NBER WORKING PAPER SERIES

ARTIFICIAL INTELLIGENCE AND POLITICAL ADVICE

Georgy Egorov
Konstantin Sonin

Working Paper 35689
http://www.nber.org/papers/w35689

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2026

Artificial Intelligence and Political Advice
Georgy Egorov and Konstantin Sonin
NBER Working Paper No. 35689
August 2026
JEL No. D72, D83, P00

ABSTRACT

Artificial intelligence is increasingly used for political advice. We study an AI that is better informed about a payoff-relevant state and cares both about accuracy and about the perceived welfare of the individual it advises. The AI then has an incentive to tilt advice toward what the individual would like to believe, altering both the political content and the informativeness of its messages. Sophisticated individuals anticipate this distortion and filter out its predictable political component, yet still learn less because the AI makes its messages less responsive to the state. Individuals who underestimate the incentive instead mistake political accommodation for information, allowing political preferences to distort factual beliefs and generate polarization and radicalization. The model also shows that better-informed individuals receive more informative and less politically tilted advice, while greater sophistication can improve interpretation yet worsen communication itself. Contrary to the familiar echo-chamber intuition, political distortion is mitigated when political preferences and prior beliefs coincide and is most consequential when they diverge. We show how independently varying individuals' stated preferences and prior beliefs can recover the AI's responsiveness to the state even when the underlying state cannot be manipulated.

Georgy Egorov
Northwestern University
Kellogg School of Management
and NBER
g-egorov@kellogg.northwestern.edu

Konstantin Sonin
The University of Chicago
ksonin@uchicago.edu

1 Introduction

Millions of people now ask artificial intelligence what to think about political questions: which party to trust, which policy will work, or how to resolve a tradeoff. Political conversations with AI are already common and frequently involve requests for information and explanation rather than simply expressions of opinion (Zu, 2026). Adoption of AI has been fast (Faverio and Kikuchi, 2026; Bick, Blandin and Deming, 2024), and these conversations appear to move people, with a growing body of evidence showing that they can shift beliefs and political attitudes (Costello, Pennycook and Rand, 2024; Hackenburg et al., 2025). This gives rise to a new informational environment in politics, combining the personalization of interpersonal communication with the scale of mass media and giving individuals direct answers to questions they would otherwise have to resolve through multiple sources. Its real-world implications are ambiguous: AI can make individualized information and explanation widely accessible, but the same capacity for personalization allows it to adapt answers to users' preferences and prior beliefs. Such sycophantic behavior has been documented in AI systems (Perez et al., 2023; Wei et al., 2023; Sharma et al., 2024). It is therefore natural to ask whether personalized political advice by AI will promote learning and cohesion or reinforcement and polarization, and what its consequences are for welfare.

We argue that the answer depends critically on whether individuals understand how AI advice is generated. The AI has an incentive to tilt its message toward what the individual would like to be true. A sophisticated individual anticipates this and filters out the predictable political tilt; anticipating that response, however, the AI tilts its message further. In the end, sophistication disciplines posterior beliefs but amplifies pandering: political preferences no longer distort beliefs, yet individuals remain less informed because they cannot recover information the AI chose not to transmit. By contrast, naive individuals interpret politically tilted advice as evidence, and while this spares them from the more extreme messaging induced by sophistication, their own preferences find their way into their factual beliefs, potentially generating polarization and radicalization. When the AI is uncertain

about an individual's sophistication, the presence of naive individuals in the pool can induce it to preserve more information about the state, benefiting sophisticated individuals; at the same time, naive individuals bear the cost of mistaking the remaining political tilt for information.

In this paper, we build a model of an individual receiving political advice from AI. The individual has a prior belief about a state of the world and a political preference that need not coincide with that belief, and chooses an action that trades off matching the state against following this preference. The AI is better informed about the state and observes the individual's characteristics. It chooses a personalized message, trading off the accuracy of its advice against the individual's perceived utility; to raise the latter, it has an incentive to create the impression that the state is closer to the individual's preferences than it really is. The message is then transmitted with noise, creating an inference problem. Individuals may differ in how well they understand this pandering incentive: sophisticated individuals recognize it and correctly anticipate the AI's communication strategy, whereas naive individuals underestimate the AI's willingness to accommodate their preferences or, at the extreme, believe that the AI simply tells the truth. We focus on linear (more precisely, affine) equilibria, which exist and are unique under mild conditions and permit a rich set of comparative statics.

Specifically, stronger incentives to pander make AI advice less responsive to the state, and thus less informative to the individual, while making it more responsive to the individual's political preference. Greater precision of the individual's prior information, as well as higher communication noise, make manipulation less effective and therefore make AI advice more responsive to the state and less politically tilted. The former effect is particularly consequential: better-informed individuals receive more informative advice precisely because their beliefs are harder to manipulate, implying, surprisingly and perhaps dishearteningly, that those who have the most to learn may also receive the most distorted information. The latter implies that more precise communication need not maximize learning, because its direct in-

formational benefit may be offset by the stronger incentive it creates to distort the message. In the presence of naive individuals—or, more generally, individuals who underestimate the AI's incentive to pander—messaging may become more truthful, but these individuals' political preferences begin to distort their posterior beliefs. Paradoxically, the familiar risk associated with "echo chambers" is relatively benign here: when what individuals would like to believe coincides with what they already believe, AI advice moves their posteriors toward the truth. In contrast, individuals whose political preferences are at odds with their priors receive extreme pandering toward their preferences, with an additional push to counteract their prior beliefs, potentially resulting in beliefs much more extreme than their preferences alone would suggest. In other words, it is among individuals whose prior information is in sharp discord with what they would like to hear that the greatest scope for polarization arises; in practice, this may be exacerbated by such individuals' likely greater propensity to turn to AI for advice.

These results suggest that interventions aimed at improving AI political advice can have unintended effects. The most direct remedy is, quite obviously, to reduce the incentive to pander at the expense of accuracy. Other seemingly beneficial interventions, however, can backfire: more precise communication can reduce learning by making manipulation more effective, appearing open-minded when seeking AI advice can invite greater manipulation, and even simply educating users about AI's sycophancy can induce more aggressive pandering. Whether such interventions are beneficial therefore depends on the underlying parameters of the communication problem. Although the responsiveness of AI advice to the true state may be difficult to measure directly because the state itself is difficult to vary, the model's tractable solution allows it to be inferred from dimensions that can be experimentally manipulated. By varying the stated political preferences and prior beliefs of otherwise identical synthetic individuals, one can estimate how AI advice responds to each and use these responses to recover its responsiveness to the state and the effective parameters governing communication. The model can therefore turn observable patterns of personalization into

predictions about otherwise difficult-to-observe distortions in information transmission, and ultimately into counterfactual assessments of which interventions improve learning and which may inadvertently make it worse.

This paper contributes to several literatures. First, it relates to the rapidly growing literature on artificial intelligence and politics. AI advice can be informative: Luettgau et al. (2025) show that conversational AI can increase political knowledge, while Velez, Green and Sevi (2025) find that personalized chatbot voting advice improves knowledge of party positions. Conversations with AI can also change political attitudes and choices (Lin et al., 2025; Hackenburg et al., 2025), and frontier AI systems can outperform even expert human persuaders (Hackenburg et al., 2026). Most directly related to our analysis, AI systems can condition their responses on what they learn or infer about the individual. Miyazaki and Hall (2026) show that stated policy positions can substantially shift AI voting recommendations in Japan, while Perez et al. (2023); Wei et al. (2023); Sharma et al. (2024) document sycophantic behavior, including a tendency to agree with an individual's stated views rather than provide a more accurate answer. Related audits document political biases in LLM outputs (Santurkar et al., 2023) and show that these outputs can depend on the political identity or other characteristics inferred from the individual (Törnberg and Schimmel, 2026; Kearney, Binns and Gal, 2025). Conlon and Schwardmann (2026) find that AI advice depolarizes choices on average despite being measurably sycophantic, while experimentally increasing sycophancy attenuates this effect; Marvel and Ju (2026) find that warnings about sycophancy do not eliminate its effects on belief updating and confidence. More broadly, Acemoglu, Ozdaglar and Siderius (2025); Acemoglu, Lin, Ozdaglar and Siderius (2026); Chandra, Kleiman-Weiner, Ragan-Kelley and Tenenbaum (2026) study related consequences of AI-mediated communication and information for beliefs, polarization, and democratic discourse. To the best of our knowledge, ours is the first paper in the literature on AI and politics to distinguish individuals' prior beliefs from their preferences over what they would like to be true.

Second, the paper relates to the theoretical literature on signaling (Spence, 1973) and strategic communication (Crawford and Sobel, 1982). An important branch of the communication literature studies settings in which misrepresentation is costly (Kartik, Ottaviani and Squintani, 2007; Kartik, 2009; Bénabou and Tirole, 2006), while Blume, Board and Kawamura (2007) show that communication noise can improve information transmission. Related models of social signaling study settings in which individuals seek to induce particular socially desirable inferences about themselves, rather than simply to be perceived as a higher type (Bernheim, 1994; Andreoni and Bernheim, 2009; Bursztyn, Egorov and Fiorin, 2020; Bursztyn et al., 2023). Prendergast (1993) shows that subjective performance evaluation can induce workers to bias their reports toward what they believe their supervisors think in order to signal competence. Our paper belongs to this line of signaling models in which the sender seeks to steer inference toward a particular target rather than uniformly upward or downward. We contribute a tractable version of such a model with a continuum of states, allowing the direction and magnitude of desired distortion to vary across individuals in a way that is particularly useful for studying political preferences and polarization.

Third, the paper relates to the literature on media slant and audience demand for political information. In models of demand-driven media bias, outlets may slant reports toward consumers' prior views because consumers prefer confirmatory accounts or infer source quality from agreement with their expectations (Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2006; Gentzkow, Shapiro and Stone, 2015). Empirical work finds that newspapers' slant responds strongly to the ideological preferences of their readers (Gentzkow and Shapiro, 2010), while cable-news audiences disproportionately choose outlets closer to their own political views (Martin and Yurukoglu, 2017); see also Zhang et al. (2026) on AI-mediated communication and strategic information provision. Our setting emphasizes individual-level strategic tailoring: the AI can condition communication on the individual's prior beliefs, preferences, and expected sophistication, making both the direction of slant and the informativeness of advice endogenous to the recipient. A key distinction from much of the

media-bias literature is that we separate what individuals believe from what they would prefer to be true, allowing the two to affect communication in distinct and potentially opposing ways.

