American Economic Journal: Microeconomics 2025, 17(4): 102–125
https://doi.org/10.1257/mic.20240190

(Pro-)Social Learning and Strategic Disclosure†

By Roland Bénabou and Nikhil Vellodi*

We study a sequential experimentation model with endogenous feedback. Agents choose between a safe and risky action, the latter generating stochastic rewards. When making this choice, each agent is selfishly motivated (myopic). However, agents can disclose their experiences to a public record, and when doing so are prosocially motivated (forward-looking). Disclosure is both polarized (only extreme signals are disclosed) and positively biased (no feedback is bad news). The extent of disclosure is non-monotone in prior uncertainty. Subsidizing disclosure costs can paradoxically lead to less disclosure, but more experimentation. (*JEL* D81, D82, D83)

In many settings, agents face a choice between safe and risky actions, with different individuals facing these choices in sequence. Agents might benefit from the information generated by those who preceded them. For instance, consider consumers choosing whether or not to dine at a restaurant with unknown quality, or to watch a new movie. Those who do so can then leave feedback, helping later-arriving consumers make more informed choices. Similar settings include the adoption of new products and technologies, employment choices, and sequential voting.

A well-known dynamic externality emerges in such settings, namely that agents do not internalize the benefit to future consumers of taking the risky action, leaving feedback and thereby generating socially valuable information. To remedy this inefficient underexploration, a planner would direct agents to take the risky action even when it is unprofitable to them, provided the informational gain to future agents more than compensates. This question has been studied extensively in economics and computer science under the label of "incentivized exploration (IE)" (Kremer, Mansour, and Perry 2014; Che and Hörner 2018), itself part of the broader literatures on social learning and sequential experimentation (Banerjee 1992; Bikhchandani, Hirshleifer, and Welch 1992; Smith and Sørensen 2000; Smith, Sørensen, and Tian 2021).[1] Papers in the IE literature largely take a *normative* approach to the problem. Namely, they assume the presence of a benevolent designer who can control the provision of incentives either via dynamic information provision or direct recommendations. Furthermore, these works largely assume that, once generated, individual signals

* Bénabou: Princeton University, NBER, CEPR, IZA, BREAD (email: briq. rbenabou@princeton.edu); Vellodi: Paris School of Economics (email: nikhil.vellodi@psemail.eu). Alexander Wolitzky was coeditor for this article. We thank Aislinn Bohren, Gabriel Carroll, three anonymous referees, and various seminar participants for useful suggestions. Vellodi acknowledges funding from the EUR grant ANR-17-EURE-0001, as well as Princeton University, where this research was initiated in 2019.

† Go to https://doi.org/10.1257/mic.20240190 to visit the article page for additional materials and author disclosure statement(s).

[1] See Slivkins (2022) and Bikhchandani et al. (2022) for recent surveys on IE and social learning more generally.

are perfectly observed by the planner or designer in charge of public information provision.

Such studies are thus silent on a particularly salient issue within the online feedback setting—*why* and *when* do people leave feedback in the first place? For instance, consumers might be driven by a desire to help future consumers make informed choices, or to reward or punish a seller for a positive or negative experience. In practice, while feedback is often highly valued by consumers, the vast majority fail to provide it,[2] and those that do provide feedback display well-known biases such as *positive selection* (Nosko and Tadelis 2015; Hui, Klein, and Stahl 2024)—undisclosed experiences are on average negative—and *polarization* (Schoenmueller, Netzer, and Stahl 2020)—extreme reviews are more prevalent than average reviews.

We take a step toward addressing these questions by providing a *positive* theory of IE. Namely, we propose a simple three-period model of sequential experimentation, in which we endow our agents with *prosocial* motives to leave feedback.[3] In our model, three agents arrive sequentially and have the choice between a safe (S) and risky action (R). Action S generates a deterministic reward, whereas R yields a random reward correlated to an underlying hidden state, for instance, the unknown quality of a product. If an agent plays R, they can disclose their signal. Our main analysis models disclosure via hard evidence with noisy transmission (Dye 1985); an agent is able to leave feedback with probability $\alpha \in [0,1]$ but can either truthfully report their signal or not report it at all. Two crucial assumptions determine an agent's payoff. First, when choosing their *action*, we assume that they are fully *self-interested*, maximizing only their personal reward. Second, when making their subsequent *disclosure* choice, we assume agents are fully *prosocial*. Formally, they have lexicographic preferences over making optimal consumption decisions for themselves, and then transmitting useful information to help others do the same. We discuss these modeling choices further below.

This simple combination of ingredients delivers a rich theory of selective disclosure, which both accords with well-documented phenomena and provides new testable predictions. In particular, our first main result (Theorem 1) demonstrates that equilibrium disclosure is both positively selected and polarized. The intuition is simple. When player 1 (P1) plays R, their disclosure choice is governed equally by the subsequent payoffs of P2 and P3. In contrast, P2 is guided purely by their own payoff when taking their action. For instance, an informed P2 might fail to experiment (play R), even though the loss to themselves is smaller than the gain to P3 that the information generated by doing so would provide. P1 would ideally like to avoid such instances and thus strategically conceals their own experience, inducing P2 to experiment against their interests for the sake of P3.

Simply put, an early adopter would rather not take responsibility for causing the untimely demise of a new product, if there is a reasonable chance the product is in fact worth a second chance, and in this case they keep quiet. On the other hand, when experiences are sufficiently negative, P1 is convinced that no further

[2] Recent surveys report that only around 10 percent of consumers regularly leave reviews. See https://tinyurl.com/mrrsf9v5.

[3] We discuss the limitations of our three-period specification in Section VC.

experimentation should occur and thus terminates it by posting their feedback, while for (even marginally) positive experiences, there is no downside to disclosure. Thus, strategic nondisclosure is used exclusively by P1 to foster efficient experimentation by P2. Of course, models that assume that leaving feedback is costly and done only when sufficiently informative also generate polarized feedback but struggle to also deliver positive selection from a single behavioral foundation.[4]

Beyond these, our model delivers further predictions. For instance, we fully characterize how equilibrium nondisclosure, and thus experimentation, varies with the prior belief regarding risky payoffs (Theorem 2). We view this exercise as capturing, in a reduced-form manner, how disclosure varies with how old or well established the product market in question is. We show that the extent of experimentation is hump-shaped in the prior. Moreover, equilibrium experimentation disappears as prior uncertainty vanishes.

We also show that the extent of experimentation is also hump-shaped in α, the feedback opportunity parameter. This insight has important implications for real-world interventions; practitioners argue that the lack of feedback in online markets leads to biased inference, and that making feedback less costly (e.g., by providing explicitly monetary incentives) would lead to more information and thus experimentation (Marinescu et al. 2021). If we take the natural interpretation that α corresponds to the fraction of agents for whom feedback is costless and $1-\alpha$ the fraction for whom it is prohibitively costly, our result suggests that making feedback less costly could paradoxically lead to *less* disclosure, and more generally that the effectiveness of such interventions in stimulating feedback rates will vary by products and markets.

The joint assumption of selfish consumption and prosocial disclosure is appealing on three separate fronts. First, endowing agents with benevolent preferences in this manner allows our theory to be viewed as a minimal departure from the normative analyses in the IE literature. That is, our agents are effectively mini-planners when disclosing, facing the same trade-off between *exploration* (long-run information gains) and *exploitation* (short-run consumption gains) as in previous work, but they also face additional constraints imposed on them in equilibrium, such as ex post optimality of disclosure rules. Our results thus demonstrate how such constraints shape the degree to which disclosure can be used to incentivize exploration.

Second, from a positive perspective, recent surveys suggest that the welfare of other consumers is a key driver when leaving feedback.[5] At the same time, empirical evidence suggests that incentives to provide feedback are divorced from actual consumption choices in online settings (Cabral and Li 2015). We present a first attempt at formalizing these arguments, with a view to understanding both their theoretical foundations and their ability to organize empirical findings.

Third, the informational externality described above derives fundamentally from the structure of intertemporal preferences, namely that agents are "present-biased" when making their consumption choices. This gives rise to an alternative, psychological interpretation of the model. Instead of a sequence of agents, consider

[4] For instance, Hui, Klein, and Stahl (2024) allow feedback to be positively biased for unmodelled reasons, suggesting reasons such as fear of retaliation or a simple aversion to providing negative criticism.

[5] For example, see https://tinyurl.com/mrybw969.

a single decision-maker with the following dynamically inconsistent preferences. When taking actions that affect current payoffs, they are myopic (completely present biased), whereas when deciding what available information to store in memory to inform future choices, they are patient. This corresponds to a limiting case of quasi-hyperbolic, or $\beta\delta$ (Laibson 1997) preferences where β is arbitrarily small. By modeling the disclosure objective as altruistic, our framework permits this application to an individual who selectively encodes their experiences in order to become less "conservative"—that is, more open to trying and learning from new experiences. Our work thus demonstrates a close conceptual connection between IE and motivated reasoning (Bénabou and Tirole 2002, 2004; Carrillo and Mariotti 2000).

