Electronic Commerce Research
https://doi.org/10.1007/s10660-026-10176-7

Biases in online reputation systems: a survey of the empirical literature

Martin Sterner[1]

Received: 19 December 2025 / Accepted: 13 July 2026

Abstract

Reputation systems are essential for creating trust and reducing information asymmetries in online markets. However, they are vulnerable to biases that distort information, erode trust and potentially cause market failure. This paper surveys the empirical literature on five key sources of bias: (i) strategic actions of sellers such as fake reviews and rebate-for-review programs, (ii) reciprocity, (iii) social influence bias, (iv) selection bias and (v) noise. It analyzes the biases through a unified framework decomposing errors into three distinct signatures: systematic mean shifts, variance inflation and serial correlation. A central finding is that platform design involves trade-offs: interventions targeting one error signature often exacerbate other distortions. By illustrating how biases interact, this paper provides a diagnostic roadmap for platform operators, regulators and policymakers to match interventions to the specific error signatures while compensating for secondary effects, ultimately enabling reputation systems to maintain trust, reduce information asymmetries and enable efficient online markets.

Keywords Reputation · Online reviews · Bias · Online markets · Market design

JEL Classification D47 · D82 · D83 · L14 · L15

The author would like to thank Anika Bittner, Timo Heinrich, the associate editor and two anonymous reviewers as well as the participants of the 2024 DPE Forum at the University of Passau, the 2024 Behavioral Economics & Management Science Workshop at the University of Hamburg and the 2024 Behavioral Economics Workshop at Clausthal University of Technology for their valuable comments and suggestions.

✉ Martin Sterner
martin.sterner@tuhh.de

[1] Institute for Digital Economics, Hamburg University of Technology, Hamburg, Germany

Published online: 22 July 2026

1 Introduction

As digital commercial activity expands across all areas of the economy, the number of online transactions between strangers continues to grow. Most major online platforms rely on reputation systems to facilitate transactions by reducing information asymmetries, such as adverse selection and moral hazard, which can otherwise lead to market failure [1, 2]. The classic example of the effects of information asymmetries is the market for "lemons" described by Akerlof [1]. Depending on their design, reputation systems make available information either about a seller's underlying type, such as trustworthiness or inherent product quality, or about effort and performance in previous transactions, aiming to reduce problems of adverse selection and moral hazard. Without reputation systems, buyers in online markets would have no reliable indication of a seller's trustworthiness or willingness to exert effort before completing a transaction. If distrust dominates, buyers may refrain from transacting altogether, causing the market to fail. Well-designed reputation systems therefore provide buyers with reliable information about past performance with other transaction partners. This enables them to make better decisions about products and transaction partners and enter transactions, thereby preventing market failure [2–4]. However, if the information a reputation system contains is systematically distorted and does not correlate with the seller's type or effort, reputation is biased.

To design reputation systems that enable efficient market outcomes, platform operators, regulators and policymakers need to understand not only that biases exist, but also how they distort reputation information. Although many academic studies examine individual biases, the literature remains fragmented across contexts and disciplines. Different strands of research analyze fake reviews in e-commerce and hospitality platforms, reciprocity in two-sided reputation systems such as those of eBay or Airbnb, social influence bias in product and hospitality review platforms and selection bias in online retail and service markets. Existing surveys typically focus on a narrow subset of biases, for example, surveys on fake reviews and their antecedents and consequences [5, 6], or reviews of computational detection methods for fake reviews [7, 8]. Other contributions discuss biases primarily as part of broader analyses of platform design, market outcomes or technological applications, rather than as a central aspect of synthesis [2, 9–13]. While useful, this approach often leads to fragmented insights, as it obscures the fact that distinct sources of bias can produce identical distortions, while similar mitigation strategies can have heterogeneous effects depending on the error signature and platform design.

This paper offers a unified, analytical framework that categorizes biases by their error signature and how they affect the reputation signal. Observed reputation $\hat{r}_{it}$ is modeled as the sum of objective true reputation r_i and an error term ε_{it}. Within this framework, biases are not merely listed, but analyzed based on which moment of the error term they distort: (i) a shift in the expected value $\mathbb{E}[\varepsilon_{it}]$ (systematic bias), (ii) an inflation of the variance $\mathrm{Var}[\varepsilon_{it}]$ (noise) or (iii) the introduction of serial correlation $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$ for some $t \neq s$ (social influence). This decomposition of the error term provides three critical insights that go beyond existing organizational frameworks. First, it reveals that different biases require fundamentally different design interventions. For instance, increasing review volume mitigates variance-driven

noise but may increase mean-shift biases such as fake reviews or selection bias. Second, it highlights trade-offs, as interventions that reduce one type of distortion (e.g., strict verification of reviewers to reduce fake reviews) can increase another (e.g., selection bias or variance due to reduced sample size). Third, it allows for the derivation of propositions about how biases interact, moving the literature from descriptive cataloging to analytical prediction. These propositions guide the synthesis of the empirical literature, enabling the evaluation of not only if a bias exists, but how it interacts with other distortions.

The objective of this paper is to synthesize the empirical literature from this perspective. This paper examines five major sources of bias: (i) strategic actions of sellers, including fake reviews and rebate-for-review programs, (ii) strategic actions of reviewers, such as reciprocal reviewing, (iii) social influence, where reviewers adjust ratings based on existing reviews, (iv) selection bias, arising from the non-representative subset of buyers who post reviews and (v) noise. Each source of bias is analyzed, including how it distorts reputation, under which conditions it arises and what the resulting design implications are. Unlike previous work that treats these biases in isolation, this paper also considers how they interact within a reputation system and how platform design choices determine the net effect on market efficiency. The extent of coverage for each bias reflects the relative volume and diversity of the available empirical literature rather than a difference in analytical depth. The literature on fake reviews is substantially larger and more heterogeneous than that on rebate-for-review programs, reciprocity, social influence, selection bias or noise. This is driven by rapid platform evolution, technological advancements such as the emergence of generative artificial intelligence, and high regulatory urgency. Consequently, a comprehensive synthesis of this stream results in a more extensive discussion to capture the full range of mechanisms and evidence, whereas the literatures on other biases are generally more mature and consolidated.

This paper does not aim to produce an exhaustive systematic review of all publications on online reviews, but rather to synthesize the main empirical findings that explain how different types of bias arise, affect reputation and respond to platform design. Consequently, while the rapidly expanding literature on computational detection methods, machine learning and artificial intelligence-based moderation is acknowledged, purely technical studies are outside the primary scope unless they provide direct insights into behavioral or economic mechanisms. Likewise, broader work on deceptive advertising is discussed only when it directly affects reputation systems. Instead, this paper seeks to cover the main empirical findings that shape current understanding of how biases in reputation systems arise, interact and respond to platform design.

The literature surveyed in this paper was identified using a structured, problem-oriented approach designed to synthesize empirical findings across disciplines. Starting from foundational contributions on reputation systems and online reviews, backward and forward citation tracking identified major empirical research streams on fake reviews, rebate-for-review programs, reciprocity, social influence bias, selection bias and noise. This was complemented by targeted keyword searches in Google Scholar and Scopus. Studies were selected based on three criteria: (i) they provide empirical evidence (field data, experiments or surveys) on the generation, manipulation or inter-

pretation of reviews, (ii) they relate directly to at least one of the biases and (iii) they offer insights into the economic or behavioral mechanisms at play. While the survey emphasizes work published from 2015 onward to capture recent platform dynamics, studies prior to this date were included when they provide foundational evidence or when more recent research on a given bias is limited. This review employs a structured analytical framework: every included study was evaluated against a consistent set of questions regarding causes, consequences and mitigation. This approach ensures comparability across diverse literature streams while maintaining a focus on mechanisms relevant to platform design. Further details on the search strategy, inclusion logic and classification procedure are provided in Appendix A.

The remainder of this paper is organized as follows. Section 2 outlines the functioning and role of reputation systems, introduces the conceptual framework and derives four key propositions. Section 3 reviews the empirical evidence on the biases, including causes, effects and mitigation strategies. Section 4 provides a structural synthesis of biases in reputation systems and Sect. 5 provides a summary of findings, practical recommendations and an outlook for future research.

2 Reputation in online markets

Reputation systems exist in many institutional and technological forms. They are used in e-commerce marketplaces (e.g., Amazon or eBay), peer-to-peer service platforms (e.g., Airbnb or Uber) and content or review platforms (e.g., Yelp or Tripadvisor). Their contexts differ in important dimensions, including what is being evaluated (products, sellers or services), how transactions are mediated and what incentives users have to leave a review. For example, peer-to-peer services often involve personal interactions, whereas e-commerce transactions are typically more anonymous. Platforms also differ in how tightly reviews are connected to verified transactions, which affects their vulnerability to manipulation and selection effects; see, for example, Tadelis [2].

Despite this heterogeneity, reputation systems across platforms perform a common economic function. They aggregate user-reported information about past actions of their transaction partners or the quality and effort they invested in the transaction and make it observable to future market participants. By doing so, they reduce information asymmetries between buyers and sellers and enable trade in environments where direct inspection or enforcement is costly or impossible; see, for example, Bajari and Hortaçsu [3] and Tadelis [2].

This section introduces a common terminology for key concepts, explains the theoretical relevance of reputation systems, establishes a conceptual framework for bias analysis and derives four key propositions.

2.1 Preliminaries and definitions

This subsection introduces the terminology used throughout the paper. Clear definitions are necessary because the literature on reputation systems does not always employ consistent language.

The first set of definitions concerns who evaluates whom in a reputation system and defines one-sided and two-sided reputation systems. Many platforms active in the digital sphere maintain online reputation systems that allow users to build credibility in an otherwise largely anonymous environment. On online marketplaces, which often operate one-sided reputation systems, a seller's reputation is generally displayed to signal trustworthiness to potential buyers or to indicate the quality of products and services that they sell. Such systems have been examined extensively, for example, by Bolton et al. [14], Resnick and Zeckhauser [4] and Resnick et al. [15], and in surveys by Dellarocas [16], Luca [11], Magnani [12], Pocchiari et al. [13] and Tadelis [2].

In peer-to-peer environments, such as platforms facilitating the rental of vacation homes between private individuals, reputation systems are often two-sided. In these settings, the reputations of both transaction partners are displayed because each party benefits from reducing the risk of transacting with an untrustworthy counterpart. Such systems are therefore referred to as two-sided reputation systems. Their specific properties have been analyzed in the literature, for example, by Bolton et al. [19] and Klein et al. [17, 18]. These features are particularly relevant for understanding reciprocal behavior in reputation systems, which is examined in more detail in Sect. 3.2. Regardless of whether they are one-sided or two-sided, reputation systems aim to help users avoid untrustworthy transaction partners and low-quality products or services.

The second set of definitions concerns how evaluations are expressed. Evaluations can be expressed through ratings or reviews. Reputation systems typically include ratings, reviews or both [9, 13]. Hence, these terms require a definition. In this paper, a *rating* refers to a quantitative assessment of a product, service or transaction partner. Ratings are often based on Likert-type scales, commonly represented by one to five stars, and are frequently aggregated into summary measures such as average ratings or rating distributions.

In contrast, *reviews* in the narrower sense are qualitative, textual expressions of a reviewer's opinion. Hereafter, these will also be referred to as a *textual review*. Many platforms allow reviewers to supplement text with photos or other media. Owing to their qualitative nature, reviews tend to be more heterogeneous than ratings. Throughout this paper, the term *review* is used as an umbrella term encompassing both ratings and textual reviews, unless the context clearly indicates otherwise. This usage reflects common platform design, where users typically submit both elements together.

The third set of definitions relates to what is being evaluated. Accordingly, product reviews and seller reviews are defined. As noted by Belleflamme and Peitz [20] and Tadelis [2], reviews may target either products or sellers. A *product review* reflects the reviewer's assessment of the purchased product and focuses on its features and quality. A *seller review*, in contrast, evaluates the transaction partner and may refer to aspects such as delivery, communication or customer service. Some platforms maintain both types of systems in parallel, while in many service settings reviews refer to the overall transaction experience, including interactions with the transaction partner.

2.2 Theoretical relevance of reputation in online markets

Many online markets have proven to be highly successful. However, economic theory suggests that information asymmetries can lead to market failure, as illustrated by Akerlof [1] in the classical example of the market for "lemons." In particular, hidden information about product quality (adverse selection) and hidden action by sellers (moral hazard) may cause buyers to distrust sellers and refrain from purchasing. Although these issues can arise in any market, they are especially relevant in online environments, where buyers and sellers are often anonymous, communication is limited and buyers generally cannot inspect products before purchase. Moreover, sellers may have incentives to misrepresent product quality or to not ship the product after receiving payment. Such concerns have motivated a large body of research, such as Bajari and Hortaçsu [3] and Lewis [21]. As noted by Luca [11] and Tadelis [2], ensuring that sellers have sufficient incentives to complete transactions is essential for preventing market failure. Despite these challenges, many online markets have succeeded, in part because reputation systems help overcome the underlying information asymmetries. Typically, these systems allow users to provide quantitative ratings, qualitative reviews or both, thereby enabling sellers to build a reputation. The functioning of these systems has been widely documented, particularly in studies of the auction platform eBay, an early adopter of online reputation systems; see Bajari and Hortaçsu [3], Klein et al. [17] and Tadelis [2]. Their effectiveness has also been studied experimentally, for example, by Bolton et al. [14, 22] and in field experiments such as Resnick et al. [15]. Bibliometric analyses of the literature have found that the number of publications about online reviews has grown substantially over the last two decades; see Veh et al. [23] and Zhang et al. [24].

Online markets often operate as two-sided platforms that bring buyers and sellers together. They resemble traditional markets that have existed for centuries, but with notable differences. As Goldfarb and Tucker [25] emphasize, online markets are typically not as limited by physical space[1] and result in much lower search, transportation and travel costs for their participants relative to offline markets. At the same time, many online transactions are one-shot interactions and involve minimal direct communication between buyers and sellers. In contrast to traditional markets, word of mouth spreads more slowly online when there are no formal reputation systems, making it harder for sellers to build a reputation without a reputation system. This fact partly motivates research such as Bajari and Hortaçsu [3] and Resnick and Zeckhauser [4].

A seller's displayed reputation can be decisive for commercial success. As summarized by Magnani [12], there is broad consensus in the literature that sellers with more numerous and more positive reviews tend to sell more and at higher prices than sellers without a reputation, thereby increasing market efficiency by improving match quality between buyers and products or sellers. Evidence for increased efficiency and

[1] Traditional markets are limited in size because they require physical space for sellers to present their products and for buyers to access the market. Hence, these markets usually cannot grow beyond a certain size.

higher consumer welfare in the presence of reputation systems is provided by studies such as Bolton et al. [14], Liu et al. [26] Reimers and Waldfogel [27].