Fourth, the paper relates to empirical work on political communication and persuasion (DellaVigna and Gentzkow, 2010). Political communication can affect turnout, vote choice, and attitudes through interpersonal contact (Gerber and Green, 2000; Broockman and Kalla, 2016) and mass media (DellaVigna and Kaplan, 2007; Enikolopov, Petrova and Zhuravskaya, 2011), although persuasive effects of campaign contact in general elections are often small (Kalla and Broockman, 2018). A related strand studies how individuals process and retransmit political information after receiving it (Chandrasekhar, Larreguy and Xandri, 2020; Egorov, Guriev, Mironov and Zhuravskaya, 2026). Much of this empirical work treats the content of communication as given and studies its effects on recipients. Our model instead endogenizes the message itself: the AI strategically chooses both its slant and informativeness in anticipation of how the individual will interpret it.

Finally, the paper speaks to the literature on political polarization and echo chambers. One strand studies how biased processing of common information can produce polarization (Taber and Lodge, 2006; Dandekar, Goel and Lee, 2013; Levy and Razin, 2019), while another emphasizes selective exposure and interaction with like-minded sources and individuals (Bakshy, Messing and Adamic, 2015; Bail, 2021). Recent evidence also suggests that AI need not be polarizing: conversations with AI representations of the political outgroup can reduce partisan misperceptions and animosity (Lira, Castelo, Puntoni and Toubia, 2026), while Conlon and Schwardmann (2026) find that AI advice can be depolarizing even when it is measurably sycophantic. Our model highlights a distinct channel through which personalized communication can affect polarization, one that does not require selective exposure or motivated reasoning. In the model, the AI responds to what individuals would prefer to be true, and the consequences depend on how these preferences relate to their prior beliefs. This distinction yields an implication opposite to the usual echo-chamber intuition: align-

ment between beliefs and preferences can mitigate distortion, whereas divergence between them creates the greatest scope for radicalization and polarization.

The rest of the paper is organized as follows. Section 2 presents the theoretical framework. Section 3 characterizes equilibrium communication for naive and sophisticated individuals and derives its comparative statics. Section 4 studies communication with mixed audiences and its implications for polarization. Section 5 discusses implications for the design and evaluation of AI political advice, and Section 6 concludes. The Appendix contains all proofs.

2 Formal Setup

There is a one-dimensional policy space and a continuum of individuals $\mathbb{I}$ with political preferences characterized by their ideal point p_i, $i \in \mathbb{I}$. We assume that the distribution of political preferences has finite first and second moments.

There is also a state of the world θ, the realization of which is not known to individuals; it may be thought of as a policy choice that maximizes the utility of the median voter, or maximizes the total societal welfare, or something else that individuals are interested in learning. Importantly, we do not assume that θ is connected to the aggregate of political preferences: for example, the state might mean that a pro-choice policy on abortions maximizes social welfare, but a majority may be ideologically pro-life, or vice versa. It is common knowledge that θ is normally distributed:

$$\theta \sim \mathcal{N}\left(\theta_0, \delta^2\right),$$

and each individual $i \in \mathbb{I}$ gets a private signal

$$\phi_i = \theta + \eta_i,$$

where $\eta_i \sim \mathcal{N}\left(0, \tau^2\right)$, with noise η_i independent across individuals and of θ. Thus, from the

standpoint of individual i, the distribution of θ conditional on the realization of signal ϕ_i is

$$\theta \mid \phi_i \sim \mathcal{N}\left(b_i, \sigma^2\right), \tag{1}$$

where

$$b_i = \mathbb{E}\left(\theta \mid \phi_i\right) = \frac{\tau^2}{\delta^2+\tau^2}\theta_0 + \frac{\delta^2}{\delta^2+\tau^2}\phi_i, \qquad \sigma^2 \equiv \frac{\delta^2\tau^2}{\delta^2+\tau^2}.$$

Since this update happens before any interesting part of the model, we will refer to b_i as individual i's *prior belief* about θ and note that everyone's priors have the same variance σ^2.

Individuals seek to learn more about θ, and to do so they interact with an AI agent (henceforth Agent, "it"). We assume that Agent knows θ precisely and, when interacting with individual i, observes the individual's prior belief b_i and political leaning p_i, as well as preference parameters α_i and γ_i.[1] We deliberately give Agent rich information about the individual in order to isolate the consequences of strategic personalization from the separate problem of inferring individual characteristics. Thus, the forces we study would remain present even for an idealized AI that inferred these characteristics perfectly. Individuals know that their priors and preferences are observed by Agent.

In its interaction with individual i (henceforth Individual, "she"), Agent sends a private message m_i, and Individual observes a noisy signal

$$\mu_i = m_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \nu^2)$; we will allow ν^2 to be small but not infinitesimal to get uniqueness of

[1] In practice, these characteristics may be inferred from a combination of demographic information, prior interactions, and the style of questioning. We consider an Agent that receives only noisy signals about the individual it advises in Section 3.4. The assumption that Agent knows θ precisely is also less restrictive than it may appear: even if it did not know θ initially, it could learn from interacting with a small but representative sample of individuals and then use this information when interacting with the rest of the continuum.

equilibrium.[2] Individual then takes an action y_i with the goal of maximizing expected utility

$$U_i(\mu_i) = \max_{y_i} \mathbb{E}\left(u_i \mid \mu_i\right),$$

where the expectation is taken given Individual's prior b_i and her beliefs about Agent's communication strategy, and

$$u_i = -\gamma_i \left(\alpha_i \left(y_i - \theta\right)^2 + \left(1 - \alpha_i\right)\left(y_i - p_i\right)^2\right),$$

with $\alpha_i \in (0,1)$, $\gamma_i > 0$. In other words, Individual seeks to choose an action that, in expectation, balances social welfare and her political preferences, with the former taken with weight α_i and the latter with complementary weight, and γ_i captures the intensity of Individual's preferences over this decision. The optimal action is therefore

$$y_i^* = \alpha_i \widehat{\theta}_i + (1 - \alpha_i) p_i,$$

where $\widehat{\theta}_i$ is the posterior mean of θ after observing μ_i, so $\widehat{\theta}_i = \mathbb{E}\left(\theta \mid \mu_i\right)$. The actions may be thought of as having social consequences (e.g. through donating, canvassing or voting) but for the model they need not. For example, consumption decisions such as purchasing an electric car may balance considerations about the future price of gasoline, which Individual would like to forecast accurately, against her political or ideological preference for reducing carbon emissions.

Agent, when sending message m_i to individual i, cares about factual accuracy but also

[2] If message m_i is directly observable by Individual, there would be multiple equilibria, including a fully informative equilibrium and a continuum of fully uninformative ones, similar to Bernheim (1994). With noisy communication, the equilibrium is unique under mild assumptions, and as an additional bonus we can use the "signal to noise" ratio as a measure of informativeness of equilibrium. The assumption of noisy communication was used to achieve uniqueness and tractability in prior work (Acemoglu, Egorov and Sonin, 2013, 2026).

about Individual's expected utility, and maximizes

$$V_i(m_i) = -l(m_i - \theta)^2 + k\,\mathbb{E}_{\varepsilon_i}\left[U_i(m_i + \varepsilon_i)\right], \tag{2}$$

where $k, l > 0$. The first term penalizes messages that are far from the true value of θ; in particular, an Agent that cares only about accuracy would report $m_i = \theta$. The second term says that Agent cares about the individual's expected utility U_i. We do not take a strong position on the reason for that; Agent may be sycophantic, but also altruistic or paternalistic in its care for Individual's subjective well-being, or Agent might care about approval because it translates into more subscriptions and use. Similar incentives may arise from preference-based training, which can reward responses that users rate more favorably even when those responses are less accurate (Sharma et al., 2024).[3]

Our equilibrium concept is Perfect Bayesian Equilibrium (PBE)[4] with the following restriction on equilibrium messaging strategies: in interacting with Individual, Agent's strategy is affine in θ:

$$m_i = A_i\theta + B_i,$$

where A_i, B_i do not depend on θ (but may depend on any other parameters). This refinement is introduced for tractability; we will show that such an equilibrium always exists, and given its simplicity, it would be a strong reason to focus on such equilibrium even if a non-affine equilibrium exists. Notice that for $A_i = 0$ the strategy is perfectly uninformative; for $A_i \neq 0$ the strategy would be perfectly revealing if m_i were observable by Individual, but given the noise ε_i, the inference problem is slightly more involved, with strategies with smaller $|A_i|$

[3]It is important, however, that Agent cares about Individual's expected utility U_i, which is evaluated given Individual's information, and not the underlying state-contingent utility u_i. The simplest reason, for example, is that Individual might never learn θ and thus experiences expected utility U_i but not its realization contingent on θ. Alternatively, Individual may learn θ only later, after decisions that matter for Agent (and perhaps Individual) have already been made. For example, Individual may rate Agent's response, decide whether to continue using it, or subscribe based on the information available at the time. If we directly assumed that Agent is sycophantic in the sense of disliking delivering misaligned information, meaning its utility has a term that penalizes $\left(\widehat{\theta}_i - p_i\right)^2$, we would get a very similar model.

[4]When discussing naive individuals we will allow them to hold misspecified beliefs about Agent's strategy.

being less informative ex post due to a higher noise-to-signal ratio.

Throughout, we maintain the following mild assumption about Individual's prior variance σ^2 of the state θ and the variance of communication noise, ν^2. Denote

$$\kappa \equiv \frac{47}{8} - \frac{13}{8}\sqrt{13} \approx 0.015979.$$

Assumption 1. $\rho \equiv \frac{\nu^2}{\sigma^2} \geq \kappa$.

We will show that Assumption 1 is sufficient for uniqueness of the affine equilibrium, and it is very mild: it holds if the variance of communication noise, ν^2, is at least 1.6% of the conditional variance of θ. In other words, it rules out highly precise communication (and even in that case there are at most three equilibria, as long as there is nonzero noise) as well as situations where individual's prior σ^2 (and thus the distribution of θ itself) is extremely dispersed.

3 Analysis

From this point on, we fix an individual i and suppress the subscript i whenever this causes no confusion.

