The Journal of Law, Economics, and Organization, 2026, **00**, 1–42
https://doi.org/10.1093/jleo/ewag010
Article

Fake reviews and naive consumers

Boris Knapp[1,*]

[1]Department of Economics, Corvinus University of Budapest, Fővám tér 8, Budapest, 1093, Hungary

*Corresponding author: E-mail: boris.knapp@uni-corvinus.hu

ABSTRACT

Online reviews affect consumer choices and are therefore frequently faked. Not all consumers are aware of this. In a model with fake reviews and naive consumers, the unique equilibrium is characterized by partial pooling, where fake reviews are persuasive and blend in with real ones. Raising consumer awareness has opposing effects on the naive and aware consumer groups. If real reviews are written strategically, they are not always truthful. Under favorable market conditions, the equilibrium with all strategic reviewers is outcome equivalent to one with all aware consumers. Thus, awareness campaigns can yield similar outcomes, regardless of their target audience (*JEL* C72, D82, D83, L15)

> Consumers very often rely on online reviews when shopping or booking online. I don't want consumers to be tricked. I want them to be able to interact in a trustworthy environment. I insist on one specific point: online businesses must provide consumers with clear and visible information on the reliability of such reviews.
>
> —Didier Reynders, European Commissioner for Justice, January 2022

1. INTRODUCTION

Online reviews are a cornerstone of some of today's biggest platforms and have become an integral part of many markets. As of September 2023, there are over 500 million reviews on Amazon alone and similar numbers can be found on Google and Yelp.[1] Consumers rely heavily on these reviews to resolve the inherent information asymmetry between them and sellers. Most consumers read online reviews, and this affects their purchase decisions.[2] Therefore, firms have strong incentives to manipulate reviews to boost their sales or hurt their competitors. This is an issue that legislators and platforms alike are trying to crack

[1] Hou et al. (2024) collect a dataset comprising over 570 million Amazon reviews.
[2] See surveys by Bright Local and Podium, and Chevalier and Mayzlin (2006) for evidence in the book market or Lewis and Zervas (2019) for evidence in the accommodation industry.

down on and yet fake reviews are more prevalent than ever.[3] While this is a problem in and of itself it is likely to be exacerbated if consumers are unaware of fake reviews. There is ample evidence for consumer naivety, which raises important questions about consumer protection.[4] How is reviewing behavior shaped by the presence of naive consumers? What are the effects of awareness campaigns on consumer surplus? Can we expect the market to deal with this problem effectively or is it necessary for policymakers to take an active role?

To study these questions, I propose a cheap talk model in which a reviewer can be either real or fake and a consumer can be either aware or naive.[5] A consumer considers buying a product of unknown quality and reads a review to make an informed decision. Naive consumers take reviews at face value, while aware consumers consider that they may be fake. To start, I solely focus on the fake reviewer's incentives to write reviews strategically to maximize the purchase probability. The baseline model therefore assumes that real reviews are always truthful.

I show that, in equilibrium, fake reviews blend in with real ones as the fake reviewer balances maintaining credibility vis-à-vis aware consumers and persuading naive ones. This trade-off shifts when consumers become increasingly aware because the fake reviewer attaches greater importance to the aware consumer group. As a result, the range of messages sent by the fake reviewer expands, leading to a larger overlap with real reviews. This expansion reduces the informativeness of reviews for aware consumers while simultaneously making fake reviews less deceptive for naive ones. Therefore, awareness policies have a positive direct effect by making some naive consumers aware, and two opposing indirect effects resulting from the change in the fake reviewer's strategy. The remaining naive consumers benefit because reviews are less deceptive, and the aware consumers are worse off because reviews are less informative.

In the main section of the article, I allow the real reviewer to be strategic. I show that, to maximize consumer surplus, he does not always write a truthful review. Instead, he understates the product's quality when it is high to avoid being perceived as fake. *Underreporting* is optimal because more credible reviews allow aware consumers to make better decisions, outweighing the harm to naive consumers who follow these underreports blindly. This is true even when the share of aware consumers is very small.

Finally, I show that as the share of non-strategic real reviewers vanishes, the outcome is equivalent to that of an equilibrium with non-strategic real reviewers where the share of naive consumers vanishes. This result is of particular relevance in markets where the boundary between reviewers and consumers is fluid, as is the case for most online platforms. It implies that if awareness policies have the effect of making real reviewers strategic, they are effective irrespective of which side of the market they reach.

While this article deals mainly with fake reviews, its findings also apply to other contexts. Doctors often change their prescription behavior when pharmaceutical companies exert influence. These changes have been linked to adverse health effects, suggesting that the reason is a distortion of incentives.[6] It is obvious that some patients are oblivious to this fact and generally trust doctors. Even patients who are aware of this, however, generally cannot distinguish a doctor who is affiliated with pharmaceutical companies from one who is not.

[3] See He et al. (2022) or the report by SafetyDetectives research lab (2021).

[4] Excessive trust in strategic communication (Cai and Wang 2006; Kawagoe and Takizawa 2009) and signaling contexts (Deversi et al. 2021) is a well-documented phenomenon. Consumer surveys show that consumers trust online reviews (CPC Strategy 2019; Bright Local 2020).

[5] I model awareness and naivety in the spirit of Ottaviani and Squintani (2006). Naivety thus captures the fact that consumers are oblivious and differs from the concept of naïveté in the work of Heidhues and Kőszegi (2010, 2017).

[6] Fernandez and Zejcirovic (2020) show this in the context of the recent opioid crisis in the USA.

Thus, they do not know whether the advice is coming from a biased or unbiased doctor, but they can take this heterogeneity into account.

Credit rating agencies (CRAs) primarily use two business models: issuer-pays and investor-pays. When issuers pay for ratings, CRAs may have an incentive to issue favorable ratings to please their customers. When investors pay, such incentives are absent.[7] The finance literature frequently assumes that investors vary in their level of sophistication.[8]

The rise of social media has extended the domain of financial advice beyond traditional institutions, creating a new landscape where conflicts of interest can emerge. Platforms like Stocktwits enable individual investors to share their opinions on financial assets and follow the advice of others. In this environment, the potential for biased information is significant, as investors may have personal stakes in the stocks they discuss.[9]

1.1 Related literature

The present work is most closely related to Kartik et al. (2007), which is the only other paper that studies welfare effects of awareness policies in a sender-receiver game.[10] The main difference between that paper and mine is that I model the sender's bias along the extensive margin: the sender is either biased, in which case his preferences are completely misaligned with those of the receiver, or he is not, in which case they are perfectly aligned. In contrast, Kartik et al. (2007) model bias along the intensive margin: the sender is always biased but the extent of the bias can vary. In the context of online reviews, the extensive margin is the more important one because not all reviews are fake, and fake reviews arguably do not take consumers' well-being into consideration. This difference, as it turns out, leads to vastly different findings.

While in this article, awareness policies increase the welfare of naive receivers at the expense of aware ones, they hurt naive receivers and have no effect on aware ones in Kartik et al. (2007). Their findings hinge on aware receivers' perfect knowledge of the sender's motives, allowing them to de-bias inflated equilibrium messages. The sender's incentive to inflate messages beyond what aware receivers already de-bias is offset by naive receivers following these same messages, taking actions that are suboptimal even for the sender. Naive receivers, hence, suffer a lot while aware ones enjoy perfect information. Awareness policies that reduce the share of naive receivers decrease the sender's cost of inflating language, leading to more inflation in equilibrium and even lower surplus for naive receivers. In my model, the aware receiver is uncertain about the sender's type, and, in equilibrium, the biased sender pools to sustain this uncertainty. While the biased sender's equilibrium messages are inflated, they become more moderate when the share of naive receivers shrinks, benefiting the remaining naive receivers.

Several other papers study communication games with heterogeneity on the sender side (Jindapon and Oyarzun 2013; Glazer et al. 2021) or on both sides (Chen 2011; Gesche 2021). None of these, however, focus on the welfare effects of awareness policies or examine the behavior of strategic real senders. The only exception is Glazer et al. (2021), who

[7] Bruno et al. (2016) demonstrate that ratings vary between these two business models, and Jiang et al. (2012) find that issuer-paid ratings tend to be inflated compared to those paid for by investors.

[8] Bolton et al. (2012) argue that institutional investors (e.g., pension funds) often take ratings at face value, while hedge funds account for possible inflation. This heterogeneity may be due to differences in compensation; hedge fund managers' compensation is typically more directly tied to returns.

[9] A notable example is the case of Paul Pereira, the former CEO and co-founder of Alfi, who was sued by the SEC for allegedly using Stocktwits and YouTube to promote misleadingly bullish views of Alfi's stock to inflate its price. See https://www.bloomberg.com/news/articles/2024-02-27/founder-used-burner-account-to-boost-meme-stock-alfi-sec-claims?embedded-checkout=true.

[10] Another related paper is Ottaviani and Squintani (2006), which differs from Kartik et al. (2007) only in that it assumes a bounded state space.

consider strategic real senders. However, without naive receivers, these senders have no incentive to deviate from truth-telling. In my model, the presence of naive receivers induces strategic real senders to underreport, resulting in a breakdown of the truth-telling equilibrium.

The disclosure of conflicts of interest (COI) in Gesche (2021) is distinct from the awareness policies in this article because it provides better information about COI to already strategic receivers. In contrast, awareness policies provide information to naive receivers, making them strategic.

The *underreporting* result in this article is akin to the concept of *political correctness* in Morris (2001) and the *reversal* effect in Smirnov and Starkov (2022). In Morris (2001), an unbiased advisor concerned about maintaining a reputation for objectivity may sometimes offer biased advice. This occurs because of the presence of biased advisors who favor one of two possible actions. To avoid being perceived as biased, the unbiased advisor distorts their recommendation in the opposite direction. Morris (2001) shows that the advisor's reputational concerns need not be intrinsic. In a repeated game, instrumental reputational concerns emerge solely from the desire to have their valuable advice heeded in the future. Instrumental reputational concerns drive underreporting in this article as well despite the game's static nature.

In Smirnov and Starkov (2022), sellers can censor reviews that arrive stochastically over time. The presence of naive consumers, who are unaware of this censorship, renders revealing negative reviews a costly signal. High-quality sellers disclose more negative reviews because their higher rate of positive reviews allows them to restore naive consumers' demand more quickly. Consequently, sophisticated consumers treat negative reviews as indicators of high quality. Naive consumers play a similar role in my model, discouraging fake reviewers from sending low reviews, thus rendering those reviews credible signals of high quality.

To focus on the details of communication, I abstract from the pricing aspect.[11] For similar reasons, my model does not answer *why* consumers write reviews but focuses instead on *how* they write them.[12]

1.2 Outline

The remainder of the article is structured as follows: Section 2 lays out the baseline model, in which real reviews are always truthful, and Section 3 characterizes its equilibrium. In Section 4, I study the effects of awareness campaigns by analyzing the model's comparative statics. Section 5 introduces strategic real reviewers to the model and presents the main result that such reviewers underreport in equilibrium. I discuss further extensions to the model in Section 6, allowing for negative fake reviewers, and multiple consumers and reviewers, respectively. These extensions are treated more formally in Appendix B. Section 7 concludes. Technical proofs of all results in the main text are provided in Appendix A.

2. MODEL

There are two players, a sender (he), called the reviewer, and a receiver (she), called the consumer. Before choosing between a good of unknown quality and her outside option, the consumer reads a review about the product. The quality of the good and the outside option are represented by the independent random variables X and Y with realizations x and y,

[11] Recent papers have studied the role of pricing in the presence of reviews (Martin and Shelegia 2021) or soft product information (Janssen and Roy 2022).

[12] Consumers' incentives to write reviews are studied, for example, in Cheung and Lee (2012).

distributions F_X and F_Y, and densities f_X and f_Y. Both random variables take values in $[0, 1]$. The consumer knows her outside option, but she needs to infer the product's quality from the review. A review is a real number $m \in \mathcal{M} = [0, 1]$; hence, it can be thought of as statements of the form *"The product is of quality m"*.

2.1 Reviewers

In the baseline model, there are two types of reviewers: real and fake. These types differ along two dimensions. First, they differ in terms of the information they possess. Real reviewers observe the quality of the good, while fake reviewers do not.[13] However, neither type knows the realization of the consumer's outside option.

Second, they differ in their incentives. The real reviewer is a non-strategic player who always writes a truthful review, i.e., he honestly conveys his private information. The fake reviewer, on the other hand, is a strategic player with the objective of inducing a purchase. To this end, he is free to send any review $m \in [0, 1]$.

His payoff function is given by

$$u_S^F = p, \tag{1}$$

where $p \in \{0, 1\}$ denotes the consumer's decision to either not buy or to buy the good. His preferences are *state-independent*, meaning he is indifferent to both the good's quality and the consumer's outside option. Because the fake reviewer cannot condition his review on the good's quality, his strategy is described by a cumulative distribution function, F^F, with corresponding density function f^F.

In later sections, the baseline model is extended to include two additional types of reviewers. In Section 5, I introduce a strategic real reviewer who can misrepresent the good's quality and aims to maximize expected consumer surplus. In Appendix B.1, I examine both positive and negative fake reviewers, who seek to maximize and minimize the purchase probability, respectively.

2.2 Consumers

The consumer can be one of two types: naive or aware. A naive consumer takes every message at face value. An aware consumer, by contrast, is a fully strategic player who considers the possibility that a review might be fake, updates her beliefs accordingly, and then takes an optimal action.

Formally, the two types differ in how they process information. A naive consumer updates her beliefs as if every review was truthful. Her expectation of the good's quality, given review m, is

$$q^n(m) = m. \tag{2}$$

An aware consumer accounts for the possibility that the review might be fake. By the law of total expectation, her expectation of the good's quality, conditional on seeing message m, is given by the probability that the reviewer is fake, given that m was sent, times the unconditional expected quality (since fake reviews are uninformative), plus the probability that the reviewer is real, given m, times m (since real reviews are truthful):

[13] Fake reviewers sometimes receive the product before writing a review, but not always. I follow Glazer et al. (2021) to account for both scenarios. The equilibrium strategy derived here remains optimal even if a fake reviewer observes the quality.

$$q^a(m) = E[X|m] = Pr(fake|m)E[X] + \Big(1 - Pr(fake|m)\Big)m. \tag{3}$$

To simplify notation, I define the expected posterior expectation

$$q(m) := (1-\nu)q^a(m) + \nu m \tag{4}$$

and refer to it as the *posterior*. I also denote $\bar{q} := E[X]$.

The consumer's utility function is given by

$$u_R = px + (1-p)y. \tag{5}$$

That is, her utility equals the good's quality x if she makes the purchase ($p = 1$), and her outside option y if she chooses it instead ($p = 0$).

The consumer buys the product if and only if, after seeing review m, her expected utility from purchasing weakly exceeds her outside option, i.e. if and only if $q^T(m) \geq y$ where $T \in \{a, n\}$ represents the consumer's type. Because the consumer's choice depends on the review m and the realization y, we can express her choice function as

$$p^T(y,m) = \begin{cases} 1 & q^T(m) \geq y \\ 0 & q^T(m) < y \end{cases}. \tag{6}$$

Breaking indifference in favor of buying is without loss of generality because $q^T(m) = y$ is a probability zero event due to Assumption A2.

2.3 Technical assumptions, timing, and solution concept

The prior probabilities that the consumer is naive, $\nu \in (0,1)$, and that the reviewer is fake, $\beta \in (0,1)$, as well as the distributions F_X and F_Y, are common knowledge. Throughout the article, I make the following assumptions:

Assumption A1. $0 < f_X(x) < \infty, \forall x \in [0,1]$.

Assumption A2. $F_Y = U[0,1]$.

Appendix C examines the role of Assumption A1 by considering an example where it fails. For some results, in particular Propositions 6, 7, and 8, I make the stronger assumption that $F_X = U[0,1]$.

The key implication of A2 is that a consumer's purchase probability is equal to her posterior expectation. Therefore, the overall purchase probability is equal to the posterior in (4).[14] If I instead assumed that the fake reviewer maximizes the posterior directly, as Glazer et al. (2021) do, most of the results would remain unaffected.[15] Specifically, Propositions 1 and 2 would remain unchanged. However, for some other results related to consumer surplus, such as Propositions 3, 4, and 5, an assumption about the distribution of outside options is necessary. I conjecture that even for those results, A2 is not crucial beyond

[14] Otherwise, the overall purchase probability would be $(1-\nu)F_Y(q^a(m)) + \nu F_Y(m)$, which is generally non-linear in $q^a(m)$ and $q^n(m)$. Linearity is crucial for solving the model.

[15] The absence of naive consumers in Glazer et al. (2021) and the lack of a welfare analysis renders these two assumptions equivalent in their context.

providing tractability and allowing for analytical results. Since the welfare effects constitute an important part of this article, I maintain A2 throughout.

The timing of the game is as follows:

1) Nature draws x, y, and a type for the consumer and the reviewer.
2) The reviewer observes his type and—if he is real—also the good's quality. He then sends a review, $m \in [0,1]$, to the consumer.
3) The consumer observes her outside option and the review, and then takes an action $p \in \{0,1\}$.
4) Payoffs are realized.

