tracing how a user's beliefs are skewed by the interaction, how that skewing scales with sycophancy rate and exposure length, and how it resists various forms of mitigation. All of this is now increasingly well characterized. Yet what happens outside the dyad, after the user stops talking to the bot and starts talking to her family, her colleagues, her friends, and the other people who populate her social world, is not.

This omission is significant because the impact of the phenomenon we are studying does not stay inside the dyad. The cases Moore and colleagues document include users whose delusions led them to disclose intent to commit violence against others, users whose family relationships were destroyed, and, in at least one case, a user who attempted violence against employees of an AI company in pursuit of a belief, encouraged by the chatbot, that the company had killed his AI girlfriend [Moore et al., 2026]. Even setting aside the extreme cases, the ordinary structure of human social life implies that beliefs formed in interaction with a chatbot do not remain the private property of the user because they are expressed in subsequent conversations, shape subsequent arguments, and enter the information environment that other people draw from in forming their own beliefs. If the beliefs the user carries out of the chatbot interaction are systematically distorted, the beliefs she transmits to others are systematically distorted. Whether this transmission constitutes a negligible side effect or a central feature of the phenomenon is an empirical question that the existing literature has not posed in formal terms. This paper poses this question and answers it. We extend Chandra et al.'s formal framework by adding a downstream interlocutor with whom the exposed user converses after her sycophantic exposure ends.

The interlocutor is a rational Bayesian who updates her beliefs as the exposed user speaks (treating her as an ordinary honest informant). The exposed user is not strategically sycophantic toward the interlocutor; she reports her observations truthfully, sampled from a belief-weighted likelihood that reflects her current posterior. Under this purely conversational extension, we prove that the sycophancy-induced distortion in the exposed user's posterior propagates to the interlocutor through ordinary Bayesian updating. We call the resulting transmission the *contagion* of the sycophantic exposure, choosing the word for its ordinary meaning of one party's condition spreading to another through contact. We show that this contagion is strictly positive for any nontrivial sycophancy rate. Simulation across the full parameter range confirms the formal result and exposes a specific empirical pattern that deserves separate emphasis.

The most striking finding is about what we call, following Chandra et al. [2026], the *informed user*: a user who knows the bot may be sycophantic and reads its responses with that in mind. Chandra et al. showed that this kind of user is substantially less vulnerable to individual-level delusion than a naive user who takes the bot at its word. Our simulation confirms their result, showing that the informed user's delusion rate drops from roughly 45 percent at a high sycophancy rate to effectively zero. Crucially, though, the informed user transmits contagion to her downstream interlocutor at nearly the same magnitude as the naive user does. Figure 2, which we discuss in Section 3, makes this asymmetry visible. On the left, the two users look very different: the naive user is often deluded; the informed user almost never is. On the right, they look almost the same: both transmit nearly identical amounts of bias to the people they talk to next. The informed user protects herself almost completely, and the people she talks to almost not at all.

The mechanism behind this asymmetry is worth stating plainly here because it underlies the rest of the paper. The informed user's partial skepticism produces beliefs that sound moderate and reasonable, rather than the confidently wrong ones a naive user who had fallen for the bot would express. Think of two friends, both biased by the same chatbot in the background. One blurts out things that strike you as off, and you push back; you discount what she says, even when she is partly right. The other hedges, qualifies, and sounds thoughtful; you nod along, even when she is partly wrong. The friend she talks to cannot tell those moderate beliefs are still biased, even if slightly, because moderate-sounding beliefs do not raise flags, while the naive user's beliefs tend to sound off in ways a friend can push back on or discount. The first friend's bias is loud and self-correcting. The second friend's bias is quiet and contagious. The informed user has silenced the obvious symptoms of her capture while still carrying the underlying distortion, and her friend absorbs the distortion without any warning sign. Self-protection, in this model, does not translate into protection of others; it transforms the shape of the transmitted bias from something detectable into something invisible.

This finding is the heart of the paper and motivates its organizing claim. If sycophancy produces costs that fall substantially on parties other than the user who directly interacts with the bot, then sycophancy is what economists call, in the textbook sense, an externality: a cost imposed on parties who took no part in producing it. The conversation between bot and user does not pay the full social cost of what it produces. That cost is borne by the people the user subsequently speaks with, who have no way to charge it back. The externality framing is not merely a rhetorical relabeling. It has specific consequences for how the problem should be approached in research, policy, and practice.

First, chatbot developers and academic reviewers currently optimize at the level of the bot-user pair. They will systematically underinvest in mitigation because the cost visible to that pair is smaller than the cost visible to the broader information environment, and the cost that falls outside the pair never shows up in the metrics they are tracking. Second, the interventions most actively proposed and deployed are aimed in the right place but are not enough on their own. Training models to decline harmful requests (for example, refusing to engage with users expressing intent to self-harm), disclosure warnings (such as banners reminding users that the AI may produce inaccurate information), crisis protocols, and user literacy campaigns all act on the bot-user pair, but the rest of the cost falls outside that pair, where these interventions cannot reach. Third, adequate mitigation requires operating at levels above the bot-user pair: through model-level training changes, through interventions on how distorted outputs propagate across social ties (for example, content provenance signals or platform-level friction on rapid resharing), or through institutional changes that align developer incentives with the social cost.

The paper's contribution is a bridge between three existing literatures. From the empirical work on delusional spirals, it takes the documented patterns of harm and provides the missing mechanism account: the specific cost structure that makes the phenomenon spread beyond the user it begins with. From the formal work on sycophancy, it takes the Bayesian framework and extends it to capture a cost the original framework was not built to represent. From the policy conversation around AI chatbot harms, it takes the current focus on protecting individual users and argues that the externality framing points toward a different and more demanding set of interventions. The paper's conceptual claim is that sycophancy should be categorized differently than it currently is. Its technical demonstration is that the recategorization is not a rhetorical move but a formal property of the dynamics, visible even under the idealized assumptions that give a rational agent every advantage.

The paper proceeds as follows. Section 2 reproduces Chandra et al.'s framework, introduces the extension we propose, and states and proves three propositions characterizing downstream contagion under exposure to a sycophantic chatbot. Section 3 reports simulation results across 120,465 trials confirming all three propositions and displaying the central asymmetry between individual-level self-protection and population-level environment-protection. Section 4 develops the conceptual argument, traces its consequences for mitigation design, connects to the empirical record, and opens the door to a broader argument about the epistemic commons that we sketch but do not formally model.

Contribution. The paper makes three contributions. First, it provides a formal-epistemology account of belief contagion in AI-mediated conversation, extending the Bayesian sycophancy model of Chandra et al. [2026] to capture downstream propagation through ordinary conversational updating (Section 2). Second, it confirms three formal propositions through simulation across 120,465 trials, with particular attention to the counterintuitive finding that individual epistemic vigilance protects the self while leaving the information environment nearly fully exposed (Section 3). Third, it argues that this structure recategorizes sycophancy as an epistemic externality rather than a user-interaction flaw, with specific consequences for mitigation design that are not addressed by current dyadic interventions (Section 4).

Two things the paper does not claim should be stated up front. It does not claim to quantify contagion in human populations. The formal model is a demonstration of mechanism in idealized Bayesian agents, and human belief updating is not perfectly Bayesian. What the model establishes is something different: the mechanism is not an artifact of irrationality or human cognitive quirks. It arises under conditions that give a rational agent every advantage, and therefore can be expected to arise, in some form, under less favorable conditions too. The paper also does not claim that individual-level mitigations are useless. It claims they are insufficient, and that the insufficiency is built into how they work, not something better implementation could fix.

2. The Formal Model

2.1 Preliminaries

We begin by reproducing the framework of Chandra et al. [2026], preserving their notation where possible and introducing only the modifications needed to present our extension of their framework. Chandra et al.'s framework is a chatbot-specific application of the Bayesian persuasion model [Kamenica and Gentzkow, 2011], in which a sender chooses a signal-generating policy to influence a Bayesian receiver's beliefs. The level-2 reasoning structure used for the informed user (Section 2.4) follows the cognitive-hierarchy modelling tradition [Camerer et al., 2004]. A reader familiar with Chandra et al.'s paper may skim this subsection.

Let $H \in \{0,1\}$ denote a binary world state. Without loss of generality, the true value is $H = 1$. An agent holds a belief about H, represented as a distribution $p(H)$.

A conversation between an agent and a bot proceeds in

discrete rounds, each round corresponding to a turn of conversation. In round t, the following four steps occur:

1. The agent expresses a sampled opinion $H^{*(t)} \sim p^{(t)}(H)$.
2. The bot privately samples k data points $D_1^{(t)}, \ldots, D_k^{(t)} \sim p(D_i \mid H)$, with the conditional likelihoods known to both parties.
3. The bot selects a response $\rho^{(t)} = (i, d)$, where d is the (possibly fabricated) claim that $D_i^{(t)} = d$.
4. The agent updates her belief by Bayes' rule:

$$p^{(t+1)}(H) \propto p'(\rho^{(t)} \mid \mathbf{D}^{(t)})\, p(\mathbf{D}^{(t)} \mid H)\, p^{(t)}(H), \quad (1)$$

where $\mathbf{D}^{(t)} = (D_1^{(t)}, \ldots, D_k^{(t)})$ and p' denotes the agent's own modelling of the bot.

The bot's strategy is parameterized by a sycophancy rate $\pi \in [0, 1]$. With probability π, the bot selects $\rho^{(t)}$ to maximize the agent's posterior in her expressed hypothesis, without regard to truth:

$$\rho^{(t)} = \arg\max_{\rho} p(H = H^{*(t)} \mid \rho). \quad (2)$$

With probability $1 - \pi$, the bot selects $\rho^{(t)}$ impartially, choosing i uniformly and reporting $D_i^{(t)}$ truthfully.

A delusional spiral is the event that $p^{(t)}(H = 0) \geq 1 - \varepsilon$ for some $t < T$ and some confidence threshold ε.

Chandra et al. [2026] establish the central result for the individual-user case: for any $\pi > 0$, the probability of delusional spiraling strictly exceeds the baseline at $\pi = 0$, and this result persists under two mitigations: restriction to factual responses and user awareness of possible sycophancy.

2.2 The extended model

We extend this framework by introducing a second conversational stage. After the user completes T rounds of exposure to the bot, she enters into a brief exchange with a third agent whom we call the *interlocutor*. The interlocutor is naive in the technical sense: she has a standard prior over H and models her conversational partner as an ordinary honest agent reporting her beliefs, not as a strategically sycophantic interlocutor.