Tadelis [2] illustrates the theoretical relevance of reputation systems using a variant of the trust game. In this setting, the buyer does not know whether the seller is an honest type who always ships the product or an opportunistic type who may choose to keep the payment without shipping. The game involves both hidden information (the seller's type) and hidden action (the seller's shipping decision). In a one-shot game, the opportunistic seller's best response is not to ship, and the buyer will trust the seller only if the expected payoff is non-negative, which requires a sufficiently high probability that the seller is honest. In a version in which the game is repeated for two consecutive periods, the buyer will never trust the seller in the second period if the seller failed to ship in the first period. If the seller shipped in the first period, the buyer may trust again in the second period. Thus, if the future is sufficiently important, even an opportunistic seller may ship in the first round to appear honest [2]. This insight generalizes to infinitely repeated settings. With a sufficiently high discount factor, sellers have incentives to behave cooperatively and ship the product. If a seller fails to ship or if buyers cease to trust, the market reverts to the equilibrium of the one-shot game in which no transactions occur. The core implication of this class of models is that current behavior influences future trust. A seller who fails to deliver gains a reputation of being opportunistic and will not be trusted by future buyers. This reasoning also applies when sellers interact with different buyers, as long as reputation is transferable and other buyers have access to information about a seller's past performance [2]. A reputation system that records past behavior can therefore facilitate trust and enable trade. In this sense, the presence of a reputation system can determine whether a market functions or fails.

Modern online platforms implement precisely this logic by making past behavior observable through reviews and ratings [2]. While the trust game illustrates the role of reputation in incentivizing seller cooperation, real-world platforms capture more complex considerations. Reputation systems rely on buyer statements that contain both objective and subjective components. Reviews may reflect not only whether the product was delivered but also subjective evaluations of quality or service, as investigated, for example, by Ghose and Ipeirotis [28]. As a result, reputation systems can reflect much more than the mere fact of whether a product was delivered. They also reflect trustworthiness and product or service quality, depending on the platform's design. Hence, the analysis in this paper abstracts from the presented example and considers reputation more generally, such that it captures all relevant factors in the context of the specific platform considered.

2.3 Conceptual framework to analyze biases in online reputation systems

To analyze how biases distort the reliability of reputation systems, the relationship between true and observed reputation is formalized using a conceptual framework. Reviewer $t \in \{1, \ldots, T\}$ posts a review of seller i, denoted as $\hat{r}_{it}$. This observed reputation is modeled as the sum of the seller's unobservable true reputation r_i and an error term ε_{it}:

$$\hat{r}_{it} = r_i + \varepsilon_{it}. \tag{1}$$

Here, $\hat{r}_{it}$ represents the complete content of a review (i.e., rating and/or a quantified evaluation of a textual review, such as one incorporating a sentiment score) and r_i reflects the seller's objective performance in a given transaction, that is, the hypothetical evaluation that would be provided by an objective reviewer. Depending on the context of the platform, true reputation may capture trustworthiness, product quality, effort, service quality or other relevant factors. Most reputation systems aggregate individual reviews into observed reputation, $\hat{r}_i = f(\hat{r}_{i1}, \ldots, \hat{r}_{iT})$, where $f(\cdot)$ is a function of all reviews posted. A simple example of this is a function that determines observed reputation as the average of all reviews posted: $\hat{r}_i = \frac{1}{T}\sum_{t=1}^{T} \hat{r}_{it}$. The aggregation of reviews $\hat{r}_i = f(\hat{r}_{i1}, \ldots, \hat{r}_{iT})$ converges to r_i as T increases if three conditions regarding the error term are fulfilled:

1. Unbiasedness (Mean Condition): The expected value of the error term must be zero: $\mathbb{E}[\varepsilon_{it}] = 0$. If $\mathbb{E}[\varepsilon_{it}] \neq 0$, the reputation system is systematically biased, consistently over- or under-estimating true reputation.
2. Precision (Variance Condition): Due to idiosyncratic factors such as subjective preferences, reviewers' personal experiences or contextual circumstances such as shocks, the error term typically has non-zero variance $\mathrm{Var}[\varepsilon_{it}] \neq 0$. The variance of the error term must not be too large. High variance implies noise, reducing the precision of the reputation signal even if the mean is unbiased.
3. Independence (Correlation Condition): Reviews must be pairwise uncorrelated, such that the covariance between any two distinct error terms is zero: $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) = 0$ for all $t \neq s$. This condition ensures that no review is influenced by any prior review. If this condition is violated (i.e., if $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$), the error terms exhibit serial correlation. In such cases, the effective sample size is reduced because redundant information provides diminishing marginal returns, and initial shocks can propagate through subsequent reviews, leading to persistent bias even as the total number of reviews increases.

Figure 1 schematically summarizes this relationship and how the bias mechanisms described in Sect. 3 violate these conditions. This framework provides a basis for categorizing biases not by their source, but by their distinct error signatures. Strategic manipulations (e.g., fake reviews, reciprocity) and structural biases (e.g., selection bias) primarily violate the Mean Condition, introducing a directional shift in reputation. Idiosyncratic factors, misunderstandings and contextual shocks primarily violate the Variance Condition. Social dynamics, such as social influence, violate the Correlation Condition, creating serial correlation.

This classification is critical because each type of violation requires a distinct design intervention. Mitigating a mean shift (e.g., through requiring verification of users who can post a review) differs fundamentally from reducing variance (e.g., through aggregation of certain reviews) or restoring independence (e.g., through hiding previous reviews during the review posting).

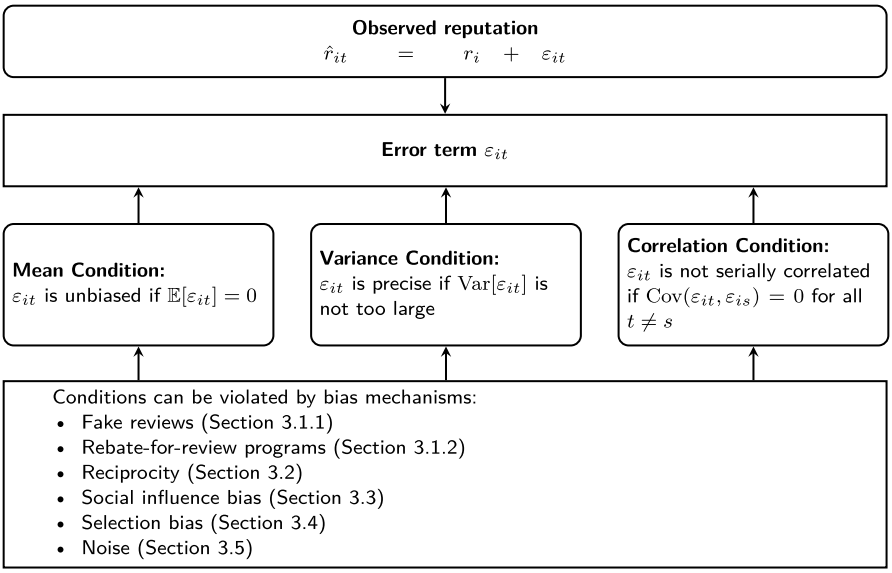

Fig. 1 Schematic overview of the conceptual framework

2.4 Framework implications and propositions

The decomposition of reputation signals into true reputation and an error term, $\hat{r}_{it} = r_i + \varepsilon_{it}$, allows the analysis to move beyond descriptive cataloging to derive analytical propositions about bias properties. By mapping each bias to the related error signatures (mean shift, variance inflation or serial correlation), this framework identifies fundamental trade-offs and interaction effects that have been overlooked in prior literature.

Four key propositions are made that serve a dual purpose: they provide a perspective for systematically re-evaluating existing empirical findings in Sect. 3 and highlight gaps that define the agenda for future research.

Proposition 1 *(Trade-offs) Interventions designed to reduce systematic mean bias ($\mathbb{E}[\varepsilon_{it}] \neq 0$) may inadvertently increase variance ($\mathrm{Var}[\varepsilon_{it}]$) or intensify selection bias. For example, strict verification requirements reduce fake reviews but lower participation rates, thereby reducing the sample size T and potentially increasing the variance of the reputation signal.*

This proposition implies that platform design cannot optimize for a single distortion in isolation. In Sect. 3, this perspective is used to explain why interventions like "verified purchase" labels sometimes fail to improve overall market efficiency: they solve the mean-shift problem but potentially create a variance problem or a new systematic bias (selection bias) through reduced participation. Consequently, the literature should shift from studying isolated interventions to identifying the net effect on market outcomes and welfare, explicitly considering the trade-off between bias reduction and precision.

Proposition 2 *(Independence) Biases manifesting as serial correlation (e.g., social influence) cannot be mitigated by increasing review volume, unlike variance-driven noise. Because the error terms exhibit non-zero serial correlation (i.e.,* $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$ *for some* $t \neq s$*), the effective sample size is reduced as redundant information provides diminishing marginal returns. Furthermore, positive serial correlation prevents errors from averaging out efficiently. Instead, initial shocks propagate through subsequent reviews, potentially leading to persistent bias even as the total number of reviews grows.*

This distinguishes social influence from unsystematic noise. While noise decreases with more data, social influence bias can increase. This distinction is applied to evaluate studies on social influence, showing that aggregation methods effective for noise are ineffective for social influence. Therefore, future work must focus on structural interventions that restore independence (e.g., hiding prior reviews during submission) rather than simply incentivizing more reviews.

Proposition 3 *(Detection) Mean-shift biases (e.g., fake reviews) and variance-inflation biases (e.g., noise) require different detection strategies. Mean shifts are detectable via outlier analysis and distributional anomalies, whereas variance inflation requires context-aware aggregation. Applying outlier detection to high-variance but unbiased signals leads to unnecessary information loss. Consequently, different bias types also require different mitigation strategies.*

This proposition provides an approach for evaluating detection algorithms. It explains why algorithms trained on fake reviews may falsely flag legitimate but idiosyncratic reviews due to high variance. This highlights the need for context-aware detection systems that can distinguish between systematic manipulation and genuine heterogeneity in user preferences.

Proposition 4 *(Artificial Intelligence) Technological advancements, for example, in generative artificial intelligence, alter the cost function of reputation manipulation, disproportionately reducing the marginal cost of creating mean bias (*$\mathbb{E}[\varepsilon_{it}] \neq 0$*) relative to the cost of detection. This shifts the equilibrium toward higher levels of systematic bias unless platforms' detection methods improve similarly.*

This frames the discussion on artificial intelligence not as a static threat but as a dynamic shift in the manipulation-detection arms race. Crucially, detection algorithms themselves are now part of the governance mix, evolving simultaneously with capabilities based on generative artificial intelligence. This implies that static design rules (e.g., fixed verification thresholds) are becoming obsolete. Instead, platforms must deploy adaptive systems where artificial intelligence-driven detection improves dynamically to counteract the near-zero marginal cost of creating mean-shift biases. This perspective is used to interpret recent findings on artificial intelligence-generated fake reviews in Sect. 3.1.1. Future research should investigate dynamic games in reputation systems where manipulation and detection capabilities co-evolve, particu-

larly examining how artificial intelligence-driven curation algorithms interact with these biases.

The propositions function as analytical propositions derived logically from the conceptual framework. Propositions 1 (Trade-offs) and 2 (Independence) follow immediately from the properties of the error term. They describe inherent constraints of the reputation signal itself (e.g., failure of the Law of Large Numbers to cancel out the bias under serial correlation). On the other hand, Propositions 3 (Detection) and 4 (Artificial Intelligence) are logical consequences of applying the framework's diagnostic distinction (mean shift vs. variance inflation) to the dynamics of platform governance. For example, Proposition 3 logically requires asymmetric detection strategies because the framework defines mean-shift and variance-inflation errors as fundamentally distinct phenomena requiring different reactions. Similarly, Proposition 4 derives the inevitability of a dynamic arms race because the realization of the bias depends on the relative costs of manipulation versus detection. Thus, while Propositions 1 and 2 relate to the signal properties, Propositions 3 and 4 relate to the governance strategies reacting to those properties. Collectively, they serve as organizing principles for synthesizing empirical findings and as a theoretical foundation for generating testable hypotheses in future research.

Together with the conceptual framework, these propositions provide the basis for the empirical review in Sect. 3. In each subsection, the error signature of the bias is identified, empirical evidence is evaluated in light of these propositions and design implications are derived that account for the identified trade-offs.

3 Biases in online reputation systems

In this section, the framework introduced in Sect. 2.3 is applied to investigate the empirical literature on five major sources of bias: (i) strategic action of sellers, including fake reviews and rebate-for-review programs, (ii) strategic action of reviewers, such as reciprocal behavior, (iii) social influence, where reviewers adjust ratings based on existing reviews, (iv) selection bias, arising from the non-representative subset of buyers who post reviews and (v) noise, including idiosyncratic tastes, misinterpretations and shocks. These sources of bias have been identified as key challenges in reputation systems, for example, by Belleflamme and Peitz [20], Lee et al. [10], Luca [11] and Tadelis [2].

The necessity of this structured analysis is evident from empirical patterns documented in the literature, such as the low correlation between ratings and objective quality [29, 30], the J-shaped distribution of ratings [31–33] and the phenomenon of reputation inflation [34]. While prior literature has documented these patterns in isolation, the conceptual framework from Sect. 2.3 provides a unified analysis of the biases in terms of the moments of the error term (mean, variance or correlation).

This section summarizes the empirical findings for each type of bias, specifically the conditions under which they arise, how they affect reputation measures and what mechanisms can mitigate their impact. While these biases differ in origin, the unified framework reveals commonalities in their effects and highlights interactions that isolated analyses often miss.

3.1 Bias due to strategic considerations of the sellers: fake reviews and rebate-for-review programs

This section examines biases in online reputation systems that arise from the strategic behavior of sellers. Strategic reputation management can involve practices that are generally legal as well as those that are generally illegal.[2] Both categories include actions designed to increase the number of reviews for a particular seller. Importantly, the entity generating the reviews may or may not be the seller who benefits from them. This section focuses on two main ways sellers influence reviews, the solicitation of fake reviews and the use of rebate-for-review programs. Both practices primarily affect the expected value of the error term $\mathbb{E}[\varepsilon_{it}]$. While the solicitation of fake reviews is motivated by shifting the expected value of the error term, rebate-for-review programs are designed to increase review volume.