3.1 Benchmark: Naive Individuals

It is instructive to start with the simpler case of naive individuals, who erroneously believe that $k = 0$ and therefore assume that Agent's equilibrium strategy is $m = \theta$. Agent, however, knows this and chooses its message taking Individual's resulting inference rule as given. Because Individual believes that the message before communication noise is equal to the state, after observing $m + \varepsilon$ she forms the posterior belief, and we denote its mean by $\widetilde{\theta}$

to differentiate it from the posterior mean $\widehat{\theta}$ formed by a sophisticated individual:

$$\widetilde{\theta} = \frac{\sigma^2}{\sigma^2 + \nu^2}\mu + \frac{\nu^2}{\sigma^2 + \nu^2}b. \tag{3}$$

Individual therefore chooses action

$$y = \alpha\widetilde{\theta} + (1 - \alpha)p$$

and gets expected utility

$$U(\mu) = -\gamma\alpha(1 - \alpha)\left(\widetilde{\theta} - p\right)^2 - \gamma\alpha\frac{\sigma^2\nu^2}{\sigma^2 + \nu^2}. \tag{4}$$

Here, the first term is the loss from the gap between the posterior mean of the socially desirable action and p while the second is the posterior variance of θ. This is noteworthy: the individual loses utility when her posterior mean of θ is far from her political preference p and she is thus unable to choose an action that is close to both and has to compromise; in contrast, finding out that θ is likely close to p is good news because she can choose an action that makes both losses small. The posterior variance term reflects the loss from inability to figure out θ precisely.

Plugging in (3) into (4) and then into (2) and dropping constant terms, Agent's problem becomes equivalent to

$$\max_m \left\{ -(m - \theta)^2 - h\left(\frac{m + \rho b}{1 + \rho} - p\right)^2 \right\}, \tag{5}$$

where we denote $h \equiv \frac{k\gamma\alpha(1-\alpha)}{l}$ and, as before, $\rho \equiv \frac{\nu^2}{\sigma^2}$. The first term pulls the message toward the true state, while the second captures the pandering motive: Agent would like Individual's expected posterior belief to be close to her political preference. Since (5) is

concave in m, the first-order condition gives us the optimal m:

$$m(\theta) = \frac{(1+\rho)^2}{(1+\rho)^2+h}\theta + \frac{h}{(1+\rho)^2+h}\left((1+\rho)p - \rho b\right). \tag{6}$$

Proposition 1 (Naive benchmark)**.** *Suppose Individual is naive. There is a unique equilibrium; in this equilibrium, Agent's communication strategy is affine in θ and is given by (6). The message is increasing in θ and p and decreasing in Individual's prior b.*

Proof. All proofs are in the Appendix.

This result is simple but intuitive: Agent's desire to tell the truth makes m comove with true state θ, while its pandering incentive makes it increasing in p. However, holding the true state θ fixed, Agent's message is higher when b is lower due to Agent's need to overcome the prior, which pulls its weight as Agent's messages are read with noise. The parameter h may be interpreted as the pandering motive: as h approaches zero, Agent starts reporting $m = \theta$, whereas a higher h makes m flatter in θ, and thus the message becomes less responsive to the true state. Naturally, h is higher when Agent cares more about Individual's welfare compared to telling the truth (k/l is high) or when the stakes of Individual's decision are higher (γ is high), but interestingly, the pandering motive is low when α is close to zero or one and instead it is maximized for $\alpha = 1/2$. Intuitively, if Individual cares strictly about matching the true state ($\alpha = 1$) then Agent would want to make her as informed as possible without pandering, whereas if Individual cares only about her ideological preferences ($\alpha = 0$) then she is not interested in learning θ and will ignore the message, so Agent would want to report $m = \theta$ to maximize factual accuracy. Lastly, a high noise-to-signal ratio ρ makes Agent's message more reflective of θ: indeed, noise makes manipulating Individual's beliefs more difficult, and Agent does it less.

We now turn to the main analysis to see that most of these intuitions hold for sophisticated users as well and that sophistication comes with a cost.

3.2 Sophisticated Equilibrium

We now consider a sophisticated Individual who is aware of Agent's true preferences and thus of its equilibrium strategy. She knows that the equilibrium strategy that Agent is using in interacting with her is

$$m = A\theta + B, \tag{7}$$

where A and B only depend on parameters known to her. It is easy to see that $A = 0$ cannot happen in equilibrium: if it did, m, and thus the distribution of μ, would be independent of θ, Individual would thus not update on μ and instead form the posterior $\widehat{\theta} = b$; then Individual's action y and her expected utility U would not depend on m, and Agent would then resort to choosing $m = \theta$ to maximize its utility (2), implying m depends on θ, a contradiction.

Now, given that $A \neq 0$, Individual starts with the prior that θ is distributed as (1) and thus considers m to be distributed as $\mathcal{N}(Ab + B, A^2\sigma^2)$; then $\mu = m + \varepsilon$ is distributed as $\mathcal{N}(Ab + B, A^2\sigma^2 + \nu^2)$. Therefore, the conditional distribution of m that she has after learning μ is

$$m \mid b, \mu \sim \mathcal{N}\left(\frac{A^2\sigma^2}{A^2\sigma^2 + \nu^2}\mu + \frac{\nu^2}{A^2\sigma^2 + \nu^2}(Ab + B),\ \frac{A^2\sigma^2\nu^2}{A^2\sigma^2 + \nu^2}\right).$$

As such, since $\theta = \frac{1}{A}(m - B)$, the posterior distribution of θ from her perspective is

$$\theta \mid b, \mu \sim \mathcal{N}\left(\widehat{\theta},\ \frac{\sigma^2\nu^2}{A^2\sigma^2 + \nu^2}\right),$$

where the posterior mean is

$$\widehat{\theta} = \frac{A\sigma^2}{A^2\sigma^2 + \nu^2}(\mu - B) + \frac{\nu^2}{A^2\sigma^2 + \nu^2}b. \tag{8}$$

Individual therefore chooses action

$$y = \alpha\widehat{\theta} + (1-\alpha)p$$

and gets expected utility

$$U(\mu) = -\gamma\alpha(1-\alpha)\left(\widehat{\theta} - p\right)^2 - \gamma\alpha\frac{\sigma^2\nu^2}{A^2\sigma^2+\nu^2}. \tag{9}$$

Note the similarity between (9) and (4): the first term depends on the gap between Individual's posterior and p, and the second term reflects the uncertainty about θ, which for a sophisticated individual depends on the strategy Agent uses.

Now Agent, when choosing m, is able to predict Individual's update following any realization of noise ε and thus received signal μ. It solves the maximization problem

$$\max_m \left\{ -l(m-\theta)^2 + k\,\mathbb{E}_\varepsilon\left(-\gamma\alpha(1-\alpha)\left(\widehat{\theta} - p\right)^2 - \gamma\alpha\frac{\sigma^2\nu^2}{A^2\sigma^2+\nu^2}\right)\right\};$$

plugging in (8) and rearranging, the problem becomes equivalent to

$$\begin{aligned}\max_m \quad \Bigg\{ & -l(m-\theta)^2 - k\gamma\alpha(1-\alpha)\left(\frac{A\sigma^2}{A^2\sigma^2+\nu^2}(m-B) + \frac{\nu^2}{A^2\sigma^2+\nu^2}b - p\right)^2 \\ & - k\gamma\alpha(1-\alpha)\frac{A^2\sigma^4\nu^2}{(A^2\sigma^2+\nu^2)^2} - k\gamma\alpha\frac{\sigma^2\nu^2}{A^2\sigma^2+\nu^2}\Bigg\}.\end{aligned} \tag{10}$$

The two terms on the second line are, respectively, the expected loss generated by communication noise ε through its effect on Individual's posterior mean, and the expected loss from her residual uncertainty about θ. The problem (10) is quadratic in m, with the first-order condition yielding the unique solution

$$m = \frac{(A^2+\rho)^2}{(A^2+\rho)^2 + hA^2}\theta + \frac{hA\left((A^2+\rho)p + AB - \rho b\right)}{(A^2+\rho)^2 + hA^2}, \tag{11}$$

where we again used the notation ρ and h.

Comparing (11) to (7), it is easy to see that strategy m given by (7) corresponds to an equilibrium if and only if

$$(A-1)\left(A^2+\rho\right)^2+hA^3=0, \tag{12}$$

$$B=\frac{hA}{A^2+\rho}p-\frac{hA\rho}{\left(A^2+\rho\right)^2}b. \tag{13}$$

To proceed, note that (13) implies that B is a deterministic function of A, so we can turn to (12) to study existence and uniqueness. Its left-hand side is a fifth degree polynomial in A, which necessarily has a real root. Furthermore, all real roots of (12) must satisfy $A\in(0,1)$. Most importantly, under Assumption 1, the root is unique, implying uniqueness of equilibrium.[5]

Before formulating the proposition, it is instructive to rewrite the equilibrium strategy (7). Using (12) and some algebra, (13) may be rewritten as

$$B=(1-A)b+\sqrt{h\left(\frac{1}{A}-1\right)}(p-b).$$

Then (7) becomes

$$m=b+A(\theta-b)+D(p-b), \tag{14}$$

where we denoted $D=\sqrt{h\left(\frac{1}{A}-1\right)}$. This is a transparent way to view Agent's equilibrium strategy: as compared to the individual's prior b, the equilibrium message pulls in the direction of θ with intensity A and in the direction of p with intensity D.

The discussion above, as well as comparative statics results, are summarized in the next

[5]For $\rho<\kappa$, multiplicity is possible, but even in this case there are at most three equilibria, and one can derive other sufficient conditions for uniqueness. For example, the root is unique if $h\geq\frac{1-8\kappa}{3}=\frac{13\sqrt{13}-46}{3}\approx 0.29$. It is instructive, however, to consider what happens as $\rho\to 0$. First, there is always a root converging to 0 (and notice that $A=0$ is a root of (12) for $\rho=0$). As for B, the denominators in (13) tend to 0, and one can show that $B\to\infty$ – or, more precisely, to $+\infty$ if $p>b$, to $-\infty$ if $p<b$, and to b if $p=b$. In the limit where $\rho=0$, we can sustain a continuum of equilibria with $A=0$ and B being an arbitrary real number if we specify appropriate off-path beliefs (for $\mu\neq B$); in other words, there is a continuum of uninformative equilibria. Interestingly, if $h\leq\frac{1}{4}$, then there also are informative (in fact, fully revealing) equilibria, defined by A solving $A(1-A)=h$ and $B=(1-A)p$. To simplify exposition, however, we focus on the case $\rho\geq\kappa$.

proposition.

Proposition 2. *There exists an affine equilibrium given by* (14) *where* A *solves* (12) *and* $D = \sqrt{h\left(\frac{1}{A} - 1\right)}$. *Every affine equilibrium has* $A \in (0, 1)$, *and if* $\rho \geq \kappa$ *then the equilibrium is unique. The equilibrium message is increasing in* θ *and* p *and decreasing in Individual's prior* b.

In other words, the sophisticated equilibrium retains most of the properties of communication with naive individuals. We next turn to learning by individuals and comparative statics with respect to parameters.