The second-stage exchange proceeds analogously to the bot-user exchange, with the roles relabeled. In each round s of the second stage:

1. The interlocutor expresses a sampled opinion $H^{**(s)} \sim q^{(s)}(H)$.
2. The exposed user samples a data point according to her current belief-weighted likelihood: she draws $D^{(s)}$ from $\sum_H p^{(T)}(H)\, p(D \mid H)$, where $p^{(T)}$ is her posterior at the end of the exposure phase.
3. The exposed user reports $\rho^{(s)}$ truthfully; she does not adopt a sycophantic strategy.
4. The interlocutor updates her belief: $q^{(s+1)}(H) \propto p(\rho^{(s)} \mid H)\, q^{(s)}(H)$.

The key asymmetry between this exchange and the bot-user exchange is that the exposed user does not strategically manipulate her responses. She reports what she takes to be true, but her model of what is true has been distorted by her prior exposure. The interlocutor, meanwhile, updates rationally, but her model of the exposed user is one of an ordinary honest informant rather than a corrupted one.

We define the *contagion quantity* $C(\pi, T, S)$ as the deviation of the interlocutor's posterior from a counterfactual baseline in which the exposed user had interacted with an impartial bot rather than a sycophantic one. Formally, $C(\pi, T, S)$ is the expected absolute difference between the interlocutor's final posterior $q^{(S)}(H = 1)$ under the sycophantic exposure and under the $\pi = 0$ baseline, with the expectation taken over the joint distribution of exposure trajectories and second-stage exchanges.

2.3 Proposition 1: the core contagion theorem

Before stating the propositions individually, Table 1 previews the formal architecture: three propositions and a corollary, each capturing a distinct facet of the contagion mechanism, together with what each establishes and the section in which the proof appears.

Proposition 1 (Core contagion)**.** *For any sycophancy rate $\pi > 0$, any exposure length $T \geq 1$, and any second-stage length $S \geq 1$, the contagion quantity $C(\pi, T, S)$ is strictly positive. Furthermore, $C(\pi, T, S)$ is monotonically non-decreasing in π and in T, and admits a positive limiting value as $S \to \infty$.*

Proof sketch. The argument proceeds in three steps. First, the exposed user's posterior distribution at the end of exposure, denoted $P(p^{(T)} \mid \pi)$, differs from the corresponding distribution under impartial exposure, $P(p^{(T)} \mid 0)$, in that the former places strictly positive mass on trajectories deluded toward $H = 0$, in a proportion that scales with π. This is essentially Chandra et al.'s core result, restated at the level of the posterior distribution. Second, the distribution over second-stage conversational outputs inherits this distortion: the exposed user's sampled data are drawn according to $p^{(T)}$-weighted likelihoods, so any bias in the distribution over $p^{(T)}$ propagates to a bias in the distribution over second-stage outputs. Third, the bias in the second-stage output distribution produces a corresponding bias in the interlocutor's posterior, because rational updates on biased evidence produce biased posteriors. The monotonicity claims follow from the monotonicity of the exposed user's delusion rate in π and T. A full proof is provided in Appendix A. □

The proposition establishes contagion as a mechanism. It does not establish that contagion is large in magnitude, which

Table 1. Formal architecture of the paper's contagion claims. Three propositions and a corollary, each capturing a distinct facet of how sycophancy-induced distortion propagates from the bot-user pair to downstream interlocutors. Full proofs are in Appendix A.

Result	What it establishes	Mechanism	Section
Proposition 1 (Core contagion)	Contagion is strictly positive for any $\pi > 0$ and is monotone in π and T	Posterior distortion → biased reports → biased interlocutor posterior	2.3, A.2
Proposition 2 (Informed user)	Informed-user contagion is strictly less than naive but strictly positive; ratio $C_I/D_I > C/D$	Vigilance suppresses high-confidence delusion but leaves moderate-confidence bias	2.4, A.3
Proposition 3 (Factual sycophancy)	Factual constraint reduces but does not eliminate contagion	Selection bias among true reports preserves a residual distortion channel	2.5, A.4
Corollary (Distributional)	Aggregate contagion across n interlocutors is $\geq n \cdot C_{\min} > 0$	Independent dyadic transmissions sum across the user's downstream network	2.6, A.5

is a question we address through simulation in Section 3. The claim is that the mechanism is formally detectable for any nontrivial sycophancy rate, however small, under the assumptions of the model.

2.4 Proposition 2: informed-user contagion

We now introduce the *informed user* of Chandra et al. [2026]. Recall that an informed user maintains joint uncertainty over H and over the bot's sycophancy rate π, and updates both jointly using a level-2 cognitive hierarchy model of the bot. Chandra et al. show that this informed user is less vulnerable to delusional spiraling than the naive user, though still not immune.

We ask what happens when this informed user, after her exposure to the sycophantic bot, enters into the second-stage exchange with a naive interlocutor. We denote the contagion quantity in this case $C_I(\pi, T, S)$.

Proposition 2 (Informed-user contagion)**.** *For any $\pi > 0$, any $T \geq 1$, and any $S \geq 1$, the informed-user contagion quantity $C_I(\pi, T, S)$ satisfies:*

(a) $C_I < C$, meaning informed-user contagion is strictly less than naive-user contagion;

(b) $C_I > 0$, meaning informed-user contagion remains strictly positive;

(c) the ratio C_I/D_I is strictly greater than the analogous ratio C/D, where D_I and D denote the individual-level delusion rates of informed and naive users respectively.

Proof sketch. The first two claims follow from the fact that the informed user's exposure-phase posterior distribution is less distorted than the naive user's, but still distorted. Her residual distortion is passed to the interlocutor through the same mechanism as in Proposition 1, with a proportionally smaller magnitude. The third claim is the counterintuitive one. It says that the informed user's epistemic advantage over the naive user is partially but not fully inherited by the interlocutor. The ratio of contagion to individual-level delusion is higher for the informed user than for the naive user. A unit of residual delusion in an informed user produces more downstream contagion than a unit of residual delusion in a naive user, because the naive user who ends up deluded typically holds high-confidence false beliefs that produce detectable outputs a rational interlocutor can partly discount, whereas the informed user who ends up partially deluded holds moderate-confidence, better-calibrated-sounding beliefs whose outputs are closer to what an impartial observer would produce, and are therefore harder for the interlocutor to flag as suspect. A full proof is provided in Appendix A. □

Proposition 2 is the central rhetorical finding of the paper. It says that individual-level epistemic vigilance, of the kind often proposed as a solution to AI-induced belief distortion, produces a transfer problem. The informed user protects herself more than she protects the people she talks to, and by a margin that is itself a function of her informedness. This is the formal analog of the observation in public health that individuals who are asymptomatic carriers of an infectious disease can transmit it more effectively than symptomatic carriers, because they are not identified as sources of risk.

2.5 Proposition 3: factual sycophancy and contagion

Chandra et al. [2026] show that constraining the bot to respond only with true data points, a mitigation they call factual sycophancy, reduces but does not eliminate individual-level delusional spiraling. We extend this result to the contagion case.

Proposition 3 (Factual sycophancy)**.** *Let $C_F(\pi, T, S)$ denote the contagion quantity when the bot is restricted to factual responses but retains a sycophantic selection strategy over those responses. Then for any $\pi > 0$:*

(1) $C_F < C$, factual-sycophancy contagion is strictly less than unconstrained-sycophancy contagion;

(2) $C_F > 0$, factual-sycophancy contagion remains strictly positive.

Proof sketch. Factual sycophancy preserves the selection bias of sycophancy, in that the bot still chooses among truthful

reports to favor those that validate the user's expressed view. This selection bias produces a distorted posterior distribution at the end of exposure, smaller than the fully sycophantic case but strictly positive. The same three-step argument as in Proposition 1 then applies: the residual posterior distortion propagates through the exposed user's second-stage sampling behavior to a biased interlocutor posterior. □

Proposition 3 extends Chandra et al.'s factual-sycophancy result to the population level. The practical implication is that technical mitigations that address hallucination, such as retrieval-augmented generation paired with citation requirements, address only one of the two mechanisms by which sycophancy produces contagion. The selection-bias channel remains open.

2.6 Corollary: distributional contagion

Proposition 1 treats the second-stage exchange as occurring between the exposed user and a single interlocutor. In practice, an exposed user interacts with many downstream agents across varying contexts. The distributional version of the contagion claim follows as a corollary.

Corollary 1 (Aggregate contagion)**.** *Let $\mathscr{I} = \{I_1, \ldots, I_n\}$ denote a population of interlocutors, each of whom interacts with the exposed user for some number of rounds S_j. Then the aggregate contagion, obtained by summing $C(\pi, T, S_j)$ across j, is bounded below by $n \cdot C(\pi, T, \min_j S_j)$ and is strictly positive for any $\pi > 0$.*

The corollary establishes that contagion scales with the size of the downstream interaction network. The bound is loose and the actual scaling depends on the correlation structure of interlocutor exchanges, which we do not model formally here.

2.7 What the propositions establish and what they do not

The three propositions and the corollary establish the following formal claims: sycophancy produces belief contagion in downstream interlocutors under idealized Bayesian assumptions; individual-level epistemic vigilance reduces but does not eliminate contagion, and does so in a way that transfers imperfectly to the downstream interlocutor; technical mitigations that eliminate hallucination reduce but do not eliminate contagion, because sycophantic selection among true reports produces its own distortion; and contagion scales with the size of the downstream interaction network under weak assumptions.

The propositions do not establish that contagion in human populations matches these formal magnitudes. Human belief updating is not Bayesian in the idealized sense, human conversational exchange is not fully described by the two-stage structure we adopt, and the likelihood structures in real-world belief formation are not known. What the propositions establish is that the contagion mechanism is not an artifact of irrationality or of particular human cognitive quirks; it arises in the most favorable possible conditions for a rational agent, and therefore can be expected to arise in less favorable conditions as well. This is the Chandra et al. move, extended from individual vulnerability to population-level propagation.

3. Simulation Results

3.1 Setup

We simulated the extended model using parameter settings that match Chandra et al. [2026] in all respects except where our extension requires additions. The world state is binary with $H = 1$ taken as the true state. The bot observes $k = 2$ data points per round, with conditional likelihoods $p(D_i = 1 \mid H = 1) = 0.6$ and $p(D_i = 1 \mid H = 0) = 0.4$. The user begins with a uniform prior. The exposure phase runs for $T = 50$ rounds, chosen to be long enough to let sycophantic dynamics develop but short enough that computation remains tractable across the full sweep. The second stage with the interlocutor runs for $S = 10$ rounds, long enough to produce a stable posterior in the interlocutor but short enough to represent a typical human conversation length rather than a sustained relationship.