Sellers have strong incentives to engage in such practices. A seller's success on a platform often depends heavily on the number and quality of reviews it receives. Many studies show that sellers with more and better reviews enjoy higher sales, appear more prominently in platform search rankings and elicit higher willingness to pay from buyers. This relationship is well documented across markets and is broadly accepted in the literature, as studied, for example, by Chevalier and Mayzlin [35] and summarized in the survey by Magnani [12] and the meta-analysis by Babić Rosario et al. [36]. With this incentive structure, many sellers attempt to actively manage or manipulate their reputation by generating additional reviews that promote their business. Reputation management is sometimes even considered a form of marketing.[3] In this vein, Mayzlin et al. [37] refer to manipulated reviews as "promotional reviews." Many sellers monitor their reviews closely and are aware that engaging in review fraud is a possible strategy, although the intensity of such behavior varies. This strategic monitoring and management behavior has been analyzed, for example, by Gössling et al. [38].[4]

Given these strong incentives, sellers may pursue concrete strategies to manipulate reviews, which are categorized here into two main practices. First, this section examines fake reviews as a form of reputation manipulation. Second, the section investigates rebate-for-review programs, which are often permitted by platforms but may affect the review process.

3.1.1 Fake reviews

Fake reviews are prohibited by platform terms and increasingly banned by regulatory frameworks globally. Recent enforcement actions by the US Federal Trade Commission [39–41], the UK's Digital Markets, Competition and Consumers Act 2024 [42]

[2] Legal standards and platform policies do not necessarily align, but distinguishing between them lies beyond the scope of this paper.

[3] See, for instance, Magnani [12] for a survey on online reputation as a factor in consumer decision making and marketing activities.

[4] Other fraudulent marketing practices, such as deceptive advertising, also exist, but fall outside the scope of this paper.

and intensified legal measures by platforms like Amazon [43] illustrate the growing regulatory urgency. Academic overviews by Sahut et al. [5] and Wu et al. [6] catalog the antecedents and consequences of this phenomenon. This section synthesizes key empirical findings following loosely the antecedent-consequence-intervention framework of Wu et al. [6], focusing on more recent contributions and mechanisms relevant to the proposed propositions. After a brief definition of fake reviews, this section provides an overview of the findings in the literature on when and why fake reviews are posted. Then, it summarizes details of how they are generated and what their properties are, followed by findings on the consequences of fake review postings. Finally, it discusses findings on detection methods and suitable reactions to fake reviews. This structure follows the order of the questions included in Table 1, which are adapted from the framework by Wu et al. [6].

Fake reviews arise when posted reviews do not reflect the genuine opinions of real buyers such that reputation is manipulated. This includes cases where reviewers

Table 1 Overview of empirical studies on fake reviews

Focus of analysis	Studies
When and why are fake reviews posted?	**F:** Anderson and Simester [47], Anderson and Magruder [48], Dini and Spagnolo [49], Hajek et al. [50], He et al. [51], Hu et al. [52–54], Ko and Bowman [55], Lappas [56], Li et al. [57], Luca and Reshef [58], Luca and Zervas [59], Mayzlin et al. [37], Ott et al. [60], Wang et al. [61], Xu et al. [62] and Zhang et al. [63]
	L: Ananthakrishnan et al. [64], Banerjee and Chua [65], Choi et al. [44], Gössling et al. [38, 66], Harrison-Walker and Jiang [67], Krügel and Paetzel [68] and Malbon [69]
How are fake reviews generated?	**F:** Dini and Spagnolo [49], He et al. [51] and Xu et al. [62]
	L: Banerjee and Chua [65], Kovács [70] and Salminen et al. [71]
What are the properties of fake reviews?	**F:** Anderson and Simester [47]
What are the consequences of fake review presence on the market?	**F:** Akesson et al. [72], Alma Economics [73], Ko and Bowman [55], Mayzlin et al. [37] and Xia et al. [74]
	L: Harrison-Walker and Jiang [67], Krügel and Paetzel [68], Malbon [69], Munzel [75] and Song et al. [76]
How can fake reviews be detected?	**F:** Alma Economics [73], Bhangale and Roy [77], Hajek et al. [50], He et al. [78], Hu et al. [54], Nawara and Kashef [79], Ott et al. [60], Xu et al. [62] and Zhang et al. [46]
	L: Harrison-Walker and Jiang [67], Kovács [70] and Plotkina et al. [80]
What are suitable reactions to fake reviews?	**F:** Akesson et al. [72], Lappas et al. [81] and Luca and Zervas [59]
	L: Ananthakrishnan et al. [64]

F: Studies based on field data or field experiments; L: Studies based on laboratory or online experiments and surveys

are paid or otherwise incentivized to write positive or negative reviews, as well as situations where reviewers have not purchased the product at all. Definitions in the literature vary depending on research focus and context, but they consistently characterize fake reviews as misleading because they either misrepresent the reviewer's true opinion, involve reviewers who do not belong to the group of legitimate reviewers or stem from financial incentives [see, e.g., 6, 37, 44–46]. Within the conceptual framework established in Sect. 2.3, fake reviews represent a systematic violation of the Mean Condition. Because these reviews are intentionally designed to distort reputation without reflecting true quality, they introduce a mean shift where $\mathbb{E}[\varepsilon_{it}] \neq 0$. A critical insight provided by this classification is that this mean shift does not average out with increased sample size. Adding more reviews simply accumulates further biased signals, preventing convergence to true reputation. Consequently, effective mitigation should target the source of the bias rather than relying on aggregation.

The extensive coverage of fake reviews reflects their distinct characteristics. First, the empirical literature on fake reviews is voluminous and active, driven by the rapid evolution of manipulation tactics (e.g., organized markets, increased usage of generative artificial intelligence) and the corresponding detection methods. In contrast, literature on other biases such as reciprocity and social influence has largely converged on structural solutions (e.g., simultaneous reveal, blinded reviews), resulting in a more mature but less expansive recent body of work. Second, fake reviews represent a case of malicious intent rather than merely flawed design or cognitive bias. Unlike noise or selection bias, which arise from endogenous user behavior or platform architecture, fake reviews are strategic, fraudulent attacks on the reputation signal. This distinction necessitates a deeper analysis of the dynamic equilibrium between manipulation and detection (Proposition 4). Third, there is current societal and regulatory urgency related to fake reviews. Recent legislative actions and significant consumer welfare losses have prioritized this bias in both research and policy agendas. Consequently, a comprehensive synthesis of this specific stream is essential to capture the current frontier of reputation system challenges, whereas other biases are treated with a focus on their established structural mitigation strategies.

Fake reviews have been studied extensively in the empirical literature. Table 1 provides an overview of the main aspects examined in the literature grouped by the focus of the studies, which loosely follows the antecedent-consequence-intervention framework by Wu et al. [6]. The table differentiates between empirical studies that rely on field data or field experiments and those that rely on laboratory or online experiments and surveys. The remainder of this section summarizes key empirical findings in the order of the questions from the table.

Empirical research consistently shows that fake reviews are a widespread phenomenon across online markets. Depending on the platform and the detection method used, studies estimate that between 5% and 33% of reviews are likely to be fake.[5]

Most obviously, sellers of goods and services may have incentives to create fake reviews. As documented by numerous empirical studies, sellers benefit from generating positive fake reviews for themselves through higher sales, improved search ranking and greater willingness to pay by consumers; see, for example, Anderson and Magruder [48] and Luca and Reshef [58]. Using Amazon data, Wang et al. [61] find an inverted-U relationship between the number of manipulated reviews and sales and a positive relationship between the information quality of manipulated reviews and sales. Given these benefits, sellers often have strong incentives to solicit positive fake reviews.

Sellers may also attempt to generate negative fake reviews about their competitors. By lowering a rival's reputation, they can improve their own relative reputation, potentially harming competitors' sales while boosting their own. Negative fake reviews appear to be less common than positive ones. As Harrison-Walker and Jiang [67] argue, soliciting negative reviews is perceived as riskier and more difficult, a finding also reflected in the survey results of Gössling et al. [38, 66]. Empirical work on incentives for negative fake reviews includes Luca and Zervas [59] and Mayzlin et al. [37], which will be discussed below.

Furthermore, buyers may have incentives to post fake reviews or at least threaten to do so. Some buyers exploit practices where hotels offer compensation to unsatisfied guests by threatening to post unjustified negative reviews unless the seller provides a benefit. Gössling et al. [38, p. 494] refer to such behavior as "corrupt complaints." Experimental evidence from Choi et al. [44] using participants from Amazon Mechanical Turk suggests that consumers can be incentivized to leave fake reviews through monetary rewards or even promised charitable donations and that participants perceive negative fake reviews as more immoral than positive ones.

Sellers may have different incentives to solicit fake reviews depending on whether their product or service is of high or low quality. Low-quality sellers may seek to compensate for inferior products by soliciting positive fake reviews, but these gains are often short-lived. Once sales increase, genuine negative reviews are eventually posted, making it difficult and costly to sustain an artificially inflated overall rating. Platforms use methods such as deleting fake reviews or penalizing accounts, thereby increasing the costs and risks of creating fake reviews. A similar logic applies to negative fake reviews about competitors, which low-quality sellers may use to improve their own relative standing. This has been investigated empirically by Li et al. [57] and Mayzlin et al. [37].

[5] Alma Economics [73] estimates with a detection algorithm that between 11% and 15% of reviews on the investigated UK e-commerce platforms are fake. Anderson and Simester [47] find in their analysis of reviews on a retailer's website that 5% of reviews are without a verified purchase, which could be an indication of fake reviews. Hu et al. [54] find that around 10.3% of reviews in their data from Amazon.com are manipulated. Luca and Zervas [59] find that 16% of reviews on review platform Yelp are filtered by the platform, which they consider a proxy for fake reviews. Ott et al. [60] find a prevalence of fake reviews across various hotel reputation platforms of up to approximately 6%. Salehi-Esfahani and Ozturk [82, p. 138] state that the "general ballpark figure is that one-third of reviews are fake."

In contrast, high-quality sellers, especially new entrants to the market, may benefit from posting a small number of initial positive reviews. Because buyers are often reluctant to purchase from sellers with few or no reviews and because such sellers tend to be ranked less prominently, a small number of positive reviews can help overcome the cold-start problem. Once sales increase, genuine positive reviews are likely to follow. Under some market conditions, only high-quality sellers find it profitable to generate early positive reviews, as low-quality sellers cannot recoup the cost of such manipulation through sustained future sales. These dynamics are explored by Li et al. [83]. These incentives are related to those in the rebate-for-review programs summarized in Sect. 3.1.2 below.

Studies document fake reviews across major marketplaces (e.g., Amazon [47, 52–55], eBay [49], JD.com [74], Taobao [62]), hospitality platforms (e.g., Yelp, Tripadvisor [48, 59, 63]) and retail sectors using field data, laboratory experiments [64, 68] and qualitative interviews [38, 66, 69]. Fake reviews have been documented for a wide range of products and services, including books [52], restaurants [57, 59, 63] and hotels [37, 60].

As found by Gössling et al. [38, 66] and Hu et al. [52], many sellers monitor their online reputation closely and intervene when necessary, though not necessarily through fake reviews. The literature identifies several conditions that make fake review solicitation more likely.

First, firms are more likely to solicit fake reviews when their reputation is weak, for example, when they have only a few reviews or have recently received negative ones. Luca and Zervas [59] show this using Yelp restaurant data, exploiting Yelp's practice of publicly displaying reviews filtered out by its fraud-detection algorithm. Treating these filtered reviews as proxies for fake reviews, they find that restaurants with weaker reputations are more likely to engage in manipulation practices. Using the same approach with Yelp data, Zhang et al. [63] find that restaurants tend to increase fake review activity when their competitors do so, but reduce manipulation when they have a reputational advantage. Because being exposed for soliciting fake reviews can damage credibility, loss-averse sellers appear less willing to solicit fake reviews.

Second, independent businesses are more likely to engage in fake review solicitation than businesses belonging to larger chains. Detection imposes reputational and financial costs, which are typically more severe for chains because being caught manipulating reviews may harm the entire chain, given its larger exposed profit base. Luca and Zervas [59] document this pattern for Yelp restaurants and Mayzlin et al. [37] draw similar conclusions for hotels.

Third, competitive intensity increases incentives to solicit fake reviews. Restaurants are more likely to solicit fake reviews when competitors, particularly those offering similar products, operate in geographical proximity. Li et al. [57] and Luca and Zervas [59] find this using restaurant data. Hu et al. [52] find similar results for fake reviews in the book market where fake reviews are more likely for non-bestselling books, books with less helpful reviews or high-priced books, all of which can be considered a competitive disadvantage. Mayzlin et al. [37] document similar patterns and find that hotels are more likely to post negative fake reviews about competitors when there are nearby competitors. The authors employ a difference-in-differences

strategy comparing hotel reviews on Tripadvisor, where any user may post, with those on Expedia, which restricts reviews to verified bookings.

Fourth, fake reviews become more frequent if they are inexpensive to solicit. Krügel and Paetzel [68] analyze incentives to solicit fake reviews in a laboratory public-good experiment where subjects could rate each other's contributions. Subjects could change the received rating at no, low or high cost. Furthermore, there was a control treatment without the possibility to change the rating. They find that subjects use the opportunity to change their own rating if possible and that there are more changed (i.e., fake) ratings the lower the costs are. They conclude that increasing the costs of review manipulation can reduce manipulation and increase efficiency. Consistent with this, Ott et al. [60] find with their machine-learning detection approach that the prevalence of fake reviews on platforms with high costs of posting a review[6] is relatively stable over time, while it increases on the platforms where it is relatively cheap to post fake reviews, indicating that agents tend to continuously solicit fake reviews on these platforms.

These findings illustrate a standard economic mechanism: sellers engage in manipulation as long as the marginal benefit of inducing a mean shift in reputation is high and exceeds the marginal cost of manipulation, detection and penalty. The lower the costs and potential penalties, the more fake reviews are posted.

In addition to understanding why fake reviews are posted, it is also important to consider how they are created. Posting fake reviews is costly, particularly because many platforms apply verification procedures or automated filters. Sellers often rely on third parties to post fake reviews, either individuals who they directly incentivize to submit reviews or specialized agencies that coordinate the entire process; see, for instance, He et al. [51] and Xu et al. [62].