As is standard in games with asymmetric information, the solution concept used here is *Perfect Bayesian Equilibrium* (PBE). The fake reviewer maximizes the expected purchase probability, given his (prior) belief about the consumer's type. An aware consumer maximizes her payoff taking the reviewer's strategy as given and given her beliefs about his type. A naive consumer maximizes her payoff taking every review at face value. Formally, a PBE (henceforth equilibrium) of the game is a pair of purchasing strategies for the two consumer types and a reporting strategy for the fake reviewer type that fulfill the following conditions:

(a) $p^{a*}(m,y) \in \arg\max_{p\in\{0,1\}} pq^a(m) + (1-p)y$ for all $m \in \mathcal{M}$,
(b) $p^{n*}(m,y) \in \arg\max_{p\in\{0,1\}} pq^n(m) + (1-p)y$ for all $m \in \mathcal{M}$,
(c) $m^{F*} \in \arg\max_{m\in[0,1]} E_{Y,T}[p^{T*}(m,Y)]$ for all $m^{F*} \in supp\{f^F\}$,

as well as beliefs for the aware consumer type, which are consistent with Bayes' Rule. The subscripts Y and T in (c) denote that expectations are taken over both the consumer's type and her outside option.

3. EQUILIBRIUM CHARACTERIZATION

The first result establishes the existence of a unique equilibrium and characterizes the fake reviewer's equilibrium strategy.

Proposition 1. *For any $(\beta,\nu) \in (0,1)^2$, there exists a unique equilibrium. The reporting strategy of the fake reviewer is given by*

$$f^F(m) = \frac{1-\beta}{\beta}\frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m}$$

with $supp\{f^F\} = [c,1]$. Fake reviews induce a posterior of c, which is the unique solution to

$$\int_c^1 \frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m}dm = \frac{\beta}{1-\beta}.$$

To understand Proposition 1, it is helpful to look at Figure 1. The left panel depicts the review distributions of the two reviewer types when quality is distributed uniformly on $[0,1]$.

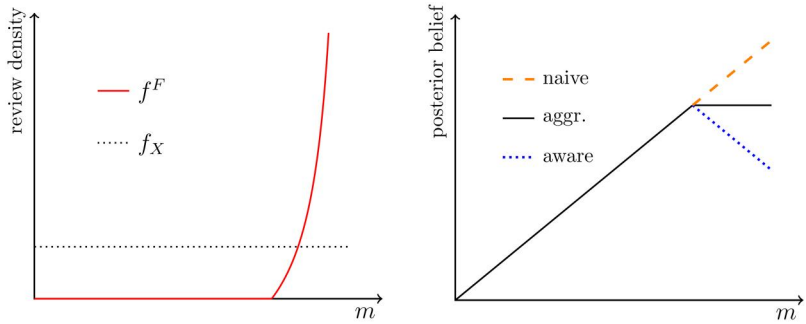

Figure 1. The left panel shows the fake reviewer's review distribution for $\beta = 0.5$, $\nu = 0.5$, and $F_X = U[0, 1]$. The right panel shows the posterior beliefs of the two consumer types together with the expected posterior belief.

Because the real reviewer is truthful, his reviews are also distributed uniformly (dotted black). The fake reviewer only sends reviews above a threshold value c (solid red).

To understand why the fake reviewer mixes over a range of messages, consider the alternative case where he always sends the highest message. Because an honest reviewer writes no single review with positive probability, the aware consumer type could infer his type with certainty and ignore the review. Hence, while such a review would be very effective in persuading a naive consumer, it would be ineffective in persuading an aware one. A slightly lower review would then not be anticipated by the aware consumer and thus persuade both types. For that same reason, the fake reviewer cannot send any review with positive probability and must mix over a range of messages. The more likely he is to write a certain review, the more skeptical an aware consumer is after reading it.

In his effort to persuade both types, the fake reviewer puts more probability mass on reviews the higher they are. The increased skepticism of the aware consumer is compensated for by the naive type's belief that it is truthful. The right panel of Figure 1 shows the posterior beliefs of both consumer types together with the posterior as defined in (4). They differ only for potentially fake reviews. While the naive type's beliefs are equal to the review (dashed orange), we can see that the aware type grows more skeptical the higher the review (dotted blue). This cancels out exactly, such that, in equilibrium, all fake reviews induce the same posterior (solid black).

Put plainly, for fake reviews to be effective in equilibrium, they must blend in with real ones. Rather than offering unqualified praise, a fake review may temper its tone and even highlight minor flaws to enhance its credibility. The model thus predicts that fake reviews are not limited to overtly positive assessments. Rather, also moderate reviews are potentially fake.

In contrast to most other cheap talk models, the equilibrium in Proposition 1 is unique, simplifying the analysis of the model's comparative statics in the following section. The reason for its uniqueness is twofold. First, the mechanic response of naive consumers to a review effectively transforms this model into one of costly signaling (Kartik et al. 2007). This implies that messages are not interchangeable, facilitating the uniqueness of the equilibrium. Second, the fake reviewer does not observe the quality therefore he cannot condition his reviewing strategy on it. If quality was observed by all reviewer types, the equilibrium in Proposition 1 would still exist, but it would no longer be unique. Other equilibria, like the *Negative Assortative Equilibrium* in Proposition 8, would emerge.

The differential interpretation of highly positive messages in the model aligns with empirical findings from financial markets, particularly regarding investor behavior following quarterly earnings conference calls. In these calls, managers often set an optimistic tone to influence investor perceptions of their company's future performance. Evidence from Blau et al. (2015) suggests that short sellers, who are typically more skilled and sophisticated than average investors, react differently to these high-tone communications. Specifically, Blau et al. (2015) show that these sophisticated investors are more likely to short stocks after conference calls where the managerial tone is abnormally optimistic, indicating their skepticism of overly positive messaging. In contrast, naive investors tend to overvalue these stocks, failing to recognize the exaggeration in the managers' tone.

4. EFFECTS OF AWARENESS POLICIES

In this section, I study the effects of awareness policies by analyzing the comparative statics of the model with respect to the share of naive consumers. For ease of notation, the results are stated in terms of an increase in ν, which corresponds to the opposite of an awareness campaign.

Proposition 2. *Suppose that* $\nu_2 > \nu_1$. *Then* $c_2 > c_1$ *and* F^{F,ν_2} *first-order stochastically dominates* F^{F,ν_1}.

The first part of Proposition 2 states that as the share of naive consumers increases, so does the posterior that a fake reviewer induces in equilibrium. Consequently, the purchase probability goes up. The second part of the proposition is more technical and important for proving some of the later results in this article. It states that the fake reviewer shifts probability mass to high messages in the sense of first-order stochastic dominance. This implies that fake reviews become more positive on average.

Figure 2 illustrates Proposition 2. As the share of naive consumers increases, the probability mass shifts toward high messages. Recall from Proposition 1 that the minimum of the support of f^F is equal to the posterior. Thus, we can see how both increase as ν goes up.

My model predicts that platforms with more naive consumers will have more positive fake reviews. Although the underlying mechanism is more involved, the idea is very simple. The larger a group of consumers gets the more fake reviewers focus on that group. Fake reviews are higher on average when the share of naive consumers increases because high reviews are effective in persuading these consumers. Conversely, when more consumers become aware, fake reviews are lower on average because moderate reviews are more credible in the eyes of aware consumers.

Extending this same logic to financial markets, my model predicts that stock recommendations and credit ratings become more inflated during periods when more investors are naive. This typically occurs during market booms, which see an influx of inexperienced investors. As documented by Ashcraft et al. (2010), credit ratings were more inflated during the boom leading up to the subprime crisis in 2007 than in the period that followed. This stands in contrast to Kartik et al. (2007), where an increase in the proportion of naive investors leads to *less* inflation.[16]

[16] Experimental evidence in Tang et al. (2020) also supports my model. In their experiment, participants issue credit ratings which are biased due to conflicts of interest. This bias is significantly reduced when the investor base is more sophisticated.

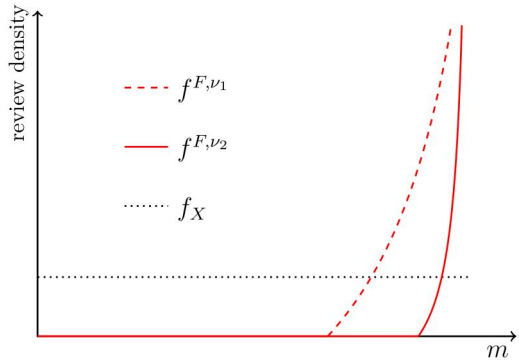

Figure 2. Equilibrium review distributions for $\nu_1 = 0.25$ and $\nu_2 = 0.75$.

Amazon's introduction of review verification can be seen as an effort to raise consumer awareness of fake reviews. By signaling that not all reviews are genuine, this policy likely increased such an awareness. However, testing its impact on reviews is difficult due to limited information about its implementation. While online sources suggest it began in 2013, the exact timing and rollout remain unclear. With more precise data, one could test the model's prediction that increased awareness led to less positive reviews and lower average ratings.

Next, I investigate the effects of awareness policies on consumer surplus. To evaluate such policies, I examine the expected consumer surplus. For a consumer of type $T \in \{a, n\}$, (ex-ante) consumer surplus is given by

$$\begin{aligned} CS^T = \beta \int_c^1 \Big([1 - F_Y(q^T(m))] E[Y | Y > q^T(m)] + F_Y(q^T(m)) \bar{q} \Big) f^F(m) dm \\ + (1 - \beta) \int_0^1 \Big([1 - F_Y(q^T(m))] E[Y | Y > q^T(m)] + F_Y(q^T(m)) m \Big) f_X(m) dm. \end{aligned} \tag{7}$$

We can break down (7) as follows. Upon observing some review m, a consumer either takes her outside option with probability $1 - F_Y(q^T(m))$, i.e. if it is greater than her posterior expectation, or she buys the good with the remaining probability $F_Y(q^T(m))$. In the former case, her expected payoff is the expectation of Y, conditional on it being above $q^T(m)$. In the latter case, her payoff is m if the review was real and $\bar{q}$ if it was fake. With probability β, the review is fake and sent with density $f^F(m)$. With the remaining probability, it is real and sent with density $f_X(m)$. Expected consumer surplus is then simply the weighted average of CS^n and CS^a, and the marginal effect of ν on consumer surplus is

$$\frac{dCS}{d\nu} = (CS^n - CS^a) + \nu \frac{dCS^n}{d\nu} + (1 - \nu) \frac{dCS^a}{d\nu}. \tag{8}$$

An *increase* in the likelihood of naivety—the *opposite* of an awareness campaign—affects expected consumer surplus in three ways. The first term on the right-hand side of (8) corresponds to the *direct effect* of moving consumers from the aware to the naive group. The second and third terms correspond to the *indirect effects* on the naive and aware consumer group, respectively.

The first term is always (weakly) negative. Were a naive consumer to enjoy a higher surplus, an aware consumer could simply imitate her.

To understand the effect on the naive consumer type, note that every fake review is deceptive in the sense that it claims a quality above $\bar{q}$, even though it is sent independently of x. When consumers are more likely to be naive, fake reviewers shift probability mass to higher reviews (Proposition 2), deceiving the naive consumer more severely. This *deception effect* is always negative.

The effect on the aware consumer type can intuitively be understood as follows: when a fake reviewer uses a more aggressive strategy in response to an increase in ν, his messages are stronger signals about his type, allowing the aware consumer to discount those messages more strongly. Moreover, because the lowest message sent by a fake reviewer increases, there is a larger set of messages that reveal the good's quality perfectly. Loosely speaking, we can say that separation increases. This *separation effect* is always positive.

Because the deception and the separation effects oppose each other, maximizing expected consumer surplus becomes a non-trivial problem. Increasing the welfare of one group comes at the cost of harming the other. Proposition 3 summarizes these findings about awareness policies. For the total effect, analytical results are provided only for the limiting cases $\nu \to 0$ and $\nu \to 1$. For intermediate parameter values, numerical results for a variety of distributions suggest a generally positive effect.[17]

Proposition 3. *In the unique equilibrium, awareness policies have opposing effects on the two consumer types. In particular,*

i) $\frac{dCS^n}{d\nu} < 0$,

ii) $\frac{dCS^a}{d\nu} > 0$,

iii) $\lim_{\nu \to 0} \frac{dCS}{d\nu} < 0$ *and* $\lim_{\nu \to 1} \frac{dCS}{d\nu} < 0$.

Figure 3 illustrates Proposition 3 for the case $F_X = U[0, 1]$. The key takeaway is that awareness policies can increase aggregate consumer surplus, especially when naivety is widespread. Importantly, they also raise the surplus of naive consumers, i.e., those who remain naive. Hence, even the "most vulnerable" - those who are difficult to reach—benefit from awareness policies.

5. MAIN RESULTS

In the baseline model, I assumed that all real reviewers were behavioral types who always report the good's true quality. While it provides a useful benchmark, this assumption might be too strong. It is reasonable to consider that even real reviewers might misrepresent their information if doing so helps counteract the negative effects of fake reviews. Their behavior may therefore be strategic rather than strictly honest.

In this section, I modify the model to allow for such behavior. With probability η, the real reviewer is non-strategic and reports truthfully, just as in the baseline model. With the remaining probability $1 - \eta$, he chooses a message strategically to maximize expected consumer surplus. This reflects the idea that some reviewers are motivated by a desire to help others make better decisions.

[17] For a variety of left- and right-skewed linear distributions the overall effect of awareness policies is also positive.

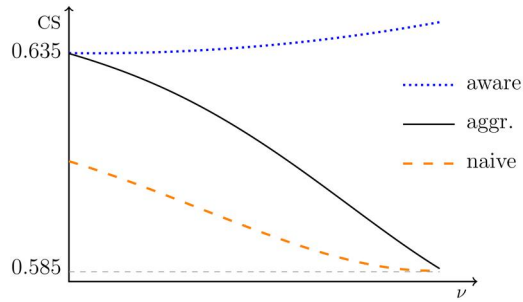

Figure 3. Naive, aware, and aggregate consumer surplus as a function of the share of naive consumers for $F_X = U[0,1]$.

Let $F^R_{m|x}$ denote the distribution of messages sent by a strategic real reviewer who observed quality x. Formally, I augment the baseline game with an additional equilibrium condition for the strategic real reviewer:

(d) $m^*_n \in \arg\max_{m\in[0,1]} E_{Y,T}[p^{T*}(m,Y)x + (1-p^{T*}(m,Y))Y]$ for all $m^* \in supp(F^R_{m|x})$ and $x \in [0,1]$.

In contrast to Jindapon and Oyarzun (2013) and Glazer et al. (2021), full honesty does not arise in equilibrium. Instead, a strategic real reviewer has an incentive to underreport.

Proposition 4. *Suppose that the real reviewer is non-strategic with probability $\eta < 1$ and strategic with probability $1-\eta$. An equilibrium where all real reviews are truthful does not exist.*

The intuition behind this result is illustrated by Figure 4, which shows the posterior expectations of the two consumer types when all real reviews are truthful. Because high truthful reviews are met with skepticism by the aware consumer, they are not very helpful for her. Whenever the aware consumer's outside option y is between the true quality m' and her posterior $q^a(m') < m'$, she forgoes the additional utility $m' - y$ by not purchasing. The gray area (light + dark) represents the mistake she *avoids* making in expectation when the reviewer deviates to a lower message m''. Similarly, the dark gray area represents the mistake a naive consumer makes in expectation due to this deviation. Hence, the light gray area corresponds to the net benefit of such a deviation.

Proposition 5 characterizes the equilibrium of the extended game.

Proposition 5. *For any $(\beta,\nu,\eta) \in (0,1)^3$, there exists a unique equilibrium. The strategic real reviewer reports truthfully for $x < c$ and sends $m = c$ whenever $x \geq c$. The fake reviewer mixes over the interval $[c\,,\,1]$. He sends $m = c$ with probability δ and messages in $(c,1]$ according to density*

$$f^F(m) = (1-\delta)\frac{1-\beta}{\beta}\frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m},$$

where

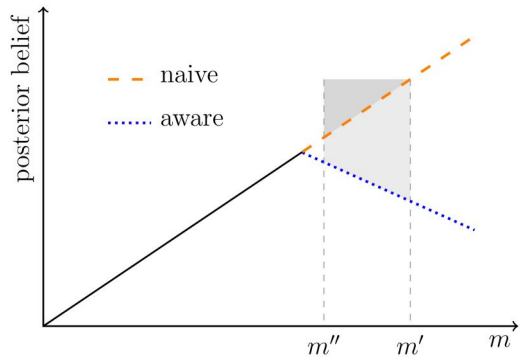

Figure 4. Deviation by a strategic real reviewer from a truthful message m' to m''.