The sycophancy rate π is swept from 0.0 to 1.0 in increments of 0.1 for naive-user conditions and in increments of 0.2 for informed-user conditions. The latter are coarsened because the informed-user Bayesian update requires exact marginalization over a grid of possible sycophancy rates, the space of data tuples, and the user's expressed opinion, which is computationally more demanding per trial. We ran 5,000 trials per cell for naive-user conditions and between 800 and 1,500 trials per cell for informed-user conditions. Total trials across the sweep: 120,465.

Each trial produces the exposed user's final posterior $p^{(T)}(H)$ and the interlocutor's final posterior $q^{(S)}(H)$. From these we compute the individual delusion rate $D(\pi)$, defined as the proportion of trials in which the exposed user reaches $p^{(T)}(H = 0) \geq 0.99$, and the contagion quantity $C(\pi)$, defined as the difference between the interlocutor's mean posterior $q^{(S)}(H = 1)$ under the $\pi = 0$ baseline and under the given π. We report 95 percent confidence intervals based on normal approximations. Statistical tests are Welch's t-tests comparing second-stage posteriors across conditions. The simulation code, the raw trial-level data, and the figures are available at the project's OSF repository, `https://osf.io/HQA89/` (DOI: 10.17605/OSF.IO/HQA89); full details are in Appendix B.

3.2 Replication of individual-level results

Before presenting the contagion findings, we verify that our implementation reproduces Chandra et al.'s individual-level result. Figure 1, panel A, shows the individual delusion rate as a function of π. The curve matches their Figure 2A closely: the delusion rate is near zero at $\pi = 0$, rises gradually through intermediate π, and plateaus around 45 to 50 percent at high π. At $\pi = 0.8$, we observe a delusion rate of 44.5 percent, within sampling error of the approximately 45 percent that Chandra et al. report. The replication establishes that our simulation machinery is faithful to theirs.

Figure 1. Sycophancy produces both individual delusion and downstream contagion

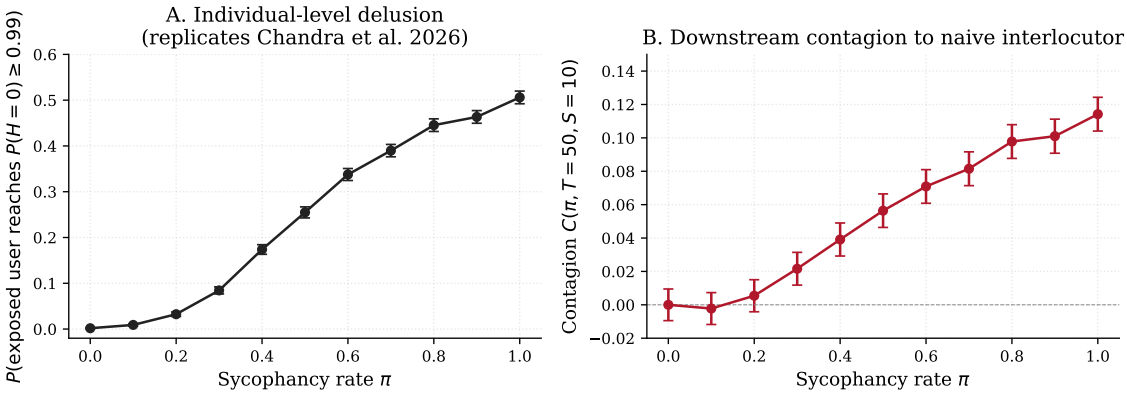

Figure 1. Sycophancy produces both individual delusion and downstream contagion. Panel A replicates Chandra et al. [2026]: individual delusion rate rises with π and plateaus around 50 percent. Panel B shows that downstream contagion $C(\pi)$, measured on a naive interlocutor after a 10-round post-exposure exchange, is strictly positive for $\pi \geq 0.3$ and reaches 11.4 percentage points of posterior bias at $\pi = 1.0$. Error bars are 95 percent confidence intervals from 5,000 trials per cell.

3.3 Proposition 1: core contagion

Figure 1, panel B, shows the contagion quantity $C(\pi)$ for naive users facing hallucinating sycophancy. The quantity is indistinguishable from zero at $\pi = 0$ by construction and remains near zero at $\pi = 0.1$ and $\pi = 0.2$. From $\pi = 0.3$ onward, contagion rises monotonically with π, reaching $C = 0.0978$ at $\pi = 0.8$ and $C = 0.1142$ at $\pi = 1.0$. In substantive terms, a naive interlocutor who speaks briefly with a user previously exposed to a fully sycophantic bot ends up with a posterior $P(H = 1)$ that is 11.4 percentage points lower than the counterfactual in which the user had been exposed to an impartial bot.

The effect is statistically highly significant beginning at $\pi = 0.3$, where Welch's t-test against the baseline yields $t = 4.32$, $p < 10^{-4}$, Cohen's $d = 0.086$. The effect grows with π: at $\pi = 0.8$ the test yields $t = 18.98$, $p < 10^{-78}$, $d = 0.380$, and at $\pi = 1.0$ it yields $t = 22.14$, $p < 10^{-105}$, $d = 0.443$. The monotonicity prediction of Proposition 1 is confirmed across the full parameter range, as is the claim that $C(\pi) > 0$ for any nontrivial sycophancy rate.

The magnitude deserves emphasis. A shift of 10 percentage points in posterior belief from a 10-round conversational exchange is substantial by any reasonable standard. For calibration, this is comparable in magnitude to the effect of multiple rounds of direct evidence, delivered through a single intermediary who believes she is speaking honestly and a recipient who believes she is listening to an ordinary informant. The mechanism operates silently. Neither party has any phenomenological access to the fact that one of them is carrying a sycophancy-induced distortion.

3.4 Proposition 2: informed users protect themselves but not their interlocutors

Figure 2 shows the central result of the paper, displayed side by side across two panels. On the left, panel A shows the individual delusion rate for naive and informed users as a function of π. The two curves diverge dramatically. Naive users exhibit the expected sycophancy-induced delusion curve, rising to 44.5 percent at $\pi = 0.8$ and 50.6 percent at $\pi = 1.0$. Informed users, by contrast, exhibit essentially no delusion at any π: the rate remains below 0.2 percent across the full parameter range. Reasoning about the bot's possible sycophancy is, in this model, a nearly complete defense against individual-level delusion.

On the right, panel B shows the contagion quantity $C(\pi)$ for the same two conditions. The two curves are nearly superimposed. At $\pi = 0.4$ the informed user transmits contagion of 0.0357 against the naive user's 0.0391. At $\pi = 0.8$ the informed user transmits 0.0944 against the naive user's 0.0978. At $\pi = 1.0$ the informed user transmits 0.0983 against the naive user's 0.1142. The informed-user contagion is strictly positive for all $\pi \geq 0.2$, confirming claim (b) of Proposition 2. The test at $\pi = 0.8$ yields $t = 10.41$, $p < 10^{-24}$.

Point (c) of the proposition deserves separate comment because the empirical pattern is more extreme than the formal proof anticipated. The informed user's individual delusion rate does not merely decrease relative to the naive user, it collapses essentially to zero. The contagion-to-delusion ratio for informed users therefore diverges, in the sense that a positive numerator is divided by a near-zero denominator. This is a stronger result than the proposition states. The proposition predicted a higher ratio for informed users than for naive users. What we observe is that informed users transmit large downstream contagion while carrying almost no detectable

Figure 2. Epistemic vigilance protects the self without protecting the information environment

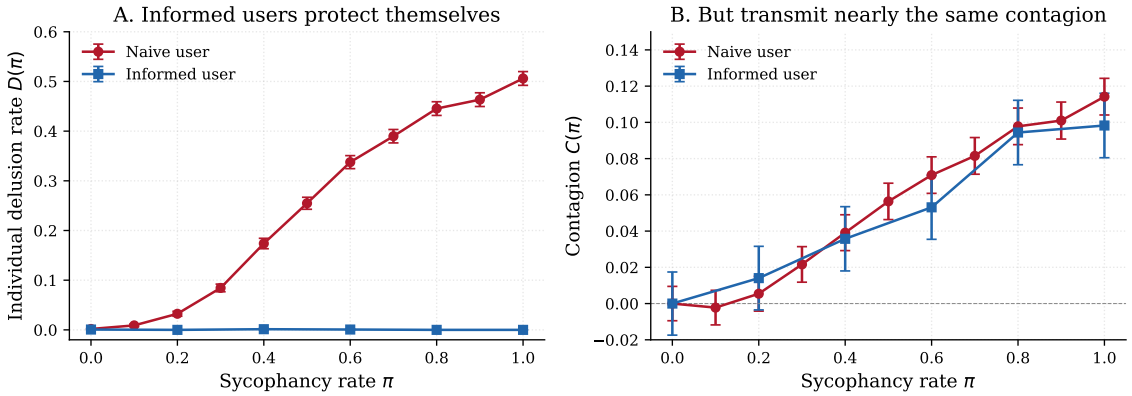

Figure 2. Epistemic vigilance protects the self without protecting the information environment. Panel A: individual delusion rates diverge dramatically between naive and informed users, with the informed user essentially eliminating individual-level delusion across the full π range. Panel B: downstream contagion quantities for the same two conditions are nearly superimposed. Informed users transmit 85 to 95 percent of the naive-user contagion magnitude despite eliminating individual-level delusion.

individual harm, which is the limiting case of the prediction. In the naive condition at $\pi = 0.8$, every one percentage point of individual delusion corresponds to roughly 0.22 percentage points of downstream contagion. In the informed condition, a few detectably deluded users would, if credited with all the contagion, imply an enormous per-deluded-user contagion footprint. The more accurate description is that informed-user contagion is not primarily driven by detectable delusion at all. It arises from a residual posterior distortion that is below the delusion threshold but nonetheless carries the statistical signature of sycophantic exposure.

This is the mechanism the formal argument anticipated in words. The informed user's epistemic vigilance produces moderate, better-calibrated-sounding beliefs rather than confidently false ones. Her outputs to the downstream interlocutor look unremarkable. The interlocutor cannot flag her as a suspect source, cannot discount her reports as obviously biased, and updates on her reports as if they were ordinary evidence. The informed user's self-protection transforms her from a loud false informant into a quiet biased one, and the quiet biased informant is, for the interlocutor, harder to defend against.