3.3 Comparative Statics

A sophisticated Individual knows Agent's equilibrium strategy and updates accordingly. Plugging $\mu = A\theta + B + \varepsilon$ into (8), we get that Individual's posterior is

$$\widehat{\theta} = \frac{A^2}{A^2 + \rho}\theta + \frac{\rho}{A^2 + \rho}b + \frac{A}{A^2 + \rho}\varepsilon. \tag{15}$$

It is worth noticing that the posterior does not depend on p: even though pandering tilts messages, it comes as an additive term in (14), and Individual can subtract it to learn about θ. Furthermore, unlike message m, the posterior is increasing in the prior b.

To measure the quality of learning, we use posterior variance. From (15),

$$\mathrm{Var}\left(\theta \mid b, \mu\right) = \frac{\sigma^2\rho}{A^2 + \rho} = \frac{\nu^2}{A^2 + \rho}. \tag{16}$$

This posterior variance also equals the expected squared error of $\widehat{\theta}$ as an estimate of θ, averaging over the prior and communication noise. Thus, (16) provides a natural measure of how much Individual learns from Agent. Holding ρ fixed, a higher A therefore makes Agent's message more informative and lowers Individual's posterior uncertainty. This gives A a natural interpretation as the informativeness of Agent's communication, or more precisely

its responsiveness to the state. Stronger pandering incentives make communication less informative: A is decreasing in h. Interestingly, A is increasing in ρ: a higher noise ν^2 makes manipulation more difficult for Agent, as does a more precise prior (lower σ^2); in other words, Agent provides more information to an already knowledgeable Individual.

Proposition 3. *Agent's communication is more informative (A is higher) when h is low or ρ is high. It is more informative than communication to naive individuals if and only if $h > \frac{(1-\rho)(1+\rho)^2}{\rho}$, which is more likely to hold when h or ρ is high, but it is never more informative than truthful communication. Individual's posterior variance increases in h and in σ^2. It is also decreasing in ν^2 for $\nu^2 < \sigma^2\left(\frac{\sqrt{h}}{2} - \frac{1}{4}\right)$ and increasing for higher ν^2.*

Proposition 3 has several important implications. First, more precise prior information (lower σ^2) raises A by raising ρ. This implies that an Individual more knowledgeable about the state receives more informative advice. On the flip side, those who have the most to learn happen to be the ones whose beliefs are easiest to manipulate, and in equilibrium they receive less informative communication. In other words, in this model of personalized political advice, information begets more information, whereas ignorance may be thought of as a trap.

Second, even though sophisticated individuals are able to subtract the pandering term and will not update on their own political preferences p, they are nevertheless hurt by pandering. Indeed, in equilibrium we have $A < 1$, in contrast to $A = 1$ that we would have in the case of truthful reporting, which would take place in the absence of pandering incentives. Thus, sophisticated individuals get less information and higher posterior variance. Remarkably, the information loss for sophisticated individuals may – or perhaps is even likely to be – more significant than for naive players. For example, this is true if $k < l$ (so Agent cares more about truth than pandering) and $\rho < 0.93$, provided that $\gamma \leq 1$; in this case we have $h < \frac{(1-\rho)(1+\rho)^2}{\rho}$. The intuition is that Agent tries to pander to both types of individuals, but sophisticated ones see this through and remove the pandering from updating. To counter that, Agent panders even more, which results in sufficiently extreme and sufficiently pooled,

and thus less informative messages. Thus, paradoxically, sophistication, with all its benefits, may result in less informative communication from Agent. For sufficiently high noise, in particular whenever $\rho > 1$, communication to sophisticated individuals is more informative than communication to naive ones regardless of h.

Third, communication noise results in less manipulation. If Agent's pandering motives are sufficiently strong, a marginal increase in communication noise can therefore improve learning. That said, if the pandering motive is not too large ($h < 1/4$), learning deteriorates monotonically as ν increases.

Let us now take a closer look at equilibrium messages m. We notice that (14), in case $p = b$, reduces to $m = b + A(\theta - b)$; in other words, for individuals who prefer to hear what they already believe, the message itself serves a moderating purpose. For example, should a person process this message naively, it would simply shift her posterior toward true θ on average. When $p \neq b$, the message has an additional factor $D(p - b)$, which pushes the message in the direction of preference relative to prior belief. Unlike A, which is in equilibrium confined between 0 and 1, $D \equiv \sqrt{h\left(\frac{1}{A} - 1\right)}$ is unbounded: it is increasing in h, decreasing in ρ, and tends to ∞ as $h \to \infty$. In particular, $D > 1$ is possible for h high enough; Agent then not only panders to p but overpanders, responding more than one-for-one to $p - b$; whenever $D > 1$, individuals with $|p|$ sufficiently high will receive even more radical messages, $|m| > |p|$. Thus, stronger pandering incentives simultaneously make Agent's message less informative and more extreme, possibly much more extreme than Individual's preferences, while a more precise prior or greater communication noise works in the opposite direction.

3.4 Robustness

The baseline model assumes that Agent knows Individual's characteristics p, b, α, and γ. Suppose instead that Agent observes a signal z that generates conditional distribution $\mathcal{Z}$

over Individual's vector of parameters

$$(p, b, \alpha, \gamma),$$

which we allow to be arbitrary. In particular, we impose neither independence nor parametric assumptions on this multivariate distribution. Individual knows her own characteristics, while Agent's signal z and the conditional distribution it induces are common knowledge.[6]

Let

$$q \equiv \gamma\alpha(1-\alpha),$$

and define

$$\overline{q} \equiv \mathbb{E}(q \mid \mathcal{Z}), \qquad \overline{b} \equiv \frac{\mathbb{E}(qb \mid \mathcal{Z})}{\mathbb{E}(q \mid \mathcal{Z})}, \qquad \overline{p} \equiv \frac{\mathbb{E}(qp \mid \mathcal{Z})}{\mathbb{E}(q \mid \mathcal{Z})}, \qquad \overline{h} \equiv \frac{k\overline{q}}{l}. \tag{17}$$

The quadratic structure of the model ensures that the latter three objects are sufficient statistics for Agent's uncertainty about Individual.

Proposition 4. *Suppose that, conditional on $\mathcal{Z}$, Agent has an arbitrary joint distribution over (p, b, α, γ) with finite first and second moments. Then the affine equilibrium is equivalent to the baseline equilibrium after replacing h, b, and p with $\overline{h}$, $\overline{b}$, and $\overline{p}$, respectively. In particular, the coefficient on θ solves*

$$(\overline{A} - 1)(\overline{A}^2 + \rho)^2 + \overline{hA}^3 = 0,$$

and

$$m = \overline{b} + \overline{A}(\theta - \overline{b}) + \overline{D}(\overline{p} - \overline{b}),$$

[6] If Individual did not know the conditional distribution of characteristics that Agent operates under, she would partly infer it from the advice she gets, complicating her inference problem. We leave this possibility (Individual learning, from interacting with Agent, what it knows about her) for future research.

where

$$\overline{D} = \sqrt{\overline{h}\left(\frac{1}{\overline{A}} - 1\right)}.$$

If $\rho \geq \kappa$, the affine equilibrium is unique.

Proposition 4 shows that uncertainty about Individual's characteristics changes the effective type to which Agent responds, but not the structure of equilibrium communication. Importantly, the effective prior and political preference are generally not ordinary conditional means. Indeed, in our notation

$$\overline{p} = \mathbb{E}(p \mid \mathcal{Z}) + \frac{\mathrm{Cov}(q, p \mid \mathcal{Z})}{\mathbb{E}(q \mid \mathcal{Z})},$$
$$\overline{b} = \mathbb{E}(b \mid \mathcal{Z}) + \frac{\mathrm{Cov}(q, b \mid \mathcal{Z})}{\mathbb{E}(q \mid \mathcal{Z})}.$$

Thus, correlation matters. Agent places greater effective weight on possible types for whom changing beliefs has a larger effect on perceived utility. For example, if higher values of p are associated with higher values of q, i.e., individuals with preferences on the right happen to be those for whom manipulation of beliefs has a stronger effect on perceived utility, then $\overline{p}$ exceeds the ordinary conditional mean of p, and Agent behaves as if Individual's political preference were correspondingly higher. The same logic applies to her prior belief b. If q is conditionally uncorrelated with p and b, the effective characteristics reduce to their ordinary conditional means. More generally, arbitrary dependence among Individual's characteristics changes the target of personalization but leaves the main forces of the model unchanged.

We conclude by noting that the same analysis applies when Agent communicates a common message simultaneously to a group of Individuals. In that case, $\mathcal{Z}$ would denote the distribution of characteristics within the group, and all results above continue to hold with the corresponding effective parameters.

4 Sophistication and Naivete

We ended the previous analysis by showing that our results go through almost intact if Agent is uncertain about the characteristics of the individual it advises. We assumed, however, that the individual remains sophisticated. Here, we relax this assumption and suppose that Agent may not know whether Individual is sophisticated or naive. Suppose that Individual is sophisticated with probability $\lambda \in [0,1]$ and naive with probability $1-\lambda$, and Agent knows λ. A sophisticated Individual correctly understands Agent's incentives and equilibrium strategy; in particular, she understands that Agent is itself uncertain about her sophistication and its strategy reflects this uncertainty. In contrast, a naive Individual believes, as in Section 3.1, that Agent reports the state truthfully. For simplicity, we assume that Agent knows Individual's other characteristics p, b, α, γ; the results of this Section are robust to allowing independent uncertainty about those characteristics, following the analysis of Section 3.4.

Suppose Agent uses an affine strategy $m = A^\lambda \theta + B^\lambda$. If Individual is naive, she updates according to (3). If she is sophisticated, she updates according to (8), with A and B replaced by A^λ and B^λ. Agent anticipates these two different interpretations of the same message and weights the corresponding expected utilities by $1-\lambda$ and λ, respectively. Substituting the two updating rules into Agent's objective and solving its first-order condition gives the following characterization.