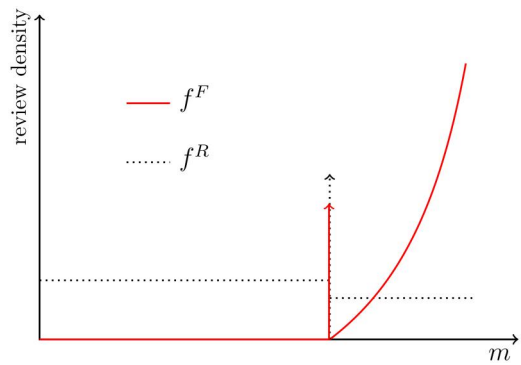

Figure 5. Density of real (dotted black) and fake (solid red) reviews in an equilibrium with strategic real reviewers.

$$\delta = (1-\eta)\frac{1-\beta}{\beta}\frac{\int_c^1 x f_X(x)dx - \frac{c}{1-F_X(c)}}{c-\bar{q}},$$

and c is given by the solution to

$$\int_c^1 \frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m}dm = \frac{\beta}{1-\beta}\frac{1-\delta}{\eta}.$$

In this equilibrium, a strategic real reviewer is *partially* truthful but underreports when the quality is above the threshold c because sending a lower review seems more credible. This credibility induces the fake reviewer to also send $m = c$ more often. In equilibrium, the probability with which he does this is precisely such that his incentive compatibility constraint is satisfied. Figure 5 illustrates the reviewers' strategies in this equilibrium.

Hence, underreporting by strategic real reviewers induces fake ones to send more moderate reviews. By an argument similar to that in Section 4, naive consumers benefit from fake reviews being less deceptive. The analogy to awareness policies is not too far off since, in the

limit, both the reduction of naive consumers and of behaviorally honest reviewers can result in the same market outcome.

The following proposition establishes this for the case where the quality is distributed uniformly.

Proposition 6. *Assume that* $F_X = U[0,1]$. *If the shares of naive consumers and/or that of fake reviewers are relatively low such that* $\nu < \frac{1-\sqrt{\beta}}{1+\sqrt{\beta}}$ *holds, then as real reviewers become strategic,* $\eta \to 0$, *the equilibrium outcome is equivalent to eliminating naive consumers,* $\nu \to 0$.

The intuition behind this result is as follows: as η approaches zero, all real reviewers become strategic and underreport, sending no messages above c. Therefore, fake reviewers also avoid sending messages above c, but it is precisely those high messages that deceive naive consumers. When they disappear from the review pool, naive consumers are no longer deceived and their decisions become as accurate as those of aware consumers. However, when the shares of naive consumers *and* fake reviewers are relatively high, this equilibrium breaks down and the equivalence no longer holds.[18]

The two limit equilibria are outcome equivalent because as $\eta \to 0$ (limit equilibrium I), the constant posterior converges to the same value as when $\nu \to 0$ (limit equilibrium II). For any pair of realizations (x, y), a consumer's purchase decision and thus her welfare and the reviewer's payoff are the same in both limit equilibria.

The implication of Proposition 6 is that awareness policies can be effective even when they fail to reach consumers if they instead make reviewers strategic. In the context of platforms where the distinction between consumers and reviewers is fluid, this result is particularly compelling. One way of thinking about the implication of Proposition 6 is the following: consumers differ in their responsiveness to awareness policies, and some might be very difficult to reach. These "most naive" consumers are not likely to consume a lot of media or interact much with the world around them. It is not far-fetched to assume that such consumers are also less likely to write reviews. Conversely, consumers who actively engage on review platforms are likely easier to reach with awareness policies. Proposition 6 then states that the full awareness outcome can be obtained even if not all consumers are susceptible to awareness policies, provided those who write reviews are.

Without the uniformity assumption, the proof of Proposition 6 becomes intractable, but I conjecture that a similar result holds more generally since the same intuition applies.[19]

6. DISCUSSION

In this section, I offer further thoughts about awareness policies and how this article contributes to the ongoing debate about regulating online platforms. I also discuss two extensions to the model, both of which are treated more formally in Appendix B.

[18] To see why, suppose that $\nu \to 1$, then sending $m = 1$ induces a posterior arbitrarily close to 1. This constitutes a profitable deviation for a fake reviewer in any equilibrium where he induces a posterior below 1. It can be shown that the equilibrium in this parameter region is characterized by fake reviewers mixing between $m = \frac{1}{1+\sqrt{\beta}}$ and $m = 1$, resulting in different beliefs for the two consumer types.

[19] It may be, however, that the posterior beliefs in the two limit equilibria do not coincide, such that the outcomes are similar but not equivalent.

6.1 Awareness policies

While this article remains agnostic about the specifics of awareness policies, I briefly discuss several possible implementations. The quote by Didier Reynders introducing this article suggests an approach where platforms inform consumers, for example by displaying prominent information about review reliability at the top of review pages. This method is likely effective and efficient, reaching affected consumers at the relevant moment. Awareness could also be raised through media campaigns. In 2016, the European Commission ran a two-week Facebook campaign on consumer rights. Similar efforts could highlight the problem of fake reviews. Finally, more indirect measures such as public support for research on fake reviews and increased media coverage could also raise awareness over time, even if their effects are less immediate.

Although this article does not directly investigate the incentives to raise consumer awareness, its findings help shed light on this question. Since aware consumers are harmed by increased awareness, it is unlikely that information about fake reviews is disseminated among consumers. Platforms that profit from increased sales also have little incentive to disclose the problem. This may explain why Amazon emphasizes its efforts to combat fake reviews but does little to inform users that such reviews remain common. Sellers who abstain from faking reviews might have such incentives but they lack credibility in claiming that their competitors use fake reviews whereas they do not. Third-party services like Fakespot provide useful information but mostly to consumers who are already aware. These findings highlight the importance for policymakers to take an active role in raising consumer awareness.

6.2 Negative fake reviewers

In the main part of the article, I have only considered positive fake reviews, i.e., fake reviews with the objective of increasing sales. In practice, however, fake reviews can be both positive and negative.[20] In Appendix B.1, I extend the baseline model to allow for both types of fake reviews. The resulting analysis is greatly facilitated by the *strategic independence* of positive and negative fake reviewers. This independence arises from the fact that positive and negative fake reviewers mix over disjoint intervals in equilibrium. Compared to the equilibrium of the baseline model, we observe a symmetric counterpart at the low end of the message space, as illustrated in Figure 6.[21] A change in the strategy of one type thus leaves the review distribution faced by the other type—and hence his strategy—unchanged.

This strategic independence allows me to fully characterize the equilibrium of this extended model, which is illustrated in Figure 6, and generalize all results from Sections 3–5 to this extended setting. I show this formally for Propositions 1–3, but the same can be done for Propositions 4–6.

6.3 Multiple reviewers and consumers

For the sake of parsimony, I formulated my model as the interaction between one reviewer and one consumer. This is a drastic simplification; in reality, multiple consumers read multiple reviews before making their purchase.[22]

[20] Mayzlin et al. (2014) not only show that hotels with higher incentives to manipulate reviews have more positive ones, but also that hotels whose competitors have higher incentives for manipulation have more negative reviews, indicating that negative fake reviews by rivals are a problem as well.

[21] This counterpart is exactly symmetric only when the shares of positive and negative fake reviews are the same. Otherwise, we have a "qualitatively symmetric" counterpart.

[22] A recent survey found that 88% of online shoppers read at least three reviews before completing a purchase, see https://digital.com/54-of-online-shoppers-read-reviews-before-every-purchase/. In other contexts, a single sender is more plausible. For example, patients often get medical advice from only one doctor.

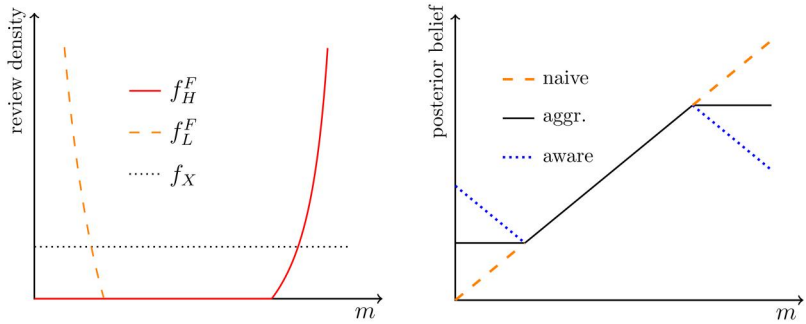

Figure 6. Equilibrium with positive and negative fake reviewers for $(\beta_H, \beta_L, \nu) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{2}\right)$ and $F_X = U[0,1]$. The left panel shows the review distributions. The right panel shows the posterior beliefs.

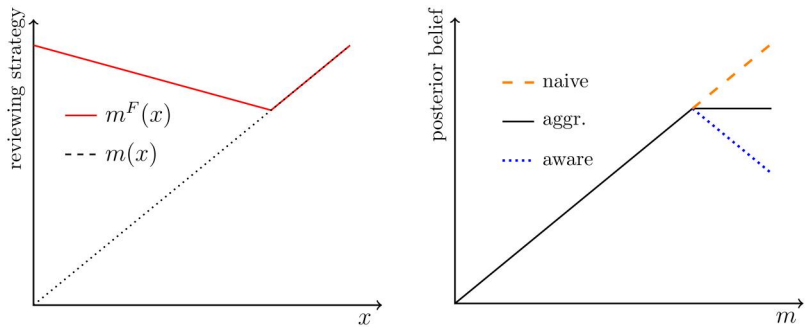

Figure 7. NAE for $(\beta, \nu, k) = \left(\frac{2}{3}, \frac{1}{2}, 3\right)$ with $\check{x} = \frac{25-4\sqrt{7}}{19} \approx 0.76$. The left panel shows the strategies of real (dotted black) and fake (solid red) reviewers. The right panel shows the posterior expectations of naive (dashed orange) and naive (dotted blue) consumers, and the posterior (solid black).

In Appendix B.2, I extend the model to allow for multiple consumers and multiple reviewers by replacing the single consumer with a unit mass of consumers and by allowing $k \geq 1$ reviewers to each post a review. Replacing the single consumer with a mass of consumers is straightforward and does not require changing the model or any additional analysis. One only needs to reinterpret the existing model.

Allowing for multiple reviewers, however, requires modifications to the model, since (2) and (3) do not specify consumers' beliefs for vectors of reviews. In addition, one must consider whether and how fake reviewers can coordinate. One possibility I consider is that coordination occurs through observing the quality of the good. I show that in this case, a *Negative Assortative Equilibrium* exists, which resembles the equilibrium of the baseline model in terms of induced beliefs (see Figure 7). In this equilibrium, fake reviewers map quality levels below a threshold $\check{x}_c$ onto high messages in a negative assortative way, while truthfully reporting quality levels above that threshold. Moreover, any set of fake reviews induces the same posterior.

7. CONCLUSION

This article contributes to the debate on regulating online platforms by theoretically analyzing commonly discussed awareness policies. I develop a model that captures two key features of the context: fake reviews and naive consumers. Awareness policies trade off the surplus of naive consumers against that of aware ones, but the positive effects typically dominate, leading to higher aggregate consumer surplus.

Unlike models without naive consumers, writing truthful reviews is not always optimal for strategic real reviewers seeking to maximize consumer surplus. Instead, underreporting arises in equilibrium, where they downplay a good's quality when it is very high. If all real reviewers act strategically, the outcome is equivalent to one where all consumers are aware, given that market conditions are not too unfavorable.

This result has practical implications. Online platforms do not have a clear boundary between reviewers and consumers. Awareness policies that target consumers will therefore also reach reviewers. If informed reviewers respond by writing strategically, awareness policies converge to the same outcome, regardless of whether consumers, reviewers, or both are targeted. Such policies, therefore, need not have a narrow focus but can be implemented broadly.

ACKNOWLEDGMENTS

I thank Maarten Janssen, Philipp Schmidt-Dengler, Daniel Garcia, Matan Tsur, Eeva Mauring, and Juha Tolvanen for support and helpful advice. This article has also benefited from a research stay at the Kelley School of Business and fruitful conversations with Rick Harbaugh during that time. I am also grateful for discussions with Carlos Oyarzun, Günter Strobl, and Alessandro De Chiara.

FUNDING

This research was financially supported by the Austrian Science Fund (FWF) and the Austrian Economic Association (NOeG). Boris Knapp is a receipient of a DOC Fellowship of the Austrian Academy of Sciences at the Vienna Graduate School of Economics (25766).

Conflict of interest statement. None declared.

APPENDIX A: TECHNICAL PROOFS

To prove Proposition 1, I will utilize Lemmata 1 and 2 which I formulate and prove below.

Lemma 1. *The fake reviewer's equilibrium strategy is characterized by an atomless distribution.*

Proof. I begin with the observation that the expected purchase probability equals the posterior:

$$\begin{aligned} E_T[Pr(Y \leq q^T(m))] &= (1-\nu)Pr(Y \leq q^a(m)) + \nu Pr(Y \leq q^n(m)) \\ &= (1-\nu)q^a(m) + \nu m = q(m). \end{aligned}$$

Therefore, the fake reviewer is maximizing the posterior and his incentive compatibility (IC) constraint is given by $q(m) = c \ \forall m \in supp\{f^F\}$. Recall that for any review, m, an aware consumer's posterior expectation is given by

$$q^a(m) = Pr(fake|m)\bar{q} + \Big(1 - Pr(fake|m)\Big)m. \tag{A1}$$

Since $q^a(m)$ is a convex combination between $\bar{q}$ and m, sending $m \leq \bar{q}$ is dominated by any $m > \bar{q}$ and so we only need to consider the latter case. Because f_X is bounded away from 0, so is the density of the real review distribution. Consequently, we have that for any m that the fake reviewer

- never sends: $Pr(fake|m) = 0$ and $q^a(m) = m$
- sends with positive probability: $Pr(fake|m) = 1$ and $q^a(m) = \bar{q}$
- sends with positive density: $0 < Pr(fake|m) < 1$ and $\bar{q} < q^a(m) < m$

I will prove a series of three claims which, taken together, prove Lemma 1.

Claim 1. *The fake reviewer's equilibrium strategy cannot contain more than one atom.*

Proof. Suppose, to the contrary, that there were two atoms, $m'_a < m''_a$. Because both of these messages are sent with positive probability, they induce posteriors $q(m'_a) = (1-\nu)\bar{q} + \nu m'_a$ and $q(m''_a) = (1-\nu)\bar{q} + \nu m''_a$. But then $q(m'_a) < q(m''_a)$ which violates the fake reviewer's IC constraint. □

Claim 2. *A message $m' < 1$ cannot be an atom in the fake reviewer's equilibrium strategy.*

Proof. Suppose to the contrary that m' is an atom. Then it induces a posterior of $q(m') = (1-\nu)\bar{q} + \nu m'$. But $q(1) \geq (1-\nu)\bar{q} + \nu > q(m')$ and therefore sending m' cannot be optimal for the fake reviewer. □

Claim 3. *The message $m' = 1$ cannot be an atom in the fake reviewer's equilibrium strategy.*

Proof. Suppose, to the contrary, that it was in which case it would induce a posterior of

$$q(1) = (1-\nu)\bar{q} + \nu. \tag{A2}$$

Let $\varepsilon > 0$ be arbitrarily small and consider the review $m'' = 1 - \varepsilon$. It cannot be the case that m'' is never sent in equilibrium by the fake reviewer because it would otherwise induce a posterior of $q(m'') = m''$. For small enough ε, $q(m'') = 1 - \varepsilon > (1-\nu)\bar{q} + \nu = q(1)$.

This implies that not only $m'' = 1 - \varepsilon$ has to be sent in equilibrium by the fake reviewer but also all $m \in (1-\varepsilon, 1)$. We can then let $f^F(m)$ denote the probability density associated with $m \in (1-\varepsilon, 1)$ and write $Pr(fake|m) = \frac{\beta f^F(m)}{(1-\beta)f_X(m) + \beta f^F(m)}$. We can use this and (A1) to express the posterior as

$$q(m) = (1-\nu)\left(\frac{(1-\beta)f_X(m)}{(1-\beta)f_X(m) + \beta f^F(m)}m + \frac{\beta f^F(m)}{(1-\beta)f_X(m) + \beta f^F(m)}\bar{q}\right) + \nu m. \tag{A3}$$

Utilizing the reviewer's IC constraint we can combine (A2) and (A3) and solve for $f^F(m)$ to obtain

$$f^F(m) = \frac{1-\nu}{\nu}\frac{1-\beta}{\beta}\frac{m-\bar{q}}{1-m}f_X(m) - \frac{1-\beta}{\beta}f_X(m). \tag{A4}$$

Since $[1-\varepsilon, 1] \subseteq supp\{f^F\}$, we must have that $\int_{1-\varepsilon}^{1} f^F(m)dm \leq 1$. Showing that $\int_{1-\varepsilon}^{1} \frac{m-\bar{q}}{1-m}dm = \infty$ and thus also $\int_{1-\varepsilon}^{1} f^F(m)dm = \infty$ contradicts the fact that $f^F(m)$ is a density. To do so, observe that $\frac{1-\varepsilon-\bar{q}}{1-m} \leq \frac{m-\bar{q}}{1-m} \forall m \in [1-\varepsilon, 1]$ and consider the integral $\int_{1-\varepsilon}^{1} \frac{1-\varepsilon-\bar{q}}{1-m}dm$, which can be integrated by substitution to obtain $\int_{1-\varepsilon}^{1} \frac{m-\bar{q}}{1-m}dm \geq \int_{1-\varepsilon}^{1} \frac{1-\varepsilon-\bar{q}}{1-m}dm = \infty$. □

Lemma 2. *In equilibrium, a fake reviewer mixes over the interval* [c , 1]*, where c is the posterior that he induces in equilibrium. Furthermore,* $c \geq \underline{c} = (1-\nu)\bar{q} + \nu > \bar{q}$.