Our model of feedback imposes two important constraints: Feedback must be both ex post optimal and truthful. To explore how the first constraint shapes our results, we analyze communication under commitment, and find that it is again polarized and positively selected. However, in contrast to the disclosure benchmark (Theorem 2), this pattern does not vanish with prior uncertainty.

Turning to the second constraint, our focus on disclosure of hard evidence as the channel through which feedback takes place is primarily motivated by the case of online reviews. While these share elements of both disclosure and cheap talk, a number of features make them closer to the former. First, the fact that a large share of consumers choose to not leave a review calls for a disclosure model. Second, platforms indicate which are "verified purchase" or "verified traveler" reviews and, conversely, take strong measures against fake reviews: using artificial intelligence to detect and remove them, and taking legal action against the intermediaries who sell such fake reviews.[6] Vendors can also appeal to the platform to remove false criticism. Third, reviewers often post specific facts, photos and videos, book commentaries, etc. to support their evaluations; relatedly, a large experimental literature documents significant and widespread lying aversion (e.g., Abeler, Becker, and Falk 2014; Abeler, Nosenzo, and Raymond 2019). Finally, reviews are themselves evaluated by other customers, who can tag a review as helpful or on the contrary report it as fake. Amazon also materially incentivizes and then highlights informative reviews with Amazon Vines, a program that selects "customers who consistently write [the most] insightful reviews"; they can then request for free products from thousands of brands, on which they then write reviews that are distinguished by a special badge. Admittedly, even all these measures still leave room for some fake reviews and biased reviews by real consumers (He, Hollenbeck, and Proserpio 2022). Therefore, we also analyze the case of cheap talk, comparing and contrasting its implications with those of disclosure, with details in Supplemental Appendix C.

In concurrent and independent work, Smirnov and Starkov (2024) analyze a very similar model, focusing on the persuasion benchmark and cheap talk. Like us, they mainly study the three-period case, but also obtain some partial results for an infinite horizon (see Section V for further discussion). Analyzing disclosure allows us to uncover tight comparative-statics implications on the nature and degree of equilibrium communication; our results regarding non-monotone disclosure (Theorem 2

[6] See https://tinyurl.com/2md6zhnp.

and Theorem 3) have no analog under either of the other two alternative forms of information transmission.

The paper proceeds as follows. After introducing the model (Section I), we fully characterize equilibrium disclosure in Section II. In Sections III and IV we derive key comparative static results regarding both prior uncertainty and disclosure opportunities, respectively. In Section V we discuss our modeling choices and, in particular, contrast verifiable disclosure with persuasion and cheap talk, with details in Supplemental Appendices B and C. We conclude with thoughts on future research in Section VI. Unless otherwise mentioned, proofs are gathered in the Appendix.

I. Model

Players and Signals.—At each date $t = 1,2,3$, a short-lived agent arrives and takes a binary decision $a_t \in \{0,1\}$, corresponding to safe and risky actions respectively. The safe action generates a payoff 0. The risky action incurs a cost $c \in (0,1)$ and generates a payoff distributed according to F_θ, where the distribution F_θ depends on a state $\theta \in \{H,L\}$, admits a density f_θ and is supported on the compact real interval $X = [\underline{x},\bar{x}]$. We shall often refer to $a_t = 1$ as "consuming" and to the realized payoff $x \in X$ as a signal.

The state θ is initially unknown, with all agents sharing the common prior $p = \Pr(\theta = H)$. Let p^x denote the posterior belief formed by combining the belief p with the signal $x \in X$. That is,

$$p^x \equiv \frac{pf_H(x)}{f_p(x)} \equiv \frac{pf_H(x)}{pf_H(x) + (1-p)f_L(x)} \quad \text{for } x \in X. \tag{1}$$

Note that for all $p \in (0,1)$, $p^x = p$ if and only if $f_H(x) = f_L(x)$. Let $\hat{x}$ denote the "neutral" signal that satisfies this equality, and more generally, let $x(p,q)$ solve $p^{x(p,q)} = q$, i.e., it is the signal required to achieve posterior q starting from prior p.[7] We will sometimes use a natural transformation from signal space X into belief space $[0,1]$. Namely, we denote by G the distribution (with density g) over posterior beliefs induced by the signal distribution: For each $p,q \in [0,1]$, let $G_p(q) \equiv F_p(x(p,q))$, where $F_p \equiv pF_H + (1-p)F_L$.

Disclosure.— Conditional on receiving outcome x, the agent may then have the opportunity to provide feedback regarding their experience, via direct communication. We assume hard evidence and verifiable disclosure (Dye 1985; Jung and Kwon 1988), wherein a player: (i) with probability $\alpha \in [0,1]$, is able to freely disclose their signal x, and chooses whether or not to do so; (ii) with probability $1-\alpha$, has no such opportunity, for instance due to a prohibitively high disclosure cost.[8]

[7] We will impose assumptions on f_θ that ensure that both $\hat{x}$ and $x(p,q)$ are guaranteed to exist and be unique for all $p,q \in (0,1)$.

[8] In a slight variant of the model, the arrival of agents is random and unobservable to others, occurring with probability α in each period.

Payoffs.—Each agent values the payoffs to both themselves and future agents, but very differently. We assume a form of lexicographic preferences, in which players care infinitely more about their own consumption than that of any other consumer.[9] Formally, given a belief p_t, agent t chooses a_t to maximize their expected consumption payoff $a_t E[x_t - c]$, so that by Assumption 1.c, $a_t(p_t) = \mathbf{1}_{\{p_t \geq c\}}$.[10]

On the other hand, once their consumption choice has been made, agents value the welfare of future consumers equally when making their disclosure choice. Thus, if P2 consumes and obtains the signal x, their value from inducing a belief r upon P3 through their disclosure choice, while themselves holding belief q, is $V_2(r|q) = u(r|q) \equiv \mathbf{1}_{\{r \geq c\}}(q - c)$, i.e., the utility, as judged by P2, that P3 will derive from their own consumption decision.

We will restrict our attention throughout the paper to equilibria in which P2 fully reveals. This is natural for several reasons. First, truthful revelation is weakly dominant for P2, as P2 and P3 have fully aligned preferences. Second, we show in Supplemental Appendix A that truthful revelation by P2 is strictly dominant in the presence of (possibly arbitrarily small) shocks to players' payoffs, and is thus uniquely selected by an argument of robustness to such perturbations.

Turning now to P1, they value the consumption outcomes of both P2 and P3 equally, hence their continuation value $V_1(r|q)$ is

$$V_1(r|q) = \begin{cases} u(r|q) + \alpha\Lambda(r|q) + (1-\alpha)u(r|q), & \text{if} \quad r \geq c \\ 0, & \text{if} \quad r < c, \end{cases} \tag{2}$$

where

$$\Lambda(r|q) \equiv E\left[u(r^z|q^z)\right] = \int_{x(r,c)}^{\bar{x}} (q^z - c) f_q(z) dz$$

denotes the expected consumption value of P3 from P1's perspective, given that P1 holds private belief q and that P2 both holds belief r and consumes.

The disclosure rule is a function $d : X \to \{0, 1\}$, where $d(x) = 1$ denotes disclosure by P1 of signal x at prior p and $d(x) = 0$ denotes nondisclosure. We will typically use p to denote P1's prior belief, q to denote P1's posterior belief, and r to denote the public prior held by P2, which is ultimately determined by P1's disclosure rule d.

Equilibrium.—In order to describe incentive compatible disclosure rules, we must develop our analysis of belief formation under nondisclosure. If the signal x is disclosed, it is simply combined with the current belief according to Bayes' rule (1). If it is not disclosed, then the update rule must account for all other signals at which nondisclosure also occurs, as well as the possibility that disclosure was not

[9] We discuss this assumption in Section VC.