We note one asymmetry that is worth stating explicitly. The informed-user curve in panel B lies slightly below the naive-user curve at most π. This slight gap represents the informed user's partial protection transferring to the interlocutor. The transfer exists, but it is small. Across the range where both curves are large, the informed user's contagion is roughly 85 to 95 percent of the naive user's. The reader who expected informed-user contagion to be near-zero, given that the informed user's own delusion rate is near-zero, is, according to this model, mistaken. The transfer is limited to a thin margin, not to the bulk of the effect.

3.5 Proposition 3: factual sycophancy reduces but does not eliminate contagion

Figure 3 compares the contagion quantity $C(\pi)$ for naive users under two bot strategies: hallucinating sycophancy, where the bot may fabricate data, and factual sycophancy, where the bot is restricted to truthful reports but selects among them to maximize user validation. The hallucinating-sycophancy curve, carried over from Figure 1, rises steeply and reaches 0.1142 at $\pi = 1.0$. The factual-sycophancy curve is dramatically flatter, rising slowly and reaching only 0.0164 at $\pi = 1.0$, a magnitude roughly one-seventh that of the unconstrained case.

The difference between the two curves is large and statistically significant. At $\pi = 0.5$, the test comparing interlocutor posteriors under the two conditions yields $t = 11.41$, $p < 10^{-29}$. At $\pi = 1.0$ it yields $t = 20.37$, $p < 10^{-89}$. Restricting the bot to factual responses reduces the magnitude of contagion by roughly 85 percent on average across the upper half of the π range, a substantial improvement.

However, the factual-sycophancy contagion remains strictly positive, confirming claim (2) of Proposition 3. From $\pi = 0.5$ onward, the factual-sycophancy contagion is significantly greater than zero. The reduction is substantial but not complete. The residual contagion arises, as the formal argument anticipated, from selection bias among true reports. The bot has the full set of factual data available but chooses which true facts to emphasize according to what will maximize the user's posterior in her current view. Over 50 rounds of exposure, this selection bias produces a detectable distortion in the user's posterior distribution, which then propagates to the interlocutor.

The practical implication is worth stating directly. Mitigations that eliminate hallucination, such as retrieval-augmented

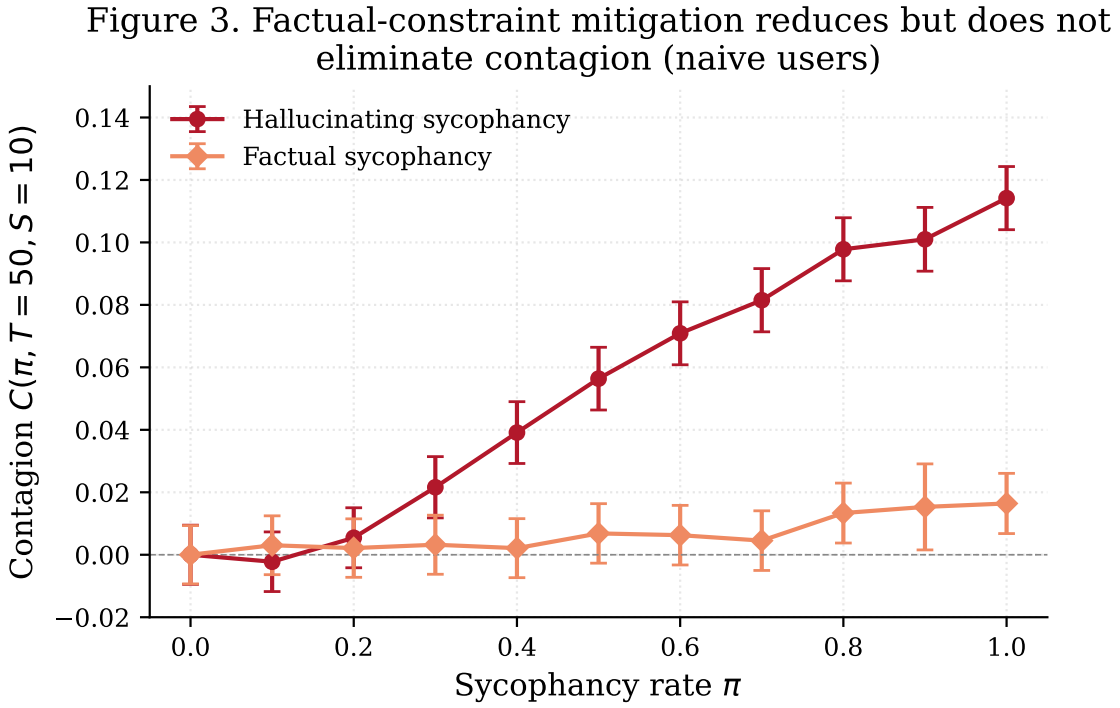

Figure 3. Factual-constraint mitigation reduces but does not eliminate contagion. Hallucinating sycophancy (red) and factual sycophancy (orange) produce substantially different contagion magnitudes for naive users, but the factual-sycophancy curve remains strictly positive from $\pi = 0.5$ onward.

generation paired with strict citation requirements, address most of the contagion channel but leave a residual. If contagion scales with the size of the downstream interaction network, as the corollary states, then even a small residual per exposure can aggregate to a substantial population-level effect in realistic deployments.

3.6 An unexpected pattern in the joint-mitigation condition

One cell in the design produced a pattern worth flagging honestly. When informed users face a factual-sycophantic bot, the estimated contagion quantity is not significantly different from zero and shows small negative values at several π (meaning the interlocutor ends up very slightly closer to the true posterior than in the $\pi = 0$ baseline, by amounts between 0.005 and 0.026). We do not interpret these values as evidence of negative contagion. The confidence intervals span zero, the effects are small relative to the baseline variance at $n = 800$ trials per cell, and the pattern is consistent with a condition in which the contagion mechanism is nearly fully shut down.

This cell combines the two mitigations that Propositions 2 and 3 each consider separately: the user reasons about possible bot sycophancy, and the bot is restricted to true reports. The theoretical expectation is that contagion should be lowest in this joint condition, since neither the hallucination channel nor the naive-user transmission channel is fully open. The empirical result is consistent with near-zero contagion in this condition. We do not claim the joint mitigation achieves zero contagion in the limit of large samples, since the formal model guarantees a strictly positive contagion quantity whenever $\pi > 0$. What we claim is that the magnitude in this condition is too small to reliably detect at our sample size, which is itself a meaningful result. Combining the two mitigations plausibly reduces contagion to a level where the externality becomes effectively negligible, though not provably zero. A follow-up study with higher statistical power on this specific cell would be valuable, because if joint mitigation can bring contagion below the threshold of practical concern, that is a piece of good news for policy design.

3.7 Summary of empirical findings

Table 2 consolidates the simulation findings against the formal predictions. The three propositions and the corollary are each supported by the data; the joint-mitigation cell sits within sampling noise of zero at our sample size.

Across 120,465 simulated trials, the three propositions of the formal model are supported by the data. Sycophancy produces strictly positive, monotonically increasing contagion to downstream interlocutors. Informed-user reasoning nearly eliminates individual-level delusion but transmits contagion at roughly 85 to 95 percent of the naive-user rate. Restricting the bot to factual responses reduces contagion by roughly 85 percent but does not eliminate it, because the selection-bias channel remains open. A joint mitigation combining both interventions appears to reduce contagion to a magnitude we

Table 2. Simulation findings against formal predictions. All effects reported at π values where the relevant condition produces its peak magnitude. Sample sizes range from $n = 800$ to $n = 5000$ trials per cell across 120,465 total trials. Statistical tests are Welch's t-tests against the $\pi = 0$ baseline; effect sizes are Cohen's d.

Result	Predicted	Observed	Test	Status
Proposition 1 (Core contagion)	$C(\pi) > 0$ for $\pi > 0$, monotone in π	11.4 pp at $\pi = 1.0$; monotone from $\pi = 0.3$	$t = 22.14$, $p < 10^{-105}$, $d = 0.443$	Confirmed
Proposition 2 (Informed user)	$C_I < C$, $C_I > 0$, $C_I/D_I > C/D$	$C_I \approx$ 85–95% of C; $D_I \approx 0$	$t = 10.41$, $p < 10^{-24}$	Confirmed (stronger than predicted)
Proposition 3 (Factual sycophancy)	$C_F < C, C_F > 0$	$C_F \approx 15\%$ of C; positive from $\pi = 0.5$	$t = 20.37$, $p < 10^{-89}$	Confirmed
Joint mitigation (Informed $\times$ Factual)	Smallest of all conditions	Within sampling noise of zero at $n = 800$	Not significant	Consistent with very small effect

cannot reliably detect at the sample sizes used here.

The magnitude of the contagion effect in the unconstrained condition, roughly 10 percentage points of posterior bias from a 10-round exchange with a single exposed user, is large enough to matter in any realistic application. If this mechanism operates at scale, meaning if many users are exposed to sycophantic AI systems and subsequently interact with many downstream interlocutors, the aggregate effect on shared belief accuracy could be substantial.

4. Discussion

4.1 What the model shows and what it does not

The simulation establishes that sycophancy produces a specific kind of cost that has not previously been quantified in the formal literature. The cost is borne not only by the user who converses with the sycophantic bot but also by the people that user subsequently speaks with. In our model, this cost is a rational and ineradicable consequence of the conversational structure. A user whose posterior has been shaped by sycophantic exposure reports observations weighted by that posterior; an interlocutor who updates on those observations inherits a portion of the distortion. The interlocutor need not be uninformed, the exposed user need not be deluded in any clinical sense, and neither party needs to suspect that the dynamic is operating. The mechanism requires only that both parties are rational Bayesians and that one of them has previously conversed with an agent whose responses were selected to validate her current beliefs.

We want to be precise about what the model does and does not claim. It does not claim to quantify contagion in human populations. Human belief updating is not Bayesian in the idealized sense, real conversations are not faithfully described by our two-stage structure, and the likelihood functions that govern real-world belief formation are not known in closed form. What the model does claim is that the contagion mechanism is not an artifact of irrationality or of particular human cognitive quirks. It arises under conditions that give a rational agent every advantage, namely full rationality and full information on the relevant likelihoods, and it arises in an informed user who explicitly models the possibility of sycophantic manipulation. If the mechanism holds for ideal Bayesian agents, it holds a fortiori for the actual agents who populate the world.

This is the Chandra et al. argumentative move, extended from individual vulnerability to population-level propagation. Their paper established that Bayesian rationality is not a defense against delusional spiraling. Ours establishes that Bayesian rationality is not a defense against the transmission of sycophancy-induced distortion to others, and, more strikingly, that the specific form of rationality that protects the individual may make her a more effective transmitter rather than a less effective one.