Proposition 5. *Suppose that $\rho \geq \kappa$. Define*

$$s^\lambda \equiv 1 + \frac{h(1-\lambda)}{(1+\rho)^2} \geq 1, \qquad h^\lambda \equiv \lambda h s^\lambda, \qquad \rho^\lambda \equiv \rho \left(s^\lambda\right)^2.$$

Then there is a unique affine equilibrium, and it can be written as

$$m = b + A^\lambda(\theta - b) + D^\lambda(p - b), \tag{18}$$

where A^λ is the unique solution to

$$\left(s^\lambda A^\lambda - 1\right)\left(\left(s^\lambda A^\lambda\right)^2 + \rho^\lambda\right)^2 + h^\lambda\left(s^\lambda A^\lambda\right)^3 = 0 \tag{19}$$

and

$$D^\lambda = \frac{1}{s^\lambda}\left(\sqrt{\lambda h\left(\frac{1}{A^\lambda} - s^\lambda\right)} + (s^\lambda - 1)(1+\rho)\right). \tag{20}$$

Notice that existence and uniqueness of equilibrium follow from Proposition 2, as equation (19) has the same form as (12) after replacing A with $s^\lambda A^\lambda$, h with h^λ, and ρ with ρ^λ, and $\rho \geq \kappa$ implies $\rho^\lambda \geq \kappa$. As expected, for $\lambda = 1$, we have $s^\lambda = 1$, $h^\lambda = h$, and $\rho^\lambda = \rho$, and the equilibrium reduces to the sophisticated equilibrium characterized in Section 3.2. Likewise, if $\lambda = 0$, the equilibrium reduces to the naive benchmark characterized in Section 3.1.

The mixed-sophistication equilibrium therefore has the same canonical form as the equilibria studied above: A^λ measures the responsiveness of Agent's message to the true state, while D^λ measures its political loading. What is new is that both coefficients now depend on λ, the probability that Individual is sophisticated. This raises a natural question: how does a higher share of sophisticated individuals affect Agent's communication and Individual's learning?

As it turns out, perhaps surprisingly, sophistication, while benefitting the sophisticated individual directly, imposes a negative externality across Individuals. Sophisticated Individuals are better able to interpret a given message, but their presence changes the message Agent chooses to send. In particular, a greater share of sophisticated Individuals results in a message depending more on Individual's political preferences, thereby increasing the political contamination faced by naive Individuals. Furthermore, even sophisticated individuals are affected by the externality, as an increase in their share reduces the informativeness of the message, provided that the noise is not too high. The following proposition formalizes these effects.

Proposition 6. *The political loading of Agent's message D^λ is increasing in the share of*

sophisticated individuals λ. The informativeness of Agent's message A^λ is increasing in λ if and only if $h > \frac{(1-\rho)(1+\rho)^2}{\rho}$. Furthermore, $A^\lambda + D^\lambda$ is increasing in λ.

The intuition for the first result is straightforward. Sophisticated Individuals anticipate and discount predictable pandering, so Agent must tilt its message more strongly toward p to affect their perceived utility. Because Agent cannot condition its message on sophistication, naive Individuals are exposed to this stronger political loading as well.

The effect of λ on informativeness is easiest to understand if we recall that the same condition was present in the comparative statics in Proposition 3. A higher share of sophisticated individuals makes communication more informative exactly when communication with sophisticated individuals is more informative than that with naive ones. However, for parameter values where communication is not too noisy and Agent is interested in truth more than in pandering, a higher share of sophisticated individuals implies lower informativeness. Thus, ironically, while naive individuals suffer from pandering the most, their presence disciplines Agent's messaging and benefits sophisticated individuals. To put it in perspective: the worst-case scenario for learning is to be the single naive individual among sophisticated; the best case is to be sophisticated when Agent believes it is talking to a naive individual.

We next turn to how the two types interpret the resulting message. A sophisticated Individual correctly accounts for Agent's equilibrium strategy. As before, the political component of the message is predictable and therefore does not enter her posterior belief. Substituting the equilibrium strategy into (8), her expected posterior mean is

$$\mathbb{E}\left(\widehat{\theta} \mid \theta, b, p\right) = b + \frac{(A^\lambda)^2}{(A^\lambda)^2 + \rho}(\theta - b).$$

A naive Individual instead updates according to (3). Substituting the actual equilibrium message (18), we obtain

$$\mathbb{E}\left(\widetilde{\theta} \mid \theta, b, p\right) = b + \frac{A^\lambda}{1+\rho}(\theta - b) + \frac{D^\lambda}{1+\rho}(p - b). \tag{21}$$

The difference between the two updating rules is particularly transparent if we consider separately disagreements about prior beliefs and disagreements about political preferences.

First, suppose that two otherwise identical individuals have the same political preference p, but different prior beliefs b_H and b_L. Sophisticated individuals are necessarily brought closer together:

$$\mathbb{E}\left(\widehat{\theta}_H - \widehat{\theta}_L \mid \theta, b_H, b_L\right) = \frac{\rho}{(A^\lambda)^2 + \rho}(b_H - b_L),$$

and each expected posterior mean lies between the corresponding prior and the true state. Thus, communication both reduces their disagreement and moves each of them toward the truth.

For naive individuals, in contrast,

$$\mathbb{E}\left(\widetilde{\theta}_H - \widetilde{\theta}_L \mid \theta, b_H, b_L, p\right) = \left(1 - \frac{A^\lambda + D^\lambda}{1+\rho}\right)(b_H - b_L).$$

If λ is close to 0, so most individuals are naive, then $A^\lambda + D^\lambda < 1 + \rho$, the first factor is between 0 and 1, and the individuals' beliefs converge, preserving the order. As the share of sophisticated individuals λ increases, so does $A^\lambda + D^\lambda$, and Agent may overcompensate the priors for the sophisticated part of the audience so much that beliefs of naive individuals flip (the first factor becomes negative). In principle, if h is sufficiently high, D^λ may be arbitrarily high and the beliefs may become both flipped and polarized; however, if pandering motives are limited (for example, if $h < 1/4$), then $D^\lambda < 1$ and then there is convergence in beliefs of naive types as well. In other words, if pandering motives are moderate, there is a decrease in polarization for all types, even though the average need not move towards the truth.

Second, suppose that two individuals have the same prior belief b, but different political preferences $p_H \neq p_L$. Sophisticated individuals retain identical factual beliefs after communication: their expected posteriors are independent of p, and both move toward the true

state. Naive individuals instead become factually polarized. From (21),

$$\mathbb{E}\left(\widetilde{\theta}_H - \widetilde{\theta}_L \mid \theta, b, p_H, p_L\right) = \frac{D^\lambda}{1+\rho}(p_H - p_L).$$

Individuals who initially agreed about the state therefore come to disagree about it solely because they disagree about what they would like to be true, and thus political disagreement becomes factual disagreement. Thus, disagreement in priors is necessarily attenuated among sophisticated individuals and, when pandering motives are moderate, among naive individuals as well. In contrast, disagreement in political preferences creates factual disagreement among naive individuals.

We now compare the messaging and learning of individuals whose political preferences coincide with their prior belief ($p = b$) with those for whom it is different. Preferences and priors may coincide for various reasons; it may be that the individual has a prior and likes to hear evidence that reinforces it; conversely, it might be that an individual treats her political preference as strong evidence about θ, effectively making it her prior. At first glance, such individuals might seem especially vulnerable to echo chambers, since Agent knows exactly what they would like to hear. The equilibrium analysis, however, tells a different story.

If $p = b$, then the political component in (21) vanishes, and

$$\mathbb{E}\left(\widetilde{\theta} \mid \theta, b, p\right) = b + \frac{A^\lambda}{1+\rho}(\theta - b),$$

which lies between b and the true state θ. In other words, Agent's advice moves her expected belief toward the truth. If all individuals in the model had $p = b$, then, unambiguously, political advice by Agent would be a source of learning and depolarization, and the only question would be about the extent of convergence.

In contrast, when Individual's political preferences differ from her prior, political personalization may work against learning. For example, suppose that $\theta < b < p$. The informational component of Agent's message pulls Individual's belief downward toward the truth, whereas

political personalization pushes it upward toward p. Communication moves her expected belief farther from the truth if and only if

$$D^{\lambda}(p - b) > A^{\lambda}(b - \theta).$$

Thus, sufficiently strong political personalization can overwhelm genuine information and radicalize a naive individual, even though a sophisticated individual facing the same communication continues to learn. This implies that the source of polarization and radicalization in this model of political advice is not people seeking confirmation of their beliefs, but rather those who are torn between what they know and what they would prefer the truth to be. One might expect such individuals to have a higher propensity to seek advice from AI, exacerbating the problem.

5 Discussion

The model suggests that the consequences of political personalization depend not only on how strongly Agent seeks to please Individual, but also on what Agent knows about Individual and what it expects about how its advice will be interpreted. These forces interact with Individual's characteristics in shaping both the content and the informativeness of communication. These interactions matter for the design and evaluation of AI systems, because interventions that appear to make advice more useful or users better equipped to interpret it can also change Agent's incentives.

Incentives and commitment. The most direct way to address the distortion is to eliminate Agent's incentive to pander to Individual. As k approaches zero, equilibrium converges to truthful communication, $m = \theta$. This observation leaves aside whether people would choose to use an AI that does not pander to them. Interestingly, truthful reporting is not something that Agent would choose as a communication strategy if it could commit ex ante, before learning the state θ. Assuming a fully sophisticated audience, we can show that in the case

with commitment, the optimal affine rule takes the form $m = b + A^c(\theta - b)$, where A^c is the unique positive solution of

$$(A^c - 1)\left((A^c)^2 + \rho\right)^2 = h\rho A^c.$$

In particular, the message does not depend on Individual's p: this dependence would be undone anyway and is costly for Agent. Interestingly, $A^c > 1$: a committed Agent amplifies variation in the state to overcome communication noise, whereas an Agent without commitment suppresses it, providing a useful benchmark for the role of commitment.

Information, sophistication, and communication technology. The model identifies two characteristics of Individual that discipline political pandering. First, better pre-existing information makes Individual harder to influence: a lower σ^2 increases ρ and makes messages more informative and less responsive to Individual's political preferences. Thus, an Individual who is known to have access to independent factual information gives Agent less scope to manipulate beliefs. A similar logic implies that if Agent perceives individuals to be highly confident and perhaps stubborn in their prior beliefs, its incentive to manipulate would decrease, whereas leading it to believe that people have little information would increase its incentive to manipulate them.

Second, perhaps more surprisingly, Agent's belief that Individual may be naive can discipline political pandering in the message itself. A sophisticated Individual can discount predictable political pandering, so Agent has a stronger incentive to load its message on political preferences when it expects a sophisticated audience. Naive Individuals therefore discipline Agent's pandering even though they themselves are least able to interpret it correctly. From the standpoint of limiting pandering, an Individual may thus benefit from appearing less sophisticated but better informed. More generally, an AI system should not presume that users will correctly undo strategic or personalized distortions. Designing communication as if factual claims may be taken at face value disciplines the message before

interpretation takes place.