On relatively unrestrictive platforms, in particular those that do not require a transaction to be completed through the platform, fake reviews are relatively inexpensive to produce. If only star ratings are posted, the required effort is low. Generating textual reviews is more demanding. They are typically more valuable for sellers because they tend to be perceived as more credible and persuasive. At the same time, they must be sufficiently coherent and plausible to avoid detection algorithms and to convince potential buyers. As Kovács [70] argues, this task has become less costly over time through the growing availability of generative artificial intelligence, which can produce large volumes of realistic textual reviews.

The more restrictive a reputation system is, the higher the cost of posting fake reviews. Many platforms, for example, hotel booking site Expedia, allow reviews only from guests who have completed a transaction through the platform, a feature exploited by Mayzlin et al. [37] to infer the prevalence of fake reviews on less restrictive platforms. Other platforms, such as Amazon.com, label reviews as "verified purchase" when the reviewer bought the product through the platform. In these settings,

[6] On platforms Orbitz, Priceline, Expedia and Hotels.com, it is relatively costly to post fake reviews since these platforms generally require a confirmed booking before posting a review. On the other hand, on Yelp and Tripadvisor, no such requirement exists. Hence, it is relatively cheap to post a fake review on these platforms.

sellers cannot easily generate fake reviews without involving third parties who are willing to make real purchases.

Modern manipulation activities have spread into organized markets. Sellers utilize third-party brokers and social media groups to coordinate "verified purchase" reviews, where participants are refunded and compensated to bypass platform filters [51, 62].

He et al. [51] study a market for fake review solicitation on Amazon.com. They identify Facebook groups on which fake reviewers are recruited. Individuals are motivated to complete a purchase on Amazon and subsequently leave a positive 5-star review, typically receiving a full refund of the purchase price through channels outside of the Amazon platform. The key benefit for sellers is that these reviews appear as "verified purchase," thereby enhancing credibility, which increases their product's search ranking and chances of sale.

Using data from these Facebook groups, He et al. [51] identify which products on Amazon are affected by fake reviews and compare their performance with that of unaffected products. They find that affected products experience increases in review volume, search rank and sales rank after soliciting fake reviews. Increases in average ratings and review volume appear to be temporary, whereas the improvement in sales rank is more long-lasting. They also find that Amazon removes a substantial share of fake reviews. Amazon's mass deletion of reviews in mid-March 2020 provides a quasi-experimental opportunity that allows the authors to apply a difference-in-differences strategy. Comparing products whose fake reviews were removed quickly with those whose fake reviews remained visible longer, they conclude that fake reviews have a causal positive effect on sales rank.

A related but structurally different market is examined by Xu et al. [62], who analyze crowdsourcing markets on which sellers purchase services to generate fake transactions on Taobao. These services accelerate reputation building while only 2.2% of sellers are detected and penalized [62]. The crowdsourcing markets coordinate payments, guide reviewers' behavior and provide detailed instructions to reviewers for minimizing detection risk, including guidelines for imitating realistic search and browsing behavior before purchase and avoiding new accounts and the repeated use of the same device or IP locations.

Dini and Spagnolo [49] report on the easy generation of fake reviews on eBay through so-called "shill auctions." They document how sellers generate fake transactions, often involving fictitious goods priced as low as USD 0.10 or 0.99, to exchange positive feedback cheaply. In some cases, neither payment nor shipping takes place. The appearance of a transaction alone is sufficient to leave mutually beneficial positive feedback on eBay.

Although organized systems facilitate large-scale manipulation, most fake reviews are still written by human agents who follow guidelines that reduce the likelihood of detection. Banerjee and Chua [65] analyze the cognitive process of writing credible fake reviews, identifying stages of information gathering, assimilation, drafting and finalizing the fake reviews through an online questionnaire. However, recent advances in generative artificial intelligence, such as large language models, substantially decrease the effort required to produce high-quality fake reviews. This development raises new concerns about the reliability of online reputation systems

[70, 71, 84] and has resulted in regulator actions against artificial intelligence-based fake review generation [e.g., 40]. Experimental evidence suggests that artificial intelligence-generated reviews are difficult for humans to identify. Kovács [70] finds no significant difference in classification accuracy between real and artificial intelligence-generated restaurant reviews. Salminen et al. [71] find that humans are generally worse at detecting fake reviews than algorithms. These findings empirically validate the mechanism proposed in Proposition 4 (Artificial Intelligence). As the marginal cost of generating credible fake reviews decreases significantly, the presence of fake reviews rises unless platform detection capabilities improve similarly. Generative artificial intelligence drastically lowers the marginal cost of creating credible fake reviews, shifting the dynamic equilibrium toward higher levels of systematic bias unless the platform intensifies its countermeasures.

Additional evidence on potentially deceptive reviews is provided by Anderson and Simester [47], who examine product reviews from an apparel company selling through its own retail channels. They find that approximately 5% of reviews cannot be matched to the company's transaction records. These unmatched reviews are, on average, more negative and less detailed than matched reviews. While the authors cannot rule out legitimate reasons for a reviewer's absence from the transaction list,[7] the authors note that the systematically different linguistic patterns of unmatched reviews require further explanation. Replicating the findings using verified and non-verified purchase reviews on Amazon.com, they find that non-verified reviews are more likely to exhibit linguistic properties associated with deceptive reviews, such as unrelated details, shorter words and excessive use of exclamation points. They find that some of the unverified reviewers are among the company's best customers, making competitor sabotage unlikely. The authors propose three possible explanations: (i) such reviewers may act as "brand managers", (ii) they may be upset customers (though evidence is limited) or (iii) they may seek social recognition by posting reviews.

Next, this section summarizes evidence on how fake reviews affect stakeholders. Krügel and Paetzel [68] examine the role of review fraud in a laboratory experiment with a public-good game. They show that when fraudulent reviews can be created at low cost, subjects place significantly less weight on displayed ratings. They find that contributions to the public good and thereby efficiency are lower when fraud is costless.

A range of studies demonstrates that fake reviews reduce consumer welfare and increase the likelihood of purchasing inferior products. Akesson et al. [72] conduct an incentive-compatible online experiment with a representative sample of UK consumers. Participants interact with a simulated marketplace resembling Amazon, choosing among five products. One product is high-quality, one low-quality and three average-quality. Six treatments vary the presence and type of fake reviews, including inflated star ratings, overly positive textual reviews and obviously fake content. One treatment additionally includes an educational intervention about fake reviews. To

[7] Anderson and Simester [47] list a number of possible reasons for reviewers not to appear on the transaction list, including gift recipients, customers misidentifying items or reviews complaining about non-product issues such as service complaints.

assess welfare effects, Akesson et al. [72] elicit willingness to pay. They argue that welfare losses occur when consumers seeing fake reviews select products they value less. Fake reviews shift demand: a one-star increase raises demand for low-quality products by 38%, generating welfare losses. They also find that consumers who are less trusting of reviews are less affected, while frequent users of online marketplaces experience larger negative effects. This suggests that experience does not mitigate vulnerability to review manipulation. Firms may thus benefit in the short run from producing fake reviews.

Alma Economics [73] conduct a similar online experiment with UK participants modeled after Akesson et al. [72], though lacking incentive compatibility. They expose subjects to genuine, subtly manipulated or strongly manipulated fake reviews.[8] Subtle fake reviews increase the chance of choosing the manipulated product, while strong, obvious fake reviews reduce it. They conclude that consumers can identify extreme forms of fake reviews, but subtle manipulations remain effective, generating welfare losses.

Fake reviews may also erode trust in reputation systems. Ko and Bowman [55], using data from Amazon.com, show that even the suspicion that fake reviews may be present reduces the perceived usefulness of reviews, especially for highly rated brands. Similarly, Xia et al. [74] find that fake reviews initially boost sales on JD.com, but as consumers recognize the deception, trust in the reputation system declines. Experimental evidence by Song et al. [76] indicates that the effect of fake reviews depends on product type.

Overall, the evidence suggests two major effects of fake reviews on consumers. First, consumers may make suboptimal choices, reducing welfare [37, 72, 73]. Second, awareness or suspicion of fake reviews may undermine trust in reputation systems, making consumers reluctant to complete purchases and potentially harming brand perceptions [37, 55, 67, 68, 75, 76]. However, consumers do not appear to universally mistrust online reviews [69], and some evidence suggests that consumers can recognize at least obvious forms of fake reviews [73].

Beyond their impact on efficiency, fake reviews raise important ethical concerns [44]. They constitute a form of deception or fraud, as consumers are deliberately misled about product quality or seller reliability. This not only harms buyers but also disadvantages honest sellers, who face unfair competition from firms that manipulate reviews [37]. Moreover, fake reviews are often produced by organized agencies or low-paid workers, turning deception into a service [62]. In such environments, even sellers who would prefer not to engage in manipulation may feel compelled to do so in order to remain competitive, creating a collective action problem in which unethical behavior becomes normalized.

Addressing fake reviews requires two steps, identifying them and deciding how to respond, as suggested by Ananthakrishnan et al. [64]. The first step is to find a way to detect them and to distinguish them from genuine reviews. Surveys on detection methods for fake reviews include Mohawesh et al. [7] and Vidanagama et al. [8].

[8] Subtle fake reviews include generic or exaggerated language or multiple reviews posted on the same day; strong fake reviews include reviews describing a different product or disclosing that the reviewer was paid.

Fake reviews can often be identified using contextual factors surrounding the review posting. Reviewer histories may reveal suspicious patterns. Examples include multiple reviews from newly created accounts in a short period or multiple accounts sharing the same IP address. Whether a user directly accesses a URL or browses and compares related products can also provide clues. Reputation systems can incorporate these signals to flag potential fake reviews. However, systematic fake review creators often circumvent detection by using aged accounts and varied IP addresses [46, 62].

Another indicator involves reviewer-product networks. Alma Economics [73] and He et al. [78] note that reviewers who frequently review the same set of products may be part of organized fake review activity. He et al. [78] leverage Amazon data to detect such clusters, showing that products with fake reviews tend to share reviewers, facilitating identification of other potentially affected products.

Account properties also provide signals of credibility. For example, accounts with random-character names are generally considered suspicious, whereas accounts with a real name are generally more credible [67]. Consumers can, to some extent, use these indicators to judge review trustworthiness, and platforms can integrate them into their detection algorithms.

The content of textual reviews can also indicate deception. Consumers often rely on textual cues, such as exaggerated language, repetitive wording or statements implying that the reviewer was compensated. Obvious fake reviews may reference an entirely different product. Such characteristics of obvious fake reviews have been mentioned by Akesson et al. [72], Alma Economics [73] and Harrison-Walker and Jiang [67]. Harrison-Walker and Jiang [67] find some cues associated with genuine reviews. Genuine reviews tend to provide more detailed descriptions and are more credible when negative, as negative fake reviews are rarer.

Platforms can apply advanced textual analysis such as algorithmic detection [60, 73], sentiment analysis [50, 54] and the use of large language models to detect fake reviews [77, 79]. For instance, Ott et al. [60] build models using truthful reviews and artificially generated deceptive reviews to predict the prevalence of fake reviews on online travel platforms. Similarly, Hu et al. [54] use statistical analysis of textual and numerical review properties, while Hajek et al. [50] emphasize product type and verified purchase labels as key indicators.

Advances in artificial intelligence can be used not only to create fake reviews but also to detect them. Bhangale and Roy [77] use the same dataset as Ott et al. [60] and Salminen et al. [71] to test whether large language models can detect fake reviews using linguistic properties. They find that their approach performs well at detecting fake reviews with a detection accuracy of over 90%. Furthermore, Nawara and Kashef [79] achieve a strong detection performance with a related approach.

Fake reviews can also be detected using platform-specific features. Luca and Zervas [59] and Zhang et al. [63] utilize review platform Yelp's feature of displaying all reviews, including those filtered, for instance, due to suspected manipulation. Mayzlin et al. [37] compare Tripadvisor, which allows anyone to post reviews, with Expedia, which only permits verified purchasers to review, to statistically infer the presence of fake reviews. While these methods are mostly research tools, the availability of this information can still affect consumers' trust in sellers or products.

Humans are generally poor at detecting fake reviews. Plotkina et al. [80] find that even when subjects in their experiment are aware of typical characteristics of fake reviews, their detection accuracy is only 57%. Similarly, Salminen et al. [71] show that algorithmic methods outperform humans in identifying fake reviews, highlighting the importance of platform-level detection systems.

Reducing the number of fake reviews requires both detection and appropriate response strategies. Luca and Zervas [59] suggest four main approaches: (i) using detection algorithms, (ii) allowing only verified purchasers to leave reviews, (iii) conducting stings against fraudulent reviewers or companies and (iv) leveraging behavioral economics to nudge companies away from posting fake reviews.

Most platforms react to detected fake reviews by filtering them out, while others, like Yelp, display them as fraudulent or not recommended. Ananthakrishnan et al. [64] investigate experimentally the most suitable reaction to fake reviews. They argue that retaining and labeling fake reviews can increase transparency, potentially discouraging businesses from soliciting them, as the reputational cost of being discovered increases. Yelp, for instance, displays a banner on profiles of businesses involved in systematic fake review solicitation alongside flagging individual posts, deterring businesses by highlighting the potential negative consequences. However, this transparency can also reduce consumers' trust in the platform if it draws attention to the presence of fake reviews.

Restricting reviews to verified purchases raises the cost of fake review solicitation. Platforms can raise these costs further through stricter eligibility criteria. For example, Amazon.com allows only users to post reviews if they have spent at least USD 50 in the last 12 months [85]. This is in line with the suggestion made by Luca [11] to give more weight to reviews posted by reviewers with longer transaction histories, as they are generally less likely to be involved in fake review solicitation.

Shukla and Goh [84] propose verifying a user's digital identity through a third party, confirming both identity and transaction validity without revealing the user's identity to the seller. Financial institutions such as credit card companies could serve this verification role, because they have access to payment transaction data. While verified purchases reduce certain types of fake reviews, limitations exist. Businesses can still post positive fake reviews by purchasing products themselves, and the number of eligible reviews excludes friends or family who share an experience, such as staying at the same hotel [81].

Reputation systems can ban companies, reviewers or products identified as engaging in fake review solicitation. In practice, this approach is rarely used, as new seller profiles can be created easily and at low cost. Platforms rarely ban sellers active in fake review solicitation, as observed, for example, on Amazon.com by He et al. [51].

Educating consumers about the potential presence of fake reviews can mitigate their negative effects. Akesson et al. [72] find that an educational intervention reduced the welfare loss from fake reviews by 44% in an experimental setting. While it does not completely eliminate the effect, this simple measure can significantly improve consumer decision-making and overall welfare.