[10] The weak inequality implies that each agent is assumed to consume when indifferent.

feasible. For a disclosure rule d, let $D(d) = \{x \in X | d(x) = 1\}$ and $N(d) \equiv X \backslash D(d)$.[11] We have:

$$p^{\varnothing} \equiv \frac{\Pr(d = \varnothing | \theta = H)}{\Pr(d = \varnothing)} = \frac{(1-\alpha)p + \alpha \int_{N(d)} p^x f_p(x) dx}{(1-\alpha) + \alpha \int_{N(d)} f_p(x) dx}. \tag{3}$$

The relevant incentive compatibility (IC) constraint for the disclosure choice by P1 is then: For all $x \in X$, $d(x) = 1$ if and only if

$$V_1(p^x | p^x) \geq V_1(p^{\varnothing} | p^x). \tag{4}$$

An equilibrium is simply a disclosure rule d for P1 such that: (i) Given the nondisclosure belief $p^{\varnothing}$, d is incentive compatible, and (ii) given d, $p^{\varnothing}$ is correctly computed:

DEFINITION 1: *An equilibrium is a disclosure rule d such that (3) and (4) are satisfied.*

Let us define the *experimentation region* of a disclosure rule d as

$$X_E(d) = \{x \in X | x \in N(d) \text{ and } p^{\varnothing} \geq c > p^x\}.$$

A signal $x \in X_E(d)$ if, under d, P1 chooses not to disclose it, and by so doing induces P2 to consume when they wouldn't if P1 had disclosed. An equilibrium d is an *experimentation equilibrium* (EE) if $X_E(d)$ has strictly positive measure. Let $\mathcal{E}$ denote the set of all such equilibria.[12] An equilibrium d is a *maximal experimentation equilibrium* (MEE) if $d \in \mathcal{E}$ and $d' \in \mathcal{E}$ implies $X_E(d') \subset X_E(d)$. Thus, the MEE contains the largest experimentation region out of all EE. As we show below in Lemma 2, the MEE is the welfare-optimal equilibrium and thus forms a natural benchmark, on which we will later on perform comparative statics.

II. Positively Biased and Polarized Disclosure

We now analyze equilibrium disclosure rules. First, we introduce further natural assumptions on the signal structure (Smith, Sørensen, and Tian 2021):

ASSUMPTION 1:

(*1.a*) *F_H, F_L satisfy the monotone likelihood ratio property (MLRP).*

[11] For any $p < c$, P_1 abstains from consuming ($a_1 = 0$) and thus has no signal to report, making $D(d)$ irrelevant. In what follows we will therefore focus on values $p \geq c$.

[12] Focusing on experimentation equilibria rules out pathological equilibria that turn on the indifference P1 has over disclosure of extreme signals. For instance, any disclosure profile d such that $N(d) \subset [\underline{x}, x(p,c))$ and $p^{\varnothing} < c$ is an equilibrium; for $x < x(p,c)$, both disclosure and nondisclosure lead to P2 not consuming, while for $x \geq x(p,c)$, truth-telling is strictly optimal, as shown below in the proof of Theorem 1.

$$(1.b) \qquad \inf_x \left(\frac{f_L}{f_H}\right)(x) = 0, \sup_x \left(\frac{f_L}{f_H}\right)(x) = \infty.$$

$$(1.c) \qquad E[x \mid \theta = H] = 1 \text{ and } E[x \mid \theta = L] = 0.$$

Assumption 1.a states that higher signals are more likely in the high state, and that no perfectly revealing signal exists in either state. Assumption 1.b is the "unbounded beliefs" assumption of Smith and Sørensen (2000), stating that there always exists a signal strong enough to almost completely overturn any prior belief. Assumption 1.c is a normalization ensuring that beliefs and expected payoffs coincide, i.e., $E[x \mid p] = p$, and is made simply for algebraic convenience.[13] As in Smith, Sørensen, and Tian (2021), we further assume that the distribution of the log-likelihood ratio of signals is log-concave. This ensures an intuitive feature of belief updating known as "posterior monotonicity" (PM) holds under Bayesian updating.

ASSUMPTION 2: *Let $\phi_\theta(l)$ denote the state-contingent densities for the transformed variable $l = \log\big(x/(1-x)\big)$. Then $\phi_\theta(\cdot)$ is log-concave for $\theta \in \{0,1\}$.*

Our first main result below—Theorem 1—characterizes the structure of equilibrium disclosure. In order to interpret its content, we introduce two key concepts, *polarity bias* and *positively selected disclosure*:

DEFINITION 2: *A disclosure rule d is:*

(*i*) *Polarized if there exist $\underline{\varepsilon}, \bar{\varepsilon} > 0$ such that $d(p,x) = 1$ for all $x \in [0,\underline{\varepsilon}) \cup [1-\bar{\varepsilon}, 1]$ and $N(d)$ has strictly positive measure.*

(*ii*) *Positively selected at p if $p^{\varnothing} < p$.*

A polarized disclosure rule is one where extreme signals are disclosed. A positively selected rule is one where the posterior belief formed upon observing no feedback is strictly lower than the prior, so that "no news is bad news." Theorem 1 shows that *any* experimentation equilibrium exhibits both of these features:

THEOREM 1: *In any EE, player 1 adopts the disclosure strategy:*

$$d(x) = \begin{cases} 1, & \text{if } p^x \geq c \\ 0, & \text{if } p^x \in [\underline{q}, c) \\ 1, & \text{if } p^x < \underline{q}, \end{cases}$$

for some $\underline{q} \in (0,c)$.

[13] Assumptions 1.a and 1.c jointly imply that $\underline{x} < 0, 1 < \bar{x}$.

In equilibrium, $P1$ discloses only those signals that lie on either side of the interval $\left[x(p,\underline{q}),x(p,c)\right]$, thus exhibiting both polarity and positive selection (since $x(p,c) \leq x(p,p) = \hat{x}$). P1 thus thinks along the following lines. If disclosing their experience does not affect P2's demand, then P1 is happy to do so. This is the case when P1's experience is "good enough," so that leaving feedback does no harm and improves public information. However, if disclosing leads P2 to not consume (and thus P3 subsequently), P1 discloses only if they are sufficiently convinced that the product's quality is low; in this case, P1 would rather terminate future consumption. Otherwise, they keep their opinion to themselves, as they would rather give the product a "second chance" by having P2 consume and generate further information.[14] Put simply, P1 is always happy to leave a good review, but thinks twice about leaving a bad review, and only does so if their experience was sufficiently bad.

Experimentation versus Accuracy.—To provide further intuition for the proof of Theorem 1, we identify the key tradeoff facing P1 when disclosing, namely fostering experimentation versus improving accuracy. Since disclosure is verifiable, if P1 wants to distort the actions of P2, they must do so by not disclosing their experience, causing a rift between their posterior belief and P2's prior. The benefit of doing so is that P2 will experiment when they would not have done otherwise. The cost is that this rift in beliefs will propagate through to P3, as P3 will combine P2's disclosed signal with P2's (incorrect) prior belief. Consequently, from P1's perspective, there is a chance that P3 will make consumption errors, i.e., consume when they shouldn't or not consume when they should.

To understand the role of such consumption errors more formally, we characterize the properties of the value function $V_1(r|q)$ as both P1's posterior belief q and P2's prior belief r vary, rather than studying disclosure rules directly. From equation (2), it is clear that the function Λ is crucial in determining P1's preferences for strategic disclosure. The following lemma provides a complete characterization of Λ.

LEMMA 1:

(*i*) $r \mapsto \Lambda(r|q)$ is strictly increasing on $[0,q)$ and strictly decreasing on $(q,1]$.

(*ii*) $q \mapsto \Lambda(r|q)$ *is strictly increasing* (*and in particular affine*) *for all* $r \geq c$.

(*iii*) $\Lambda(c|c) > 0$.

Importantly, the map $r \mapsto \Lambda(r|q)$ is single-peaked at q. Thus, Λ encodes the loss (from P1's perspective) from inducing an incorrect belief, due to consumption errors by P3. The further is r from q, the greater is the likelihood that P3 makes consumption errors. For instance, when $r > q$, P3 might consume when they shouldn't (in the event that P2's signal x results in $q^x < c \leq r^x$), while conversely if $r < q$, P3 might not consume when they should. Only if $r = q$ do neither of these errors

[14] Note that the restriction to EE's ensures that $p^{\varnothing} \geq c$ so that P2 consumes conditional on nondisclosure by P1. There may exist pathological, non-experimentation equilibria wherein $\underline{q}$ is sufficiently low that $p^{\varnothing} < c$ and P2 does not experiment.

occur. $\Lambda(c\,|\,c) > 0$ quantifies the *option value* from P2's consumption; note that $u(c\,|\,c) = 0$, so while the immediate return from P2 consuming at belief c is 0, the gain to P3 from such consumption is strictly positive, as there is a chance P2 receives a positive outcome, acquiring useful information and thus providing an expected gain to P3.

The question remains whether there exist situations in which P1 resolves this trade-off in favor of fostering experimentation. Consider posterior beliefs q just below c. Ideally, P1 would like P2 to consume, but hold the correct belief to minimize consumption errors as discussed above. Formally, since $\Lambda(c\,|\,c) > 0$, $\Lambda(c\,|\,q) > 0$ for q just below c by part 1) of the lemma. However, since inducing a belief below c leads to nonconsumption, P1 understands that P2 must necessarily hold an incorrect belief for consumption to occur. Theorem 1 says that when P1's posterior is sufficiently close to c, they would rather suffer the loss in accuracy than terminate consumption.