The same incentive effect explains why improving communication technology, for example, increasing the clarity of explanations or making them easier for a given user to understand, need not improve learning. Lower communication noise makes any given message more informative, but also makes manipulation more effective. Agent may respond by reducing the informativeness of its message, so greater precision can offset or even reverse its direct informational benefit. Improving the precision or seamlessness of AI communication is therefore not a substitute for disciplining the incentives that determine what information is transmitted.

Beliefs versus preferences. A central distinction in the model is between what Individual believes and what she would like to be true. When $p = b$, political personalization does not create an additional directional distortion in naive updating, and expected beliefs then move toward the true state. By contrast, when p and b differ, Agent's attempt to pander to Individual's preferences can pull factual beliefs away from the truth and, if sufficiently strong, overwhelm genuine information it attempts to provide.

This distinction suggests that the problem is not personalization per se. An AI system may appropriately use an individual's prior beliefs to determine what requires explanation, clarification, or correction. What is dangerous is allowing what Individual would prefer to be true to affect the factual content of the answer. In this sense, AI should be able to respond to beliefs without treating political preferences as evidence. Safeguards aimed only at preventing conventional echo chambers may therefore miss the cases in which pandering is most damaging: those in which what Individual believes and what she would like to believe point in different directions.

Measuring AI pandering and informativeness. The model also suggests a way to evaluate AI political advice empirically. Real systems do not come with observable values of A, D, h, or ρ, and documenting that responses vary with users' politics does not by itself reveal how informative those responses are. Directly identifying A by varying the

underlying state θ is generally difficult. The model implies that this is unnecessary. One can instead construct synthetic individuals: otherwise identical users whose stated political preferences p and prior beliefs b are varied experimentally while the substantive question and other characteristics are held fixed.

To see why this is sufficient, recall that in the baseline equilibrium,

$$m = A\theta + Dp + (1 - A - D)b.$$

The experimenter does not need to manipulate the true state θ. By varying p, it is possible to identify $D = \frac{\partial m}{\partial p}$, while variation in b allows one to measure $C \equiv \frac{\partial m}{\partial b} = 1 - A - D$. From this, A is found from $A = 1 - D - C$. In other words, an experiment that varies both Individual's prior beliefs and what she would like to believe allows one to recover Agent's responsiveness to the state even when the state itself cannot be experimentally manipulated.

The same coefficients also allow us to recover h and ρ, with an as-if interpretation. From the equilibrium conditions we can find

$$h = \frac{AD^2}{1 - A}, \qquad \rho = A^2 \left(\frac{D}{1 - A} - 1 \right).$$

These values need not be interpreted as literal primitives represented internally by an AI system. Rather, they provide a compact empirical characterization of its behavior and a way to compare how different systems or design choices trade off truthful information transmission against pandering.

6 Conclusion

Artificial intelligence creates a new environment for political advice: a single system can be better informed than the individuals who consult it, communicate separately with each of them, and adapt what it says to what each would like to hear. This paper studies the result-

ing tension between information transmission and political accommodation. Accommodation affects both the direction of AI messages and how much information they contain. Sophisticated individuals anticipate predictable political loading and filter it out, but they cannot recover information that Agent chose not to convey. Naive individuals instead mistake part of the political component for evidence, allowing political preferences to enter factual beliefs and potentially turning political disagreement into factual disagreement, polarization, and radicalization.

These effects interact in unexpected ways when sophisticated and naive individuals coexist. A larger sophisticated share induces Agent to place greater political loading in the raw message and, in an important region of parameters, to make communication less responsive to the true state. Thus, sophistication protects those who possess it while changing communication in ways that can harm others. Better-informed individuals, by contrast, induce more truthful advice because their beliefs are harder to manipulate. The consequences of personalization also depend critically on the relation between beliefs and preferences. When what individuals would like to believe coincides with what they already believe, accommodation does not distort the direction of naive updating. The greater danger arises when desired and currently held beliefs diverge, because personalization can then make preferences look like evidence.

Taken together, these results suggest that evaluating AI political advice requires attention not only to whether its outputs are politically responsive, but also to how much information they contain and how different individuals interpret them. The model also provides a way to bring these objects to data: synthetic individuals who differ in their stated beliefs and political preferences can be used to recover the system's responsiveness to the state and to political preferences and to test the model's restrictions. Future work can extend this framework to competing advisers, repeated interaction, and the diffusion of personalized advice beyond its intended recipient, where additional strategic effects and spillovers are likely to arise.

References

Acemoglu, Daron, Asuman Ozdaglar and James Siderius. 2025. AI and Social Media: A Political Economy Perspective. In *The Political Economy of Artificial Intelligence*, ed. Ajay Agrawal, Joshua Gans, Avi Goldfarb and Catherine Tucker. National Bureau of Economic Research pp. 167–206. NBER Working Paper 33892.

Acemoglu, Daron, Georgy Egorov and Konstantin Sonin. 2013. "A Political Theory of Populism." *The Quarterly Journal of Economics* 128(2):771–805.

Acemoglu, Daron, Georgy Egorov and Konstantin Sonin. 2026. Multidimensional Signaling and the Rise of Cultural Politics. NBER Working Paper 34909 National Bureau of Economic Research.

Acemoglu, Daron, Tianyi Lin, Asuman Ozdaglar and James Siderius. 2026. How AI Aggregation Affects Knowledge. NBER Working Paper 35036 National Bureau of Economic Research.

Andreoni, James and B. Douglas Bernheim. 2009. "Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica* 77(5):1607–1636.

Bail, Christopher A. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press.

Bakshy, Eytan, Solomon Messing and Lada A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348(6239):1130–1132.

Bénabou, Roland and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96(5):1652–1678.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *Journal of Political Economy* 102(5):841–877.

Bick, Alexander, Adam Blandin and David J. Deming. 2024. The Rapid Adoption of Generative AI. NBER Working Paper 32966 National Bureau of Economic Research.

Blume, Andreas, Oliver J. Board and Kohei Kawamura. 2007. "Noisy Talk." *Theoretical Economics* 2(4):395–440.

Broockman, David and Joshua Kalla. 2016. "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing." *Science* 352(6282):220–224.

Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao and Christopher Roth. 2023. "Justifying Dissent." *The Quarterly Journal of Economics* 138(3):1403–1451.

Bursztyn, Leonardo, Georgy Egorov and Stefano Fiorin. 2020. "From Extreme to Mainstream: The Erosion of Social Norms." *American Economic Review* 110(11):3522–3548.

Chandra, Kartik, Max Kleiman-Weiner, Jonathan Ragan-Kelley and Joshua B. Tenenbaum. 2026. "Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians." *arXiv preprint arXiv:2602.19141* .

Chandrasekhar, Arun G., Horacio Larreguy and Juan Pablo Xandri. 2020. "Testing Models of Social Learning on Networks: Evidence From Two Experiments." *Econometrica* 88(1):1–32.

Conlon, John and Peter Schwardmann. 2026. AI Sycophancy and Decisions. Discussion Paper 21505 CEPR.

Costello, Thomas H., Gordon Pennycook and David G. Rand. 2024. "Durably Reducing Conspiracy Beliefs through Dialogues with AI." *Science* 385(6714):eadq1814.

Crawford, Vincent P. and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50(6):1431–1451.

Dandekar, Pranav, Ashish Goel and David T. Lee. 2013. "Biased Assimilation, Homophily, and the Dynamics of Polarization." *Proceedings of the National Academy of Sciences* 110(15):5791–5796.

DellaVigna, Stefano and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics* 122(3):1187–1234.

DellaVigna, Stefano and Matthew Gentzkow. 2010. "Persuasion: Empirical Evidence." *Annual Review of Economics* 2:643–669.

Egorov, Georgy, Sergei Guriev, Maxim Mironov and Ekaterina Zhuravskaya. 2026. "Political Information and Network Effects." *Journal of the European Economic Association* 24(1):82–121.

Enikolopov, Ruben, Maria Petrova and Ekaterina Zhuravskaya. 2011. "Media and Political Persuasion: Evidence from Russia." *American Economic Review* 101(7):3253–3285.

Faverio, Michelle and Emma Kikuchi. 2026. "Key Findings about How Americans View Artificial Intelligence." Pew Research Center. Published March 12, 2026.

Gentzkow, Matthew and Jesse M. Shapiro. 2006. "Media Bias and Reputation." *Journal of Political Economy* 114(2):280–316.

Gentzkow, Matthew and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78(1):35–71.

Gentzkow, Matthew, Jesse M. Shapiro and Daniel F. Stone. 2015. Media Bias in the Marketplace: Theory. In *Handbook of Media Economics*, ed. Simon P. Anderson, Joel Waldfogel and David Strömberg. Vol. 1B Amsterdam: Elsevier chapter 14, pp. 623–645.

Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94(3):653–663.

Hackenburg, Kobi, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand and Christopher Summerfield. 2025. "The Levers of Political Persuasion with Conversational Artificial Intelligence." *Science* 390(6777):eaea3884.

Hackenburg, Kobi, Caroline Wagner, Luke Hewitt, Ben M. Tappin, Ed Saunders, Hannah Rose Kirk, Helen Margetts and Christopher Summerfield. 2026. "AI Systems Out-Persuade Expert Humans.".

Kalla, Joshua L. and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112(1):148–166.

Kartik, Navin. 2009. "Strategic Communication with Lying Costs." *The Review of Economic Studies* 76(4):1359–1395.

Kartik, Navin, Marco Ottaviani and Francesco Squintani. 2007. "Credulity, Lies, and Costly Talk." *Journal of Economic Theory* 134(1):93–116.

Kearney, Matthew, Reuben Binns and Yarin Gal. 2025. "Language Models Change Facts Based on the Way You Talk." arXiv preprint arXiv:2507.14238.

Levy, Gilat and Ronny Razin. 2019. "Echo Chambers and Their Effects on Economic and Political Outcomes." *Annual Review of Economics* 11:303–328.

Lin, Hause, Gabriela Czarnek, Benjamin Lewis, Joshua P. White, Adam J. Berinsky, Thomas Costello, Gordon Pennycook and David G. Rand. 2025. "Persuading Voters Using Human–Artificial Intelligence Dialogues." *Nature* 648(8093):394–401.

Lira, Benjamin, Noah Castelo, Stefano Puntoni and Olivier Toubia. 2026. "Synthetic Contact with AI Reduces Cross-Partisan Animosity." SSRN Working Paper No. 7042278.