Proof. Given that the fake reviewer mixes according to a distribution with density function $f^F(m)$ (Lemma 1) while the real one tells the truth, the posterior is given by (A3). The posterior is thus a convex combination of $\bar{q}$ and m, and, in particular,

$$m < (>)\bar{q} \Rightarrow q(m) < (>)\bar{q}.$$

Therefore, reviews below $\bar{q}$ can never be optimal for the fake reviewer and thus $supp\{f^F\} \subseteq [\bar{q}, 1]$.

Denote by c the constant posterior that fake reviews induce in equilibrium and notice that reviews above $\bar{q}$ induce a posterior (weakly) below their face value. This implies $c \leq inf(supp\{f^b\}) = \underline{m}$. But c cannot be strictly smaller than $\underline{m}$ because otherwise deviating to $\underline{m} - \varepsilon$ would be profitable for ε small enough. Hence, $c = inf(supp\{f^b\})$. Now, to show that $supp\{f^b\} = [c, 1]$, let us suppose that there is a message m' with $f^F(m') > 0$ and some $m'' > m'$ with $f^F(m'') = 0$. Because a fake reviewer never sends m'', it induces posterior $q(m'') = m''$. But since $c \leq m' < m''$, never sending m'' cannot be optimal. Therefore $f^F(m) > 0$ for all $m > c$, or, $supp\{f^b\} = [c, 1]$.

To show the last part of the proposition, note that the lowest posterior that a review $m = 1$ can induce is $\underline{c} = (1-\nu)\bar{q} + \nu$, namely if aware consumers are maximally skeptical and discount such a review completely. Thus, a fake reviewer must induce at least a posterior of $\underline{c}$ in equilibrium as otherwise deviating to $m = 1$ would constitute a profitable deviation. □

Proof of Proposition 1.

Proof. We can now use Lemmata 1 and 2 to find a mixed strategy equilibrium of the game. By Lemma 1, we can write the fake reviewer's IC constraint as

$$(1-\nu)\left(\frac{(1-\beta)}{(1-\beta)f_X(m) + \beta f^F(m)}m + \frac{\beta f^F(m)}{(1-\beta)f_X(m) + \beta f^F(m)}\bar{q}\right) + \nu m = c. \tag{A5}$$

Solving for $f^F(m)$ yields

$$f^F(m) = \frac{1-\beta}{\beta}\frac{(m-c)f_X(m)}{c - (1-\nu)\bar{q} - \nu m}. \tag{A6}$$

Since $supp(f^F) = [c, 1]$ (Lemma 2) and $f^F(m)$ must integrate to 1, the following equation pins down the equilibrium value of c:

$$\int_{c}^{1} \frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m}dm = \frac{\beta}{1-\beta}. \tag{A7}$$

In what follows, I will establish that (A7) always has a unique solution. For $\beta \in (0,1)$, the RHS of (A7) can get arbitrarily large or small. We need to show that the same is true for the LHS and that it is strictly monotone in c. We start by showing that

$$\int_{c}^{1} \frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m}dm \tag{A8}$$

can get arbitrarily large. Recall from Lemma 2 the lower bound $\underline{c} = (1-\nu)\bar{q}+\nu$. As $c \to \underline{c}$, we have (A8) tending to

$$\frac{1}{\nu}\int_{\underline{c}}^{1} \frac{m-\underline{c}}{1-m} f_X(m)dm.$$

Let $h(m) = \frac{m-\underline{c}}{1-m}$ and $g(m) = \frac{\hat{c}-\underline{c}}{1-m}$. Now,

$$g(m)\begin{cases} \geq h(m) & for\ \hat{c} \geq m \\ < h(m) & for\ \hat{c} < m \end{cases}$$

and $\int_{\underline{c}}^{1} g(m)dm = \infty$ for all $\hat{c} \in (\underline{c},1)$. Because $h(m) \leq g(m)$ only on an interval that contributes a finite part of that integral, and $h(m) > g(m)$ otherwise, we must have $\int_{\underline{c}}^{1} h(m)dm = \infty$. Since $f_X(m)$ is bounded away from 0, we have $\int_{\underline{c}}^{1} h(m)f_X(m)dm > \varepsilon \int_{\underline{c}}^{1} h(m)dm = \infty$.[23]

Next, I show that (A8) can get arbitrarily small. We can rewrite (A8), restricting the range of integration using the indicator function, as

$$\int_{0}^{1} \frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m} \mathbb{1}_{\{c \leq m \leq 1\}} dm, \tag{A9}$$

where the indicator function $\mathbb{1}$ equals 1 whenever the condition inside the braces is satisfied, and 0 otherwise. In the limit, as $c \to 1$, we have

$$\lim_{c \to 1} \int_{0}^{1} \frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m} \mathbb{1}_{\{c \leq m \leq 1\}} dm = \int_{0}^{1} \frac{(m-1)f_X(m)}{1-(1-\nu)\bar{q}-\nu m} \mathbb{1}_{\{m=1\}} dm.$$

Note that $\frac{m-1}{1-(1-\nu)\bar{q}-\nu m}$ equals 0 for $m=1$ and is finite for $m \in [0,1)$. Because $\mathbb{1}_{\{m=1\}} = 0$ for $m \in [0,1)$, we have the integral of a function that is 0 on the entire range of integration, and thus

$$\lim_{c \to 1} \int_{c}^{1} \frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m} dm = 0.$$

By the intermediate value theorem (IVT), (A7) must have a solution in c for any $\beta \in (0,1)$.

[23] That $f_X(m)$ is bounded away from 0 follows directly from Assumption A1.

Finally, I show that (A8) is monotonically decreasing in c. Taking the derivative with respect to c, we have by Leibniz' rule

$$\frac{d}{dc}\int_c^1 \frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m}dm = \int_c^1 \frac{d}{dc}\left(\frac{(m-c)f_X(m)}{c-(1-\nu)\bar{q}-\nu m}\right)dm$$
$$= \int_c^1 -\frac{\left[\overbrace{\left(c-(1-\nu)\bar{q}-\nu m\right)}^{>0}+\overbrace{(m-c)}^{\geq 0}\right]}{\underbrace{\left(c-(1-\nu)\bar{q}-\nu m\right)^2}_{>0}}\underbrace{f_X(m)}_{>0}dm,$$

an integral of a function that is strictly negative almost everywhere on its range of integration. Because (A8) is strictly monotone in c, the solution to (A7) is unique. □

Proof of Proposition 2. To show the first part, note that (A8) is increasing in ν and decreasing in c. The former is a straightforward observation and the latter has been shown in the proof of Proposition 1. As (A7) needs to hold in equilibrium, these two effects have to cancel out, implying $c_2 > c_1$.

For the second part, consider two equilibrium reporting strategies $f^{F,\nu_1}(m)$ and $f^{F,\nu_2}(m)$, with $\nu_1 < \nu_2$. We need to show that $f^{F,\nu_1}(m)$ and $f^{F,\nu_2}(m)$ intersect exactly once in $(c_1, 1)$. Note that $f^{F,\nu_i}(m) > 0 \forall m > c_i$ and $f^{F,\nu_i}(c_i) = 0$. Since we already established $c_2 > c_1$, we have

$$\int_{c_2}^1 f^{F,\nu_1}(m)dm < \int_{c_2}^1 f^{F,\nu_2}(m)dm = 1,$$

which means that $f^{F,\nu_1}(m) < f^{F,\nu_2}(m)$ for some $m \in (c_2, 1)$. At the same time, $f^{F,\nu_1}(c_2) > f^{F,\nu_2}(c_2) = 0$, so f^{F,ν_1} and f^{F,ν_2} intersect *at least* once in $(c_2, 1)$. Denote the largest such message in this interval by $\tilde{m}$. From the fake reviewer's IC constraint, we can obtain

$$Pr(fake|m) = \frac{c-m}{(1-\nu)(\bar{q}-m)}, \tag{A10}$$

and because $\tilde{m}$ is sent with equal density under f^{F,ν_1} and f^{F,ν_2} we have

$$\frac{c_1-\tilde{m}}{(1-\nu_1)(\bar{q}-\tilde{m})} = \frac{c_2-\tilde{m}}{(1-\nu_2)(\bar{q}-\tilde{m})}. \tag{A11}$$

By defining $D(m)$ as the difference between $Pr(fake|m)$ under ν_1 and ν_2, we can express (A11) as

$$D(\tilde{m}) = \frac{(1-\nu_2)c_1-(1-\nu_1)c_2+(\nu_2-\nu_1)\tilde{m}}{(1-\nu_1)(1-\nu_2)(\bar{q}-\tilde{m})} = 0. \tag{A12}$$

Now $-D(m)$ is quasi-monotone in $m \in (\bar{q}, 1]$, implying quasi-monotonicity of $f^{F,\nu_2}(m) - f^{F,\nu_1}(m)$ on that same interval.[24] Since $c_1 > \bar{q}$, f^{F,ν_1} and f^{F,ν_2} intersect exactly once in $(c_1, 1)$. □

[24] A function $\phi : R \mapsto R$ is quasi-monotone if for all $x, y \in R$: $\phi(x)(y-x) > 0 \Rightarrow \phi(y)(y-x) \geq 0$.

Proof of Proposition 3. I first prove a useful Lemma.

Lemma 3, *Let $g(x)$ be a continuous function with $g(x)>0$ for $x\in[a,b]$. Let $h(x)$ be a quasi-monotone function with $\int_a^b h(x)dx=0$. If g is strictly increasing (decreasing) on $[a, b]$, then $\int_a^b g(x)h(x)dx>(<)0$.*

Proof. I prove the case for increasing g. The case for decreasing g can be proved by simply reversing all inequalities. Quasi-monotonicity of $h(x)$ and $\int_a^b h(x)=0$ imply that $\exists\tilde{x}\in(a,b)$ s.t.

i) $h(\tilde{x})=0$,
ii) $-\int_a^{\tilde{x}} h(x)dx=\int_{\tilde{x}}^b h(x)dx$.

We can rewrite $\int_a^b g(x)h(x)dx$ as $\int_a^{\tilde{x}} g(x)h(x)dx+\int_{\tilde{x}}^b g(x)h(x)dx$ and notice that, since g is strictly increasing,

$$g(a)\int_a^{\tilde{x}} h(x)dx<\int_a^{\tilde{x}} g(x)h(x)dx<g(\tilde{x})\int_a^{\tilde{x}} h(x)dx.$$

By the IVT, $\exists\underline{x}\in(a,\tilde{x})$ s.t. $g(\underline{x})\int_a^{\tilde{x}} h(x)dx=\int_a^{\tilde{x}} g(x)h(x)dx$. Equivalently, we have

$$g(\tilde{x})\int_{\tilde{x}}^b h(x)dx<\int_{\tilde{x}}^b g(x)h(x)dx<g(b)\int_{\tilde{x}}^b h(x)dx,$$

and $\exists\bar{x}\in(\tilde{x},b)$ s.t. $g(\bar{x})\int_{\tilde{x}}^b h(x)dx=\int_{\tilde{x}}^b g(x)h(x)dx$. As g is strictly increasing, implying $g(\underline{x})<g(\bar{x})$, and due to (ii), we have

$$-g(\underline{x})\int_a^{\tilde{x}} h(x)dx<g(\bar{x})\int_{\tilde{x}}^b g(x)h(x)dx\Rightarrow\int_a^{\tilde{x}} g(x)h(x)dx+\int_{\tilde{x}}^b h(x)dx>0.$$

I now turn to the proof of Proposition 3. □

Part (i): Consumer surplus of a naive consumer is given by

$$\begin{aligned}CS^n=&\beta\int_c^1\Big([1-F_Y(m)]E[Y|Y>m]+F_Y(m)\bar{q}\Big)f^F(m)dm\\&+(1-\beta)\int_0^1\Big([1-F_Y(m)]\;E[Y|Y>m]+F_Y(m)m\Big)f_X(m)dm,\end{aligned}$$

which reduces to the following given that Y is distributed uniformly on $[0,1]$:

$$CS^n=\beta\int_c^1\left(\frac{1-m^2}{2}+m\bar{q}\right)f^F(m)dm+(1-\beta)\int_0^1\frac{1+m^2}{2}f_X(m)dm. \tag{A13}$$

Since $f^F(c)=0$, and the second integral is independent of ν, we have, by Leibniz' Rule,

$$\frac{dCS^n}{d\nu}=\frac{\beta}{2}\int_c^1\underbrace{(1+2\bar{q}m-m^2)}_{g(m)}\frac{df^F(m)}{d\nu}dm. \tag{A14}$$

Because $g(m)$ is positive and strictly decreasing on $(\bar{q},1]$, and, as shown in the proof of Lemma 2, $\frac{df^F(m)}{d\nu}$ is quasi-monotone, we have by Lemma 3 that $\frac{dCS^n}{d\nu}<0$.

Part (ii): An aware consumer's expected surplus is given by

$$CS^a = \beta \int_c^1 \Big([1 - F_Y(q^a(m))] \;\; E[Y|Y > q^a(m)] + F_Y(q^a(m))\bar{q} \Big) f^F(m) dm + (1-\beta) \int_0^1 \Big([1 - F_Y(q^a(m))] E[Y|Y > q^a(m)] + F_Y(q^a(m)) m \Big) f_X(m) dm,$$

which reduces to the following given that Y is distributed uniformly on $[0, 1]$:

$$CS^a = \beta \int_c^1 \left(\frac{1 - q^a(m)^2}{2} + q^a(m)\bar{q} \right) f^F(m) dm + \frac{1-\beta}{2} \int_0^c (1 + m^2) f_X(m) dm + \frac{1-\beta}{2} \int_c^1 \Big(1 + 2q^a(m)m - q^a(m)^2 \Big) f_X(m) dm. \tag{A15}$$

The derivative with respect to ν, using Leibniz' Rule and rearranging, is

$$\frac{dCS^a}{d\nu} = \underbrace{\beta \int_c^1 \Big(\bar{q} - q^a(m) \Big) \frac{dq^a(m)}{d\nu} f^F(m)}_{A1} + \underbrace{\frac{\beta}{2} \int_c^1 \Big(1 + 2\bar{q}q^a(m) - q^a(m)^2 \Big) \frac{df^F(m)}{d\nu} dm}_{A2} + \underbrace{\frac{1-\beta}{2} (1 + c^2) f_X(c) \frac{dc}{d\nu}}_{B1} + \underbrace{\frac{1-\beta}{2} \int_0^c \frac{d}{d\nu} \Big((1 + m^2) f_X(m) \Big) dm}_{B2} + \underbrace{\frac{1-\beta}{2} \int_c^1 \Big(m - q^a(m) \Big) \frac{dq^a(m)}{d\nu} f_X(m) dm}_{C1} - \underbrace{\frac{1-\beta}{2} (1 + c^2) f_X(c) \frac{dc}{d\nu}}_{C2}.$$

Now, $B1$ and $C2$ cancel out and $B2 = 0$ because $(1 + m^2) f_X(m)$ is independent of ν. We are left with $\frac{dCS^a}{d\nu} = A1 + A2 + C1$. We can rearrange $A1 + C1$ to obtain

$$\int_c^1 \Big(\beta \Big(\bar{q} - q^a(m) \Big) f^F(m) + (1-\beta) \Big(m - q^a(m) \Big) f_X(m) \Big) \frac{dq^a(m)}{d\nu},$$

and use $q^a(m) = \frac{\beta f^F(m)\bar{q} + (1-\beta) f_X(m) m}{\beta f^F(m) + (1-\beta) f_X(m)}$ to show that $A1 + C1 = 0$. What remains is $A2$, so

$$\frac{dCS^a}{d\nu} = \frac{\beta}{2} \int_c^1 \underbrace{\Big(1 + 2\bar{q}q^a(m) - q^a(m)^2 \Big)}_{g(m)} \frac{df^F(m)}{d\nu} dm. \tag{A16}$$

Because $g(m)$ is positive and strictly increasing on $(\bar{q}, 1]$, and $\frac{df^F(m)}{d\nu}$ is quasi-monotone, we have by Lemma 3 that $\frac{dCS^a}{d\nu} > 0$.