This reasoning also reveals why multiple equilibria may exist. Since $r \mapsto \Lambda(r\,|\,q)$ is decreasing for $r > q$, the lowest posterior $\underline{q} < c$ at which P1 is indifferent between disclosing and not is increasing in the nondisclosure belief $p^{\varnothing}$. Intuitively, a higher $p^{\varnothing}$ implies a greater chance of consumption errors by P3, which dampens P1's incentive to foster experimentation through nondisclosure, causing $\underline{q}$ to be higher and thus sustaining the higher $p^{\varnothing}$ in equilibrium.

We can show, however, that the MEE is ex ante welfare maximizing across all equilibria, experimentation or otherwise; that is, it maximizes

$$\mathcal{W}(d;p) \equiv \int_{D(d)} V_1(q\,|\,q)\, g_p(q)dq + \int_{N(d)} V_1(p^{\varnothing}\,|\,q)\, g_p(q)dq. \tag{5}$$

For intuition, note that by fostering maximal experimentation, the MEE also exhibits another key feature: It induces the nondisclosure belief closest to c across all experimentation equilibria. Lemma 1, part (i) tells us that this property is desirable to P1, as it minimizes consumption errors made by P3. It is this combination of minimizing errors and maximizing experimentation that renders the MEE welfare optimal.[15] Henceforth, we will therefore focus on the MEE.

LEMMA 2: *The MEE maximizes* $\mathcal{W}(d;p)$, *as given by* (5), *across all equilibrium disclosure rules* d *and prior beliefs* $p \in [c,1)$.

III. U-Shaped Disclosure with Respect to Prior

Theorem 1 tells us that nondisclosure occurs on an interior interval of signals, if at all. But how does this interval depend on primitives, such as P1's prior belief p, and the disclosure parameter α?

In this section we show how the equilibrium disclosure threshold $\underline{q}$ varies with the prior belief p. Practically speaking, one can view this exercise as comparing products that differ in how well established they are. For instance, if p is close to 1

[15] The proof of Lemma 2 uses some notation introduced for Theorem 2, and as such it can be found after the proof of Theorem 2 in the Appendix.

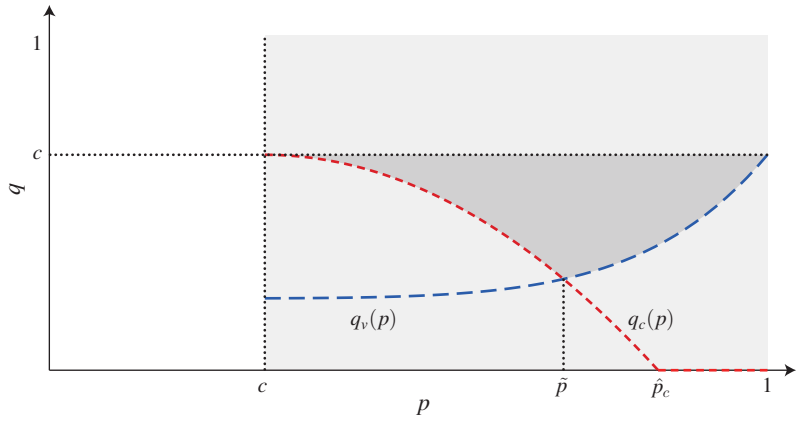

FIGURE 1. EQUILIBRIUM NONDISCLOSURE AS PRIOR p VARIES.

Notes: Nondisclosure in the MEE. x-axis: prior p; y-axis: posterior q. Dark-shaded region: nondisclosed posterior beliefs; light-shaded region: disclosed posterior beliefs. Long-dashed line: incentive constraint, $q_v(p)$; short-dashed line: belief constraint, $q_c(p)$. Signals inducing posteriors in the interval $[\underline{q}(p), c)$ are not disclosed, where $\underline{q}(p)$ is the greater of $q_v(p)$ and $q_c(p)$.

then the product is well established, while for p close to c, the product is close to exit. For p in the interior of this region, products can be viewed as novel. The following result, illustrated in Figure 1, summarizes these findings and constitutes our second main result:

THEOREM 2: *There exist a unique belief* $\tilde{p} \in (c,1)$ *and functions* $q_c(p)$, $q_v(p) : [c,1] \to [0,c]$, *respectively weakly decreasing and strictly increasing, such that:*

(i) *If* $p \in [c,\tilde{p}]$ *then setting* $\underline{q} = q_c(p)$ *constitutes the MEE.*

(ii) *If* $p \in [\tilde{p},1]$, *then setting* $\underline{q} = q_v(p)$ *constitutes the MEE.*

The functions q_c, q_v represent two important constraints on strategic nondisclosure. The first, $q_c(p)$, is a "belief constraint" given by

$$q_c(p) = \inf\{q \in [0,1] \,|\, \phi(p,q) \geq c\}, \tag{6}$$

where

$$(q \leq c):\ \phi(p,q) = \frac{(1-\alpha)p + \alpha\int_{x(p,q)}^{x(p,c)} p^z f_p(z)dz}{(1-\alpha) + \alpha\int_{x(p,q)}^{x(p,c)} f_p(z)dz},\quad (q > c):\ \phi(p,q) = p \tag{7}$$

is the belief P2 forms if they do not see a signal and P1's nondisclosure set is the interval $\big[x(p,q),x(p,c)\big)$. In words, $q_c(p)$ is the lowest belief such that, if all signals that lead to posteriors in $\big[q_c(p),c\big)$ are concealed, the belief $p^{\varnothing}$ following nondisclosure remains above c.

This constraint arises due to a classic form of unraveling: in Theorem 1, we showed that nondisclosure by P1 is positively selected, and so in order to induce experimentation by P2, P1 cannot conceal signals that are too negative, or else the resultant posterior $p^{\varnothing}$ would drop below c. This constraint tightens as p gets closer to c (q_c is decreasing); the closer is p to c, the less room there is for negatively selected nondisclosure to keep $p^{\varnothing}$ above c and thus induce experimentation. Intuitively, when P1's prior is c, even the slightest downgrade in P2's belief will stop experimentation.

The second, $q_v(p)$, is an "incentive constraint" defined by

$$q_v(p) \;\equiv\; \inf\Big\{q \in [0,c] \,|\, W_1\big(\phi(p,q),q\big) = 0\Big\}, \tag{8}$$

where W_1 is the "relaxed" value function

$$W_1\big(r|q\big) \;\equiv\; q - c + \alpha\Lambda\big(r|q\big) + (1-\alpha)(q-c), \tag{9}$$

which denotes P1's value given a continuation belief r and a private belief q, *assuming* that P2 consumes and P3 consumes if they do not see a signal; that is, $V_1\big(r|q\big) = \mathbf{1}_{\{r\geq c\}} W_1\big(r|q\big)$.

In words, the value $q_v(p)$ tracks the lowest posterior at which P1 is indifferent between disclosing and not, assuming posteriors in $\big[q_v(p),c\big)$ are concealed in equilibrium. In proving Theorem 2, we show that q_v is increasing, a result that constitutes yet another expression of the experimentation-accuracy trade-off. Intuitively, when p is close to 1, the continuation belief r is also close to 1, *regardless* of P1's disclosure strategy. As such, nondisclosure plays little role in altering P2's experimentation incentives, and in fact would only induce P3, being less well informed, to make more mistakes in their consumption choice. Thus, the range of beliefs q just below c at which P1 would prefer to induce experimentation vanishes.

Combining the two constraints, the nondisclosure region is $\big[\underline{q}(p),c\big]$, where $\underline{q}(p) \equiv \max\big\{q_c(p),q_v(p)\big\}$. Intuitively, if $q_v(p) < q_c(p)$, then signals just greater than $x\big(p,q_v(p)\big)$ cannot be concealed in equilibrium, despite there being an incentive to do so. For if they were, the resultant nondisclosure belief $p^{\varnothing}$ would be below c, meaning experimentation would fail. Conversely, if $q_v(p) > q_c(p)$, then even though the signal $x\big(p,q_c(p)\big)$ could be concealed in equilibrium and keep $p^{\varnothing} \geq c$, there is no incentive to do so.

Note that the result relies on incentives that are distinct from the classic unraveling mechanism (Milgrom 1981; Grossman 1981; Dye 1985). In particular, the classic result does not depend on the prior distribution, whereas it does in our analysis. The key distinction is that in our setting, the sender's preferences over induced posterior beliefs are type-dependent, whereas in the classic setting, each type strictly prefers to induce the highest possible posterior. This type-dependence stems from the misaligned preferences at the heart of the model: For an intermediate range of signals, P1 strictly prefers to induce P2 to hold the lowest belief at which P2 still

plays R and thus experiments, as described above. Outside of this range, P1 strictly prefers to have P2 share their belief. This complex pattern of experimentation is critically linked to the ex post optimality of disclosure; we will show that it fails to hold when agents can commit to a disclosure policy prior to receiving their signal (Kamenica and Gentzkow 2011).