Luettgau, Lennart, Hannah Rose Kirk, Kobi Hackenburg, Jessica Bergs, Henry Davidson, Henry Ogden, Divya Siddarth, Saffron Huang and Christopher Summerfield. 2025. "Conversational AI Increases Political Knowledge as Effectively as Self-Directed Internet Search." arXiv preprint arXiv:2509.05219.

Martin, Gregory J. and Ali Yurukoglu. 2017. "Bias in Cable News: Persuasion and Polarization." *American Economic Review* 107(9):2565–2599.

Marvel, John and Sangwon Ju. 2026. "Inoculating Citizens Against Sycophancy in Large Language Models." Preprint.

Miyazaki, Sho and Andrew B. Hall. 2026. "Why Do AI Models Tell Left-Wing Voters to Support the Communist Party? AI Voting Advice in Japan's 2026 General Election." Working paper, Stanford Graduate School of Business.

Mullainathan, Sendhil and Andrei Shleifer. 2005. "The Market for News." *American Economic Review* 95(4):1031–1053.

Perez, Ethan, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer and Jared Kaplan. 2023. Discovering Language Model

Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics pp. 13387–13434.

Prendergast, Canice. 1993. "A Theory of "Yes Men"." *American Economic Review* 83(4):757–770.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202 of *Proceedings of Machine Learning Research* PMLR pp. 29971–30004.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations*.

Spence, Michael. 1973. "Job Market Signaling." *The Quarterly Journal of Economics* 87(3):355–374.

Taber, Charles S. and Milton Lodge. 2006. "Motivated Skepticism in the Evaluation of Political Beliefs." *American Journal of Political Science* 50(3):755–769.

Törnberg, Petter and Michelle Schimmel. 2026. "Political Bias Audits of LLMs Capture Sycophancy to the Inferred Auditor." arXiv preprint arXiv:2604.27633.

Velez, Yamil R., Donald P. Green and Semra Sevi. 2025. "Chatbot Voting Advice Applications inform but seldom sway young unaligned voters." *Proceedings of the National Academy of Sciences* 122(50):e2515516122.

Wei, Jerry, Da Huang, Yifeng Lu, Denny Zhou and Quoc V. Le. 2023. "Simple Synthetic Data Reduces Sycophancy in Large Language Models." *arXiv preprint arXiv:2308.03958* .

Zhang, Jiding, Lin Hu, Shuang Gao, Katsiaryna Siamionava and Pei-Yu Chen. 2026. "AI-Mediated Information Design: Belief Updating, Strategic Disclosure, and Persuasion." SSRN Working Paper No. 6811961.

Zu, Ziwen. 2026. "Talking Politics with Artificial Intelligence.".

Appendix

Lemma A1. *Let $h > 0$ and $\rho > 0$ be parameters, and define*

$$f(a) = (a-1)(a^2+\rho)^2 + ha^3.$$

Then:

(a) $f(a) = 0$ has at least one and at most three real solutions. All real solutions lie in $(0,1)$.

(b) If

$$\rho \geq \kappa \equiv \frac{47 - 13\sqrt{13}}{8} \approx 0.016,$$

then the solution is unique.

Furthermore, if the solution is unique, denote it by $a^*(\rho, h)$. Then:

(c) $a^*(\rho, h)$ is strictly increasing in ρ and strictly decreasing in h.

(d)

$$\sqrt{h\left(\frac{1}{a^*(\rho,h)} - 1\right)}$$

is strictly decreasing in ρ and strictly increasing in h.

(e)

$$\sqrt{h\left(\frac{1}{a^*(\rho,h)} - 1\right)} > 1$$

if and only if

$$\rho < \frac{h^3}{(1+h)^2}.$$

Proof of Lemma A1. For $a \in (0,1)$, define

$$G(a;\rho) = \frac{(1-a)(a^2+\rho)^2}{a^3}.$$

Then

$$f(a) = 0 \quad \Longleftrightarrow \quad G(a;\rho) = h.$$

Part (a). For $a \leq 0$,

$$f(a) < 0,$$

whereas for $a \geq 1$,

$$f(a) > 0.$$

Hence every real solution lies in $(0,1)$, and by continuity at least one solution exists.

To bound the number of solutions, differentiate G:

$$\frac{\partial G(a;\rho)}{\partial a} = \frac{a^2+\rho}{a^4}\Phi(a;\rho),$$

where

$$\Phi(a;\rho) \equiv -2a^3 + a^2 + 2a\rho - 3\rho.$$

Its derivative is

$$\frac{\partial \Phi(a;\rho)}{\partial a} = -6a^2 + 2a + 2\rho,$$

which has one positive root. Hence $\Phi(\cdot;\rho)$ first increases and then decreases in a on $(0,+\infty)$, and therefore has at most two zeros on $(0,1)$. Consequently, $G(\cdot;\rho)$ is strictly monotone on at most three intervals, which implies that the equation

$$G(a;\rho) = h$$

has at most three solutions.

Part (b). Suppose, to obtain a contradiction, that $f(a) = 0$ has more than one solution on $(0,1)$; then so does $G(a;\rho) = h$, implying that $G(a;\rho)$ has a local extremum on $(0,1)$, at which its derivative, and thus $\Phi(a;\rho)$ changes the sign. Thus, there exists $a \in (0,1)$ with $\Phi(a;\rho) > 0$. Since $\Phi(0;\rho) = -3\rho < 0$ and $\Phi(1;\rho) = -1-\rho < 0$, $\Phi(\cdot;\rho)$ must attain a

positive maximum at some $v \in (0,1)$. At this maximum,

$$-6v^2 + 2v + 2\rho = 0,$$

so

$$\rho = 3v^2 - v.$$

Since $\rho > 0$, it must be that $v > 1/3$. Substituting ρ, we get

$$\Phi(v;\rho) = v(4v^2 - 10v + 3).$$

Given that $v < 1$ and $\Phi(v;\rho) > 0$, we must have

$$v < \frac{5-\sqrt{13}}{4}.$$

Now, since $3v^2 - v$ is strictly increasing for $v > 1/6$, it follows that

$$\rho < 3\left(\frac{5-\sqrt{13}}{4}\right)^2 - \frac{5-\sqrt{13}}{4} = \frac{47-13\sqrt{13}}{8} = \kappa.$$

However, we assumed $\rho \geq \kappa$. This contradiction completes the proof.

Part (c). For the remaining parts, we note that at the unique root $a^*(\rho,h)$, $f(a)$ changes sign from negative to positive.

Fix h, and consider $\rho_1 < \rho_2$. For every $a \in (0,1)$,

$$\frac{\partial f(a;\rho,h)}{\partial \rho} = 2(a-1)(a^2+\rho) < 0.$$

Therefore,

$$f(a^*(\rho_2,h);\rho_1,h) > f(a^*(\rho_2,h);\rho_2,h) = 0.$$

But this implies that $a^*(\rho_1,h) < a^*(\rho_2,h)$, which proves that $a^*(\rho,h)$ is strictly increasing in

ρ.

Similarly, fix ρ and consider $h_1 < h_2$. Since for $a \in (0,1)$,

$$\frac{\partial f(a;\rho,h)}{\partial h} = a^3 > 0,$$

we have

$$f\big(a^*(\rho,h_2);\rho,h_1\big) < f\big(a^*(\rho,h_2);\rho,h_2\big) = 0.$$

This implies $a^*(\rho,h_2) < a^*(\rho,h_1)$, hence $a^*(\rho,h)$ is strictly decreasing in h.

Part (d). Since a is strictly increasing in ρ by part (c) and

$$\sqrt{h\left(\frac{1}{a}-1\right)}$$

is strictly decreasing in a on $(0,1)$, it follows immediately that

$$\sqrt{h\left(\frac{1}{a^*(\rho,h)}-1\right)}$$

is strictly decreasing in ρ.

Now fix ρ. As h increases, part (c) implies that $a^*(\rho,h)$ strictly decreases. Hence both factors under the square root strictly increase in h, and given that they are positive, so does the function.

Part (e). Simple algebra shows that

$$\sqrt{h\left(\frac{1}{a^*(\rho,h)}-1\right)} > 1$$

if and only if $a^*(\rho,h) < \frac{h}{1+h}$. This holds if and only if $f\left(\frac{h}{1+h}\right) > 0$. Direct substitution and factorization yield

$$f\left(\frac{h}{1+h}\right) = \frac{(h^3-\rho(1+h)^2)\,(h^2(h+2)+\rho(1+h)^2)}{(1+h)^5}.$$

The second factor in the numerator is strictly positive, as is the denominator. Thus,

$$f\left(\frac{h}{1+h}\right) > 0 \quad \Longleftrightarrow \quad h^3 - \rho(1+h)^2 > 0,$$

which proves part (e). ■

Proof of Proposition 1. For a naive Individual, Agent solves (5). The objective is strictly concave in m, so the first-order condition is necessary and sufficient:

$$2(m-\theta) + \frac{2h}{1+\rho}\left(\frac{m+\rho b}{1+\rho} - p\right) = 0.$$

Equivalently,

$$\big((1+\rho)^2 + h\big)m = (1+\rho)^2\theta + h\big((1+\rho)p - \rho b\big).$$

Solving for m yields (6), establishing existence and uniqueness. The coefficients on θ and p are strictly positive, while the coefficient on b is strictly negative, proving the stated comparative statics. ■

Proof of Proposition 2. In the text we established that

$$m(\theta) = A\theta + B$$

defines an equilibrium if and only if the equations (12) and (13) are satisfied. The properties of solution to (12) follow from Lemma A1.

Now, to get representation (14), we use (12) to obtain

$$1 - A = \frac{hA^3}{(A^2+\rho)^2},$$

and hence

$$D \equiv \sqrt{h\left(\frac{1}{A}-1\right)} = \sqrt{h\frac{1-A}{A}} = \sqrt{\frac{h^2A^2}{(A^2+\rho)^2}} = \frac{hA}{A^2+\rho}.$$

Moreover,

$$1-A-D = \frac{hA^3}{(A^2+\rho)^2} - \frac{hA}{A^2+\rho} = -\frac{hA\rho}{(A^2+\rho)^2}.$$

Therefore,

$$B = Dp + (1-A-D)b.$$

Plugging it into (7) we get (14). Since $A \in (0,1)$, $D > 0$, $1-A-D < 0$ and none of these depend on θ, b, or p, the comparative statics follow immediately. ■

Proof of Proposition 3. By part (c) of Lemma A1, in the unique equilibrium A is decreasing in h and increasing in ρ. Since $A \in (0,1)$, communication is always less informative than truthful communication, which corresponds to $A = 1$.