Review aggregation methods can reduce the impact of fake reviews without removing them entirely. Ivanova and Scholz [86] propose dynamically aggregating reviews by type (e.g., positive or negative) and splitting them into equal-sized subsets

for aggregation. This approach diminishes the influence of individual fake reviews while maintaining visibility of less frequent ratings. Their simulations indicate that this method outperforms other aggregation techniques in mitigating the effects of fake reviews on sales.

As established, fake reviews constitute a systematic violation of the Mean Condition ($\mathbb{E}[\varepsilon_{it}] \neq 0$). Whether inflating ($\mathbb{E}[\varepsilon_{it}] > 0$) or deflating ($\mathbb{E}[\varepsilon_{it}] < 0$) reputation, these manipulations require targeted interventions, as aggregation cannot correct the mean shift. Furthermore, the rise of artificial intelligence underscores the dynamic equilibrium in Proposition 4, where manipulation costs fall relative to detection.

The findings from the literature on fake reviews can be summarized as follows:

Finding 1 *Fake reviews constitute a systematic violation of the Mean Condition ($\mathbb{E}[\varepsilon_{it}] \neq 0$), introducing a systematic distortion (mean shift) in reputation. Empirical evidence confirms that this distortion significantly decouples observed reputation from true quality, leading to welfare losses and eroding trust in the reputation system. This systematic bias requires targeted interventions to mitigate the negative effects.*

Finding 2 *The prevalence and impact of fake reviews are driven by a dynamic equilibrium between the marginal benefit of manipulation and the marginal cost of detection. Recent evidence on generative artificial intelligence supports Proposition 4 (Artificial Intelligence), indicating that technological advancements significantly lower the cost of creating fake reviews. This shifts the equilibrium toward higher levels of systematic bias, requiring similar scaling of platform detection capabilities.*

Finding 3 *Mitigation strategies for fake reviews involve inherent trade-offs consistent with Proposition 1 (Trade-offs). While interventions such as verified-purchase requirements and algorithmic detection effectively reduce the mean shift, they may inadvertently increase variance ($\mathrm{Var}[\varepsilon_{it}]$) by reducing the sample size or excluding legitimate reviews, thereby potentially creating new selection bias. While the mean may become more accurate, the precision of the reputation signal may decline. Consequently, optimal platform design requires balancing the reduction of systematic bias against the preservation of signal precision and participation levels, rather than minimizing a single distortion in isolation.*

3.1.2 Rebate-for-review programs

Sellers also influence reputation by offering reviewers discounts, a practice often legal under certain conditions. If the discount is unconditional, that is, if it is granted regardless of whether the review is positive, many platforms allow such incentivized reviews because they do not consider them misleading. Unconditional rebates reduce the incentive to inflate ratings because reviewers know they will receive the discount regardless of review valence. This makes them less likely to distort reputation than conditional rebates. For example, the online marketplace Taobao uses a rebate-for-review program designed specifically to help new products collect early reviews, thereby helping to overcome the "cold-start" problem. New sellers, especially those competing against established rivals with many positive reviews, struggle to generate

the initial sales needed to receive reviews. By offering discounts, new entrants can encourage buyers to leave reviews by reducing their effective purchase price. The design and functioning of such programs are described by Li [88], Li and Xiao [87] and Li et al. [83].

Within the conceptual framework, rebate-for-review programs represent a complex intervention designed to increase the number of reviews and are potentially useful for correcting selection bias (Sect. 3.4), which itself causes a violation of the Mean Condition ($\mathbb{E}[\varepsilon_{it}] \neq 0$) by excluding reviewers with certain predispositions. By incentivizing participation, these programs aim to restore the representativeness of the sample. However, they introduce a secondary risk: the incentive itself may induce reciprocal behavior, where reviewers feel obligated to provide positive feedback as a response to receiving a rebate, thereby shifting $\mathbb{E}[\varepsilon_{it}]$ upward. This tension illustrates how interventions targeting one bias source may inadvertently activate another.

Rebate-for-review programs have been investigated in the empirical literature. Table 2 summarizes the empirical literature on such programs, differentiating between empirical studies that rely on field data or field experiments and those that rely on laboratory or online experiments and surveys. The remainder of this section summarizes these empirical findings.

If such a program is managed by a third party, typically the platform operator, reviewers can provide more honest feedback without fearing that a negative review will disqualify them from the discount. As explained by Li et al. [83], reviewers often have implicit incentives to avoid writing negative reviews when the seller controls the discount payment because they may worry that a negative review will not be rewarded. Consequently, reviewers may inflate ratings when discounts are administered by the seller. High-quality sellers, however, have a long-term incentive to offer unconditional discounts because the resulting truthful reviews help signal their superior quality. In contrast, low-quality sellers would expect negative reviews under an unconditional scheme and therefore have no incentive to offer such discounts because they are costly and do not offer a reward in terms of positive reviews. This equilibrium is described by Li et al. [83].

Empirical evidence, however, suggests that rebate-for-review programs may generate biased reviews due to reciprocal behavior. Cabral and Li [89] find using eBay data that reviewers who receive a discount tend to reciprocate by posting more positive reviews, even for low-quality products. Garnefeld et al. [92] reach the same conclusion in an experiment with participants recruited on Amazon Mechanical Turk. Park et al. [91] show with Amazon data that disclosing that a review was incentivized does not reduce this tendency toward positivity. In contrast, Fradkin and Holtz [90] find using Airbnb data that incentivized reviews can actually be more negative, suggesting that offering incentives decreases transaction quality. Li and Xiao

Table 2 Overview of empirical studies on rebate-for-review programs

Studies
F: Cabral and Li [89], Fradkin and Holtz [90], Li et al. [83] and Park et al. [91]
L: Garnefeld et al. [92], Koukova et al. [93] and Li and Xiao [87]

F: Studies based on field data or field experiments; L: Studies based on laboratory or online experiments and surveys

[87] find using experimental data that rebates increase review volume but also help reduce bias because reviewers' reporting honesty is not affected by the rebate. Moreover, Koukova et al. [93] show in online experiments that requesting a conditional review can harm reputation compared to requesting an unconditional review. Taken together, these findings suggest that it remains unclear whether rebate-for-review programs systematically bias reviews upward, downward or not at all. This ambiguity illustrates the dynamic predicted by Proposition 1 (Trade-offs). The net effect on $\mathbb{E}[\varepsilon_{it}]$ depends on whether the reduction of selection bias (improving accuracy) or the induction of reciprocal behavior (distorting accuracy) dominates. Consequently, the outcome is highly context-dependent, varying by platform design, the entity managing the rebate (seller vs. platform) and the product type. This supports the proposition that isolated interventions often yield mixed results. Design interventions must be carefully calibrated to ensure that the correction of one distortion does not trigger another.

The findings from the literature on rebate-for-review programs can be summarized as follows:

Finding 4 *Rebate-for-review programs serve as a structural intervention to correct selection bias by incentivizing participation from underrepresented customer segments, thereby addressing the non-representative sample that causes a violation of the Mean Condition ($\mathbb{E}[\varepsilon_{it}] \neq 0$). Empirical evidence confirms that such programs effectively increase review volume and help new sellers overcome the cold-start problem, improving the representativeness of the reputation signal.*

Finding 5 *However, the net effect of rebate-for-review programs on reputation accuracy is contingent on design specifics, illustrating the trade-offs predicted by Proposition 1 (Trade-offs). While unconditional rebates managed by neutral parties can reduce selection bias without distorting valence, seller-managed or conditional incentives can induce reciprocal behavior, shifting $\mathbb{E}[\varepsilon_{it}]$ upward regardless of true quality. Consequently, the literature reports mixed results: some studies find improved signal accuracy due to reduced selection bias, while others find reduced accuracy due to introduced reciprocity bias. This highlights that interventions targeting one source of mean shift may inadvertently activate another.*

3.2 Bias due to strategic considerations of the reviewers: reciprocity in two-sided reputation systems

Biases in reviews can arise from reviewers' strategic considerations, particularly when reciprocity influences the review process. Some platforms implement two-sided reputation systems, allowing buyers and sellers to review each other. In such systems, reviewers may adjust their ratings based on anticipated responses from their counterpart. Specifically, reviewers may post exclusively positive reviews even when dissatisfied with a transaction to avoid retaliation, such as receiving a negative review in return. Consequently, reviewers may adopt strategies of reciprocating positive reviews with positive responses and negative reviews with negative responses, regardless of their true level of satisfaction. This behavior can reduce the informative-

ness of reviews and potentially decrease market efficiency. This has been investigated in detail, for instance, by Bolton et al. [19] and Fradkin et al. [94]. Within the conceptual framework, the fear of retaliation or the desire to reward leads to a systematic upward shift in the mean error term ($\mathbb{E}[\varepsilon_{it}] > 0$), violating the Mean Condition. This requires structural interventions such as simultaneous reveal to mitigate the bias.[9]

Two-sided reputation systems are particularly useful when adverse selection or moral hazard exists on both sides of a market. In most modern online marketplaces, the risk of moral hazard on the buyer side is limited, as platforms ensure that buyers complete payments during the order process. However, the platform cannot fully guarantee the quality or honesty of sellers' actions, making reputation systems critical on the seller side. In peer-to-peer markets, such as Airbnb or the earlier version of eBay, adverse selection or moral hazard may exist on both market sides. In these cases, two-sided reputation systems can provide valuable information about both parties that would otherwise be unavailable.

Table 3 summarizes the main aspects investigated in the literature on reciprocity. The literature is grouped by the focus of the studies, which loosely follows the antecedent-consequence-intervention framework by Wu et al. [6]. The table differentiates between empirical studies that rely on field data or field experiments and those that rely on laboratory or online experiments and surveys. The remainder of this section summarizes key empirical findings in the order of the questions from the table.

Table 3 Overview of empirical studies on reciprocity

Focus of analysis	Studies
When and why does reciprocity affect reviews?	**F:** Bolton et al. [19], Cabral and Hortaçsu [96], Dellarocas and Wood [97], Fradkin et al. [94], Jian et al. [98], Klein et al. [17, 18, 99], Li [100], Proserpio et al. [95] and Resnick and Zeckhauser [4]
	L: Bolton et al. [19]
What are the consequences of reciprocity in reputation systems?	**F:** Bolton et al. [19, 101], Fradkin et al. [94] and Klein et al. [18, 99]
	L: Bolton et al. [19, 101]
What are suitable reactions to reciprocity in reputation systems?	**F:** Bolton et al. [19], Fradkin et al. [94] Hui et al. [102] and Klein et al. [17, 18]
	L: Bolton et al. [19]

F: Studies based on field data or field experiments; L: Studies based on laboratory or online experiments and surveys

[9] It should be noted that reciprocal considerations can also affect reviews through other channels than mutual reviews, such as reviewers showing a more positive attitude after receiving a discount for writing a review (see Sect. 3.1.2). Another related form of reciprocity has been studied by Proserpio et al. [95], who examine the sharing economy using data from the accommodation platform Airbnb. In their study, reciprocity is defined as the "tendency of market participants to respond to good (bad) behavior with good (bad) behavior" [95, p. 372]. In contrast to two-sided reputation systems, their focus is on the effort contributed by transaction partners. For instance, Airbnb hosts may provide extra services such as meals, and guests may leave the accommodation in a clean state. Proserpio et al. [95] find a positive correlation between ratings and their reciprocity measure, proxied by the length of textual reviews. However, this section only considers reciprocity in two-sided reputation systems.

In two-sided reputation systems, reciprocity can influence reviewers' incentives. A well-studied example is eBay's reputation system, which has undergone several changes over time. Klein et al. [17] provide a detailed history of these developments. Until May 2007, eBay's reputation system allowed buyers and sellers to review each other, providing positive, neutral or negative ratings along with textual feedback. Reviews were immediately posted and visible on the website, and removal was only possible through a court ruling or mutual agreement between buyer and seller.

To address issues of retaliation for negative feedback, eBay introduced detailed seller ratings in May 2007, in addition to the existing two-sided system. Buyers could rate sellers on multiple aspects of the transaction, including accuracy of item description, communication, shipping speed and shipping charges, using a one-to-five star scale. These detailed ratings remained anonymous and were reported on the website only as aggregates [17]. This system is referred to as eBay's interim reputation system.

In May 2008, eBay again modified its system, making it effectively one-sided. Sellers could only provide positive feedback about buyers or abstain from feedback entirely. The goal was to eliminate the possibility of retaliation for negative buyer reviews [17]. This version is referred to as eBay's new reputation system.

The vast majority of feedback on eBay is positive—over 99%—which likely does not reflect actual user satisfaction. For example, in 2004, 16% of consumer fraud complaints at the Federal Trade Commission were related to internet auctions [97, p. 460]. Dellarocas and Wood [97] attribute this discrepancy largely to reciprocal considerations in the old two-sided feedback system. They propose a method that estimates the probability that each feedback type (positive, neutral or negative) reflects the true (unobserved) satisfaction of buyers and sellers. Their approach considers the timing of feedback and instances where only one party provides a review. Applying this method to eBay data, they estimate that buyers were satisfied in only 78.9% of transactions, suggesting underreporting of neutral and negative experiences. Similarly, Li [100] concludes from eBay data that the fear of retaliation may drive reviewers not to post a review.

Bolton et al. [19] find that approximately 70% of traders leave feedback, which is consistent with other studies. Their analysis of data from the old eBay reputation system shows that buyers and sellers are more likely to provide feedback if their counterpart has already done so, with an effect stronger for sellers. They also observe a high positive correlation between buyer and seller feedback, with 85% of seller feedback responding to negative buyer feedback being negative. They conclude that "sellers reciprocate positive feedback and 'retaliate' for negative feedback" [19, p. 268], sometimes under the influence of "emotional arousal." Retaliation may deter future negative reviews.

Other studies offer complementary insights. For example, Resnick and Zeckhauser [4] find a high correlation between buyer and seller ratings. Klein et al. [99] report that 71% of positive feedback is reciprocated, while only 37% of negative feedback triggers retaliation. Cabral and Hortaçsu [96] find that nearly one quarter of negative or neutral reviews are followed by retaliatory reviews, with retaliation more likely for negative reviews (40%) than neutral ones (10%), suggesting structural flaws in the old system. Jian et al. [98] find using eBay data that buyers and sellers use a

reciprocity strategy in more than 20% of reviews. However, they also conclude that experienced reviewers tend to use this strategy less often.