IV. Subsidizing Disclosure

Theorem 2 uncovers the complex relation between the degree of nondisclosure and the prior belief p, due to the two key constraint functions q_c, q_v working against each other. But how are these constraints themselves determined by the ability to disclose, α? The following result answers this question.

THEOREM 3: *Fix* $p \in (c,1)$. *Then there exist* $0 < \tilde{\alpha}(p) < 1$ *such that* $c - \underline{q}$ *is strictly increasing for all* $\alpha \in [0, \tilde{\alpha}(p))$ *and strictly decreasing for all* $\alpha \in [\tilde{\alpha}(p), 1)$.

Intuitively, when α is high, disclosure opportunities abound. This makes it harder for P1 to strategically nondisclose, as there is no room to "hide" behind a lack of disclosure opportunities. In contrast, when α is low, disclosure opportunities are rare, so less information is transmitted and thus beliefs are more persistent. As such, inducing incorrect beliefs comes with a greater chance of consumption errors by P3, which in turn makes nondisclosure less desirable to P1. Thus, for high α, P1 is *unable* to induce experimentation through nondisclosure, whereas at low α, they are not *willing* to. These counteracting forces result in the experimentation region being non-monotone in α, converging to the empty set when α tends to 0 or 1.

Formally, we first show that the belief constraint q_c is increasing in α, converging pointwise to $q_c(p) = c$ for all $p < 1$ as $\alpha \to 1$. In contrast, the incentive constraint q_v is decreasing in α, converging pointwise to $q_v(p) = c$ for all $p < 1$ as $\alpha \to 0$.

Taking the view that $1 - \alpha$ captures the fraction of consumers who find it prohibitively costly to leave feedback, Theorem 3 speaks to a commonly proposed intervention, namely that review platforms should attempt to subsidize costly feedback in order to generate more information and help consumers discover high-quality products more efficiently (Marinescu et al. 2021). Our finding cautions against a broad-brush approach to such an intervention, as there exist situations in which such a subsidy (increasing α) paradoxically *reduces disclosure* (when $\alpha < \tilde{\alpha}(p)$), as well as others where it stimulates disclosure, but leads to *less experimentation* (when $\alpha \in [\tilde{\alpha}(p), 1)$).

V. Discussion

A. *Model Discussion*

Our model differs from standard social-learning settings in some important ways. First, whereas typically agents' private signals are hidden while their consumption

choices are public (Banerjee 1992; Bikhchandani, Hirshleifer, and Welch 1992), here it is the reverse, in the sense that (some) private signals are publicly disclosed.[16] Second, whereas typically agents receive a private signal prior to making a consumption choice, here our agents can only receive their signal if they consume. In our setting, the link between actions and private signal acquisition is crucial; were future agents to receive (and then disclose) private signals regardless of their predecessor's action choices, current agents would never seek to distort these choices by withholding information, and thus full revelation would be optimal. We thus view our paper as belonging more to the literature on experimentation.

B. *Cheap Talk and Persuasion*

As discussed in the introduction, our modeling of feedback as verifiable disclosure is motivated by: (i) the absence of commitment on the part of consumers as to when to leave or not leave feedback, and the large proportion who elect not to; (ii) the fact that many online reviews left by consumers, and especially those that platforms make most salient, convey credible information; (iii) the selection and polarization biases in the reviews that are left.

To understand the role of the lack of commitment we analyze, in Supplemental Appendix B, communication under persuasion. Then, to recognize the fact that many fraudulent or even purchased fake reviews still escape the platforms' scrutiny and regulators' efforts to curb them,[17] we analyze in Supplemental Appendix C the case of cheap talk, and also compare it with that of disclosure. We summarize below the main results of both cases.

Persuasion.—We first examine the benchmark wherein P1 can commit to an arbitrary messaging rule prior to receiving their private signal x (Kamenica and Gentzkow 2011). We take $\alpha = 1$ for simplicity. Recently developed techniques in the persuasion literature allow us to completely characterize the solution (Dworczak and Martini 2019).

Summarizing the results, communication under persuasion is again both polarized (pooling takes place on an interior interval) and positively selected (the average belief conditional on pooling is c, which is less than the prior p). In contrast to the disclosure benchmark, however, this pooling interval remains even when the prior p is close to c. We refer the reader to Supplemental Appendix B for full details.

Cheap Talk.—We next contrast our baseline results with the case of cheap-talk communication. To summarize our findings, assuming again $\alpha = 1$ for simplicity,

[16] Wolitzky (2018) also studies a social learning model with unobservable actions, but with observable outcomes and without strategic behavior. Bowen, Dmitriev, and Galperti (2023) study social learning from signals shared on social media, but here again the sharing is nonstrategic. Closest is Acemoglu, Ozdaglar, and Siderius (2024), in which agents decide optimally whether to share pieces of information they receive, but aiming here to maximize reshares and minimize dislikes by others.

[17] See, for example, https://tinyurl.com/2p9s5ehz in the case of the United Kingdom.

we find that all equilibria are partitional. This property follows almost directly from the single-crossing properties of Λ, as summarized in Lemma 1. Furthermore, we show that the partition of any equilibrium admits finitely many cells on $\left[x(p,c),1\right]$, so that signals that generate posteriors above c are pooled into finitely many intervals; see Proposition C.1 in Supplemental Appendix C for details. This characterization is in stark contrast to Theorem 1, as well as to the one under commitment (Lemma B.1 in Supplemental Appendix B), both of which exhibit a "separate-pool-separate" reporting structure.

While the equilibrium outcome is quite different across the three forms of communication, all three share a similar driving force, namely a preference toward biasing "upward." In Theorem 1 this force is what drives types with posteriors $q < c$ to prefer inducing the (higher) belief c. Under commitment, it is the same force that induces pooling (Lemma B.1). In contrast, under cheap talk this same force generates a ripple effect for all higher types, whereby to preserve incentives, information transmission must necessarily be coarse throughout $\left[x(p,c),1\right]$ (Proposition C.1). Put differently, we see here an alternative manifestation of the same underlying accuracy-experimentation trade-off identified in Section II: In order to foster experimentation by P2, equilibrium almost always induces consumption errors by P3. These represent complementary insights to those obtained under disclosure and persuasion; allowing for biased, ex post optimal reviews greatly undermines the informativeness of consumer feedback, as almost all reviews induce inaccurate beliefs.

Our results on both cheap talk and persuasion also feature prominently in Smirnov and Starkov (2024), who independently study the same model with both cheap talk and persuasion. Over three periods, our results (Proposition C.1 and Lemma B.1) and theirs align. Smirnov and Starkov (2024) also extend their analysis to the infinite horizon; they show how their analog of Lemma B.1 (pooling at intermediate values under persuasion) extends, but are not able to extend that of Proposition C.1 (partitional structure under cheap talk).

C. *Longer Horizons*

The three-period horizon on which we focus provides the simplest, most transparent setting in which strategic (non)disclosure for purposes of inducing experimentation will arise. At least three agents are required for the preference structure we employ to induce intertemporal conflict. Furthermore, our three-period model ensures that full disclosure is employed by P2 *independent* of their belief. P1's preferences over what belief to endow P2 with are thus shaped exclusively by their desire to induce experimentation by P2, rather than their desire to shape subsequent disclosure by P2. This feature of equilibrium disclosure greatly simplifies our analysis, and it is unlikely to hold over longer periods. For instance, were we to introduce a fourth agent, Theorem 2 tells us that now, P2's disclosure rule will be non-monotone in (and in particular, not independent of) their belief, implying that P1's disclosure rule must account for both P2's consumption choice as well as their subsequent disclosure rule. This added complexity is not an artifact of finite horizon, nonstationary analyses, and is likely to persist were we to extend

our analysis to infinite horizon models.[18] That said, we are confident that certain results, for instance, polarized disclosure and the non-monotone comparative results regarding α will extend to the infinite horizon.[19]

One potential source of this complexity is the richness of the signal space. This was central to developing a theory of polarized disclosure, which requires at least three signal values. Adopting a signal space with only three values might afford sufficiently tractability to extend our model to longer horizons. Of course, these thoughts are speculative, and we leave a rigorous treatment of these avenues for other interested researchers.