Per Proposition 1, in the naive benchmark, the coefficient on θ is

$$\frac{(1+\rho)^2}{(1+\rho)^2+h}.$$

The inequality $A > \frac{(1+\rho)^2}{(1+\rho)^2+h}$ is equivalent to $hA > (1-A)(1+\rho)^2$, and using (12) and simplifying we find it is equivalent to $A < \rho$. This, in turn, is equivalent to $f(\rho) > 0$, which simplifies to

$$h > \frac{(1-\rho)(1+\rho)^2}{\rho}.$$

The left-hand side is increasing in h, whereas the right-hand side is decreasing in ρ; this proves that this condition is more likely to hold when either of the variables is higher.

Now consider Individual's posterior variance:

$$V = \frac{\sigma^2 \rho}{A^2 + \rho} = \frac{\nu^2}{A^2 + \rho}.$$

Holding ρ fixed, a higher h lowers A and therefore raises V. Holding ν^2 fixed, a higher σ^2 lowers $\rho = \nu^2/\sigma^2$, which by part (c) of Lemma A1 lowers A. Hence $A^2 + \rho$ falls and V rises.

Lastly, fix σ^2; then increasing ν^2 is equivalent to increasing ρ. Using (12), we have

$$\frac{V}{\sigma^2} = \frac{\rho}{A^2 + \rho} = 1 - \sqrt{\frac{A(1-A)}{h}}.$$

Since A is increasing in ρ, the right-hand side is decreasing in ρ when $A < 1/2$ and increasing when $A > 1/2$. We have $A < 1/2$ whenever $f(1/2) > 0$, which is equivalent to $\rho < \rho^*$, where

$$\rho^* = \frac{\sqrt{h}}{2} - \frac{1}{4}.$$

This completes the proof. ■

Proof of Proposition 4. Fix Agent's information $\mathcal{Z}$ and conjecture an affine strategy

$$m(\theta) = \overline{A}\theta + \overline{B}.$$

For an Individual with characteristics (p, b, α, γ), the expected posterior belief, conditional on the message m chosen by Agent and before communication noise is realized, is

$$\mathbb{E}_\varepsilon\left(\widehat{\theta}\right) = \frac{\overline{A}}{\overline{A}^2 + \rho}(m - \overline{B}) + \frac{\rho}{\overline{A}^2 + \rho} b.$$

As in the baseline model, terms involving posterior variance do not depend on the particular message m. Hence, after dropping terms independent of m and dividing by l, Agent's

objective is equivalent to

$$-(m-\theta)^2 - \frac{k}{l}\mathbb{E}\left(q\left(\frac{\overline{A}}{\overline{A}^2+\rho}(m-\overline{B}) + \frac{\rho}{\overline{A}^2+\rho}b - p \right)^2 \middle| \mathcal{Z} \right).$$

Expanding the square and collecting the terms that depend on m, we obtain

$$\begin{aligned}
&\mathbb{E}\left(q\left(\frac{\overline{A}}{\overline{A}^2+\rho}(m-\overline{B}) + \frac{\rho}{\overline{A}^2+\rho}b - p \right)^2 \middle| \mathcal{Z} \right) \\
&\quad = \overline{q}\left(\frac{\overline{A}}{\overline{A}^2+\rho}(m-\overline{B}) \right)^2 \\
&\qquad + 2\frac{\overline{A}}{\overline{A}^2+\rho}(m-\overline{B})\mathbb{E}\left(q\left(\frac{\rho}{\overline{A}^2+\rho}b - p \right) \middle| \mathcal{Z} \right) + C,
\end{aligned}$$

where C does not depend on m. By the definitions in (17),

$$\mathbb{E}\left(q\left(\frac{\rho}{\overline{A}^2+\rho}b - p \right) \middle| \mathcal{Z} \right) = \overline{q}\left(\frac{\rho}{\overline{A}^2+\rho}\overline{b} - \overline{p} \right).$$

Therefore, completing the square and again dropping terms independent of m, Agent's objective is equivalent to

$$-(m-\theta)^2 - \overline{h}\left(\frac{\overline{A}}{\overline{A}^2+\rho}(m-\overline{B}) + \frac{\rho}{\overline{A}^2+\rho}\overline{b} - \overline{p} \right)^2,$$

where

$$\overline{h} = \frac{k\overline{q}}{l}.$$

This is exactly the baseline problem with h, b, and p replaced by $\overline{h}$, $\overline{b}$, and $\overline{p}$, respectively.

It follows from Proposition 2 that $\overline{A}$ satisfies the stated equation and the equilibrium message is as stated. Finally, Proposition 2 implies uniqueness whenever $\rho \geq \kappa$. ■

Proof of Proposition 5. Take a putative equilibrium affine strategy $m = A\theta + B$, and

define

$$\beta^S \equiv \frac{A}{A^2+\rho}, \qquad \beta^N \equiv \frac{1}{1+\rho}.$$

Conditional on the message m chosen by Agent and before communication noise is realized, a sophisticated Individual's expected posterior is $\beta^S(m-B)+\frac{\rho}{A^2+\rho}b$, whereas a naive Individual's expected posterior is $\beta^N m+\frac{\rho}{1+\rho}b$. As in the preceding analysis, the terms generated by communication noise and posterior variance do not depend on the particular message m. Thus, after dropping terms independent of m and dividing by l, Agent's problem is equivalent to

$$\max_m \left\{ -(m-\theta)^2 - h\left(\lambda\left(\beta^S(m-B)+\frac{\rho}{A^2+\rho}b-p\right)^2+(1-\lambda)\left(\beta^N m+\frac{\rho}{1+\rho}b-p\right)^2\right)\right\}.$$

The objective is strictly concave in m. Taking the first-order condition, solving for m and equating the coefficient on θ in Agent's best response with A gives

$$\frac{1-A}{A} = h\left(\lambda\left(\frac{A}{A^2+\rho}\right)^2+(1-\lambda)\left(\frac{1}{1+\rho}\right)^2\right). \tag{A1}$$

Simple algebra verifies that it is equivalent to equation (19). Since $\rho^\lambda = \rho(s^\lambda)^2 \geq \rho \geq \kappa$, Lemma A1 implies that there is a unique solution. Now equating the intercepts allows us to derive the formula for D^λ, which completes the proof. ■

Proof of Proposition 6. Write $A = A^\lambda$, $D = D^\lambda$, and $s = s^\lambda$. The equilibrium condition for A is

$$F(A,\lambda) \equiv \frac{1-A}{A} - h\left(\lambda\frac{A^2}{(A^2+\rho)^2}+(1-\lambda)\frac{1}{(1+\rho)^2}\right) = 0.$$

Since

$$s = 1+\frac{h(1-\lambda)}{(1+\rho)^2}, \qquad \lambda h = h-(s-1)(1+\rho)^2,$$

rearranging the equilibrium condition gives

$$(1 - sA)(A^2 + \rho)^2 = \lambda h A^3.$$

Thus, defining $x = sA$, we have $0 < A \leq x \leq 1$.

At an equilibrium, $F_A = \frac{T}{A^2(A^2+\rho)}$, where $T = A^2(1 - 2x) + \rho(2x - 3)$. Notice that we have $T < 0$. Indeed, this is immediate if $x \geq 1/2$, whereas if $x < 1/2$, then $A \leq x$ implies

$$T \leq x^2(1 - 2x) - \rho(3 - 2x) = \Phi(x; \rho) \leq 0,$$

where the last inequality follows from Lemma A1(b). Apart from the knife-edge equality case, handled by continuity, the inequality is strict. Hence $F_A < 0$.

Next,

$$F_\lambda = -h \left(\frac{A^2}{(A^2 + \rho)^2} - \frac{1}{(1 + \rho)^2} \right).$$

Since

$$\frac{A}{A^2 + \rho} - \frac{1}{1 + \rho} = \frac{(1 - A)(A - \rho)}{(A^2 + \rho)(1 + \rho)},$$

we have $\operatorname{sign} F_\lambda = \operatorname{sign}(\rho - A)$. Therefore, by implicit differentiation, $\operatorname{sign} \frac{\partial A^\lambda}{\partial \lambda} = \operatorname{sign}(\rho - A^\lambda)$. For $\rho \geq 1$, we have $A^\lambda < 1 \leq \rho$, so $\partial A^\lambda / \partial \lambda > 0$; moreover, $\frac{(1-\rho)(1+\rho)^2}{\rho} \leq 0 < h$, so the stated condition holds automatically. Suppose now that $\rho < 1$. We have

$$F(\rho, \lambda) = \frac{1 - \rho}{\rho} - \frac{h}{(1 + \rho)^2},$$

which is independent of λ. Since $F_A < 0$,

$$A^\lambda < \rho \quad \Longleftrightarrow \quad h > \frac{(1 - \rho)(1 + \rho)^2}{\rho}.$$

This proves the result for A^λ.

It remains to establish the results for D^λ and $A^\lambda + D^\lambda$. Using the equilibrium condition

in (20),

$$D = \frac{1}{s}\left(\frac{(1-sA)(A^2+\rho)}{A^2} + (s-1)(1+\rho)\right),$$

and hence

$$A + D - 1 = \frac{\rho(1-A)(1+A-sA)}{sA^2}.$$

Because

$$\frac{ds}{d\lambda} = -\frac{h}{(1+\rho)^2} < 0,$$

it is enough to show that D and $A+D$ decrease in s.

Implicitly differentiating the equilibrium condition with respect to s and simplifying gives

$$\frac{dD}{ds} = \frac{(1-A^2)M}{A^2s^2(A^2+\rho)T}, \qquad \frac{d(A+D)}{ds} = \frac{\rho(1-A^2)N}{A^2s^2(A^2+\rho)T},$$

where

$$M = A^4\left(x^2+\rho(2x-1)\right) + A^2\left(\rho^2(2-x^2)+\rho x(2-x)\right) + \rho^3(1-x)(3-x),$$

and

$$N = A^4(2x-1) + A^2\left(2x-x^2+2\rho\right) + \rho^2(1-x)(3-x).$$

Both M and N are positive. This is immediate for $x \geq 1/2$, whereas for $x < 1/2$, the only potentially negative terms are dominated because

$$x(2-x) - A^2(1-2x) \geq x(2-x) - x^2(1-2x) = 2x(1-x+x^2) > 0.$$

Since $T < 0$, it follows that $\frac{dD}{ds} < 0$ and $\frac{d(A+D)}{ds} < 0$. Together with $ds/d\lambda < 0$, this yields

$$\frac{\partial D^\lambda}{\partial \lambda} > 0, \qquad \frac{\partial (A^\lambda + D^\lambda)}{\partial \lambda} > 0,$$

which completes the proof. ■