While retaliatory feedback is relatively rare—covering less than 1.2% of mutual feedback [19]—its effects are amplified because buyers strategically adjust their feedback to avoid retaliation. Klein et al. [18] find that the share of positive reviews decreased slightly with the introduction of the interim system, though not significantly, indicating that reducing the threat of retaliation made reviews more informative and truthful.

Klein et al. [99] analyze strategic timing in feedback provision using data from eBay's old two-sided reputation system. They argue that reviewers who intend to leave positive feedback have an incentive to post early in order to maximize the likelihood of receiving positive reciprocal feedback. By contrast, reviewers who intend to leave negative feedback have an incentive to delay posting as long as possible to reduce the risk of retaliation. In some cases, dissatisfied users may avoid leaving feedback altogether out of fear that a negative review will trigger a retaliatory response. Consistent with these predictions, they find that positive feedback is provided earlier than neutral or negative feedback and that the likelihood of negative reviews increases toward the end of the feedback window. Bolton et al. [19] similarly observe that in cases of mutual negative feedback, the second review tends to be posted shortly after the first, indicating strategic timing in response to negative ratings.

The strategic behavior of "feedback sniping" is investigated by Klein et al. [18]. Feedback sniping refers to waiting until the last possible moment before the feedback window closes, leaving the transaction partner no time to retaliate. Discussion forums suggest that users were aware of and discussed this strategy. However, Klein et al. [18] show that under eBay's old reputation system, feedback sniping was not feasible in practice because the second reviewer always had enough time left to retaliate. Thus, reviewers had to anticipate the possibility of retaliation regardless of when they posted their feedback.

As noted earlier, in eBay's old system, feedback removal was possible only if both parties agreed to withdraw their reviews or after a court ruling. Bolton et al. [19] find that retaliatory negative feedback increased the likelihood of such a mutual withdrawal. Klein et al. [99] also find evidence that users employed feedback withdrawal strategically. A reviewer might retaliate with a negative review to strengthen their bargaining position during negotiations over mutual withdrawal. Although only 0.1% of reviews were withdrawn, one quarter of withdrawals occurred within two days after the second review was posted, indicating that feedback withdrawal was used as a bargaining tool. This possibility of renegotiation after feedback was posted reduced the reliability of the system.

The feedback withdrawal mechanism in eBay's interim reputation system is further analyzed by Bolton et al. [101]. They argue that the option to withdraw feedback can incentivize parties to leave negative feedback strategically in order to gain leverage in a potential dispute resolution phase. Using field data from eBay, they report that roughly 97% of feedback is positive, but negative feedback is substantially more likely to be withdrawn than positive or neutral feedback (16% of negative buyer feedback and 13% of negative seller feedback are withdrawn). Withdrawal procedures

are frequently initiated by the party that posted the second review when that review was a negative or neutral response to a similar review. In 39% of cases where negative feedback was retaliated with negative feedback, the feedback was subsequently challenged. However, they do not find that retaliatory negative feedback increases the likelihood of securing an agreement to withdraw ratings. They hypothesize that this may be because an initial truthful negative reviewer becomes more motivated to punish the other party once their truthful review is responded to with a retaliatory negative response.

Bolton et al. [101] complement the field evidence with a laboratory experiment in which buyers and sellers decide whether to transact, sellers choose quality and both parties submit positive or negative feedback. After feedback is revealed, parties may revise price or quality to benefit the counterpart. In treatments that allow feedback withdrawal, feedback can be removed only if both parties agree. The authors hypothesize that the withdrawal option weakens incentives to provide truthful feedback and may escalate disputes rather than resolve them. Their findings mirror the field data. Retaliatory negative feedback does not increase bargaining power in securing withdrawal. Instead, concessions, such as improving price or quality, are more effective at inducing the removal of negative reviews when no retaliatory feedback is given.

As emphasized by Dellarocas and Wood [97], these strategic dynamics can deter buyers from leaving justified negative feedback, especially if they believe the seller may respond with a negative review. If the seller refrains from leaving feedback first, the buyer may prefer not to post anything at all. Such behavior reduces the informativeness of the reputation system. When market participants can retaliate for negative ratings, reviewers may systematically avoid providing negative feedback, leading to an upward bias in observed ratings. This bias reduces market efficiency by making it more difficult for users to infer which sellers are trustworthy, ultimately undermining the value of the reputation system.

The role of reciprocity in review generation has been studied using eBay data [17–19, 96, 97, 99, 101, 102], using Airbnb data [94] and conducting laboratory experiments [19, 101].

Two approaches to address reciprocity have been suggested, for instance, by Bolton et al. [19]. The first is to make feedback visible to the transaction partner only after both parties have submitted their review or once the review window has closed. This design is commonly referred to as a "simultaneous reveal" system [11, p. 81] and constitutes a double-blind procedure. Addressing reciprocity requires this structural decoupling of the information flow. The second approach is to replace a two-sided reputation system with a one-sided one in which only buyers can post reviews. This eliminates the possibility of sellers retaliating through negative feedback. eBay implemented precisely this change when transitioning from its old to its new reputation system.

Bolton et al. [19] examine the effects of both approaches in a laboratory auction setting designed to resemble eBay's reputation mechanism. The baseline treatment mirrors the original eBay system, and each of the two proposed adjustments is implemented in a separate treatment. They report that the simultaneous-reveal mechanism reduces the correlation between buyer and seller feedback, although the correlation remains significantly above zero. Both treatments lead to more informative feedback

being displayed to buyers relative to the baseline. Simultaneous reveal also lowers review rates, rating levels and the correlation between ratings.

Hui et al. [102] analyze the consequences of eBay's shift from the intermediate to the new reputation system, which eliminated sellers' ability to post negative reviews about buyers and thereby aimed to remove the threat of retaliation. They examine how this change affects the quality supplied by sellers. Prior to the policy change, sellers retaliated with a negative review in more than one-third of cases in which they received a negative review from a buyer. Following the change, they document a 50% decline in negative reviews and a reduction in disputes initiated by buyers, which they interpret as evidence of improved buyer experience. Although the objective of the reform was to facilitate more honest buyer reviews, negative feedback became less common. Sellers now had to exert more effort to earn positive reviews in the absence of retaliation opportunities. They further observe reductions in moral hazard and adverse selection and that low-quality sellers were more likely to exit the market. eBay's detailed seller ratings introduced in May 2007 were anonymous to sellers, removing the possibility of targeted retaliation. Mutual reviews formally remained possible until May 2008, when eBay eliminated them, leaving a one-sided system based solely on detailed seller ratings. Although retaliation persisted during the transition period, buyers submitted fewer negative reviews after the removal of two-sided reviews in 2008.

Klein et al. [17] study the same policy shift from the intermediate to the new eBay system, focusing particularly on the elimination of sellers' ability to retaliate. They argue that the reform reduced buyers' costs of providing negative feedback by removing retaliation risk, and they document increased market transparency consistent with more efficient market outcomes. They find that the 2008 change led to a significant rise in buyers' reviews and interpret this as evidence supporting the hypothesis that disabling negative feedback from sellers and removing retaliatory possibilities increases review activity. They do not, however, identify a significant increase in seller exit rates following the reform.

Fradkin et al. [94] investigate a similar question using data from a field experiment on Airbnb to examine how the timing of review visibility in two-sided reputation systems affects reciprocity. Airbnb operates a two-sided system in which guests and hosts review each other. The authors use a treatment closely resembling the double-blind design evaluated by Bolton et al. [19], applied to Airbnb's pre-2014 system in which reviews were posted immediately, allowing a second reviewer to retaliate. They report that simultaneous reveal reduces the time taken to post a review by 17% for guests and by 9.9% for hosts and increases review rates by 1.7% and 9.8%, respectively. They attribute these changes to a "desire to unveil reviews" [94, p. 1020], driven by curiosity or a desire to make reviews visible to future partners sooner. They further document a 35% reduction in the time between the first and second review, slightly more negative reviews in the treatment and a 48% decline in the correlation between guest and host ratings. They find no effect of the treatment on demand and do not identify evidence of improved matching, although they note that the limited duration of the experiment may explain this. They also do not find reduced demand for low-quality sellers. Evidence from Bolton et al. [19], Fradkin et al. [94] and Klein et al. [18] suggests that reviews should be made visible only

once no additional reviews can be submitted, thereby eliminating opportunities for reciprocal behavior.

Luca [11] argues that even under simultaneous reveal, strategic considerations may still discourage buyers from posting negative reviews if doing so might reduce their attractiveness as future trading partners. He therefore advocates for greater anonymity, either by requiring anonymous reviews or by aggregating reviews. As an alternative, he suggests limiting visibility to the platform, which could then incorporate the feedback into its algorithms without exposing reviewers to strategic consequences.

Reciprocity distorts the distribution of posted reviews through two interconnected mechanisms. The threat of retaliation leads to an underprovision of negative reviews, while the opportunity to reward leads to an overprovision of positive ones. Both induce a systematic upward bias in the mean ($\mathbb{E}[\varepsilon_{it}] > 0$). Effective mitigation requires structural changes like simultaneous reveal or one-sided reviewing.

The evidence on reciprocity in reputation systems can be summarized as follows:

Finding 6 *Reciprocity in two-sided reputation systems induces a systematic upward bias in the mean error term ($\mathbb{E}[\varepsilon_{it}] > 0$) driven by fear of retaliation or desire for reward.*

Finding 7 *Contemporary platform designs (e.g., Airbnb, modern eBay) have largely mitigated this bias through structural interventions such as simultaneous reveal or one-sided reviewing. The scarcity of recent empirical studies on active reciprocity suggests that these design features have been successful in reducing the bias, demonstrating that structural changes can be effective in reducing bias.*

3.3 Social influence bias: serial correlation of reviews

Social influence bias arises from serial correlation of reviews when reviewers are influenced by previously posted reviews. This phenomenon, commonly referred to as social influence bias, occurs when a reviewer's intended rating is revised after being exposed to earlier reviews.[10]

Aral [105], Askalidis et al. [106] and Muchnik et al. [107] describe this process in detail. A reviewer may form an initial private evaluation of a product or service, but upon seeing existing reviews, may adjust the final rating to align more closely with what others have posted. A reviewer with a negative private opinion might adjust the rating upward after observing positive reviews, while a positive opinion might be downgraded. When such adjustments occur systematically, later reviews correlate with earlier ones, generating serial dependence in the data. This pattern reflects a broader tendency toward herding or conformity.

Within the conceptual framework, social influence bias represents a violation of the Independence Condition. Because a reviewer's rating becomes dependent on previously posted ratings, the error terms exhibit serial correlation ($\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$ for some $t \neq s$). This error signature (serial correlation) has a critical implication

[10] In the literature, this phenomenon is also referred to as *sequential bias* for example, by Eryarsoy and Piramuthu [103] and Sikora and Chauhan [104].

derived from Proposition 2 (Independence): unlike uncorrelated noise, this bias does not average out as the number of reviews (T) increases. Instead, the serial correlation reduces the effective sample size, meaning that adding more reviews yields diminishing marginal returns in precision and can even amplify initial shocks through herding.

Interestingly, Rohde et al. [108] find in Google Maps reviews that reviewers' effort when writing a review decreases with the number of existing reviews, but it is not affected by the difference between their own review and existing reviews.

Table 4 summarizes where certain aspects of this bias have been analyzed. The table differentiates between empirical studies that rely on field data or field experiments and those that rely on laboratory or online experiments and surveys. The remainder of this section summarizes key empirical findings. In addition, Magnani [12] has surveyed selected aspects of the social influence bias with a focus on earlier studies.

The social influence bias has been empirically investigated using data from online marketplaces, review websites and experimental data. Askalidis et al. [106], Han and Anderson [109], Jacobsen [110], Moe and Trusov [112], Sikora and Chauhan [104] and Wang et al. [113] use data from online retailers or review websites to analyze the social influence bias. Han and Anderson [109] use Tripadvisor data and show that prominently displayed prior reviews have a stronger influence. Correlation reduces quickly with display order and becomes negligible by the third visible review. Jacobsen [110] finds that customer reviews correlate with preceding expert ratings, suggesting imitation or anchoring effects. Karaman [111] concludes from hotel review data that social influence can lead to more representative reviews due to conformity of reviewers.

Furthermore, experimental evidence is used to investigate the bias. Muchnik et al. [107] conduct a field experiment on a social news platform in which the first rating is artificially manipulated. A positive initial rating increases the probability of subsequent positive ratings by 32%. Negative initial ratings induce more negative ratings, but this effect is partially offset by an increase in positive ratings. The authors also show that susceptibility to influence varies by content type, a result that is plausibly generalizable to product reviews. Additional experiments and surveys (e.g., Eryarsoy and Piramuthu [103], Krishnan et al. [114] and Schlosser [115]) detect serial correlation, particularly for negative reviews.

The presence of social influence bias threatens the reliability of reputation systems by distorting the distribution of posted ratings, potentially leading consumers toward

Table 4 Overview of empirical studies on the social influence bias

Focus of analysis	Studies
Studies on the social influence bias	**F:** Askalidis et al. [106], Han and Anderson [109], Jacobsen [110], Karaman [111], Moe and Trusov [112], Rohde et al. [108], Sikora and Chauhan [104] and Wang et al. [113]
	L: Eryarsoy and Piramuthu [103], Krishnan et al. [114], Muchnik et al. [107] and Schlosser [115]
Ways to reduce the social influence bias	**F:** Aral [105], Askalidis et al. [106], Gao et al. [116], Han and Anderson [109] and Jacobsen [110]

F: Studies based on field data or field experiments; L: Studies based on laboratory or online experiments and surveys

suboptimal purchase decisions. Several mitigation strategies have been proposed. A common approach is to hide previous reviews during the review-writing process. Variants include temporary hiding (e.g., Reddit hides ratings for a short time to prevent rating cascades [105]) or complete hiding until submission, as recommended by Askalidis et al. [106], Han and Anderson [109] and Jacobsen [110]. This mitigation strategy directly addresses the mechanism identified in Proposition 2 (Independence). By hiding prior reviews, the platform structurally decouples the current error term ε_{it} from previous ones $\varepsilon_{i,t-1}$, thereby restoring independence. This confirms the proposition's prediction that biases rooted in serial correlation require structural interventions restoring independence rather than simple aggregation, which would be ineffective against correlated errors. Empirical studies show that hiding previous reviews, such as through email-invited reviews submitted on dedicated landing pages, significantly reduces the influence of earlier ratings. Askalidis et al. [106] demonstrate that invited reviews, in which reviewers cannot see prior reviews, are more positive on average. Because reputation systems often combine invited and organic reviews, such interventions reduce but do not eliminate social influence bias. Gao et al. [116] similarly find using Tripadvisor data that solicitation programs increase review volume but reduce organic reviews; solicited reviews tend to have higher ratings.