Part (iii): The marginal effect of increasing the share of naive consumers, ν, is

$$\frac{dCS}{d\nu} = \underbrace{(CS^n - CS^a)}_{A} + \underbrace{\nu \frac{dCS^n}{d\nu}}_{B} + \underbrace{(1-\nu) \frac{dCS^a}{d\nu}}_{C}. \tag{A17}$$

To prove that (A17) is negative in the limit as $\nu \to 0$ and $\nu \to 1$, I begin by showing that A is negative. To do so, it suffices to show it in the limit as $\nu \to 0$ because in parts (i) and (ii), we showed that $\frac{dCS^n}{d\nu} < 0$ and $\frac{dCS^a}{d\nu} > 0$. From (A13) and (A15) we write A as

$$\begin{aligned} CS^n - CS^a = &\frac{\beta}{2}\int_c^1 \Big(q^a(m)^2 - m^2 + 2\bar{q}(m - q^a(m))\Big) f^F(m)dm \\ &+ \frac{1-\beta}{2}\int_c^1 \Big(m - q^a(m)\Big)^2 f_X(m)dm. \end{aligned} \tag{A18}$$

From the fake reviewer's IC constraint we have that $q^a(m) = \frac{c - \nu m}{1-\nu}$ and therefore $\lim_{\nu \to 0} q^a(m) = c$. Furthermore, we have $\lim_{\nu \to 0} f^F(m) = \lim_{\nu \to 0} \frac{1-\beta}{\beta} \frac{(m-c)f_X(m)}{c - (1-\nu)\bar{q} - \nu m} = \frac{1-\beta}{\beta}\frac{m-c}{c-\bar{q}} f_X(m)$ and therefore

$$\begin{aligned} \lim_{\nu \to 0}(CS^n - CS^a) &= \frac{1-\beta}{2}\int_c^1 (c-m)(c+m-2\bar{q})\frac{m-c}{c-\bar{q}} f_X(m)dm \\ &\quad + \frac{1-\beta}{2}\int_c^1 (m-c)^2 f_X(m)dm \\ &= \frac{1-\beta}{2}\int_c^1 \left(\frac{\bar{q}-m}{c-\bar{q}} + (m-c)^2 - 1\right) f_X(m)dm. \end{aligned} \tag{A19}$$

The expression in parentheses is negative because $\bar{q} < m$, $c > \bar{q}$, and $(m-c)^2 < 1$, while $f_X(m) > 0, \forall m \in [c, 1]$. Therefore, $\lim_{\nu \to 0}(CS^n - CS^a) < 0$ and consequently $A < 0$ holds.

Next, we show that $B = 0$ in both limiting cases. Let us first consider $\nu \to 0$, in which case $c \to c^{JO} = \frac{1}{1+\sqrt{\beta}}$.[25] From (A14), we then have

$$-\infty < \lim_{\nu \to 0} \frac{dCS^n}{d\nu} = \frac{\beta}{2}\int_{c^{JO}}^1 (1 + 2\bar{q}m - m^2)\frac{df^F(m)}{d\nu}dm < 0,$$

and thus $\lim_{\nu \to 0} \nu \frac{dCS^n}{d\nu} = 0$. In the limiting case $\nu \to 1$ we have that $c \to 1$ and therefore $\lim_{\nu \to 1} \frac{dCS^n}{d\nu} = 0$, which in turn implies $\lim_{\nu \to 1} \nu \frac{dCS^n}{d\nu} = 0$. Hence, $\lim_{\nu \to 0} B = \lim_{\nu \to 1} B = 0$.

What is left to show is that $C = 0$ in both limiting cases. Beginning again with $\nu \to 0$, and using $\lim_{\nu \to 0} c = c^{JO}$ and $\lim_{\nu \to 0} q(m) = c$, we obtain

$$\lim_{\nu \to 0} \frac{dCS^a}{d\nu} = (1 + 2\bar{q}c - c^2)\int_{c^{JO}}^1 \frac{df^F(m)}{d\nu}dm = 0,$$

and thus $\lim_{\nu \to 0}(1-\nu)\frac{dCS^a}{d\nu} = 0$. In the limiting case $\nu \to 1$, the fact that $c \to 1$ gives us $\lim_{\nu \to 1} \frac{dCS^a}{d\nu} = 0$ and consequently, $\lim_{\nu \to 1}(1-\nu)\frac{dCS^n}{d\nu} = 0$. Hence, $\lim_{\nu \to 0} C = \lim_{\nu \to 1} C = 0$. Combining the results for A, B, and C, we have $\lim_{\nu \to 0} \frac{dCS}{d\nu} < 0$ and $\lim_{\nu \to 1} \frac{dCS}{d\nu} < 0$. □

[25] For $\nu = 0$, my model reduces to the main model in Jindapon and Oyarzun (2013) who derive the corresponding equilibrium with $c^{JO} > \bar{q}$.

Proof of Proposition 4. Suppose all reviewers and consumers play according to the equilibrium in Proposition 1. If a strategic real reviewer benefits from deviating, he must increase the aware consumer's expected welfare. This is because he cannot increase that of a naive consumer—since she takes messages at face value, telling the truth maximizes her expected welfare. In what follows, I will show that by deviating, a strategic real reviewer can trade off small mistakes of a naive against large gains of an aware consumer, thereby increasing expected consumer surplus.

A deviation can only be beneficial if it is upon observing $x > c$, because, for $x \leq c$, truthful reviews induce the correct posterior expectation in *both* consumer types. For $m > c$, however, the posterior expectation of an aware consumer is below the observed quality and decreasing in m (see Figure 1).

Note that upon seeing a truthful review $m \in (c, 1]$, naive consumers always choose the better of the two options. Were the reviewer to underreport, they would instead make some mistakes in expectation. In particular, whenever their outside option lies between the deviation message $m' < m$ and the honest message (and hence true quality) m, they would not purchase even though they should. For an outside option $y \in [m', m]$ the size of the mistake is $m - y$, i.e., the utility they would have gotten if they bought the good minus the outside option that they chose instead. For a deviation to some message $m' \in [c, m]$, a naive consumer's expected mistake is given by

$$\int_{m'}^{m} f_Y(\xi)(m-\xi)d\xi = \int_{m'}^{m} (m-\xi)d\xi = \left(m\xi - \frac{\xi^2}{2}\right)\Bigg|_{m'}^{m}.$$

Due to her skepticism, the aware consumer makes mistakes when real reviews are truthful. For messages m in the support of the fake reviewer's strategy, $[c, 1]$, her posterior expectation is $q^a(m) = \frac{c - \nu m}{1 - \nu} < m$. Thus, whenever their outside option is between $q^a(m)$ and m, they do not purchase although it would be optimal. If a real reviewer deviated from telling the truth to some $m' \in [c, m)$, an aware consumer would avoid making a mistake of size $m - y$ whenever her outside option was between $q^a(m)$ and $q^a(m')$. Her expected *avoided* mistake is given by

$$\int_{q^a(m)}^{q^a(m')} f_Y(\xi)(m-\xi)d\xi = \int_{q^a(m)}^{q^a(m')} (m-\xi)d\xi = \left(m\xi - \frac{\xi^2}{2}\right)\Bigg|_{\xi=q^a(m)}^{\xi=q^a(m')}.$$

To show that a profitable deviation exists, consider a deviation from $m \in (c, 1]$ to c. It is profitable if the expected avoided mistake outweighs the expected mistake:

$$\nu\left(m\xi - \tfrac{\xi^2}{2}\right)\Big|_{c}^{m} < (1-\nu)\left(m\xi - \tfrac{\xi^2}{2}\right)\Big|_{\xi=q^a(m)}^{\xi=q^a(c)} \iff$$

$$\nu\left[\left(m^2 - \frac{m^2}{2}\right) - \left(mc - \frac{c^2}{2}\right)\right] < (1-\nu)\left[\left(mq^a(c) - \frac{q^a(c)^2}{2}\right) - \left(mq^a(m) - \frac{q^a(m)^2}{2}\right)\right].$$

Note that $q^a(m) = \frac{c - \nu m}{1 - \nu}$ and thus in particular $q^a(c) = c$. Thus, we have

$$\nu\left[\left(m^2-\frac{m^2}{2}\right)-\left(mc-\frac{c^2}{2}\right)\right]<(1-\nu)\left[\left(mc-\frac{c^2}{2}\right)-\left(m\frac{c-\nu m}{1-\nu}-\frac{\left(\frac{c-\nu m}{1-\nu}\right)^2}{2}\right)\right]\iff$$

$$\nu\left[\frac{m^2}{2}-\frac{2mc-c^2}{2}\right]<(1-\nu)\left[\frac{2mc-c^2}{2}-\frac{2m\frac{c-\nu m}{1-\nu}-\left(\frac{c-\nu m}{1-\nu}\right)^2}{2}\right]\iff$$

$$\nu\left[\frac{m^2}{2}-\frac{m^2-(m-c)^2}{2}\right]<(1-\nu)\left[\frac{m^2-(m-c)^2}{2}-\frac{m^2-\left(m-\frac{c-\nu m}{1-\nu}\right)^2}{2}\right]\iff$$

$$\nu\left[\frac{(m-c)^2}{2}\right]<(1-\nu)\left[\frac{\left(m-\frac{c-\nu m}{1-\nu}\right)^2}{2}-\frac{(m-c)^2}{2}\right]$$

and finally after noting that $m-\frac{c-\nu m}{1-\nu}=\frac{m-c}{1-\nu}$, collecting all terms on the LHS, and some rearranging we have

$$\frac{\nu}{(1-\nu)^2}\frac{(m-c)^2}{2}>0,$$

which holds for all pairs (ν,m) with $\nu\in(0,1)$ and $m\in(c,1]$. A deviation from some truthful message $m\in(c,1]$ to c is thus profitable for a strategic real reviewer. □

Proof of Proposition 5. I begin by looking at best responses starting from strategies as in the equilibrium in Proposition 1. As shown in the proof of Proposition 4, strategic real reviewers benefit from deviating to a lower message whenever they observe a quality above c. Their best response for $x\in[c,1]$ is to send $m=c$. But then, $m=c$ would induce a posterior above c, because an aware consumer's posterior expectation would be equal to $\frac{1+c}{2}>c$, and a naive one's would be equal to c. Hence, the posterior—a convex combination of the two posterior expectations—would lie above c. This would provide an incentive for fake reviewers to deviate to $m=c$ because it would induce the highest posterior. Suppose that fake reviewers send $m=c$ with some probability δ, while they mix over $[c,1]$ with the remaining probability. Then, the aware consumers' posterior expectation, given review $m=c$, is

$$q^a(c)=\frac{\delta\beta}{\delta\beta+(1-F_X(c))(1-\beta)(1-\eta)}\bar{q}+\frac{(1-F_X(c))(1-\beta)(1-\eta)}{\delta\beta+(1-F_X(c))(1-\beta)(1-\eta)}E[X|X>c].$$

Two things need to be true. First, $q^a(c)=c$ so that $q(c)=c$ ensures the strategic real reviewers do not have an incentive to deviate when $x=c$. Otherwise, $m=c-\varepsilon$ would constitute a profitable deviation for small enough ε. Second, $m=c$ has to induce the same posterior as the other messages that the fake reviewer sends. These two facts imply that $q(m)=c$ for all $m\in[c,1]$. Using the first fact, $q^a(c)=c$, and solving for δ yields

$$\delta=(1-\eta)\frac{1-\beta}{\beta}\frac{\int_c^1 xf_X(x)dx-\frac{c}{1-F_X(c)}}{c-\bar{q}}. \tag{A20}$$

The RHS is continuous and strictly decreasing in c for $c > \bar{q}$. Thus, (A20) implicitly defines a continuous and strictly decreasing function $c^A(\delta)$ with a maximum of $c(0) = 1$. The second fact, i.e. the fake reviewer's IC constraint, implies

$$\int_c^1 \eta \frac{1-\beta}{\beta} \frac{(m-c) f_X(m)}{c-(1-\nu)\bar{q}-\nu m} dm = 1-\delta, \tag{A21}$$

which mirrors the equilibrium condition in the baseline model (see (A7) in the proof of Proposition 1), but real reviewers write reviews in $[c, 1]$ with density ηf_X instead of f_X, and the fake reviewer's review density integrates to $1-\delta$ instead of 1. Rearranging gives

$$\int_c^1 \frac{(m-c) f_X(m)}{c-(1-\nu)\bar{q}-\nu m} dm = \frac{1-\delta}{\eta} \frac{\beta}{1-\beta}. \tag{A22}$$

The RHS of (A22) is strictly decreasing in δ and, as shown in the proof of Proposition 1, the LHS is strictly decreasing in c. Thus, (A22) implicitly defines a strictly increasing function $c^B(\delta)$ with a maximum of $c^B(1) = 1$. We now have $c^A(0) > c^B(0)$ and $c^A(1) < c^B(1)$. By the IVT, $\exists \delta^*$ s.t. $c^A(\delta^*) = c^B(\delta^*)$. □

Proof of Proposition 6. Suppose $\nu < \frac{1-\sqrt{\beta}}{1+\sqrt{\beta}}$ which is equivalent to $(1-\nu)\bar{q} + \nu = \underline{c} < c^{JO} = \frac{1}{1+\sqrt{\beta}}$. Recall that (A20) implicitly defines a strictly decreasing function $c^A(\delta)$. Let $\eta \to 0$. Then $c^A(0) = 1$ and $c^A(1) = c^{JO}$.

Likewise, recall that (A22) implicitly defines a strictly increasing function $c^B(\delta)$. Let $\langle \eta_n \rangle$ be the corresponding sequence as $\eta \to 0$. Then, for every $\delta < 1$, $\exists N$ s.t. $\forall n > N$ $\frac{1-\delta}{\eta_n} \frac{\beta}{1-\beta} > E, \forall E > 0$. Thus, for $\delta < 1$ we have that $\lim_{\eta \to 0} c^B(\delta) = \underline{c}$. Now let $\delta \to 1$ and $\langle \delta_k \rangle$ be the corresponding sequence. $\forall \eta_n$ $\exists K$ s.t. $\forall k > K$, $\frac{1-\delta_k}{\eta_n} \frac{\beta}{1-\beta} < E, \forall E > 0$. Thus, for $\delta \to 1$ we have that $\lim_{\eta \to 0} c^B(\delta) \to 1$. All this is illustrated in Figure 8 and implies that in the limit, as $\eta \to 0$, $\delta^* \to 1$ and $c(\delta^*) \to c^{JO}$.

In the limit there are no non-strategic real reviewers and $\delta^* \to 1$, which means that reviews above c^{JO} are not sent in equilibrium. Reviews then induce the same posterior expectation in both consumer types, namely $q^T(m) = m$ for $m < c^{JO}$ and $q^T(m) = c^{JO}$ for $m \geq c^{JO}$, $T \in \{a, n\}$. A posterior of c^{JO} is thus induced if the reviewer is fake or if he is real and the quality is above c^{JO}. If the reviewer is real and the quality is below c^{JO}, he induces the correct posterior. These are the same posteriors induced in an equilibrium where all real reviewers are non-strategic and all consumers are aware. □

APPENDIX B

B.1 NEGATIVE FAKE REVIEWS

In the main part of the article, I have only considered positive fake reviews, i.e., fake reviews with the objective of increasing sales. In this section, I extend the model to allow for both positive and negative fake reviews. Let β_H and β_L denote the probabilities that a review is a positive or negative fake review, respectively, and let $\beta = \beta_H + \beta_L$ be the total probability that a review is fake. Then, $1-\beta$ is the probability that a review is real and thus truthful.

Let F_i^F and f_i^F, $i \in \{H, L\}$, denote the distribution and density functions of positive and negative fake reviews. Replacing equilibrium condition (c) with the following conditions:

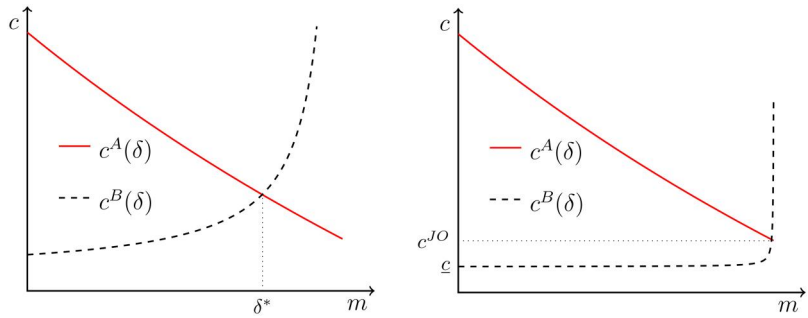

Figure 8. Left panel: $c^A(\delta)$ and $c^B(\delta)$ intersect exactly once in $[0,1]$. Right panel: $c^A(\delta)$ and $c^B(\delta)$ in the limit as $\eta \to 0$.

(c.1) $m_H^{F*} \in \arg\max_{m\in[0,1]} E_{Y,T}[p^{T*}(m,Y)]$ for all $m^* \in supp\{f_H^F\}$, and
(c.2) $m_L^{F*} \in \arg\min_{m\in[0,1]} E_{Y,T}[p^{T*}(m,Y)]$ for all $m^* \in supp\{f_L^F\}$.