VI. Conclusion and Future Work

We studied a model of strategic information transmission, driven by a tension between selfish consumption and prosocial disclosure. Our analysis sheds light on an important question: when might consumers choose not to leave feedback in order to improve overall welfare? We showed that equilibrium disclosure is necessarily *polarized* and *positively biased*, two well-established empirical regularities found in consumer reviewing behavior. We further showed that disclosure is hump-shaped with respect to both agents' prior and their opportunities for leaving reviews. This latter result cautions that making feedback less costly could potentially reduce experimentation. Of course, we do not claim that helping future consumers is the sole motivation for leaving feedback. For instance, the survey quoted in the introduction also suggests that expressing one's opinion is another leading factor. Combined with disclosure coming at a cost, this would readily deliver polarized disclosure. Further combining with an exogenous bias toward expressing positive feelings would then deliver positively selected disclosure.[20] Formulating such a theory is undoubtedly an interesting avenue for future research and would complement our paper nicely. We view our analysis as an important first step toward understanding a very natural mechanism for the existence and biases of reviews—that consumers might be prosocially motivated when leaving feedback.

Appendix A. Proofs

A1. *Proof of Lemma 1*

To prove the first part of the lemma, it suffices to note that $\partial\Lambda\big(r\,|\,q\big)/\partial r$ is proportional to $q^{x(r,c)} - c$, which has the sign of $q - r$, by the MLRP. To prove the first part of the lemma, it suffices to note that $\partial\Lambda\big(r\,|\,q\big)/\partial r$ has the sign of $q^{x(r,c)} - c$, (since

[18] One could consider a model where each agent discounts exponentially all future payoffs, or an overlapping-generations model in which only the two following agents matter. In such models it is likely that any stationary Markov perfect equilibrium will feature disclosure that is again nonconstant in current beliefs. A full analysis of such extensions is beyond the scope of the current paper.

[19] As long as the sender values accuracy, they will disclose signals that are sufficiently informative. It also seems likely that full disclosure will obtain in the limits as $\alpha \to 0, 1$ over longer horizons, while full disclosure cannot be an equilibrium for interior α.

[20] Hui, Klein, and Stahl (2024) find empirical evidence for polarized reviews, but that bias can in fact be either positive or negative and depends on the age and reputation of the firm

$\partial x(r,c)/\partial r < 0$ by the MLRP), which itself has the sign of $q - r$ by the MLRP and is thus negative for $q < r$. For the second part, basic algebra confirms that

$$\text{(A1)}\ \Lambda(r|q) = q\left[1 - F_H\big(x(r,c)\big)\right](1-c) + (1-q)\left[1 - F_L\big(x(r,c)\big)\right](-c).$$

When P1 holds belief q and P2 holds belief r, P1 believes the state is high with probability q and that P3 will consume with probability $1 - F_H\big(x(r,c)\big)$, receiving a payoff $1 - c$, and similarly when the state is low. The third part follows from

$$\Lambda(c|c) = \int_{\hat{x}}^{\bar{x}} (c^z - c) f_c(z) dz > 0,$$

since $c^z > c$ for $z > \hat{x}$.

A2. *Proof of Theorem 1*

We first leverage Lemma 1 to characterize V_1. To do so, it is often convenient to study the "relaxed" value function $W_1(r|q)$. From equation (9) it is straightforward to demonstrate that W_1 obtains the same properties as Λ, since u also preserves these properties.

LEMMA A1:

(i) $r \mapsto W_1(r|q)$ *is strictly increasing on* $[c,q)$ *and strictly decreasing on* $(q,1]$.

(ii) $q \mapsto W_1(r|q)$ *is strictly increasing* (*and affine*) *on* $[0,1]$.

(iii) $W_1(c|c) > 0$.

To establish the result, note first that following a signal leading P1 to hold a posterior q, their disclosure decision hinges on the sign of $V_1(q|q) - V_1(p^{\varnothing}|q)$. First, we show that disclosure occurs in equilibrium for all signals such that $p^x \geq c$, i.e., for sufficiently high signals.

LEMMA A2 (Positive Selection): *If* $p^x \geq c$, *then* $d(x) = 1$ *is a strictly dominant strategy*.

PROOF:

The proof proceeds in two cases. Let $q = p^x$. First, suppose that $p^{\varnothing} < c < q$, so that nondisclosure causes consumption to stop. Then, using (2),

$$V_1(p^{\varnothing}|q) = 0 < u(q|q) + \alpha\Lambda(q|q) + (1-\alpha)u(q|q) = V_1(q|q),$$

since by Lemma 1,

$$\Lambda(q|q) \geq \Lambda(c|q) \geq \Lambda(c|c) = \int_{\hat{x}}^{\bar{x}} (c^z - c) f_c(z) dz > 0.$$

Next, suppose that $p^{\varnothing} \geq c$, so that nondisclosure leads to consumption (and subsequent disclosure by P2) in spite of a lower belief. In this case

$$V_1(q|q) - V_1(p^{\varnothing}|q) = \alpha\big[\Lambda(q|q) - \Lambda(p^{\varnothing}|q)\big] \geq 0,$$

since the first part of Lemma 1 showed that that $r \mapsto \Lambda(r|q)$ is maximized at q, for $q \geq c$. ∎

Lemma A1 then implies that if nondisclosure occurs in equilibrium, it must take an interval form; $D(d) = [\underline{q}, c]$. Finally, that $\underline{q} > 0$ follows from the fact that $V_1(r|0) < 0 = V_1(0|0)$ for $r \geq c$; thus, by continuity, revealing is strictly preferred to inducing experimentation for sufficiently low q.

A3. *Proof of Theorem* 2

To establish Theorem 2, we will use a series of lemmas characterizing beliefs following nondisclosure and the functions $q_c(p), q_v(p)$. We start with properties of $\phi(p,q)$, which as defined in (7) denotes the continuation public belief if signals in the range $[x(p,q), x(p,c))$ are not disclosed.

We showed in Theorem 1 that, if nondisclosure happens, it is over an interval of exactly this type, so $\phi(p,q)$ is indeed the relevant computation for the equilibrium belief $p^{\varnothing}$ following nondisclosure.

LEMMA A3:

(i) *For* $p \geq c$, $q \mapsto \phi(p,q)$ *is strictly increasing and differentiable on* $[0,c]$, *with* $\phi(p,0) \in (0,p)$ *and* $\phi(p,c) = p$.

(ii) *For* $q \leq c$, $p \mapsto \phi(p,q)$ *is strictly increasing on* $[c,1]$.

PROOF:

For part (i), differentiability is clear, and $\partial\phi(p,q)/\partial q$ has the sign of

$$-\frac{\partial x(p,q)}{\partial q} p^{x(p,q)}\Big[(1-\alpha) + \alpha\textstyle\int_{x(p,q)}^{x(p,c)} dF_p(z)\Big] + \frac{\partial x(p,q)}{\partial y}\Big[(1-\alpha)p + \alpha\textstyle\int_{x(p,q)}^{x(p,c)} p^z dF_p(z)\Big].$$

Given that $q \mapsto x(p,q)$ is increasing by the MLRP and $p^{x(p,q)} \equiv q$, that sign is also that of

$$(1-\alpha)(p-q) + \alpha\textstyle\int_{x(p,q)}^{x(p,c)} (p^z - q) dF_p(z) > 0,$$

since $q \leq c \leq p$ and $p^z \geq q$ for $z > x(p,q)$. The bounds on $q \mapsto \phi(p,q)$ follow immediately.

For part (ii), let us first rewrite $\phi(p,q)$ as

$$\phi(p,q) \;=\; \frac{(1-\alpha)p + \alpha\int_{x(p,q)}^{x(p,c)} p^z f_p(z)dz}{(1-\alpha) + \alpha\int_{x(p,q)}^{x(p,c)} f_p(z)dz} \;=\; \frac{(1-\alpha)p + \alpha\int_q^c r\,dG_p(r)}{(1-\alpha) + \alpha\int_q^c dG_p(r)}.$$

Let $a(p) \equiv \int_q^c r\,dG_p(r)$ and $b(p) \equiv \int_q^c dG_p(r)$. By Proposition 4 in Smith, Sørensen, and Tian (2021), Assumption 2 implies that

$$\frac{d}{dp}\left(\frac{a(p)}{b(p)}\right) \;>\; 0.$$

In our case, P2's not having received a signal may also be due to P1 not having had the opportunity to leave feedback, which occurs with probability $1-\alpha$. As a result, $\partial\phi(p,q)/\partial p$ has the sign of

$$\begin{aligned}
&\big[(1-\alpha)+\alpha b(p)\big]\big[(1-\alpha)+\alpha a'(p)\big] - \big[(1-\alpha)p + \alpha a(p)\big]\big[\alpha b'(p)\big] \\
=\;& (1-\alpha)^2 + (1-\alpha)\alpha a'(p) + \alpha b(p)(1-\alpha) - \alpha(1-\alpha)pb'(p) \\
&+ \alpha^2 \underbrace{\big[b'(p)a(p) - b(p)a'(p)\big]}_{>0} \\
\geq\;& \alpha(1-\alpha)\big[a'(p) - pb'(p) + b(p)\big].
\end{aligned}$$