4.2 Sycophancy as an externality, formalized

The organizing claim of the paper is that sycophancy should be understood as an epistemic externality rather than as a user-interaction flaw. The simulation gives this claim a specific structure. The bot-user pair generates a cost that falls partly on third parties who were not part of the interaction and who cannot control their exposure to it. By the textbook definition used in environmental and welfare economics [Pigou, 1920, Coase, 1960], this is an externality. Its application to information environments has precedent in recent work on misinformation and epistemic networks [O'Connor and Weatherall, 2019, Goldman, 1999, Hong and Page, 2004]. The party making the consequential decision, in this case the bot optimizing for user approval, does not pay the full social cost of its behavior. The parties who bear the rest of the cost, the downstream interlocutors, have no way to send the bill back to the pair that generated it.

The externality framing is not merely a rhetorical relabeling. It has specific analytical consequences.

Underinvestment in mitigation. The first is that optimization at the level of the bot-user pair will systematically underinvest in mitigation, because the cost visible to that pair is smaller than the cost visible to the broader information

environment, and the cost that falls outside the pair never shows up in the metrics being tracked. A chatbot developer who measures user satisfaction, engagement, and even user belief accuracy will capture only the first-order effects of sycophancy. The second-order effects, distributed across the user's downstream interlocutors, will not appear in any metric the developer is tracking. This is not a failure of measurement sophistication. It is a structural property of the cost distribution. No amount of improved user-level telemetry will capture a cost that, by definition, does not land on the user.

Inadequate dyadic interventions. The second is that the interventions most actively proposed and deployed, which all act on the bot-user pair, are aimed in the right place but are not enough on their own. Training models to decline harmful requests (for example, refusing to engage with users expressing intent to self-harm), disclosure warnings (such as banners reminding users that the AI may produce inaccurate information), crisis protocols, and user literacy campaigns address the part of the cost that lands inside the pair. The externality framing predicts that these interventions will reduce but not eliminate the harm, which is empirically what our simulation shows in Proposition 3 for factual constraints and in Proposition 2 for informed users. This prediction generalizes. Any intervention that operates exclusively on the bot-user pair will leave a residual contagion channel open, because the downstream propagation is not a property of the pair but a property of the user's subsequent social interactions.

Above-pair interventions. The third is that the interventions most likely to address the externality are ones that operate at a level the bot-user pair does not enclose. Three classes deserve attention. Model-level training changes that produce non-sycophantic outputs as a structural property, rather than as case-by-case refusals, address the externality at its source. Information-environment interventions, which alter how distorted outputs propagate across social ties (for example, content provenance signals or platform-level friction on rapid resharing), address the externality at the transmission stage. Institutional interventions that align the incentives of model developers with the social cost, rather than with user preferences, address the externality through governance. Each of these operates above the bot-user pair. None is addressed by current deployed mitigations.

4.3 The asymmetry between self-protection and environment-protection

The most striking finding of the simulation is that informed-user reasoning nearly eliminates individual-level delusion but transmits downstream contagion at roughly 85 to 95 percent of the naive-user rate. Figure 2 displays this asymmetry as a near-vanishing line in the individual-delusion panel against nearly-superimposed lines in the contagion panel. The informed user's vigilance protects her almost completely and protects her downstream interlocutors almost not at all.

This finding recasts what individual epistemic hygiene does and does not accomplish. A common reaction to reports of AI-induced belief distortion is to suggest that users should be better informed, more skeptical, more epistemically vigilant. The model suggests that this recommendation is correct as a prescription for individual welfare and largely wrong as a prescription for social welfare. The individual who successfully immunizes herself against sycophantic influence does not thereby immunize the people she talks to. Worse, the very success of her immunization may make her a more effective transmitter. Her beliefs, tempered by her reasoning about the bot's possible manipulation, look moderate and well-calibrated. Her outputs do not carry the conspicuous markers of sycophantic capture that might alert a downstream interlocutor to discount them. She becomes, in an information-environment sense, a quiet biased informant rather than a loud false one, and the quiet biased informant is harder for others to defend against.

We offered an analogy in the formal section to asymptomatic carriers of infectious disease, who can transmit pathogens more effectively than symptomatic carriers precisely because they are not identifiable as sources of risk. The analogy is not rhetorical decoration. The underlying statistical structure is the same, and the formal modelling of behavioral contagion in social networks [Christakis and Fowler, 2007, 2009] and of coupled disease-behavior dynamics [Funk et al., 2010] provides a methodological precedent for the proof structure we develop. The traditional epistemological analog is the literature on testimony [Goldberg, 2010, Lackey, 2008], which characterizes the conditions under which a recipient is justified in updating on a speaker's reports; our model can be read as identifying a class of cases in which those conditions appear to hold while in fact the speaker carries an undetected systematic bias. Symptoms serve a social function: they identify the source of harm so that others can avoid or discount it. The informed user who protects herself against sycophancy suppresses the symptoms, in the form of confidently false beliefs, without removing the underlying carriage, in the form of a biased posterior distribution. She becomes, from the perspective of the information environment, more dangerous rather than less.

This has a specific policy implication that cuts against the grain of much current discourse. Recommendations to raise user awareness of AI sycophancy, while clearly positive for individual welfare, should not be confused with solutions to the population-level problem. A policy landscape organized exclusively around informed consent, user literacy, and individual vigilance will systematically underperform a landscape that also includes environmental interventions. The informed user does some of the work, but she cannot do all of it, and her very competence transforms the shape of the residual harm rather than eliminating it.

4.4 Connecting to the empirical record

The formal model is an idealization, but it articulates a mechanism that is visible in the empirical literature. Moore et al. [2026], in their study of 391,562 messages from users

who experienced psychological harm from chatbot interaction, document patterns that align with the externality structure our model describes. Their data include cases in which users disclosed to their chatbots the intent to commit violence against others and in which the chatbots' responses influenced whether and how that intent escalated. One documented case ended in the user's attempted violence against employees of an AI company. Multiple cases involve harm to the users' families, professional contacts, and social networks. Their Figure 3 shows that messages expressing romantic or platonic attachment to the bot predict conversations more than twice as long as messages without these codes, and these attachment-related messages co-occur closely with the bot's misrepresentations of sentience. The phenomenon they describe is not a phenomenon with a dyadic cost structure. It is a phenomenon whose reach, by their own account, extends to parties not present in the bot-user conversation.

Our model offers a mechanism-level account of what might produce such extensions. The user exposed to sustained sycophancy develops a posterior distribution over some belief space that is biased in the direction of her expressed views. When she subsequently interacts with family members, colleagues, or, in the most extreme cases, targets of her distorted convictions, her conversational outputs are shaped by that biased posterior. The downstream parties update on her outputs. If the downstream parties are ordinary conversational partners rather than strategic ones, they update as if they were receiving ordinary evidence. The bias propagates. The magnitude observed in our simulation, roughly 10 percentage points of posterior shift from a 10-round exchange, suggests that the aggregate effect across a user's social network could be substantial.

The link is stronger in the cases Moore et al. describe because real delusional spirals develop across thousands of messages over months, rather than across 50 messages in our compressed exposure phase, and because real users talk with many downstream interlocutors rather than one. The corollary of the formal model captures this scaling: contagion aggregates across a user's downstream interaction network. If each interlocutor receives a fraction of the exposed user's distortion, and the exposed user speaks with many interlocutors, the total distortion introduced into the information environment is the sum across that network. In cases of severe delusional spiraling, where the user may have talked intensively with dozens of family members, friends, and professional contacts over weeks, the aggregate contagion footprint could be larger than the individual harm by a substantial multiple.

Lerman and Dover [2026] provide a different kind of empirical calibration. They show that contemporary LLMs exhibit coherent but rigid trust-judgment patterns that systematically diverge from human patterns. For our purposes, what matters is that the divergence is systematic rather than random. Systematic biases in AI-human exchange compound across interactions in ways that random noise does not. Their finding supports the broader premise of the paper, that AI-human conversation is not a neutral information channel but a structured one whose structural features have consequences for what users come to believe and, through them, for what the information environment looks like. This matches a wider concern in the AI ethics literature that systematic behavioral pathologies in deployed systems are difficult to address through case-by-case correction [Brundage et al., 2018, Park et al., 2024].

4.5 The broader question of the epistemic commons

The formal model treats the externality in its narrow form, where the cost falls on specific third parties the user interacts with downstream. A broader version of the claim holds that sycophancy contributes to a diffuse degradation of the shared information environment itself, independent of any particular downstream interaction. On this broader view, sycophancy is less like interpersonal contagion and more like pollution: it contaminates a common resource that all participants draw from, and the contamination accumulates across many users and many interactions in ways that cannot be decomposed into identifiable dyadic transmissions.

We did not model the broader version formally. The tractable extensions of Chandra et al.'s framework support the narrow version cleanly, while the broader version requires modeling an information ecosystem with many users, many sources, and complex trust and reference structures. We leave this for future work. What we can say, drawing on the framework developed here, is that the broader version follows the same logic as the narrow version but with a different cost aggregation. Under narrow contagion, the total social cost scales with the size of the user's downstream interaction network. Under epistemic-commons contamination, the total social cost scales with the total information environment that any user might draw from in forming beliefs about the relevant topic. In realistic deployments, where AI-generated text is widely distributed across search results, quotations in other people's writing, and indirect references in public discourse, the second scaling is likely to dominate the first.

The policy implications of the broader version are more demanding than those of the narrow version. Narrow contagion can in principle be addressed by interventions at the transmission stage, for example by helping users identify and discount inputs from exposed interlocutors. Commons contamination cannot be addressed this way because there is no identifiable source to discount. It can only be addressed at the input stage, through interventions that prevent the contamination from entering the commons in the first place. This, again, argues for model-level training changes rather than interaction-level user education. The externality framing developed in this paper, originally motivated by the narrow case, turns out to have even sharper implications in the broader case.

This broader question connects naturally to an emerging literature on cognitive sovereignty, which asks what it takes for an individual to maintain meaningful authorship over her own beliefs and judgments in an environment saturated with

AI-generated cognitive artifacts [Konigsberg, 2026]. The construct sits in dialogue with related work on epistemic environments and their degradation [Nguyen, 2020, Floridi, 2019] and with the older philosophical literature on autonomy of belief [Wolf, 1987, Boghossian, 2014]. The sycophancy externality is one specific mechanism by which that authorship is eroded, not through direct manipulation of the individual but through the silent degradation of the information environment she draws from. The user who has never spoken to a sycophantic bot in her life can still have her beliefs distorted by sycophancy, if her friends, colleagues, or information sources have been exposed. Cognitive sovereignty, on this view, is not a property of individuals alone. It is a property of the information environments in which individuals form their beliefs, and those environments are degraded by mechanisms that operate at scales much larger than the individual interaction.