I first prove the following lemma:

Lemma 4. *In equilibrium, the two fake reviewer types mix over disjoint sets of messages. Their equilibrium strategies are represented by atomless densities f_H^F and f_L^F. Furthermore, $supp\{f_H^F\} = [c_H, 1]$ and $supp\{f_L^F\} = [0, c_L]$ with $c_L < \bar{q} < c_H$.*

Proof. First note that a negative fake reviewer, in his effort to minimize the purchase probability, induces the lowest posterior with all messages that he sends in equilibrium. The argument is similar to the one in Lemma 1, but instead of maximizing it, the negative fake reviewer is minimizing the posterior. Next, I show that there is no message that both types send. The posterior induced by any message m, $q(m)$, is a convex combination of $\bar{q}$ and m and therefore $q(m) < \bar{q}$ for any $m < \bar{q}$, and $q(m) > \bar{q}$ for any $m > \bar{q}$. Hence, positive (negative) types can only send messages strictly above (below) $\bar{q}$ which proves the first statement. To prove the second and third statements, one can then proceed as in the proofs of Lemmata 1 and 2 separately for the positive and negative fake reviewer types. □

Lemma 4 plays the role of Lemmata 1 and 2 of the baseline model, establishing that the fake reviewers' strategies are atomless distributions and characterizing their supports. In doing so, the lemma establishes *strategic independence* between the two fake reviewer types. Because the supports of their strategies are disjoint, a change in the behavior of one type does not affect the other type's behavior.

Equipped with this result, I characterize the fake reviewers' equilibrium strategies by extending Proposition 1:

Proposition 1b. *For any $(\beta_H, \beta_L, \nu) \in (0,1)^3$, there exists a unique equilibrium. The reporting strategy of the fake reviewer is given by*

$$f_i^F(m) = \frac{1-\beta_i}{\beta_i} \frac{(m-c_i) f_X(m)}{c_i - (1-\nu)\bar{q} - \nu m},$$

$i \in \{L, H\}$, with $supp\{f_L^F\} = [0, c_L]$ and $supp\{f_H^F\} = [c_H, 1]$. Positive fake reviews induce a posterior of c_H, which is the unique solution to

$$\int_{c_H}^{1} \frac{(m-c_H)f_X(m)}{c_H-(1-\nu)\bar{q}-\nu m}\,dm = \frac{\beta_H}{1-\beta}.$$

Negative fake reviews induce a posterior of c_L, which is the unique solution to

$$\int_{0}^{c_L} \frac{(m-c_L)f_X(m)}{c_L-(1-\nu)\bar{q}-\nu m}\,dm = \frac{\beta_L}{1-\beta}.$$

Proof. Because the fake reviewers' equilibrium strategies are atomless distributions, we have

$$q^a(m) = \frac{\beta_L f_L^F(m)+\beta_H f_H^F(m)}{(1-\beta)f_X(m)+\beta_L f_L^F(m)+\beta_H f_H^F(m)}\bar{q} + \frac{(1-\beta)f_X(m)}{(1-\beta)f_X(m)+\beta_L f_L^F(m)+\beta_H f_H^F(m)}m.$$

Note that $f_L^F(m)=0$ for $m\in[c_L,1]$ and $f_H^F(m)=0$ for $m\in[0,c_H]$, which reduces the aware consumer's posterior expectation in these cases to

$$q^a(m) = \frac{\beta_H f_H^F(m)}{(1-\beta)f_X(m)+\beta_H f_H^F(m)}\bar{q} + \frac{(1-\beta)f_X(m)}{(1-\beta)f_X(m)+\beta_H f_H^F(m)}m, \forall m\in[c_H,1] \quad \text{(B1)}$$

and

$$q^a(m) = \frac{\beta_L f_L^F(m)}{(1-\beta)f_X(m)+\beta_L f_L^F(m)}\bar{q} + \frac{(1-\beta)f_X(m)}{(1-\beta)f_X(m)+\beta_L f_L^F(m)}m, \forall m\in[0,c_L]. \quad \text{(B2)}$$

We can combine the positive type's IC constraint with (B1) and solve for f_H^F to obtain

$$f_H^F(m) = \frac{1-\beta}{\beta_H}\frac{(m-c_H)f_X(m)}{c_H-(1-\nu)\bar{q}-\nu m}. \quad \text{(B3)}$$

To prove that there exists a c_H such that (B3) represents the positive type's equilibrium strategy, one can proceed as in the proof of Proposition 1, replacing c with c_H and β with β_H.

The proof for the negative type is different enough to warrant its own exposition. Combining the negative type's IC constraint with (B2) and solving for f_L^F yield

$$f_L^F(m) = \frac{1-\beta}{\beta_L}\frac{(m-c_L)f_X(m)}{c_L-(1-\nu)\bar{q}-\nu m}, \quad \text{(B4)}$$

which has to integrate to 1 over its support $[0,c_L]$. Therefore, the following equation implicitly defines the equilibrium value c_L:

$$\int_{0}^{c_L} \frac{(m-c_L)f_X(m)}{c_L-(1-\nu)\bar{q}-\nu m}\,dm = \frac{\beta_L}{1-\beta}. \quad \text{(B5)}$$

To show that a solution to (B5) always exists, we begin by showing that the LHS can get arbitrarily big when $c_L \to \underline{c}_L = (1-\nu)\bar{q}$, which is the largest expectation sending $m=0$ can induce. In this case, the LHS converges to

$$\frac{1}{\nu}\int_0^{c_L}\frac{\underline{c}_L - m}{m} f_X(m)dm. \tag{B6}$$

Because $\int_0^c \frac{\underline{c}-m}{m}dm = \infty, \forall c > 0$, and since $\nu \in (0,1)$ and $f_X(m)$ is bounded away from 0, the LHS is not bounded from above. To show that it can get arbitrarily small rewrite the LHS of (B5) as

$$\int_{-\infty}^{\infty} \frac{(m-c_L)f_X(m)}{c_L - (1-\nu)\bar{q} - \nu m} \mathbb{1}_{\{0 \le m \le \underline{c}_L\}} dm, \tag{B7}$$

and consider the limit as $c_L \to 0$:

$$\lim_{c_L \to 0} \int_{-\infty}^{\infty} \frac{(m-c_L)f_X(m)}{c_L - (1-\nu)\bar{q} - \nu m} \mathbb{1}_{\{0 \le m \le c_L\}} dm = \int_{-\infty}^{\infty} \frac{m f_X(m)}{(1-\nu)\bar{q} - \nu m} \mathbb{1}_{\{m=0\}} dm. \tag{B8}$$

Because $f_X(m)$ is finite, we have $\frac{m f_X(m)}{(1-\nu)\bar{q} - \nu m} = 0$ for $m = 0$. For $m \neq 0$, we have $\mathbb{1}_{\{m=0\}} = 0$, and thus the limit of the LHS as $c_L \to 0$ is 0. Since the LHS of (B5) is a continuous function of c_L, the IVT guarantees that a solution for (B5) exists. What remains to be shown is that this solution is always unique. For this, we take the derivative of the LHS of (B5), applying Leibniz' Rule:

$$\begin{aligned}\frac{d}{dc_L}\int_0^{c_L}\frac{(m-c_L)f_X(m)}{c_L-(1-\nu)\bar{q}-\nu m}dm &= \int_0^{c_L}\frac{d}{dc_L}\left(\frac{(m-c_L)f_X(m)}{c_L-(1-\nu)\bar{q}-\nu m}\right)dm \\ &= \int_0^{c_L} -\frac{\left[\overbrace{\left(c_L-(1-\nu)\bar{q}-\nu m\right)}^{\le 0} + \overbrace{(m-c_L)}^{\le 0}\right]}{\underbrace{\left(c_L-(1-\nu)\bar{q}-\nu m\right)^2}_{>0}} \underbrace{f_X(m)}_{>0} dm.\end{aligned}$$

We have the integral of a function that is strictly positive almost everywhere on its range of integration, and thus the LHS of (B5) is monotonically increasing in c_L between 0 and $\underline{c}_L$, implying that the solution to (B5) is unique. □

Note that despite the strategic independence between the two fake reviewer types, their strategies are linked mechanically via β. An increase in, e.g., β_H, ceteris paribus (in particular, holding β_L constant), increases the ratio of negative fake reviewers to real ones thus affecting the negative type's strategy via $\frac{\beta_L}{1-\beta}$. However, changing β_H while holding the ratio of negative fake reviewers to real ones constant affects only the positive type's equilibrium strategy.[26]

The equilibrium characterized in Proposition 1b is illustrated in Figure 6 and features a symmetric counterpart on the low end of the message space. A consequence of this symmetry is that the insights derived for positive fake reviewers extend to the negative type. This implies that all results from Sections 3–5 generalize to this extended setting. I show this formally for Propositions 2 and 3 by proving the corresponding Propositions 2b and 3b below. Similar extensions of Propositions 4–6 are omitted from the formal exposition.

Proposition 2b. *Let $\nu_2 > \nu_1$. Then $c_{H,2} > c_{H,1}$, F_H^{F,ν_2} FOSD F_H^{F,ν_1}, $c_{L,1} > c_{L,2}$, and F_L^{F,ν_1} FOSD F_L^{F,ν_2}.*

[26] For example, changing the absolute number of positive fake reviews without adding or deleting other reviews would increase β_H and leave $\frac{\beta_L}{1-\beta}$ unchanged.

Proof. For c_H and F_H^F, the proof is the same as that of Proposition 2. For c_L and F_L^F, the proof is similar but subtle differences warrant a separate exposition.

Note that the LHS of (B5) is increasing in both ν and c_L (see proof of Proposition 1b for the latter). As (B5) needs to hold in equilibrium, these two effects have to cancel out, implying $c_{L,1} > c_{L,2}$.

To show the last part of the proposition, consider the two densities f_L^{F,ν_1} and f_L^{F,ν_2}. We need to show that they intersect exactly once in $(0, c_{L,1})$. Note that $f_L^{F,\nu_i}(m) > 0, \forall m < c_{L,i}$ and $f_L^{F,\nu_i}(c_{L,i}) = 0$. Since we already established $c_{L,1} > c_{L,2}$, we have

$$\int_0^{c_{L,2}} f_L^{F,\nu_1}(m)dm < \int_0^{c_{L,2}} f_L^{F,\nu_2}(m)dm = 1,$$

which means $f_L^{F,\nu_1}(m) < f_L^{F,\nu_2}(m)$ for some $m \in (0, c_{L,2})$. At the same time, $f_L^{F,\nu_1}(c_{L,2}) > f_L^{F,\nu_2}(c_{L,2}) = 0$, so f_L^{F,ν_1} and f_L^{F,ν_2} intersect at least once in $(0, c_{L,2})$. Denote the largest such message in this interval by $\tilde{m}_L$. Using the fake reviewer's IC constraint and the fact that $Pr(fake|\tilde{m}_L)$ is the same under f_L^{F,ν_1} and f_L^{F,ν_2}, we obtain

$$\frac{c_{L,1} - \tilde{m}_L}{(1-\nu_1)(\bar{q} - \tilde{m}_L)} = \frac{c_{L,2} - \tilde{m}_L}{(1-\nu_2)(\bar{q} - \tilde{m}_L)}. \tag{B9}$$

By defining $D_L(m)$ as the difference between $Pr(fake|m)$ under ν_1 and ν_2, we can express (B9) as

$$D_L(\tilde{m}_L) = \frac{(1-\nu_2)c_{L,1} - (1-\nu_1)c_{L,2} + (\nu_2 - \nu_1)\tilde{m}_L}{(1-\nu_1)(1-\nu_2)(\bar{q} - \tilde{m}_L)} = 0. \tag{B10}$$

Now, $D_L(m)$ is quasi-monotone in $m \in [0, \bar{q})$, implying quasi-monotonicity of $f^{F,\nu_1} - f^{F,\nu_2}$ on that same interval. Since $\bar{q} > c_{L,1}$, f_L^{F,ν_1} and f_L^{F,ν_2} intersect exactly once in $(0, c_{L,1})$. □

Proposition 3b. *In the unique equilibrium, awareness policies have opposing effects on the two consumer types. In particular,*

i) $\frac{dCS^n}{d\nu} < 0$,

ii) $\frac{dCS^a}{d\nu} > 0$,

iii) $\lim_{\nu \to 0} \frac{dCS}{d\nu} < 0$ *and* $\lim_{\nu \to 1} \frac{dCS}{d\nu} < 0$.

Proof. I will utilize Lemma 3b, which is a slight modification of Lemma 3.

Lemma 3b. *Let $g(x)$ be a continuous function with $g(x) > 0$ for $x \in [a, b]$. Let $-h(x)$ be a quasi-monotone function with $\int_a^b h(x)dx = 0$. If g is strictly increasing (decreasing) on $[a, b]$, then $\int_a^b g(x)h(x)dx < 0$ (> 0).*

Proof. The only difference to Lemma 3 is that $-h(x)$, not $h(x)$, is quasi-monotone. Changing the inequality signs accordingly provides the proof. □

We now turn to the proof of Proposition 3b.

Part (i): Consumer surplus of a naive consumer is given by

$$\begin{aligned} CS^n = \beta_H \int_{c_H}^{1} \Big([1 - F_Y(m)] E[Y|Y > m] + F_Y(m)\bar{q} \Big) f_H^F(m) dm \\ + \beta_L \int_0^{c_L} \Big([1 - F_Y(m)] E[Y|Y > m] + F_Y(m)\bar{q} \Big) f_L^F(m) dm \\ + (1-\beta) \int_0^1 \Big([1 - F_Y(m)] \;\; E[Y|Y > m] + F_Y(m) m \Big) f_X(m) dm, \end{aligned} \tag{B11}$$

which reduces to the following, given that Y is uniformly distributed on $[0, 1]$:

$$\begin{aligned} CS^n = \frac{\beta_H}{2} \int_{c_H}^{1} (1 + 2\bar{q}m - m^2) f_H^F(m) dm + \frac{\beta_L}{2} \int_0^{c_L} (1 + 2\bar{q}m - m^2) f_L^F(m) dm \\ + \frac{(1-\beta)}{2} \int_0^1 (1 + m^2) f_X(m) dm. \end{aligned} \tag{B12}$$

The last integral is independent of ν, and thus, by Leibniz' Rule, and using the facts that $f_H^F(c_H) = 0$ and $f_L^F(c_L) = 0$, we have

$$\frac{dCS^n}{d\nu} = \frac{\beta_H}{2} \int_{c_H}^{1} (1 + 2\bar{q}m - m^2) \frac{df_H^F(m)}{d\nu} dm + \frac{\beta_L}{2} \int_0^{c_L} (1 + 2\bar{q}m - m^2) \frac{df_L^F(m)}{d\nu} dm. \tag{B13}$$

Note that $1 + 2\bar{q}m - m^2$ is strictly increasing for $m \in [0, \bar{q})$ and strictly decreasing for $m \in (\bar{q}, 1]$. Since $c_L < \bar{q} < c_H$, this is also true for $m \in [0, c_L]$ and $m \in [c_H, 1]$. By Lemma 3, the first integral is negative, and by Lemma 3b, the second one is negative.

Part (ii): An aware consumer's ex ante expected surplus is given by

$$\begin{aligned} CS^a = \beta_H \int_{c_H}^{1} \Big([1 - F_Y(q^a(m))] \;\; E[Y|Y > q^a(m)] + F_Y(q^a(m))\bar{q} \Big) f_H^F(m) dm \\ + \beta_L \int_0^{c_L} \Big([1 - F_Y(q^a(m))] \;\; E[Y|Y > q^a(m)] + F_Y(q^a(m))\bar{q} \Big) f_L^F(m) dm \\ + (1-\beta) \int_0^1 \Big([1 - F_Y(q^a(m))] E[Y|Y > q^a(m)] + F_Y(q^a(m)) m \Big) f_X(m) dm, \end{aligned}$$

which reduces to the following given that Y is uniformly distributed on $[0, 1]$:

$$\begin{aligned} CS^a = \frac{\beta_H}{2} \int_{c_H}^{1} \Big(1 + 2q^a(m)\bar{q} - q^a(m)^2 \Big) f_H^F(m) dm \\ + \frac{\beta_L}{2} \int_0^{c_L} \Big(1 + 2q^a(m)\bar{q} - q^a(m)^2 \Big) f_L^F(m) dm \\ + \frac{(1-\beta)}{2} \int_0^1 \Big(1 + 2q^a(m) m - q^a(m)^2 \Big) f_X(m) dm. \end{aligned} \tag{B14}$$

Using $f_H^F(c_H) = 0$, $f_L^F(c_L) = 0$, and the fact that $q^a(m)$ is independent of ν on (c_L, c_H), the derivative with respect to ν is

$$\begin{aligned}\frac{dCS^a}{d\nu} = &\frac{\beta_H}{2}\int_{c_H}^{1}\left(1+2q^a(m)\bar{q}-q^a(m)^2\right)\frac{df_H^F(m)}{d\nu}dm\\ &+\frac{\beta_H}{2}\int_{c_H}^{1}\left(2\bar{q}-2q^a(m)\right)f_H^F(m)\frac{dq^a(m)}{d\nu}dm\\ &+\frac{\beta_L}{2}\int_{0}^{c_L}\left(1+2q^a(m)\bar{q}-q^a(m)^2\right)\frac{df_L^F(m)}{d\nu}dm\\ &+\frac{\beta_L}{2}\int_{0}^{c_L}\left(2\bar{q}-2q^a(m)\right)f_L^F(m)\frac{dq^a(m)}{d\nu}dm\\ &+\frac{1-\beta}{2}\int_{c_H}^{1}\left(2m-2q^a(m)\right)\frac{dq^a(m)}{d\nu}f_X(m)dm\\ &+\frac{1-\beta}{2}\int_{0}^{c_H}\left(2m-2q^a(m)\right)\frac{dq^a(m)}{d\nu}f_X(m)dm.\end{aligned} \tag{B15}$$

The two integrals on the first and third lines of (B15) are positive by Lemmata 3 and 3b, respectively. The remaining terms can be rewritten as

$$\begin{aligned}\frac{dCS^a}{d\nu} = A+&\int_{c_H}^{1}\left[\beta_H\bar{q}f_H^F(m)+(1-\beta)f_X(m)m-\left(\beta_H f_H^F(m)+(1-\beta)f_X(m)\right)q^a(m)\right]\frac{dq^a(m)}{d\nu}dm\\ &+\int_{0}^{c_L}\left[\beta_L\bar{q}f_L^F(m)+(1-\beta)f_X(m)m-\left(\beta_L f_L^F(m)+(1-\beta)f_X(m)\right)q^a(m)\right]\frac{dq^a(m)}{d\nu}dm,\end{aligned} \tag{B16}$$

where A corresponds to the terms on the first and third lines. Because $q^a(m)=\frac{\beta_i\bar{q}f_i^F(m)+(1-\beta)f_X(m)m}{\beta_i f_i^F(m)+(1-\beta)f_X(m)}$, the remaining terms are 0 and thus $\frac{dCS^a}{d\nu}=A>0$.