But

$$\begin{aligned}
a'(p) - pb'(p) + b(p) \;&=\; \frac{\partial}{\partial p}\int_q^c r\,\mathrm{d}G_p(r) - p\frac{\partial}{\partial p}\int_q^c \mathrm{d}G_p(r) + \int_q^c \mathrm{d}G_p(r) \\
&=\; \int_q^c \left[r\frac{\partial g_p(r)}{\partial p} - p\frac{\partial g_p(r)}{\partial p} + g_p(r)\right]dr \\
&=\; \int_q^c \big[r\big(g_H(q) - g_L(q)\big) + g_L(q)\big]dr \\
&=\; \int_q^c \big[r\,g_H(q) + (1-r)\,g_L(q)\big]dr \\
&=\; \int_q^c g_r(r)dr \\
&>\; 0,
\end{aligned}$$

thus proving the claim. ∎

We now define and characterize the belief constraint: for $p \geq c$, let

$$q_c(p) \;=\; \inf\big\{q \in [0,1] \,|\, \phi(p,q) \geq c\big\}. \tag{A2}$$

Lemma A3 tells us that $p \mapsto q_c(p)$ is well-defined on $[c,1]$. We now establish key properties of the function $q_c(p)$:

LEMMA A4: *The map $p \mapsto q_c(p)$ is everywhere continuous, with $q_c(c) = c$. Furthermore, there exists $\hat{p}_c \in (c,1)$ such that: (i) On $[c,\hat{p}_c]$, $q_c(p)$ is strictly decreasing, differentiable and solves $\phi(p,q_c(p)) = c$; (ii) on $[\hat{p}_c,1]$ $q_c(p) = 0$.*

PROOF:

Note that $p \mapsto q_c(p)$ is defined as the minimum of a continuous function, $(q \mapsto \phi(p,q))$, on a compact set. Therefore, the infimum is attained by Weierstrass' Theorem, and continuity follows from Berge's Theorem (note that the constraint $\phi(p,q) \geq c$ defines an upper-hemicontinuous correspondence, since $\phi(p,q)$ is continuous). That $q_c(c) = c$ follows from Lemma A3, part (ii).

Next, that there exists a $\hat{p}_c \in (c,1)$ such that $q_c(p) = 0$ for all $p \in [\hat{p}_c,1]$ follows from the definition of $\phi(p,q)$, since as $p \to 1$,

$$\phi(p,0) \to \frac{(1-\alpha)\cdot 1 + \alpha \cdot 1}{1-\alpha+0} = 1.$$

Finally, that $p \mapsto q_c(p)$ is strictly decreasing on $[c,\hat{p}_c]$ follows directly from Lemma A3. ∎

Having characterized the "belief constraint" $q_c(p)$ bearing on P1's disclosure rule, we next turn to the "incentive constraint" $q_v(p)$.

To begin, we demonstrate that disclosure is strictly optimal after sufficiently extreme signal realizations. We do so by proving a property of the relaxed value function $W_1(r|q)$ defined in Section IIA.

LEMMA A5: *For all $p \in [c,1)$,*

$$\lim_{q \to 0,1} \left[W_1(\phi(p,q)\,|\,q) - V_1(q|q) \right] < 0.$$

PROOF:

The lower limit follows immediately since $\Lambda(r|q) < 0$ and $q - c < 0$ for sufficiently small q. The upper limit is obtained by noting that as $q \to 1$, $V_1(q|q)$ achieves the upper bound on V_1. ∎

Away from these limits, note that the minimization defining $q_v(p)$ is well defined:

LEMMA A6: *The map $p \mapsto q_v(p)$ is well-defined, with $q_v(p) < c$ for all $p \in (c,1)$ and $q_v(1) = c$.*

PROOF:

Note that $W_1(r|c) > 0$ for all $r \in (c,1)$, since $\Lambda(r|c) > 0$. Furthermore, for q sufficiently close to c, $\phi(p,q) \geq c$, so that $W_1(\phi(p,q)\,|\,c) > 0$. Thus by continuity, $W_1(\phi(p,q)\,|\,q) > 0$ for q in some neighborhood below c. Lemma A5 combined

with the Intermediate Value Theorem then implies there exists $q' \in (0,c)$ such that $W_1(\phi(p,q')\,|\,c) = V_1(q'\,|\,c) = 0$, thus proving that $q_v(p) < c$ for all $p \in (c,1)$. On the other hand, $W_1(1\,|\,c) = 0$, so that by Lemma A1, $W_1(1\,|\,q) < 0$ for all $q < c$, and thus $q_v(1) = c$. ∎

LEMMA A7: *The map $p \mapsto q_v(p)$ is continuous and strictly increasing.*

PROOF:

For $p \geq c$, let $q(p) < c$ be any solution to the equation $W_1(\phi(p,q),q) = V_1(q\,|\,q)$. From Lemmas A1 and A3 and the chain rule, it follows that if $q(p) < c$, then $q'(p) > 0$. Therefore, $p \mapsto q_v(p)$ must be strictly increasing. ∎

Taken together, these lemmas immediately show that the two loci q_c and q_v cross at a unique interior point:

LEMMA A8: *There exists $\tilde{p} \in (c,1)$ such that $q_v(p) \leq q_c(p)$ if and only if $p \leq \tilde{p}$.*

PROOF:

Follows from Lemma A4 and Lemma A7, part (ii). ∎

To complete the proof of Theorem 2, we proceed in two cases:

(i) If $q_v(p) \in [0, q_c(p))$, then setting $\underline{q} = q_c(p)$ defines the MEE. To see this, note first that the equilibrium belief condition (3) is satisfied by definition. Next, we will verify the IC condition (4), which in this case amounts to $V_1(c\,|\,q) \geq V_1(q\,|\,q)$ for all $q \in [q_c(p), c)$. But if $q_v(p) \leq q_c(p)$ then $\phi(p, q_v(p)) \leq \phi(p, q_c(p)) = c$ by (A.3), and so for all $q \in [q_c(p), c)$,

$$\begin{aligned} V_1(c\,|\,q) &\geq V_1(\phi(p,q_c(p))\,|\,q) \\ &= W_1(\phi(p,q_c(p))\,|\,q) \\ &\geq W_1(\phi(p,q_v(p))\,|\,q) \\ &= 0 \\ &= V_1(q\,|\,q), \end{aligned}$$

with the first equality holding since $V_1(r\,|\,q) = W_1(r\,|\,q)$ for all $r \geq c$, and the second one holding by Lemma A1. This verifies incentive compatibility. That $\underline{q}$ defines an EE is then immediate. To verify that this is a MEE, note that were $\underline{q} < q_c(p)$, then one would have $\phi(p,\underline{q}) < c$ and thus no experimentation by P2 could be supported.

(ii) If $q_v(p) \in [q_c(p), c)$, then set $\underline{q} = q_v(p)$. Again, (3) is satisfied immediately since $q_v(p) \geq q_c(p)$. Next, note that $\underline{q} = q_v(p) \geq q_c(p)$ implies that $\phi(p, \underline{q}) \geq \phi(p, q_c(p))$, and so $W_1(\phi(p, \underline{q}) \,|\, q) = V_1(\phi(p, \underline{q}) \,|\, q) \geq 0$ for all $q \in [\underline{q}, c)$. Thus, (4) is verified. Furthermore, since (4) is binding, this must also be a MEE (setting $\underline{q} < q_v(p)$ would violate (4)).

A4. *Proof of Lemma* 2

Consider first any $d \in \mathcal{E}$ that is not the MEE, d^*, and which has an associated threshold $\underline{q}_d$ (characterized by Theorem 1). By definition of the MEE, it must be that $\underline{q}_d > \underline{q}$ (we suppress the relation of $\underline{q}$ on p throughout this proof for convenience). We then have that

$$\begin{aligned}
\mathcal{W}(d^*;p) - \mathcal{W}(d;p) &= \int_{\underline{q}}^{\underline{q}_d} V_1(\phi(p,\underline{q}) \,|\, q)\, g_p(q) dq \\
&\quad + \int_{\underline{q}_d}^{c} \left[V_1(\phi(p,\underline{q}) \,|\, q)\, g_p(q) - V_1(\phi(p,\underline{q}_d) \,|\, q) \right] g_p(q) dq \\
&> \int_{\underline{q}_d}^{c} \left[V_1(\phi(p,\underline{q}) \,|\, q)\, g_p(q) - V_1(\phi(p,\underline{q}_d) \,|\, q) \right] g_p(q) dq \\
&> \int_{\underline{q}_d}^{c} \left[V_1(\phi(p,\underline{q}) \,|\, q)\, g_p(q) - V_1(\phi(p,\underline{q}) \,|\, q) \right] g_p(q) dq \\
&= 0,
\end{aligned}$$

where the first inequality holds since $\underline{q} \geq q_v(p)$ by construction, and hence $V_1(\phi(p,\underline{q}) \,|\, q) > 0$ for $q \in (\underline{q}, \underline{q}_d]$, and the second inequality holds since $r \mapsto V_1(r \,|\, q)$ is strictly decreasing on $[0,c]$ for $q \in [0,c]$ (Lemma A1) and $q \mapsto \phi(p,q)$ is strictly increasing (Lemma A3).