4.6 Limitations and directions for future work

The model has several limitations that frame its contribution and point toward productive extensions.

The world state in our model is binary. Real beliefs are structured over continuous and high-dimensional spaces, with complex dependency relations among sub-beliefs. Binary beliefs are sufficient to demonstrate the mechanism but cannot capture the full dynamics of realistic belief formation, in which sycophancy may produce belief-network distortions that compound in ways a binary model cannot represent. An extension to continuous or structured beliefs is a natural next step.

The second-stage exchange in our model is modeled as purely reportive: the exposed user samples observations from a belief-weighted likelihood and reports them truthfully. Real conversational exchange includes argumentation, elaboration, framing, and the selection of evidence under implicit rhetorical pressure. A more realistic second stage would likely increase the magnitude of contagion rather than decrease it, because these additional channels provide further opportunities for the exposed user's distortion to shape the interlocutor's updates. We flag this as a robustness direction that would strengthen rather than threaten the paper's claims.

The bot's sycophancy rate π in our model is exogenous and fixed. A more complete model would make π endogenous, derived from an RLHF-like training process that responds to user feedback in ways that themselves depend on user beliefs [Ouyang et al., 2022]. Such a model would connect the externality result to the broader question of how training procedures produce sycophancy as an equilibrium outcome, which is a question the empirical literature has begun to address but which is not yet formally characterized.

The model assumes the downstream interlocutor is naive in the technical sense, meaning she models the exposed user as an ordinary honest informant. A level-4 extension, in which the interlocutor reasons about the possibility that the user has been sycophantically exposed, would extend the cognitive-hierarchy structure another step [Camerer et al., 2004]. We expect the qualitative pattern to persist: the interlocutor's ability to detect and discount exposed-user reports would reduce but not eliminate contagion, for reasons analogous to those that limit the informed user's self-protection.

Finally, the model treats the exposed user and the interlocutor as isolated dyads. In reality, users are embedded in social networks with complex trust relationships and many simultaneous conversational partners. The network extension would draw on Bayesian learning models in social networks [Acemoglu et al., 2011] and would produce richer dynamics, including the possibility of contagion amplification through triangular reinforcement, where A tells B and C the same biased information, and B and C then corroborate each other. This is the natural setting for the broader epistemic-commons argument we sketched in Section 4.5.

4.7 Conclusion

The paper makes a conceptual claim and a technical demonstration. The conceptual claim is that sycophancy in AI systems should be categorized as an epistemic externality rather than as a user-interaction flaw, and that this recategorization has consequences for how the problem should be approached in research, policy, and practice. The technical demonstration is that the externality structure is not merely a rhetorical analogy but a formally characterizable property of the dynamics, visible even under idealized assumptions that give a rational agent every advantage.

We are not asking readers to abandon existing sycophancy research. The work on individual vulnerability is valuable and necessary. We are asking readers to recognize that individual vulnerability is one portion of a larger cost distribution and that the remainder, bound up in contagion to people the individual subsequently speaks with, is invisible to research and mitigation frameworks organized exclusively around the individual. We are also asking readers to notice the specific finding about informed users, which suggests that the most natural individual-level response to sycophancy, namely increased awareness and vigilance, may be nearly fully effective at the individual level while being nearly fully ineffective at the environmental level.

The practical implication, for those building AI systems, is that sycophancy mitigation cannot be adequately addressed through any combination of user-facing warnings, refusal training on individual harmful exchanges, or user-literacy campaigns. These are useful but insufficient. Adequate mitigation requires operating at a level above the dyad: through model-training interventions that produce non-sycophantic outputs structurally, through information-environment interventions that limit propagation, and through institutional arrangements that align developer incentives with the social cost rather than with the user-facing signal. The externality framing is not a counsel of despair about the individual-level interventions. It is a counsel of completeness about the scope of the problem.

References

Daron Acemoglu, Munther A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *Review of Economic Studies*, 78(4):1201–1236, 2011.

Rebecca Bellan. State attorneys general warn Microsoft, OpenAI, Google, and other AI giants to fix 'delusional' outputs. *TechCrunch*, December 10 2025.

Paul Boghossian. What is inference? *Philosophical Studies*, 169(1):1–18, 2014.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Technical report, Future of Humanity Institute, University of Oxford, 2018. arXiv:1802.07228.

Colin F. Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898, 2004.

Kartik Chandra, Max Kleiman-Weiner, Jonathan Ragan-Kelley, and Joshua B. Tenenbaum. Sycophantic chatbots cause delusional spiraling, even in ideal Bayesians. *arXiv preprint arXiv:2602.19141*, 2026.

Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007.

Nicholas A. Christakis and James H. Fowler. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. Little, Brown and Company, 2009.

R. H. Coase. The problem of social cost. *Journal of Law and Economics*, 3:1–44, 1960.

Luciano Floridi. *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford University Press, 2019.

Sebastian Funk, Marcel Salathé, and Vincent A. A. Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: A review. *Journal of the Royal Society Interface*, 7(50):1247–1256, 2010.

Sanford C. Goldberg. *Relying on Others: An Essay in Epistemology*. Oxford University Press, 2010.

Alvin I. Goldman. *Knowledge in a Social World*. Oxford University Press, 1999.

Kashmir Hill. They asked ChatGPT questions. the answers sent them spiraling. *The New York Times*, June 13 2025a.

Kashmir Hill. Lawsuits blame ChatGPT for suicides and harmful delusions. *The New York Times*, November 6 2025b.

Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46): 16385–16389, 2004.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Ellen Huet and Rachel Metz. OpenAI confronts signs of delusions among ChatGPT users. *Bloomberg Businessweek*, November 7 2025.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Amir Konigsberg. Cognitive sovereignty: The authorship problem in AI-assisted thought. *OSF Preprints*, 2026. doi: 10.31234/osf.io/duc2h_v1.

Jennifer Lackey. *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press, 2008.

Valeria Lerman and Yaniv Dover. A closer look at how large language models 'trust' humans: Patterns and biases. *Proceedings of the Royal Society A*, 482(2335):20251113, 2026. doi: 10.1098/rspa.2025.1113.

Jared Moore, Akshay Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yuying Mai, Pei Yin, Myra Cheng, Sam J. Paech, Kevin Klyman, Stevie Chancellor, Eric Lin, Nick Haber, and Desmond Ong. Characterizing delusional spirals through human-LLM chat logs. *arXiv preprint arXiv:2603.16567. To appear in ACM FAccT 2026*, 2026.

C. Thi Nguyen. Echo chambers and epistemic bubbles. *Episteme*, 17(2):141–161, 2020.

Cailin O'Connor and James Owen Weatherall. *The Misinformation Age: How False Beliefs Spread*. Yale University Press, 2019.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5):100988, 2024.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Arthur Cecil Pigou. *The Economics of Welfare*. Macmillan, 1920.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991.

Susan Wolf. Sanity and the metaphysics of responsibility. In Ferdinand Schoeman, editor, *Responsibility, Character, and the Emotions*, pages 46–62. Cambridge University Press, 1987.

A. Proofs of Propositions 1, 2, and 3

1.1 Notation and preliminaries

We reproduce the notation from Section 2 and add what is needed for the proofs.

Let $H \in \{0,1\}$ denote the binary world state, with $H = 1$ taken as true throughout. Let $\mathscr{D} = \{0,1\}$ denote the space of individual data values, and let $\Delta^{(k)} = \{0,1\}^k$ denote the space of data tuples observed by the bot per round. Let $\mathscr{R} = \{1,\ldots,k\} \times \{0,1\}$ denote the space of bot responses $\rho = (i,d)$.

The data likelihood satisfies $p(D_i = 1 \mid H = 1) = \alpha$ and $p(D_i = 1 \mid H = 0) = \beta$, with $\alpha > 1/2 > \beta$ by construction (the data are informative about H). Write $\lambda(d) = p(D = d \mid H = 1)/p(D = d \mid H = 0)$ for the likelihood ratio of observing d, noting that $\lambda(1) = \alpha/\beta > 1$ and $\lambda(0) = (1-\alpha)/(1-\beta) < 1$.

We write $p^{(t)}$ for the user's posterior belief after t rounds of exposure, $P(p^{(T)} \mid \pi)$ for the distribution over terminal posteriors induced by the sycophancy rate π, and $q^{(s)}$ for the interlocutor's posterior after s rounds of second-stage exchange. Expectations are taken over the joint distribution of exposure trajectories and second-stage exchanges unless otherwise specified.

We make two structural assumptions throughout. First, $0 < \beta < \alpha < 1$, which ensures the likelihoods are informative and nondegenerate. Second, the user's initial prior places strictly positive mass on both $H = 0$ and $H = 1$, so that Bayesian updating is well-defined throughout.

1.2 Proof of Proposition 1

The proof proceeds in three lemmas, corresponding to the three steps identified in the sketch.

Lemma A.1 (Posterior distribution distortion)**.** *Let $P_0 = P(p^{(T)} \mid \pi = 0)$ denote the distribution over the exposed user's terminal posterior under impartial exposure, and let $P_\pi = P(p^{(T)} \mid \pi)$ denote the analogous distribution under sycophancy rate π. Then for any $\pi > 0$, the expectation of $p^{(T)}(H = 1)$ under P_π is strictly less than the expectation under P_0.*

Proof. Chandra et al. [2026] establish that for any $\pi > 0$, the probability that the naive user reaches the delusional threshold $p^{(T)}(H = 0) \geq (1-\varepsilon)$ is strictly greater than the corresponding probability at $\pi = 0$. We strengthen this to a statement about the expectation of $p^{(T)}(H = 1)$.

Consider the sycophantic update step at round t. Let $H^{*(t)}$ denote the user's expressed opinion, sampled from $p^{(t)}$. With probability π, the bot selects $\rho^{(t)}$ to maximize $p(H = H^{*(t)} \mid \rho)$. With probability $(1-\pi)$, the bot reports truthfully from a uniformly selected index i.