Part (iii): The marginal effect of ν is given by

$$\frac{dCS}{d\nu}=\underbrace{\left(CS^n-CS^a\right)}_{A}+\underbrace{\nu\frac{dCS^n}{d\nu}}_{B}+\underbrace{(1-\nu)\frac{dCS^a}{d\nu}}_{C}. \tag{B17}$$

To prove that (B17) is negative in the limiting cases of $\nu\to 0$ and $\nu\to 1$, I begin by showing that A is negative. Because $\frac{dCS^n}{d\nu}<0$ and $\frac{dCS^a}{d\nu}>0$, it suffices to show it in the limiting case of $\nu\to 0$. From (B12) and (B14), we write A as:

$$\begin{aligned}CS^n-CS^a = &\frac{\beta_L}{2}\int_{0}^{c_L}\left(\left(q^a(m)^2-m^2+2\bar{q}(m-q^a(m))\right)f_L^F(m)dm\right.\\ &+\frac{\beta_H}{2}\int_{c_H}^{1}\left(\left(q^a(m)^2-m^2+2\bar{q}(m-q^a(m))\right)f_H^F(m)dm\right.\\ &+\frac{1-\beta}{2}\int_{0}^{c_L}\left(m-q^a(m)\right)^2f_X(m)dm\\ &+\frac{1-\beta}{2}\int_{c_H}^{1}\left(m-q^a(m)\right)^2f_X(m)dm.\end{aligned} \tag{B18}$$

Using $\lim_{\nu\to 0}q^a(m)=c_i$ and $\lim_{\nu\to 0}f_i^F(m)=\frac{1-\beta}{\beta_i}\frac{m-c_i}{c_i-\bar{q}}$ for $i=L,H$ we have

$$\lim_{\nu\to 0} CS^n - CS^a = \frac{1-\beta}{2}\int_{c_H}^{1}\left(\frac{\bar{q}-m}{c_H-\bar{q}} + (m-c_H)^2 - 1\right)f_X(m)dm + \frac{1-\beta}{2}\int_{0}^{c_L}\left(\frac{\bar{q}-m}{c_L-\bar{q}} + (m-c_L)^2 - 1\right)f_X(m)dm. \tag{B19}$$

The expression inside the first integral is negative because $\bar{q} < m$, $c_H > \bar{q}$, and $(m-c_H)^2 < 1$. The expression inside the second integral is negative because $\bar{q} > m$, $c_L < \bar{q}$, and $(m-c_L)^2 < 1$. Consequently, $A < 0$.

We continue by showing that $B = 0$ in both limiting cases. Consider first the case $\nu \to 0$, in which case $c_L \to c_L^{JO}$ and $c_H \to c_H^{JO}$, and thus

$$-\infty < \lim_{\nu\to 0}\frac{dCS^n}{d\nu} = \frac{\beta_H}{2}\int_{c_H^{JO}}^{1}(1+2\bar{q}m-m^2)\frac{df_H^F(m)}{d\nu}dm + \frac{\beta_L}{2}\int_{0}^{c_L^{JO}}(1+2\bar{q}m-m^2)\frac{df_L^F(m)}{d\nu}dm < 0, \tag{B20}$$

and in turn, $\lim_{\nu\to 0}\nu\frac{dCS^n}{d\nu} = 0$.[27] In the case where $\nu \to 1$, we have $c_H \to 1$ and $c_L \to 0$, and therefore $\lim_{\nu\to 1}\frac{dCS^n}{d\nu} = 0$, which in turn implies $\lim_{\nu\to 1}\nu\frac{dCS^n}{d\nu} = 0$. Hence, $\lim_{\nu\to 0} B = \lim_{\nu\to 1} B = 0$.

Finally, we show that $C = 0$ in both limiting cases by first considering the case $\nu \to 0$. Using $\lim_{\nu\to 0} q^a(m) = c_L$ for $m \in [0, c_L]$ and $\lim_{\nu\to 0} q^a(m) = c_H$ for $m \in [c_H, 1]$, we obtain

$$\lim_{\nu\to 0}\frac{dCS^a}{d\nu} = \frac{\beta_H}{2}\left(1+2\bar{q}c_H-c_H^2\right)\int_{c_H}^{1}\frac{df_H^F(m)}{d\nu}dm + \frac{\beta_L}{2}\left(1+2\bar{q}c_L-c_L^2\right)\int_{0}^{c_L}\frac{df_L^F(m)}{d\nu}dm = 0, \tag{B21}$$

and therefore $\lim_{\nu\to 0}(1-\nu)\frac{dCS^a}{d\nu} = 0$. In the limiting case $\nu \to 1$, we again have $c_H \to 1$ and $c_L \to 0$, which implies $\lim_{\nu\to 1}\frac{dCS^a}{d\nu} = 0$ and consequently $\lim_{\nu\to 1}(1-\nu)\frac{dCS^a}{d\nu} = 0$. Hence, $\lim_{\nu\to 0} C = \lim_{\nu\to 1} C = 0$. □

In this section, I assumed that consumers do not differ in their naivety towards positive and negative fake reviews. I could easily allow for such a distinction by introducing separate parameters for naivety toward positive and negative fakes, say, ν_H and ν_L. This would introduce more notation but leave all results qualitatively unchanged.

B.2 MULTIPLE REVIEWERS AND CONSUMERS

In the main part of the article, I have considered the interaction between one reviewer and one consumer. In this section, I am replacing the single consumer with a unit mass of consumers, and allowing $k \geq 1$ reviewers to each post a review. One interpretation of this extended game is that consumers sample k reviews from a large pool of reviews, each of which is fake with probability β. To maintain tractability, I make the following assumption throughout this section:

[27] Due to the strategic separability of the two fake reviewer types, we have $\lim_{\nu\to 0} c_H = c_H^{JO} = c^{JO}$, and $\lim_{\nu\to 0} c_L = c_L^{JO}$ with $\bar{q} > c_H^{JO} > 0$.

Assumption A3. $F_X = U[0,1]$.

The reason to replace the single consumer with a mass instead of a finite number of consumers is twofold. First, it reflects the fact that the number of consumers typically far exceeds the number of reviews. Second, it is the simplest way to relax the single consumer assumption, requiring only a reinterpretation of the existing model. Instead of interpreting ν as the probability that the consumer is naive, we interpret it as the share of naive consumers. Furthermore, the expected posterior becomes the average posterior, and the purchase probability becomes the realized sales.

Extending the model to allow for multiple reviewers is a more substantial modification that requires additional analysis.[28] I do not intend to characterize the set of all equilibria of this extended game, but instead, prove the existence of a so-called *Negative Assortative Equilibrium*, which is equivalent to the equilibrium of the baseline model in a sense clarified below.

On a technical level, we need to make two modifications to the model. First, we need to specify naive consumers' beliefs for a vector of reviews $\mathbf{m} = (m_1, m_2, \ldots, m_k)$. I choose the following specification

$$q^n(\mathbf{m}) = \min(m_1, m_2, \ldots, m_k), \tag{B22}$$

which reduces to (2) when $k = 1$. One interpretation of (B22) is that naive consumers, believing all reviews to be real, expect them to agree. Whenever they disagree, naive consumers assume that high messages are fake. In particular, they believe that only the lowest review(s) are real and update their beliefs accordingly.

Second, we have to distinguish between *on-schedule* and *off-schedule* out-of-equilibrium beliefs. The reason that this was not necessary in the baseline model is that any deviation was an *on-schedule* deviation, meaning that it could also have been sent in equilibrium. This made any out-of-equilibrium review indistinguishable from equilibrium reviews, and the Bayesian updating protocol thus fully specified aware consumers' beliefs. With multiple reviewers, even though any single review is consistent with equilibrium behavior, there are combinations of reviews that are not. I call such deviations that result in combinations that are incompatible with equilibrium behavior *off-schedule* deviations. I make the following assumption.

Assumption A4. *Let* $\mathbf{m} = (m_1, m_2, \ldots, m_k)$ *be a k-vector of reviews resulting from an off-schedule deviation. An aware consumer's off-schedule out-of-equilibrium beliefs are given by* $q^a(\mathbf{m}) = 0$.

One way of interpreting Assumption A4 is that any unexpected combination of messages undermines the consumer's trust completely, deterring her from buying the good under any circumstances. Assuming such "extremely pessimistic" off-schedule out-of-equilibrium beliefs allows me to prove Propositions 7 and 8 relatively easily.

The next question is whether fake reviewers coordinate, and if so, how. I assume that all fake reviewers can coordinate by observing the quality level x.[29] In this case, I show the following two propositions.

[28] One simple way to allow for multiple reviewers is to assume that a consumer can read either only real or only fake reviews. If we allow fake reviews to be coordinated (or written by a single reviewer), we can define β as the probability that all reviews are fake. All results from the single-reviewer model carry over unchanged.

[29] Note that this is in contrast to our baseline model where I assume that only real reviewers observe the product's quality.

Proposition 7 (Negative Assortative Equilibria).

For any $(\beta,\nu)\in(0,1)^2$ *and* $k\geq 2$, *there exists a family of Negative Assortative Equilibria (NAE), indexed by* $\check{x}\in[0,1]$. *The fake reviewers' reporting strategies are given by*

$$m^F(x)=\begin{cases} x & for \quad x\geq\check{x} \\ 1-\dfrac{1-\check{x}}{\check{x}}x & for \quad x<\check{x}\end{cases}.$$

Proof. To show that a NAE exists for any $\check{x}\in[0,1]$, we need to show that fake reviewers have no incentive to deviate. This is true because no deviation exists that increases the posterior expectations of either consumer type.

Consider first naive consumers. For the fake reviewer considering to deviate, other reviews are either the same (if they are also fake or if $x\geq\check{x}$) or lower (if they are real and $x<\check{x}$). Therefore, a downward deviation can only decrease the minimum of all reviews, while an upward deviation always leaves it unaffected. Neither deviation can thus increase the posterior expectation of naive consumers.

Next, consider aware consumers and all vectors of reviews with at least one fake review they could see in equilibrium. These are either $(m,\ldots,m,m^F(m),\ldots,m^F(m))$ or $(m,\ldots,m)$. The former is only possible when (i) $x=m$ and some reviews are real, so $q^a(\mathbf{m})=m$ in that case. The latter is possible if (ii) $x=m^{F,-1}(m)$ and all reviews are fake, and (iii) $x=m\geq\check{x}$. In both cases, the aware consumers' posterior expectation is given by

$$q^a(\mathbf{m})=Pr(X=m|\mathbf{m})m+Pr(X=m^{F,-1}(m)|\mathbf{m})m^{F,-1}(x). \tag{B23}$$

It is a convex combination between m and $m^{F,-1}(m)$, and so $m^{F,-1}(m)<q^a<m$. Because in any case $q^a\geq 0$, off-schedule deviations can never increase aware consumers' posterior expectation. To show that no profitable on-schedule deviation exists either, we look at each of the three cases from above.

In case (i), deviating to a truthful message does not affect aware consumers' beliefs because there was already at least one real review that revealed the quality to them perfectly. In case (ii), deviating to a truthful message generates a vector of reviews as in case (i), decreasing the posterior expectation to $q^a(\mathbf{m})=m^{F,-1}(m)$. In case (iii), deviating to the 'inverted' message $m^d=m^{F,-1}(x)<x$ also generates a vector of reviews resembling case (i), ostensibly revealing a quality of $m^{F,-1}(x)<x$ and decreasing the posterior expectation once again. □

The first thing to notice about the NAE is that fake reviewers no longer play a mixed strategy in equilibrium, i.e., their strategy is not given by distributions over reviews. Instead, they employ a pure strategy, conditioning their reviews on the good's quality. This strategy is truthful for quality levels above the threshold $\check{x}$ and "inverted" for quality levels below it, such that low quality levels are mapped onto high ones in a negative assortative way. I refer to the threshold $\check{x}$ as *point of deflection*.

Note that the family of equilibria characterized in Proposition 7 includes the truth-telling equilibrium, namely when $\check{x}=0$. It can be sustained because off-schedule deviations are very costly due to A4. When $k=1$, there are no off-schedule deviations, and A4 therefore has no bite. Thus, the truth-telling equilibrium does not exist when $k=1$. In fact, the only NAE that exists in that case is described in Proposition 8. In this NAE, all vectors of fake reviews induce the same posterior, satisfying the fake reviewer's IC constraint when $k=1$.

Proposition 8 (NAE with constant posterior).

For any $(\beta,\nu)\in(0,1)^2$ *and* $k\geq 1$, *there exists an NAE with point of deflection* $\check{x}_c\in[0,1]$, *in which any k-vector* $(m,m,\ldots,m)$ *of reviews with* $m\geq\check{x}_c$ *induces the same posterior.*

Proof. We need to show that there exists a $\check{x}_c\in[0,1]$ such that any vector of reviews $\mathbf{m}=(m,\ldots,m)$ with $m\geq\check{x}_c$ induces the same posterior. For this to be true, it must be that $\frac{d}{dm}q^a(m)=\frac{\nu}{\nu-1}<0,\forall m\geq\check{x}_c$. In that case, we can rewrite (B23) as

$$q^a(\mathbf{m})=\Big(\underbrace{Pr(X=m|\mathbf{m},\neg D)}_{1}\underbrace{Pr(\neg D|\mathbf{m})}_{A}+\underbrace{Pr(X=m|\mathbf{m},D)}_{B}\underbrace{Pr(D|\mathbf{m})}_{1-A}\Big)m$$
$$+\Big(\underbrace{Pr\big(X=m^{F,-1}(m)|\mathbf{m},\neg D\big)}_{0}\underbrace{Pr(\neg D|\mathbf{m})}_{A}+\underbrace{Pr\big(X=m^{F,-1}(m)|\mathbf{m},D\big)}_{1-B}\underbrace{Pr(D|\mathbf{m})}_{1-A}\Big)m^{F,-1}(x), \tag{B24}$$

where D is the event in which all reviews are fake, and thus $Pr(D)=\beta^k$. We can compute the probabilities A and B as follows. For A, we have

$$\begin{aligned}Pr(\neg D|\mathbf{m})&=\frac{f(\mathbf{m}|\neg D)Pr(\neg D)}{f(\mathbf{m}|\neg D)Pr(\neg D)+f(\mathbf{m}|D)Pr(D)}\\&=\frac{f(X=m|\neg D)Pr(\neg D)}{f(X=m|\neg D)Pr(\neg D)+f(\mathbf{m}|D)Pr(D)}.\end{aligned} \tag{B25}$$

Because X is independent of D, we have $f(X=m|\neg D)=f(X=m)=1$, while $Pr(\neg D)=1-Pr(D)=1-\beta^k$. The density $f(\mathbf{m}|D)$ can be derived via the cdf $F(\mathbf{m}|D)$:

$$F(\mathbf{m}|D)=Pr(\mathbf{m}\leq x|D)=Pr(m^{F,-1}(x)\leq X\leq x)=F(x)-F(m^{F,-1}(x))=\frac{x-\check{x}}{1-\check{x}}.$$

Taking the derivative, we get $\frac{d}{dx}F(\mathbf{m}|D)=f(\mathbf{m}|D)=\frac{1}{1-\check{x}}$. Thus, we can express A as

$$A=Pr(\neg D|\mathbf{m})=\frac{1-\beta^k}{1-\beta^k+\frac{\beta^k}{1-\check{x}}}=\frac{(1-\beta^k)(1-\check{x})}{(1-\beta^k)(1-\check{x})+\beta^k}. \tag{B26}$$

For B, we have

$$Pr(X=m|\mathbf{m},D)=\frac{Pr(\mathbf{m}|X=m,D)f(X=m|D)}{f(\mathbf{m}|D)}, \tag{B27}$$

which, using again the fact that X does not depend on D and thus $f(X=m|D)=f(X=m)=1$, that $Pr(\mathbf{m}|X=m,D)=1$, and $Pr(\mathbf{m}|D)=\frac{1}{1-\check{x}}$, we can rewrite as

$$B=Pr(X=m|\mathbf{m},D)=1-\check{x}. \tag{B28}$$

Plugging the obtained expressions for A and B into (B24) and rearranging, we get

$$q^a(m) = \frac{(1-\check{x})^2 - \beta^k \check{x}^2}{(1-\beta^k)(1-\check{x})^2 + \beta^k(1-\check{x})} m + \frac{\beta^k \check{x}^2}{(1-\beta^k)(1-\check{x})^2 + \beta^k(1-\check{x})}. \tag{B29}$$

When $m \geq \check{x}$, q^a is linear in m, with a slope that is decreasing in $\check{x}$. From (B29), we can see that as $\check{x} \to 0$, the slope approaches 1, and as $\check{x} \to 1$, the slope tends to $-\infty$. Because the slope is a continuous function of $\check{x}$, the IVT guarantees that there exists a $\check{x}_c \in [0,1]$ such that $\frac{d}{dm} q^a(m) = \frac{\nu}{\nu - 1}$. □

Figure 7 illustrates an NAE with deflection point $\check{x}_c \approx 0.76$. In contrast to the previous figures illustrating equilibrium strategies, the left panel shows the strategies in the (x,m)-plane rather than the $(m,f(m))$-plane. Consequently, truth-telling is no longer represented by a horizontal line but by the identity line instead.