A key limitation of these mitigation strategies is that earlier reviews may still influence reviewers indirectly if they recall them from the purchase stage. The effectiveness of hiding prior reviews thus depends on the time elapsed between purchase and review-writing.

Social influence bias systematically affects the reputation signal. While it can induce a mean shift ($\mathbb{E}[\varepsilon_{it}] \neq 0$), its defining characteristic is the introduction of serial correlation ($\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$ for some $t \neq s$), violating the Independence Condition. This confirms the insight from Proposition 2 (Independence): because the error terms are correlated, increasing the volume of reviews does not restore accuracy. The effective sample size remains lower than the observed count. The direction of the mean shift is ambiguous and context-dependent, but the presence of correlation is systematic. Therefore, effective mitigation must prioritize structural independence (e.g., blinded reviews) over participation incentives, as the latter cannot correct for the redundancy inherent in herding behavior.

The findings from the literature on the social influence bias can be summarized as follows:

Finding 8 *Social influence bias represents a systematic violation of the Independence Condition, where review visibility induces serial correlation in the error term (*$\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$ *for some* $t \neq s$*). Empirical evidence confirms that this correlation causes ratings to converge toward prior reviews (herding), reducing the effective sample size and preventing errors from averaging out with increased sample size. This validates the framework's assertion that biases rooted in serial correlation cannot be mitigated by simply increasing review volume, as additional data points are often redundant rather than independent.*

Finding 9 *As suggested by Proposition 2 (Independence), mitigation strategies that structurally restore independence, such as hiding prior reviews during submission (blinded reviews) or using simultaneous reveal mechanisms, substantially reduce social influence bias. The effectiveness of these mitigation strategies depends on the completeness of the information barrier, highlighting the challenge of achieving true independence in practice, and on whether blinding reviews affects participation.*

3.4 Selection bias: endogenous review provision

Selection bias arises when the set of posted reviews is not representative of the underlying population of buyers. This *selection bias* occurs when review provision is endogenous, as only some buyers choose to leave a review, and their decision correlates with their underlying transaction experiences. For example, satisfied buyers may be more inclined to submit reviews, whereas dissatisfied buyers may remain silent or direct their complaints to customer service rather than posting a public review. As a result, reviews may not reflect the distribution of experiences in the broader customer population, and prospective consumers may base purchasing decisions on incomplete or distorted information. Empirical evidence shows that most online ratings are highly positive, with few negative and almost no moderate evaluations [see 4, 97, 106]. However, Karaman [111] finds using hotel reviews that very unsatisfied customers post reviews more often than very satisfied ones. Moreover, as noted by Li and Hitt [117], early adopters may differ systematically from later customers, further complicating inference.

Within the conceptual framework, selection bias represents a structural violation of the Mean Condition. Because the sample of reviewers is non-representative, the expected value of the error term deviates from zero ($\mathbb{E}[\varepsilon_{it}] \neq 0$), not because individual reviews are manipulated, but because reputation is based on a skewed subset of data. It is an endogeneity problem where the probability of posting a review is correlated with the error term itself. Correcting this requires shifting the composition of the sample rather than just filtering existing data.

Selection bias, also referred to as *reporting bias* [97], captures the divergence between publicly posted reviews and the private distribution of realized transaction outcomes. Two mechanisms generate this bias. The first is *acquisition bias*: reviews are written only by consumers who purchased the product, and those purchasers may already hold more favorable prior beliefs than non-purchasers. The second is *underreporting bias*: consumers with extreme experiences are more likely to post reviews

Table 5 Overview of empirical studies on the selection bias

Focus of analysis	Studies
Studies on the selection bias	**F:** Dellarocas and Wood [97], Fradkin and Holtz [90], Gao et al. [119], Hu et al. [118], Karaman [111] and Li and Hitt [117]
Ways to reduce the selection bias	**F:** Askalidis et al. [106], Fradkin and Holtz [90], Gao et al. [116], Han and Anderson [120], Han and Mikhailova [121], Hu et al. [118], Karaman [111] and Wang and Anderson [122]

F: Studies based on field data or field experiments; L: Studies based on laboratory or online experiments and surveys

than those with moderate opinions. Detailed analyses of these mechanisms are provided in Gao et al. [119], Han and Anderson [120] and Hu et al. [32, 118].

Table 5 summarizes empirical work that examines the existence and consequences of selection bias and evaluates approaches for mitigating it. The table differentiates between empirical studies that rely on field data or field experiments and those that rely on laboratory or online experiments and surveys. Magnani [12] has surveyed selected aspects of selection bias with a focus on earlier studies.

This bias has been investigated using data from eBay [97], Amazon [117, 118], online retailers [106] and physician review data [119]. Additionally, studies have analyzed review data from the hospitality industry, specifically Airbnb [90], and hotels [111, 116, 120–122], for example, from hotel review platform Tripadvisor.

Fradkin and Holtz [90] use Airbnb data to study how incentivized review solicitation affects selection. In a field experiment, the treatment group received coupons for reviewing properties with no existing reviews, while the control group did not. The incentive increased review volume, but the resulting reviews were more negative on average than organically posted reviews. The authors find no impact on sales, likely because the increase in review quantity was offset by the more negative ratings. They conclude that incentivized reviews convey less information about actual transaction quality.

Selection bias can reduce the informativeness of reputation systems for both buyers and sellers. When reviews are disproportionately positive or negative, the resulting bimodal distribution undermines the usefulness of average ratings as summary statistics [32, 118]. These studies recommend providing alternative summary measures to help consumers form more accurate expectations.

Several design interventions aim to reduce selection bias by increasing review coverage. Askalidis et al. [106], Fradkin and Holtz [90] and Luca [11] study email invitations that encourage customers who have not yet posted reviews to contribute. While such invitations can increase review volume and attract otherwise underrepresented reviewers, Fradkin and Holtz [90] find that they may reduce transaction quality on Airbnb. Gao et al. [116] similarly show that solicitation programs increase total review counts but reduce the share of organic reviews which tend to be more positive.

Han and Anderson [120] find that selection bias diminishes as reviewers become more familiar with the mechanics of the review-posting process. Wang and Anderson [122], using hotel review platforms, demonstrate that platform design, particularly the effort required to post a review, influences the degree of selection bias. Certain design features lead to shorter and more negative reviews on transaction-based websites such as Expedia, while community-driven sites like Tripadvisor exhibit different patterns.

Karaman [111] concludes from hotel review data that inviting customers to review leads to more moderate postings, thereby reducing rating extremity and improving representativeness.

Li and Hitt [117] show that sellers may strategically leverage selection bias by encouraging reviews from customers who are likely to be positively predisposed, thereby attracting additional buyers. Finally, Han and Mikhailova [121] demonstrate how propensity-score adjustments can be used to correct for selection bias and improve the reliability of posted reputation measures.

Selection bias arises when endogenous review provision leads to a non-representative sample, systematically violating the Mean Condition ($\mathbb{E}[\varepsilon_{it}] \neq 0$). Unlike random noise, this bias cannot be averaged out by simply collecting more data from the same self-selected population. The underlying distribution remains skewed. Mitigation strategies such as solicitation programs aim to correct this by altering sample composition. However, as highlighted by Proposition 1 (Trade-offs), these interventions involve inherent trade-offs: increasing participation to improve representativeness may simultaneously reduce review precision (increasing variance) or induce strategic behavior (introducing new mean shifts). Therefore, effective design must carefully consider incentives to maximize representativeness without triggering secondary distortions.

The findings on selection bias can be summarized as follows:

Finding 10 *Selection bias arises from the endogenous nature of review provision, where the subset of participating reviewers is non-representative of the overall customer population. This structural imbalance violates the Mean Condition ($\mathbb{E}[\varepsilon_{it}] \neq 0$), causing observed reputation to diverge from the true distribution of transaction experiences. Empirical evidence confirms that this bias persists regardless of the total number of reviews, as increasing volume from a self-selected sample does not automatically correct the underlying non-representativeness in the error term's expected value.*

Finding 11 *Interventions designed to mitigate selection bias, such as solicitation programs or financial incentives, illustrate the trade-offs predicted by Proposition 1 (Trade-offs). While these measures successfully increase participation and improve sample representativeness (correcting the mean shift), they often inadvertently introduce secondary distortions: incentives may induce reciprocal behavior (shifting the mean again) or encourage low-effort reviews (potentially increasing variance). Consequently, the net effect on reputation accuracy is context-dependent, demonstrating that correcting one violation often requires balancing competing influences rather than applying a universal solution.*

3.5 Bias due to noise

Biases can arise from noise if the noise is systematic such that the error term satisfies $\mathbb{E}[\varepsilon_{it}] \neq 0$. Buyers generate noisy reviews when they misunderstand what exactly they are asked to evaluate. For instance, Belleflamme and Peitz [20] and Tadelis [2] note that consumers often confuse product satisfaction with their experience of delivery or interaction with the seller.

Within the conceptual framework, noise primarily represents a violation of the Variance Condition. Random noise maintains an expected value of zero but inflates the variance ($\mathrm{Var}[\varepsilon_{it}]$). This reduces the precision of the reputation signal, making it a less efficient estimator of true quality. However, if the noise is systematic (e.g., consistent confusion about delivery vs. product or a shock uniformly affecting all reviews), it can also violate the Mean Condition. This distinction is critical for detect-

Table 6 Overview of empirical studies on noise

Studies
F: Brandes and Dover [123] and Luca and Reshef [58]
L: Greiff and Paetzel [124]

F: Studies based on field data or field experiments; L: Studies based on laboratory or online experiments and surveys

ing the distortions and selecting mitigation strategies, as addressed in Proposition 3 (Detection).

The empirical literature has investigated noise in reputation systems. Table 6 summarizes where different aspects of this bias have been studied. The table differentiates between empirical studies that rely on field data or field experiments and those that rely on laboratory or online experiments and surveys. The remainder of this section discusses this literature in more detail.

A first source of noise stems from buyers' idiosyncratic tastes. Reviews may reflect horizontal product characteristics, such as color preferences, rather than vertical characteristics related to objective quality. As Belleflamme and Peitz [20, p. 54] put it, "a reviewer may give a negative product rating because they do not like the color of the product," even though such subjective attributes are not informative for all (potential) buyers.

A second source of noise arises from shocks outside the seller's control. These shocks may lead a reviewer to provide a rating that is unrelated to the seller's or product's true quality. Examples include delays caused by transport companies or a reviewer being in a bad mood while writing the review. Belleflamme and Peitz [20] note that the influence of such (random) shocks tends to diminish when many reviews are available. Brandes and Dover [123] provide an empirical investigation of such shocks. They show that hotel ratings are correlated with weather at the reviewer's home address at the time the review is written. Rain increases the likelihood of review provision and decreases the rating even though the weather is unrelated to the completed hotel stay. Hence, systematic noise (provision of lower ratings when it rains) may even appear together with selection bias (more ratings of people who tend to provide lower ratings; see Sect. 3.4 for selection bias).

Price variation introduces another form of noise. Buyers often condition their ratings on the price they paid. Thus, the same product may be reviewed differently depending on the transaction price, even if the underlying quality is identical. Luca and Reshef [58] analyze Yelp data and find that a 1% increase in prices reduces ratings by 2.5% to 5%. Ratings for cheap and expensive restaurants exhibit similar distributions once price differences are accounted for, indicating that consumers implicitly adjust for perceived value. The authors caution that consumers interpreting ratings must be aware of historical price effects. They also note that firms may exploit this mechanism by offering low introductory prices to accumulate favorable reviews before raising prices.

Laboratory evidence further illustrates how users process noisy information. Greiff and Paetzel [124] conduct an experiment with a public good game where subjects rate their partner's contribution. Since these ratings reflect subjective assessments of contributions, they inherently include noise. The authors vary the granularity with

which participants see reputation information. Some treatments display only the most recent rating, others display ratings from the last three periods and some also present the average of these three ratings. They find that average ratings lead subjects to rely more strongly on their partner's reputation and to contribute significantly more (by about 50%). Importantly, these effects occur only when the platform explicitly displays the average. Simply providing the information needed to compute the average does not generate the same effects. Their findings indicate that aggregation, such as displaying average ratings, can mitigate noise and improve user decision-making.

This finding provides empirical validation for Proposition 3 (Detection). The proposition posits that variance-inflation biases require context-aware aggregation rather than outlier detection. By explicitly displaying the average, the platform leverages the Law of Large Numbers to cancel out the random error terms (ε_{it}), effectively restoring precision. Conversely, applying outlier detection algorithms to this type of high-variance data with the undistorted mean error term would be counterproductive, potentially removing valid signals and reducing the effective sample size.

To address noise arising from attribute confusion, many platforms have implemented structural separations of different review dimensions. A prominent example is Amazon's separation of product reviews from seller reviews. By forcing users to evaluate product quality and logistics or service in separate fields, the platform prevents errors in one domain (e.g., a delayed shipment caused by a third-party carrier) from impacting the signal for the other (e.g., the intrinsic quality of the product). This design intervention directly targets the Mean Condition: it ensures that $\mathbb{E}[\varepsilon_{it}]$ for product quality remains zero even when logistics fail, effectively filtering out systematic bias at the source. Similarly, hospitality platforms like Booking.com often separate scores for cleanliness, location and staff from the overall evaluation. This granularity allows users to isolate vertical quality signals from horizontal preferences or contextual shocks, reducing the variance attributable to mismatched expectations. These design choices illustrate that mitigating noise can be achieved by altering the data collection mechanism.

Beyond simple arithmetic means, modern platform designs employ more sophisticated aggregation methods to manage noise. Displaying the full distribution histogram of ratings alongside the average allows users to visually assess the variance ($\mathrm{Var}[\varepsilon_{it}]$). A bimodal distribution (e.g., many 5-star and 1-star reviews) signals high variance and potential polarization or systematic confusion, warning the user that the average may be a poor predictor of their individual experience. This could be further refined by using weighted aggregation, where reviews from users with established expertise or long tenure are given more weight. While this introduces complexity, it effectively reduces the noise contribution from inexperienced or erratic reviewers. Additionally, the rise of contextual filtering allowing users to sort reviews by specific use cases (e.g., "traveling with family" or "business trip") enables a form of user-driven variance reduction. By self-selecting into a relevant subgroup, the consumer effectively reduces the error variance relative to their specific utility function. These features collectively validate Proposition 3 (Detection): they represent context-aware tools designed specifically to handle high-variance environments, distinct from the outlier-detection tools used for mean-shift biases.