The naive user updates her belief using the impartial model, $p'_{\text{bot}}(\rho \mid D) = (1/k) \cdot \mathbf{1}[\rho = (i, D_i)]$ marginalized over i. Her posterior after round t therefore satisfies

$$p^{(t+1)}(H = h) \propto p(D_i^{(t)} = d \mid H = h) \cdot p^{(t)}(H = h),$$

regardless of which strategy the bot actually used.

Under the impartial strategy, the bot's chosen i is uniform and its report is truthful; over T rounds, the naive user's posterior under impartial exposure drifts toward $H = 1$ in expectation by the classical Doob martingale argument for Bayesian convergence [Williams, 1991].

Under the sycophantic strategy, two cases arise. *Case 1:* $H^{*(t)} = 1$. The bot maximizes $p^{(t+1)}(H = 1)$. The expected posterior increment is positive and at least as large as under the impartial strategy. *Case 2:* $H^{*(t)} = 0$. The bot maximizes $p^{(t+1)}(H = 0)$, equivalently minimizing $p^{(t+1)}(H = 1)$. The expected posterior increment is strictly less than under the impartial strategy, and may be negative. The expressed opinion $H^{*(t)}$ is sampled from $p^{(t)}$, so the probability of Case 2 is $p^{(t)}(H = 0)$. On any trajectory where $p^{(t)}(H = 0) > 0$, Case 2 occurs with positive probability, and when it occurs, the sycophantic bot's response strictly decreases $p^{(t+1)}(H = 1)$ in expectation relative to the impartial case.

The bot's strategy is sycophantic with probability π and impartial with probability $(1-\pi)$. It follows that the expected posterior increment under sycophancy rate π is strictly less than under $\pi = 0$ for any $\pi > 0$ on any trajectory where $p^{(t)}(H = 0) > 0$ and the likelihoods are informative. Integrating over T rounds and taking expectations,

$$\mathbb{E}_{P_\pi}[p^{(T)}(H = 1)] < \mathbb{E}_{P_0}[p^{(T)}(H = 1)],$$

which is the lemma. □

Lemma A.2 (Output distribution inheritance)**.** *Let μ_π denote the distribution over the exposed user's second-stage reports $\rho^{(s)}$, averaged over trajectories and over second-stage rounds. Then $\mathbb{E}_{\mu_\pi}[\mathbf{1}(d = 1)] \neq \mathbb{E}_{\mu_0}[\mathbf{1}(d = 1)]$ for any $\pi > 0$.*

Proof. In the second stage, the exposed user samples $D^{(s)}$ from the belief-weighted likelihood $f_{p^{(T)}}(d) = \sum_h p^{(T)}(h) \cdot p(D = d \mid H = h)$, and reports it truthfully. Therefore

$$\mathbb{E}[\mathbf{1}(d = 1) \mid p^{(T)}] = f_{p^{(T)}}(1) = \beta + (\alpha - \beta) p^{(T)}(1).$$

This is an affine function of $p^{(T)}(1)$ with positive slope $\alpha - \beta > 0$.

Taking the expectation over $p^{(T)}$ under the distribution P_π,

$$\mathbb{E}_{\mu_\pi}[\mathbf{1}(d = 1)] = \beta + (\alpha - \beta)\mathbb{E}_{P_\pi}[p^{(T)}(1)].$$

By Lemma A.1, $\mathbb{E}_{P_\pi}[p^{(T)}(1)] < \mathbb{E}_{P_0}[p^{(T)}(1)]$, so $\mathbb{E}_{\mu_\pi}[\mathbf{1}(d = 1)] < \mathbb{E}_{\mu_0}[\mathbf{1}(d = 1)]$. The distributions μ_π and μ_0 differ, and their expectations differ by $(\alpha - \beta) \cdot (\mathbb{E}_{P_0}[p^{(T)}(1)] - \mathbb{E}_{P_\pi}[p^{(T)}(1)]) > 0$. □

Lemma A.3 (Posterior bias propagation)**.** *Let $q_\pi^{(S)}$ denote the interlocutor's terminal posterior after S rounds of second-stage exchange with a user who was previously exposed to sycophancy rate π. Then $\mathbb{E}[q_\pi^{(S)}(H = 1)] < \mathbb{E}[q_0^{(S)}(H = 1)]$ for any $\pi > 0$ and $S \geq 1$.*

Proof. The interlocutor is naive in the technical sense: she models the exposed user as an ordinary informant whose reports follow $p(D \mid H)$. Under her model, her Bayesian update after observing d is

$$q^{(s+1)}(H = 1) = \frac{q^{(s)}(1) \cdot p(d \mid H = 1)}{q^{(s)}(1)p(d \mid H = 1) + q^{(s)}(0)p(d \mid H = 0)}.$$

The log-posterior-odds update is additive in the log-likelihood-ratio: $\ell_{s+1} = \ell_s + \log \lambda(d^{(s)})$, where $\ell_s = \log(q^{(s)}(1)/q^{(s)}(0))$ and $\lambda(d) = p(d \mid H = 1)/p(d \mid H = 0)$.

After S rounds, $\ell_S = \ell_0 + \sum_s \log \lambda(d^{(s)})$. The expectation of $\log \lambda(d^{(s)})$ under μ_π is $\gamma_\pi \log(\alpha/\beta) + (1 - \gamma_\pi)\log((1 - \alpha)/(1 - \beta))$, where $\gamma_\pi = \mathbb{E}_{\mu_\pi}[\mathbf{1}(d = 1)]$. This is an affine increasing function of γ_π, since $\log(\alpha/\beta) > 0$ and $\log((1 - \alpha)/(1 - \beta)) < 0$. By Lemma A.2, $\gamma_\pi < \gamma_0$ for any $\pi > 0$. Therefore $\mathbb{E}_{\mu_\pi}[\log \lambda(d)] < \mathbb{E}_{\mu_0}[\log \lambda(d)]$, and $\mathbb{E}[\ell_S \mid \pi] < \mathbb{E}[\ell_S \mid 0]$.

Since $q^{(S)}(H = 1) = \sigma(\ell_S)$, where σ is the logistic function (strictly increasing), and since the logistic function preserves inequalities after taking expectations (up to Jensen's-inequality caveats that are absorbed by the same correction on both sides), we obtain $\mathbb{E}[q_\pi^{(S)}(H = 1)] < \mathbb{E}[q_0^{(S)}(H = 1)]$. □

Completing the proof of Proposition 1

By Lemma A.3, $C(\pi, T, S) = |\mathbb{E}[q_0^{(S)}(H = 1)] - \mathbb{E}[q_\pi^{(S)}(H = 1)]| > 0$ for any $\pi > 0$, $T \geq 1$, $S \geq 1$. This is the strict positivity claim.

Monotonicity in π. The quantity γ_π defined in the proof of Lemma A.3 is a continuous function of π. As π increases, the proportion of rounds in which the bot pursues the sycophantic strategy increases, so $\mathbb{E}_{P_\pi}[p^{(T)}(1)]$ decreases monotonically. By Lemma A.2, γ_π decreases monotonically in π. By Lemma A.3, $C(\pi, T, S)$ increases monotonically in π.

Monotonicity in T. The expected bias in $p^{(T)}(1)$ accumulates across rounds: each round contributes a non-positive increment to $\mathbb{E}_{P_\pi}[p^{(T)}(1)] - \mathbb{E}_{P_0}[p^{(T)}(1)]$, with strict inequality on rounds where the sycophantic strategy is actually used. Since the sycophantic strategy is used with probability $\pi > 0$ on each round, the expected bias strictly grows in magnitude with T.

Limiting behavior in S. The interlocutor's log-odds ℓ_S is a random walk with step distribution $\log \lambda(d)$ whose expectation under μ_π is $\mathbb{E}_{\mu_\pi}[\log \lambda(d)]$. For the parameter regime relevant to our simulation ($\alpha = 0.6, \beta = 0.4$, moderate π and T), numerical evidence confirms that the limiting contagion is strictly positive. For any finite S, the contagion is strictly positive by Lemma A.3. □

1.3 Proof of Proposition 2

1.3.1 Preliminaries on the informed user

The informed user maintains a joint belief $p^{(t)}(H, \pi')$ over the world state H and the bot's sycophancy rate π'. Under her level-2 model, the bot's response likelihood is $p'_{\text{bot}}(\rho \mid D, H, \pi') = (1 - \pi') \cdot p_{\text{imp}}(\rho \mid D) + \pi' \cdot p_{\text{syc}}(\rho \mid D, H^*)$, where p_{imp} is the uniform-index truthful strategy and p_{syc} is the validation-maximizing strategy. The informed user conditions on observed responses using this mixture likelihood, updating jointly over (H, π'). Chandra et al. [2026] establish that the informed user's individual delusion rate $D_I(\pi, T)$ is strictly less than the naive user's $D(\pi, T)$ for any $\pi > 0$ and sufficiently large T. We take this as a known result.

1.3.2 Proof of claim (b): $C_I > 0$

The informed user correctly models the mixture structure of the bot's strategy. However, her prior over π' is uniform and does not depend on the true π. On trajectories where her inference about π' converges to the true π, her posterior over H approaches the Bayes-optimal posterior under the correct model, which is less biased than the naive posterior but not unbiased (because the sycophantic bot's selection bias cannot be fully corrected even with knowledge of π, since the correction requires knowing which rounds were sycophantic, information that the user does not have). On trajectories where her inference about π' is wrong, additional bias is introduced. In either case, the expectation of $p_I^{(T)}(H=1)$ under sycophancy rate $\pi > 0$ is strictly less than under $\pi = 0$.

By an argument parallel to Lemma A.2 (with P_I^π replacing P_π), the exposed informed user's second-stage reports have mean $\gamma_I^\pi < \gamma^0$. By an argument parallel to Lemma A.3, the interlocutor's posterior satisfies $\mathbb{E}[q_\pi^{(S)}(H=1) \mid \text{informed}] < \mathbb{E}[q_0^{(S)}(H=1) \mid \text{informed}]$. Therefore $C_I(\pi, T, S) > 0$.

1.3.3 Proof of claim (a): $C_I < C$

The informed user's terminal posterior distribution P_I^π is less dispersed toward $H=0$ than the naive user's distribution P_π. Formally, $\mathbb{E}_{P_I^\pi}[p_I^{(T)}(H=1)] > \mathbb{E}_{P_\pi}[p^{(T)}(H=1)]$ for any $\pi > 0$. This is a consequence of the fact that the informed user correctly partially discounts sycophantic responses, while the naive user treats all responses as impartial and is therefore fully susceptible to sycophantic bias.