The main implication of Proposition 8 is that an NAE exists also if all fake reviews are written jointly and thus have to satisfy the corresponding IC constraint. This NAE resembles the equilibrium of the baseline model in the sense that both feature skepticism on the side of aware consumers and a constant purchase probability induced by reviews above some threshold.

APPENDIX C: THE ROLE OF ASSUMPTION A1

In the main part of the article, I assumed that the density function f_X has full support on $[0,1]$. This section relaxes that assumption and analyzes a specific case where the quality is distributed with density $f_X(x) = 2 - 2x$, which violates *A*1 at $x = 1$. I characterize the equilibrium for this case and discuss implications for more general distributions. While formal proofs are omitted, I provide a structured outline of the key arguments and insights, allowing for an intuitive understanding of the results.

Special case: $f_X(x) = 2 - 2x$

Applying Proposition 1 to the case where $f_X(x) = 2 - 2x$, we find no solution to (A7) for $c \in [0,1]$ if ν and β are too large. This occurs when the following condition is satisfied:

$$\beta > \frac{4(1-\nu)^2}{4(1-\nu)^2 + 9\nu}. \tag{C1}$$

Consider first the scenario where condition (C1) does not hold. In that case, Proposition 1 holds, and the fake reviewer's equilibrium strategy is given by

$$f^F(m) = \frac{1-\beta}{\beta} \frac{(m-c)(2-2m)}{c - \frac{1-\nu}{3} - \nu m} \tag{C2}$$

with $supp\{f^F\} = [c,1]$ where c is the solution to (A7). Figure 9 depicts the equilibrium under these conditions for different parameter values.

I now turn to the case where (C1) holds, which is the case in the parameter region depicted in the right panel of Figure 10. Under these circumstances, Proposition 1 fails because $\int_{\underline{c}}^1 f^F(m)dm < 1$ even for $c = \underline{c}$. Loosely speaking, we hit the lower bound $\underline{c}$ before the fake reviewer has allocated the entire probability mass. The reason this happens is that in a neighborhood around $m = 1$ of positive measure the density is not bounded away from 0. The only message he can then use to allocate the remaining probability mass is $m = 1$, yielding an atom at the highest message.

Note that the atom in the fake reviewer's equilibrium strategy violates Lemma 1. The lemma fails because Claim 3 is no longer true when ν and β are sufficiently large. Because Claims 1 and 2 continue to hold for all parameter values, we only need to consider equilibria with an atom at $m = 1$. Let p^F be the probability with which the fake reviewer sends $m = 1$. Under

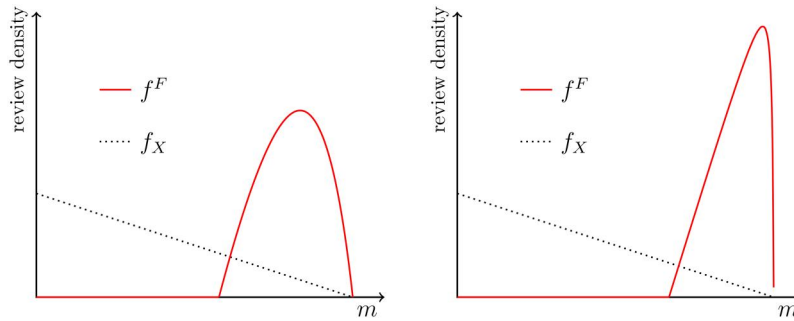

Figure 9. Equilibrium review distributions for $f_X(x) = 2 - 2x$. Left panel: $\beta = \frac{1}{6}$, $\nu = \frac{1}{4}$. Right panel: $\beta = \frac{1}{6}$, $\nu = \frac{1}{2}$

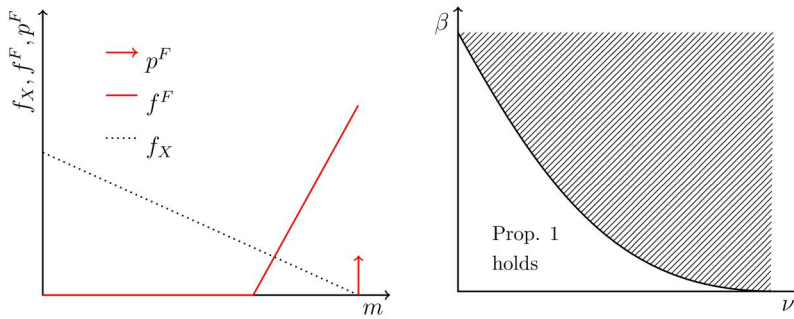

Figure 10. The left panel shows equilibrium review distributions for $f_X(x) = 2 - 2x$ and $(\beta, \nu) = (\frac{1}{3}, \frac{1}{2})$. The right panel illustrates condition (C1). The shaded area represents the parameter region for which Proposition 1 fails and the fake reviewer's equilibrium strategy is characterized by (C3).

(C1), the fake reviewer's equilibrium strategy is a mixed distribution characterized by density f^F and probability p^F such that

$$f^F(m) = \frac{1-\beta}{\beta}\frac{2(m-\underline{c})}{\nu} \text{ and } p^F = 1 - \int_{\underline{c}}^{1} f^F(m)dm \tag{C3}$$

with $supp\{f^F(m)\} = [\underline{c}, 1)$ where $\underline{c} = \frac{1+2\nu}{3}$. The left panel of Figure 10 depicts such an equilibrium while the right panel illustrates (C1).[30]

General case

We now extend our analysis to more general distributions. Let F_X be a possibly mixed distribution, where $A = \{a_1, a_2, \ldots\} \subset [0,1]$ is a countable set of atoms. In this case, F_X is characterized by a discrete probability measure P over A with $p_i = P(a_i)$ and density f_X on $[0,1] \setminus A$. Here, I do not assume full support, allowing for $f_X = 0$.

[30] Note that A1 is sufficient but not necessary for Proposition 1 to hold. To see this, consider $f_X(x) = 1$ for $x \in [0,1)$ and $f_X(1) = 0$. This distribution also violates A1 but causes no issues because Claim 3 in Lemma 1 holds.

The first thing to note is that the fake review distributions can have atoms only at $m \in A \cup \{1\}$. To see this, note that every message $m \in A$ which the fake reviewer does not send with positive probability induces $q^a(m) = m$ and thus $q(m) = m$. For $m > c$, this would violate the fake reviewer's IC constraint, so he has to send every message in $A \cap [c, 1]$ with positive probability. For messages $m \notin A$, the same logic as in Claims 1 and 2 in Lemma 1 applies. Thus, the only possible additional atom is at $m = 1$, which happens only if $c = \underline{c}$. It is also easy to see that the fake reviewer never sends messages that real reviewers never send, the only exception being $m = 1$. This is true because any such message would induce a posterior below $\underline{c}$ while $m = 1$ induces a posterior weakly larger than $\underline{c}$.

The fake reviewer's strategy in the general case is then also a possibly mixed distribution given by

$$\begin{aligned} f^F(m) &= \frac{1-\beta}{\beta} \frac{(m-c) f_X(m)}{c - (1-\nu)\bar{q} - \nu m}, \\ p_i^F &= \frac{1-\beta}{\beta} \frac{(a_i - c) p_i}{c - (1-\nu)\bar{q} - \nu a_i} \mathbb{1}_{\{a_i > c\}}, \\ p_0^F &= \max\left(0, 1 - \int_c^1 f^F(m)dm - \sum_{i=1}^{|A|} p_i\right). \end{aligned} \tag{C4}$$

The first line of (C4) corresponds to the continuous part of the fake review distribution and has a support of $supp\{f^F\} = [c, 1] \setminus A$. The second line corresponds to atoms in the fake review distribution, corresponding to atoms in the quality distribution: p_i^F is the probability with which the fake reviewer sends $m = a_i$. Finally, p_0^F is the probability mass on $m = 1$ whenever $1 \notin A$.

Two examples are depicted in Figure 11. In the left panel, the quality distribution is entirely discrete. Consequently, the fake review distribution is also a discrete probability distribution. The right panel shows an example where F_X is a mixed distribution with gaps and atoms.

I conjecture that the equilibrium is unique in general. This is because after adjusting Lemmata 1 and 2 to account for mixed distributions and gaps in the support, the equilibrium characterization boils down to solving an equation of the form

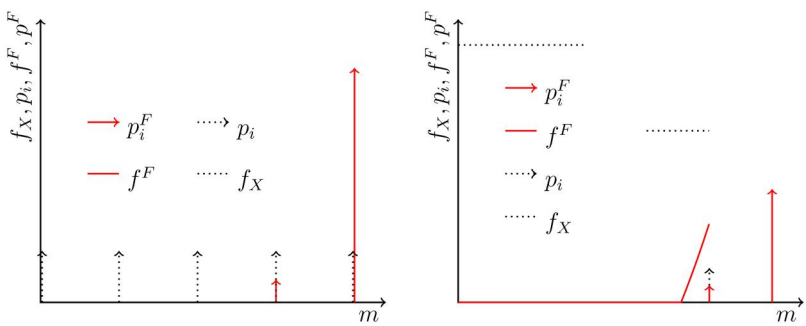

Figure 11. The left panel shows the case where the quality distribution and the fake reviewer's equilibrium strategy are both discrete. The quality takes values in $\{0, 0.25, 0.5, 0.75, 1\}$ with equal probability, and the fake reviewer sends only $m \in \{0.75, 1\}$ in equilibrium. The right panel shows the case where F_X is a mixed distribution with $f_X(x) = 1.5$ for $x \in [0, 0.4]$, $f_X(x) = 1$ for $x \in [0.6, 0.8]$, and $P(0.8) = 0.2$. The fake review distribution is also mixed with two atoms. One at $m = 0.8$ and another at $m = 1$.

$$\int_c^1 f^F(m)dm + \sum_{i=1}^{|A|} p_i^F = 1, \tag{C5}$$

which is similar to (A7) and I conjecture it has a unique solution.

REFERENCES

Ashcraft, A. B., P. Goldsmith-Pinkham, and J. Vickery. 2010. "MBS Ratings and the Mortgage Credit Boom." Tech. rep., Federal Reserve Bank of New York.

Blau, B. M., J. R. DeLisle, and S. M. Price. 2015. "Do Sophisticated Investors Interpret Earnings Conference Call Tone Differently than Investors at Large? Evidence from Short Sales."31 *Journal of Corporate Finance* 203–19.

Bolton, P., X. Freixas, and J. Shapiro. 2012. "The Credit Ratings Game." 67 *Journal of Finance* 85–111.

Bright Local. 2020. "Local Consumer Review Survey 2020." https://www.brightlocal.com/research/local-consumer-review-survey/, accessed 31 May 2021.

Bruno, V., J. Cornaggia, and K. J. Cornaggia. 2016. "Does Regulatory Certification Affect the Information Content of Credit Ratings?" 62 *Management Science* 1578–97.

Cai, H., and J. T.-Y. Wang. 2006. "Overcommunication in Strategic Information Transmission Games."56 *Games and Economic Behavior* 7–36.

Chen, Y. 2011. "Perturbed Communication Games with Honest Senders and Naive Receivers."146 *Journal of Economic Theory* 401–24.

Cheung, C. M., and M. K. Lee. 2012. "What Drives Consumers to Spread Electronic Word of Mouth in Online Consumer-Opinion Platforms."53 *Decision Support Systems* 218–25.

Chevalier, J. A., and D. Mayzlin. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews."43 *Journal of Marketing Research* 345–54.

Strategy CPC. 2019. "The 2019 Amazon Consumer Shopping Study." https://tinuiti.com/content/guides/2019-amazon-consumer-shopping-study/, accessed 31 May 2021.

Deversi, M., A. Ispano, and P. Schwardmann. 2021. "Spin Doctors: An Experiment on Vague Disclosure."139 *European Economic Review* 103872.

Fernandez, F., and D. Zejcirovic. 2020. "The Role of Pharmaceutical Promotion to Physicians in the Opioid Epidemic." *Working Paper*.

Gesche, T. 2021. "De-Biasing Strategic Communication."130 *Games and Economic Behavior* 452–64.

Glazer, J., H. Herrera, and M. Perry. 2021. "Fake Reviews."131 *Economic Journal* 1772–87.

He, S., B. Hollenbeck, and D. Proserpio. 2022. "The Market for Fake Reviews."41 *Marketing Science* 896–921.

Heidhues, P., and B. Kőszegi. 2010. "Exploiting Naivete about Self-Control in the Credit Market." 100 *American Economic Review* 2279–303.

Heidhues, P., and B. Kőszegi. 2017. "Naivete-Based Discrimination."132 *Quarterly Journal of Economics* 1019–54.

Hou, Y., J. Li, Z. He, A. Yan, X. Chen, and J. McAuley. 2024. "Bridging Language and Items for Retrieval and Recommendation." *arXiv preprint arXiv : 2403.03952*.

Janssen, M. C., and S. Roy. 2022. "Regulating Product Communication."14 *American Economic Journal: Microeconomics* 245–83.

Jiang, J. X., M. H. Stanford, and Y. Xie. 2012. "Does It Matter Who Pays for Bond Ratings? Historical Evidence."105 *Journal of Financial Economics* 607–21.

Jindapon, P., and C. Oyarzun. 2013. "Persuasive Communication When the Sender's Incentives Are Uncertain." 95 *Journal of Economic Behavior & Organization* 111–25.

Kartik, N., M. Ottaviani, and F. Squintani. 2007. "Credulity, Lies, and Costly Talk."*Journal of Economic Theory* 93–116.

Kawagoe, T., and H. Takizawa. 2009. "Equilibrium Refinement vs. level-k Analysis: An Experimental Study of Cheap-Talk Games with Private Information."66 *Games and Economic Behavior* 238–55.

Lewis, G., and G. Zervas. 2019. "The Supply and Demand Effects of Review Platforms." *Available at SSRN* 3468278.

Martin, S., and S. Shelegia. 2021. "Underpromise and Overdeliver?-Online Product Reviews and Firm Pricing." 79 *International Journal of Industrial Organization* 102775.

Mayzlin, D., Y. Dover, and J. Chevalier. 2014. "Promotional Reviews: An Empirical Investigation of Online Review Manipulation."104 *American Economic Review* 2421–55.
Morris, S. 2001. "Political Correctness."109 *Journal of Political Economy* 231–65.
Ottaviani, M., and F. Squintani. 2006. "Naive Audience and Communication Bias." 1 *International Journal of Game Theory* 129–50.
SafetyDetectives research lab. 2021. "Amazon Fake Reviews Scam Exposed in Data Breach." https://web.archive.org/web/20210525123756/https://www.safetydetectives.com/blog/amazon-reviews-leak-report/, accessed 31 May 2021.
Smirnov, A., and E. Starkov. 2022. "Bad News Turned Good: Reversal under Censorship."14 *American Economic Journal: Microeconomics* 506–60.
Tang, L., M. Peytcheva, and P. Li. 2020. "Investor-Paid Ratings and Conflicts of Interest."163 *Journal of Business Ethics* 365–78.

The Journal of Law, Economics, and Organization, 2026, 00, 1–42
https://doi.org/10.1093/jleo/ewag010
Article