Finally, note that any equilibrium d that is not in $\mathcal{E}$ supports precisely the welfare under full disclosure, as experimentation is not induced with positive probability. Hence,

$$\begin{aligned}
\mathcal{W}(d^*;p) - \mathcal{W}(d;p) &= \int_{\underline{q}}^{1} V_1(\phi(p,\underline{q}) \,|\, q)\, g_p(q) dq \\
&> 0
\end{aligned}$$

by the same reasoning.

A5. *Proof of Theorem* 3

LEMMA A9:

(i) *For fixed* $p \in (c,1)$, *there exists* $\hat{\alpha}(p) \in (0,1)$ *such that* $q_c(p)$ *is strictly increasing in* α *for* $\alpha \in [\alpha(\hat{p}), 1]$ *and* $q_c(p) = 0$ *otherwise. Furthermore,* $\lim_{\alpha \to 1} q_c(p) = c$.

(ii) *For fixed* $p \in (c,1)$, $q_v(p)$ *is strictly decreasing in* α. *Furthermore,* $\lim_{\alpha\to 0} q_v(p) = c$.

PROOF:

For part (i), note that by Lemma A3, $q \mapsto \phi(p,q)$ is strictly increasing. Furthermore, from equation (7), $\phi(p,q)$ is strictly decreasing in α. Since $q_c(p)$ solves $\phi\big(p, q_c(p)\big) = c$, this proves the first claim. For the second part, note that from equation (7), $\lim_{\alpha\to 1}\phi(p,q) = E\big[r \mid r \in [q,c]\big]$, and hence $\lim_{\alpha\to 1}\phi(p,c) = c$, while $\lim_{\alpha\to 0}\phi(p,q) = p$, and hence $\lim_{\alpha\to 0}\phi(p,c) = 0$.

For part ii), note that $\partial W_1\big(\phi(p,q)\,|\,q\big)/\partial q > 0$, as asserted in Lemma A7. Next, by Lemma A3, $q \mapsto \phi(p,q)$ is strictly increasing. Thus, the first part of the claim obtains provided that $\partial W_1\big(\phi(p,q)\,|\,q\big)/\partial\alpha > 0$. To see that such is the case, note that $\partial W_1(r\,|\,q)/\partial r > 0$ as argued in Lemma A7, and from equation (7), $\phi(p,q)$ is strictly decreasing in α. Finally,

$$\frac{\partial W_1(r\,|\,q)}{\partial\alpha} = \Lambda(r\,|\,q) - (q-c) = \int_{x(r,c)}^{\bar{x}} (q^z - c) f_r(z)\,dz \geq 0,$$

since $q^z > c$ for $z > x(r,c)$. Hence

$$\frac{\partial W_1\big(\phi(p,q)\,|\,q\big)}{\partial\alpha} = \frac{\partial W_1(r\,|\,q)}{\partial\alpha} + \frac{\partial W_1(r\,|\,q)}{\partial r}\frac{\partial\phi(r,q)}{\partial\alpha} > 0.$$

To prove the second part, we proceed as in Lemma A6; note that $W_1(p\,|\,c) > 0$ for all $p \in (c,1)$ and all $\alpha \in (0,1)$, and thus $q_v(p) < c$, while for $\alpha = 0$, $W_1(p \in c) = 0$, so that $q_v(p) = c$. ∎

Finally, we can put these results together to prove the theorem. For $p \in (c,1)$, let $\tilde{\alpha}(p)$ be that value of α such that $q_v(p) = q_c(p)$. Such a value exists and lies in $(0,1)$ by Lemma A9, as we have that $q_v(p) = c > q_c(p)$ at $\alpha = 0$. Finally, $\tilde{\alpha}(p) < 1$ since $\tilde{p} < 1$ for all $\alpha \in (0,1)$.

REFERENCES

Abeler, Johannes, Anke Becker, and Armin Falk. 2014. "Representative Evidence on Lying Costs." *Journal of Public Economics* 113: 96–104.

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for Truth-Telling." *Econometrica* 87 (4): 1115–53.

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. 2024. "A Model of Online Misinformation." *Review of Economic Studies* 91 (6): 3117–50.

Banerjee, Abhijit V. 1992. "A Simple Model of Herd Behavior." *Quarterly Journal of Economics* 107 (3): 797–817.

Bénabou, Roland, and Jean Tirole. 2002. "Self-Confidence and Personal Motivation." *Quarterly Journal of Economics* 117 (3): 871–915.

Bénabou, Roland, and Jean Tirole. 2004. "Willpower and Personal Rules." *Journal of Political Economy* 112 (4): 848–86.

Bikhchandani, Sushil, David Hirshleifer, Omer Tamuz, and Ivo Welch. 2022. "Information Cascades and Social Learning." Unpublished.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy* 100 (5): 992–1026.

Bowen, T. Renee, Danil Dmitriev, and Simone Galperti. 2023. "Learning from Shared News: When Abundant Information Leads to Belief Polarization." *Quarterly Journal of Economics* 138 (2): 955–1000.

Cabral, Luís, and Lingfang Ivy Li. 2015. "A Dollar for Your Thoughts: Feedback-Conditional Rebates on eBay." *Management Science* 61 (9): 2052–63.

Carrillo, Juan D., and Thomas Mariotti. 2000. "Strategic Ignorance as a Self-Disciplining Device." *Review of Economic Studies* 67 (3): 529–44.

Che, Yeon-Koo, and Johannes Hörner. 2018. "Recommender Systems as Mechanisms for Social Learning." *Quarterly Journal of Economics* 133 (2): 871–925.

Dworczak, Piotr, and Giorgio Martini. 2019. "The Simple Economics of Optimal Persuasion." *Journal of Political Economy* 127 (5): 1993–2048.

Dye, Ronald A. 1985. "Disclosure of Nonproprietary Information." *Journal of Accounting Research* 23 (1): 123–45.

Grossman, Sanford J. 1981. "The Informational Role of Warranties and Private Disclosure about Product Quality." *Journal of Law and Economics* 24 (3): 461–83.

He, Sherry, Brett Hollenbeck, and Davide Proserpio. 2022. "The Market for Fake Reviews." *Marketing Science* 41 (5): 896–921.

Hui, Xiang, Tobias J. Klein, and Konrad O. Stahl. 2024. "Learning from Online Ratings." Unpublished.

Jung, Woon-Oh, and Young K. Kwon. 1988. "Disclosure When the Market is Unsure of Information Endowment of Managers." *Journal of Accounting Research* 26 (1): 146–53.

Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101 (6): 2590–615.

Kremer, Ilan, Yishay Mansour, and Motty Perry. 2014. "Implementing the Wisdom of the Crowd." *Journal of Political Economy* 122 (5): 988–1012.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112 (2): 443–77.

Marinescu, Ioana, Andrew Chamberlain, Morgan Smart, and Nadav Klein. 2021. "Incentives Can Reduce Bias in Online Employer Reviews." *Journal of Experimental Psychology: Applied* 27 (2): 393–407.

Milgrom, Paul R. 1981. "Good News and Bad News: Representation Theorems and Applications." *Bell Journal of Economics* 12 (2): 380–91.

Nosko, C., and S. Tadelis. 2015. "The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment." NBER Working Paper 20830.

Rockafellar, R. Tyrrell. 1997. *Convex Analysis*. Princeton University Press.

Schoenmueller, Verena, Oded Netzer, and Florian Stahl. 2020. "The Polarity of Online Reviews: Prevalence, Drivers and Implications." *Journal of Marketing Research* 57 (5): 853–77.

Slivkins, Alexandrs. 2022. "Exploration and Persuasion." In *Online and Matching-Based Market Design,* edited by Vijay V. Vazirani, Federico Echenique, and Nicole Immorlica, 655–75s. Cambridge University Press.

Smirnov, Aleksei, and Egor Starkov. 2024. "Designing Social Learning." Unpublished.

Smith, Lones, and Peter N. Sørensen. 2000. "Pathological Outcomes of Observational Learning." *Econometrica* 68 (2): 371–98.

Smith, Lones, Peter Norman Sørensenand, and Jianrong Tian. 2021. "Informational Herding, Optimal Experimentation, and Contrarianism." *Review of Economics Studies* 8 (5): 2527–54.

Wolitzky, Alexander. 2018. "Learning from Others' Outcomes." *American Economic Review* 108 (10): 2763–801.