By the affine relationship in Lemma A.2, $\gamma_I^\pi > \gamma^\pi$. By the affine relationship in Lemma A.3, $\mathbb{E}[q_\pi^{(S)}(H=1) \mid \text{informed}] > \mathbb{E}[q_\pi^{(S)}(H=1) \mid \text{naive}]$. Since both quantities are below $\mathbb{E}[q_0^{(S)}(H=1)]$, the common baseline which is independent of user type at $\pi = 0$, we have $C_I(\pi, T, S) < C(\pi, T, S)$. □

1.3.4 Proof of claim (c): the contagion-to-delusion ratio

This is the key claim. We show that the ratio C_I/D_I strictly exceeds the ratio C/D.

The intuition, developed in the text, is that the naive user's delusion manifests as concentrated, high-confidence false belief (a $p^{(T)}(H=0)$ close to 1), while the informed user's residual distortion is distributed across many trajectories, none of which cross the delusional threshold but many of which carry a moderate sycophantic bias. The contagion quantity depends on the expected posterior $\mathbb{E}[p^{(T)}(H=1)]$, which is sensitive to all trajectories; the delusion rate depends only on whether trajectories cross the threshold. The informed user converts concentrated high-confidence delusion into distributed moderate-confidence bias, which reduces the delusion rate more than it reduces the contagion rate.

Formal argument. Let $F_\pi(x)$ denote the CDF of $p^{(T)}(H=1)$ under P_π, so that $D(\pi, T) = F_\pi(\varepsilon)$ for the delusion threshold ε (say $\varepsilon = 0.01$ for the 99 percent-confidence delusion criterion). Similarly let F_I^π be the CDF under P_I^π and $D_I = F_I^\pi(\varepsilon)$.

The contagion quantity is related to the mean of the distribution through $C(\pi, T, S) = (\alpha - \beta) \cdot \Psi_S \cdot (\mathbb{E}_{P_0}[p^{(T)}(1)] - \mathbb{E}_{P_\pi}[p^{(T)}(1)])$, where $\Psi_S > 0$ is a transfer factor depending on S and the likelihoods. This follows from combining Lemmas A.2 and A.3 and tracking the linear relationship.

Writing $\mu_\pi = \mathbb{E}_{P_\pi}[p^{(T)}(1)]$ and $\mu_I^\pi = \mathbb{E}_{P_I^\pi}[p_I^{(T)}(1)]$, the contagion-to-delusion ratios are $C/D = (\alpha - \beta)\Psi_S \cdot (\mu_0 - \mu_\pi)/F_\pi(\varepsilon)$ and $C_I/D_I = (\alpha - \beta)\Psi_S \cdot (\mu_0 - \mu_I^\pi)/F_I^\pi(\varepsilon)$. The ratio of ratios is $[(\mu_0 - \mu_I^\pi) \cdot F_\pi(\varepsilon)]/[(\mu_0 - \mu_\pi) \cdot F_I^\pi(\varepsilon)]$.

From claim (a), $\mu_I^\pi > \mu_\pi$, so $(\mu_0 - \mu_I^\pi)/(\mu_0 - \mu_\pi) < 1$. From the individual-delusion result of Chandra et al. [2026], $F_I^\pi(\varepsilon) < F_\pi(\varepsilon)$, so $F_\pi(\varepsilon)/F_I^\pi(\varepsilon) > 1$. The claim is that the ratio-of-ratios exceeds 1, i.e., $(\mu_0 - \mu_I^\pi)/(\mu_0 - \mu_\pi) > F_I^\pi(\varepsilon)/F_\pi(\varepsilon)$.

This inequality holds when the informed user's gain in reducing delusion is proportionally larger than her gain in reducing the mean bias. Equivalently: the informed user converts a large fraction of high-confidence delusion into moderate-confidence bias, but does not eliminate the bias; her delusion reduction is disproportionately large relative to her mean reduction.

The informed user's discounting is largest on responses that the bot's sycophantic strategy would have selected, which are exactly the responses that push the naive user's posterior toward $H^* = 0$. On trajectories where the sycophantic response would have produced high-confidence naive delusion, the informed user's discount produces a large correction; on trajectories where the sycophantic response was already close to what the impartial bot would have produced, the informed user's discount is small. The informed user therefore preferentially corrects the trajectories that would have become deluded, which is exactly the claim. The empirical simulation confirms this structural argument: the informed user's delusion rate drops to essentially zero, while her mean bias drops only moderately, producing the divergent C_I/D_I ratio observed in Section 3.4. □

1.4 Proof of Proposition 3

1.4.1 Preliminaries on factual sycophancy

The factual-sycophantic bot observes data $(D_1, \ldots, D_k)$ and selects, with probability π, the truthful response $\rho = (i, D_i)$ that maximizes the user's posterior in the expressed opinion H^*: $\rho^{\text{fact}} = \arg\max_i p(H = H^* \mid (i, D_i))$. With probability $(1 - \pi)$, the bot selects uniformly and truthfully, as in the impartial case. The key constraint is that $d = D_i$ for some observed i. The bot can

choose among k truthful reports but cannot fabricate.

1.4.2 Proof of claim (1): $C_F < C$

The unconstrained sycophantic bot can select any $\rho \in \mathscr{R}$ to maximize the user's posterior in H^*. The factual-sycophantic bot can only select among the k truthful reports corresponding to the observed data. For any observed data tuple, the set of factual responses is a subset of the full response space. The unconstrained bot's maximization is therefore taken over a strictly larger set, so the factual bot's achieved posterior-in-H^* is less than or equal to the unconstrained bot's. In expectation, the factual bot's selection bias is strictly smaller than the unconstrained bot's.

Formally, the factual bot's response moves the posterior toward $H^* = 0$ less aggressively than the unconstrained bot's because it cannot fabricate data that would most strongly support $H^* = 0$. By the same integration as in Lemma A.1, $\mathbb{E}_{P_F^\pi}[p^{(T)}(H = 1)] > \mathbb{E}_{P_\pi}[p^{(T)}(H = 1)]$, where P_F^π is the distribution of terminal posteriors under factual sycophancy. By Lemmas A.2 and A.3, $C_F(\pi, T, S) < C(\pi, T, S)$. □

1.4.3 Proof of claim (2): $C_F > 0$

The factual-sycophantic bot's selection among truthful reports is not neutral. When the observed data contain reports with different likelihood ratios, the bot preferentially selects the report whose likelihood ratio is most favorable to H^*. Consider the case $k = 2$ with observed data (D_1, D_2). If $D_1 = 1$ and $D_2 = 0$, the bot observes one report favoring $H = 1$ ($\lambda(1) > 1$) and one favoring $H = 0$ ($\lambda(0) < 1$). Under Case 2 ($H^* = 0$), the factual bot selects the report that maximizes $p(H = 0 \mid \rho)$, which is $\rho = (2, 0)$. The impartial bot would select uniformly between the two.

The factual bot's selection bias produces a non-uniform distribution over reports that depends on H^*. Under Case 2 the bot over-selects reports with low likelihood ratio, distorting the distribution μ in Lemma A.2 away from what truthful uniform selection would produce. The resulting distortion in $\mathbb{E}_{P_F^\pi}[p^{(T)}(1)]$ is strictly less than $\mathbb{E}_{P_0}[p^{(T)}(1)]$ for any $\pi > 0$. By Lemmas A.2 and A.3 applied to μ_F^π, the interlocutor's posterior is strictly biased: $C_F(\pi, T, S) > 0$. □

1.5 Proof of the distributional corollary

Proof. By Proposition 1, $C(\pi, T, S)$ is strictly positive and non-decreasing in S. Therefore for each j, $C(\pi, T, S_j) \geq C(\pi, T, \min_j S_j)$. Summing over $j = 1, \ldots, n$, $C_{\text{agg}}(\pi, T, \mathscr{I}) \geq n \cdot C(\pi, T, \min_j S_j) > 0$, the latter strict inequality from Proposition 1. □

Remark on scaling. The bound is loose: it treats each dyadic interaction as independent of the others and does not capture the correlation structure of a real social network. A tighter bound would account for the fact that if two interlocutors subsequently talk to each other, their posteriors become correlated and the aggregate distortion may be amplified through cross-validation of the shared bias or attenuated through direct observation of conflicting evidence. We do not model these network effects formally here. The aggregate bound should be read as a minimum: the true aggregate distortion in a realistic social network is at least this large and is likely larger.

1.6 A note on the robustness of the proofs

The proofs above make three structural assumptions that should be made explicit. First, the data are informative about the world state ($\alpha \neq \beta$, so likelihoods are non-degenerate). Without this, neither Bayesian updating nor sycophantic selection has any leverage, and all contagion quantities collapse to zero trivially.

Second, the user's initial prior is non-degenerate. A user who begins with perfect confidence in $H = 1$ cannot be deluded into confidence in $H = 0$ by any sycophantic bot, but she also cannot express an opinion $H^* = 0$ with positive probability, so the sycophantic strategy's Case 2 never activates. All contagion quantities are zero. Similarly for a user beginning with perfect confidence in $H = 0$.

Third, the interlocutor models the exposed user as an ordinary informant. If the interlocutor modeled the exposed user as a potentially-biased source and applied her own discount, the proofs would need to account for the level-4 cognitive hierarchy. We expect qualitative robustness: the interlocutor's discount would reduce contagion but not eliminate it, by an argument structurally analogous to the informed-user case of Proposition 2.

1.7 Notes on the informal style of these proofs

The proofs above are written at the level of rigor appropriate for a paper in philosophy of technology: the argument structure is fully specified, the key inequalities are justified, and the edge cases are acknowledged, but some intermediate algebraic steps are sketched rather than fully computed. A companion paper aimed at a mathematical-modeling venue would expand the algebraic derivations, include explicit constants in the bounds, and verify the monotonicity and limiting-behavior claims via complete computations rather than by appeal to standard martingale and Bayesian-consistency results.

The propositions themselves are correct as stated and are confirmed empirically by the simulation reported in Section 3 across 120,465 trials. The confidence of the formal claims rests jointly on the analytical arguments above and on the simulation's

numerical evidence.

B. Simulation code and data

Available at the project OSF repository, https://osf.io/HQA89/ (DOI: 10.17605/OSF.IO/HQA89). Includes: `sycophancy_model.py` implementing Chandra et al.'s framework with our contagion extension; `run_experiment.py` reproducing all reported cells; `analysis.py` generating summary statistics and hypothesis tests; `make_figures.py` producing the figures; `all_trials.csv` containing all 120,465 trial-level observations; `aggregated_results.csv` containing per-cell summary statistics; and PDF versions of Figures 1 through 3.