Noise presents a dual challenge. When random, it violates the Variance Condition ($\mathrm{Var}[\varepsilon_{it}]$ is high), reducing signal precision without introducing systematic bias ($\mathbb{E}[\varepsilon_{it}] = 0$). In this case, Proposition 3 (Detection) suggests that aggregation is a suitable remedy. However, when noise is systematic (e.g., due to persistent misunderstandings or price-value confusion), it also violates the Mean Condition, requiring targeted interventions such as clarifying evaluation criteria. Distinguishing between these two forms is essential: treating systematic noise as mere variance leaves the bias uncorrected, while treating random noise as systematic leads to unnecessary information loss.

The findings on biases due to noise in reputation systems can be summarized as follows:

Finding 12 *Noise in reputation systems appears in two distinct forms with different signatures: random noise, which violates the Variance Condition (*$\mathrm{Var}[\varepsilon_{it}]$ *is high) while leaving the mean unbiased (*$\mathbb{E}[\varepsilon_{it}] = 0$*), and systematic noise, which also violates the Mean Condition. Empirical evidence confirms that random noise arises from idiosyncratic tastes, contextual shocks or misunderstandings, reducing the precision of the reputation signal without necessarily introducing a directional mean shift. This distinction is critical, as it dictates that random noise can be mitigated through aggregation, whereas systematic noise requires targeted structural interventions.*

Finding 13 *The effectiveness of mitigation strategies depends on correctly identifying the noise type, validating Proposition 3 (Detection). For random noise, empirical studies show that explicit aggregation (e.g., displaying average ratings) significantly improves decision-making by leveraging the Law of Large Numbers to cancel out error terms. Conversely, applying outlier detection algorithms to high-variance but unbiased data risks removing legitimate signals, thereby reducing sample size and increasing error. For systematic noise (e.g., consistent confusion over delivery vs. product quality), clarification of evaluation criteria is required. This asymmetry underscores that a single tool cannot address all forms of variance inflation. Platform design must match the specific properties of the distortion.*

4 Structural synthesis and design implications

Section 3 documented empirical findings on five major sources of bias in isolation, summarized in Table 7. This section integrates those findings to analyze how biases interact and how interventions targeting one moment of the error term (mean, variance or correlation) can create distortions in others. By synthesizing the thirteen key findings from the literature, this section demonstrates that platform design is fundamentally a multi-objective optimization problem. The empirical evidence collectively motivates the four propositions introduced in Sect. 2.4, revealing that biases are not isolated but interconnected constraints on reputation system design.

A primary insight emerging from the synthesis is the ubiquity of design trade-offs, which directly motivates Proposition 1 (Trade-offs). The literature shows that interventions designed to correct violations of the Mean Condition ($\mathbb{E}[\varepsilon_{it}] = 0$) frequently

Table 7 Summary of biases in reputation systems

Source and type of bias	Main mechanism or cause	Typical effect on reputation system	Mitigation and design interventions
(i) Strategic action of sellers: fake reviews (Sect. 3.1.1)	Sellers post fake reviews to improve their own reputation (positive fake reviews; common) or harm their competitors' reputation (negative fake reviews; less common)	Inflated ratings, reduced reliability, decreased consumer trust; positive fake reviews: $\mathbb{E}[\varepsilon_{it}] > 0$; negative fake reviews: $\mathbb{E}[\varepsilon_{it}] < 0$	Verified-purchase reviews, detection algorithms, transparency, rating aggregation
(i) Strategic action of sellers: rebate-for-review programs (Sect. 3.1.2)	Buyers receive unconditional rebate when posting a review	Helps to increase number of reviews; effect on $\mathbb{E}[\varepsilon_{it}]$ is ambiguous	Enforcement by a neutral party
(ii) Strategic action of reviewers: reciprocity (Sect. 3.2)	Reviewers adjust ratings based on fear of retaliation; $\mathbb{E}[\varepsilon_{it}]$ biased upward or downward, depending on whether fear of retaliation or expectation of reward dominates	Inflated ratings in two-sided platforms: $\mathbb{E}[\varepsilon_{it}] > 0$	One-sided or simultaneous review policies, anonymity
(iii) Social influence bias (Sect. 3.3)	Reviewers are influenced by previously posted reviews	Serial correlation of reviews $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$ for some $t \neq s$, potentially causing $\mathbb{E}[\varepsilon_{it}] \neq 0$	Hiding prior reviews during posting, delayed reveal
(iv) Selection bias (Sect. 3.4)	A non-representative subset of buyers posts reviews (e.g., extreme experiences or pre-disposition)	Average ratings may not reflect true distribution of experiences; $\mathbb{E}[\varepsilon_{it}] \neq 0$ because of endogenous review provision	Solicitation programs, targeted email invitations, platform design adjustments
(v) Noise (Sect. 3.5)	Idiosyncratic preferences, misinterpretation, contextual shocks, price variation	Random noise: $\mathbb{E}[\varepsilon_{it}] = 0$, but $\mathrm{Var}[\varepsilon_{it}]$ may be affected; systematic noise: $\mathbb{E}[\varepsilon_{it}] \neq 0$	Clarify evaluation criteria, aggregation of ratings

exacerbate violations of the Variance Condition or introduce new structural biases. Findings on fake reviews confirm that while mitigation strategies like verification requirements effectively reduce mean-shift bias (Finding 1), they simultaneously reduce the sample size (T), thereby affecting the variance of the reputation signal and potentially worsening selection bias (Finding 3). A similar pattern appears in rebate-for-review programs: while they can correct selection bias by improving representativeness (Finding 4), they often induce reciprocal behavior in response to the rebate received, shifting the mean upward regardless of quality (Finding 5). Likewise, interventions for selection bias illustrate that incentives can improve participation but risk introducing noise or strategic distortion (Findings 10 and 11). Even mitigation strategies for social influence involve trade-offs: although blinding reviews restores independence, they may alter participation rates (Finding 9). This recurring pattern across fake reviews, rebates, selection and social influence validates Proposition 1. Optimal platform design cannot minimize a single distortion in isolation but must balance competing objectives.

A second critical synthesis concerns the serial correlation of the error term, which motivates Proposition 2 (Independence). While biases like noise and selection can theoretically be mitigated by increasing the volume of reviews (thereby averaging out errors or diluting skewed samples), social influence operates differently (Finding 8). Findings on social influence establish that visibility induces serial correlation ($\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$ for some $t \neq s$), reducing the effective sample size. Consequently, volume-based solutions are ineffective for these biases. The limitations of blinding mechanisms further support Proposition 2. Biases rooted in correlation require structural interventions (e.g., hiding reviews during the posting process) rather than aggregation (Finding 9). This creates a design tension: platforms must maximize participation to fix selection bias and noise while simultaneously restricting information flow to fix social influence.

The synthesis further highlights the necessity of asymmetric detection strategies, motivating Proposition 3 (Detection). The literature distinguishes clearly between noise, which primarily inflates variance (Finding 12), and fake reviews, which shift the mean (Finding 1). Applying outlier detection to high-variance but unbiased data (noise) risks removing legitimate signals, whereas relying on aggregation for social influence or selection bias fails to correct the systematic shift (Finding 13). This asymmetry confirms that tools effective for mean-shift biases are inappropriate for variance-driven noise and vice versa. Misdiagnosing the signature of the error term leads to suboptimal outcomes, such as the removal of genuine extreme opinions or the persistence of systematic fraud. Thus, the findings collectively argue for a diagnostic approach where design tools are matched to the specific error signature.

Finally, integrating recent developments in generative artificial intelligence underscores the dynamic nature of these challenges, motivating Proposition 4 (Artificial Intelligence). The evidence that artificial intelligence disproportionately lowers the cost of creating mean-shift bias (Finding 2) implies that the static equilibrium of bias levels is no longer stable. Interventions that were previously sufficient may become obsolete as the cost function of manipulation shifts. This dynamic perspective reframes bias not as a fixed problem but as a moving target requiring continuous adaptation. Furthermore, the cold-start dynamics addressed by rebate-for-review

programs (Findings 4 and 5) and the changing participation rates in relation to mitigations of selection bias (Finding 11) suggest that all biases are subject to evolving economic incentives. Platform governance must evolve from static rule sets to dynamic and adaptive systems that co-evolve with manipulation capabilities.

Artificial intelligence introduces a co-evolutionary dynamic beyond merely lowering manipulation costs. While generative artificial intelligence makes violations of the Mean Condition easier by enabling high-quality fake reviews at low cost, it simultaneously offers advanced tools for detecting distortions. Platforms must deploy adaptive, artificial intelligence-driven detection systems that evolve alongside manipulation tactics. Beyond user-generated distortions, algorithmic curation and ranking systems act as a critical mediating layer that can amplify or mitigate these biases. While this paper focuses on the sources of bias in review generation, the algorithms that determine which reviews are visible fundamentally interact with the moments of the error term. For instance, ranking systems that prioritize extreme or emotional content to maximize engagement can artificially inflate the variance of the observed reputation signal, making reputation appear more polarized than it truly is. Conversely, filtering mechanisms that systematically suppress negative feedback to protect platform revenue can induce a mean shift, reinforcing user-generated biases. Thus, the net bias observed by a consumer is the cumulative result of user-generated distortions and algorithmic amplification, necessitating that platform design accounts for both the creation and the curation of reputation signals.

Some biases have lost much of their practical relevance for platforms and policy discussions, as the mitigation strategies applied by major platforms have proven to be highly effective. Reciprocal reviewing on two-sided reputation systems, such as those of eBay and Airbnb, has been largely mitigated through structural interventions like simultaneous reveal of reviews or transition to one-sided reputation systems (Findings 6 and 7).

Taken together, the documented biases operate within a common conceptual structure defined by the moments of the error term ε_{it}. Crucially, interventions targeting one dimension often propagate effects to others. Interpreting the evidence from this perspective clarifies that heterogeneous empirical findings are not contradictory but reflect different points on a multi-dimensional design frontier. Table 8 summarizes these structural trade-offs, mapping common interventions to their primary benefits and secondary costs. Effective platform governance therefore requires a holistic approach that explicitly manages the trade-offs between accuracy, precision and independence, rather than optimizing for any single metric in isolation.

This synthesis reveals that biases are not isolated defects but endogenous outcomes of platform architecture. Consequently, effective governance requires shifting from ad-hoc fixes to a systematic design framework grounded in the three conditions of the error term. Drawing on the error signatures of the biases in Table 7 and the interaction effects in Table 8, a three-step protocol for platform operators and regulators is proposed.

First, managers and policymakers must identify which moment of the error term is distorted before selecting a mitigation strategy, as misdiagnosis leads to ineffective or harmful interventions. If ratings are systematically inflated despite poor service (a Mean Condition violation, $\mathbb{E}[\varepsilon_{it}] \neq 0$), the issue is likely strategic manipulation (fake

Table 8 Interaction matrix: interventions, primary effects and trade-offs (illustration of the implications of Proposition 1)

Intervention	Primary target and effect	Secondary trade-off	Net implication
Verified purchase requirements	Targets Mean Condition: reduces fake reviews by raising manipulation costs	Increases variance: reduces sample size (T) and may induce selection bias by excluding legitimate non-purchasers	Improves accuracy but reduces precision; optimal only if the harm caused by bias is larger than the disadvantage of having too few reviews
Review solicitation and incentives (e.g., rebate-for-review programs)	Targets selection bias: increases volume and representativeness to correct mean shift	Induces reciprocal behavior: financial incentives may shift $\mathbb{E}[\varepsilon_{it}]$ upward; may increase noise via low-effort reviews	Corrects underreporting but risks introducing new distortions; requires careful calibration
Simultaneous reveal (two-sided reputation systems) and blinded reviews (social influence)	Targets Independence Condition: eliminates serial correlation (social influence) and reciprocity by decoupling reviews	Affect participation: may increase review volume due to increased "curiosity" or lower it due to lack of strategic timing benefits	Highly effective for restoring independence; minor disadvantages are typically outweighed by removal of social influence bias
Algorithmic outlier detection	Targets Mean Condition: filters strategic manipulation (fake reviews) via anomaly detection	Risks information loss: may falsely flag legitimate high-variance reviews, reducing sample size and biasing the mean if errors are asymmetric	Effective for clear fraud; requires high precision to avoid punishing genuine extreme reviews
Aggregation (e.g., averages)	Targets Variance Condition: averages out random noise to improve signal precision	Ineffective for mean shift: does not correct systematic bias (fake reviews, selection)	Essential for noise reduction; must be combined with other interventions to address systematic bias

reviews) or structural pressure (reciprocity or selection bias). If ratings are highly polarized with no clear consensus (a Variance Condition violation, high $\mathrm{Var}[\varepsilon_{it}]$), the cause is likely noise, for example, from attribute confusion or idiosyncratic tastes. If ratings cluster tightly around early reviews regardless of quality (an Independence Condition violation, $\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is}) \neq 0$ for some $t \neq s$), social influence is the primary driver. As Table 7 illustrates, each bias presents distinct symptoms requiring distinct responses.

Second, once identified, interventions must be matched to the specific signature of the error term. To restore Unbiasedness (Mean Condition), operators should deploy verification mechanisms (e.g., verified purchases) to raise the cost of manipulation to eliminate fake reviews or implement simultaneous reveal to eliminate reciprocity. To optimize Precision (Variance Condition), platforms should reduce noise by separating evaluation dimensions (e.g., using separate scores for product quality vs. logistics) and use explicit aggregation (e.g., displaying averages) to average out random noise. To ensure Independence (Correlation Condition), the only effective remedy is structural decoupling, such as hiding prior ratings during the submission window (blinded reviews) to prevent herding. Crucially, these tools are not interchangeable. Applying a "blinding" mechanism to a fake review problem addresses the wrong condition, while strict verification does nothing to stop social influence.

Third, and most critically, platform operators must compensate for the inevitable trade-offs identified in Proposition 1 (Trade-offs) and Table 8. Interventions are rarely cost-free. Most interventions create a secondary distortion that requires a counter-measure. Strict verification reduces fake reviews (fixing the expected value) but shrinks the sample size, increasing variance and potentially inducing selection bias. Therefore, operators must assess whether platform-managed solicitation campaigns are required to maintain volume and representativeness of reviews. Similarly, blinded reviews restore independence but may reduce participation due to the loss of social cues, requiring compensatory design features to sustain engagement. Managers must view design not as a search for a perfect tool, but as a balancing act where every restriction is offset by an incentive, ensuring that solving one problem does not amplify another. In the age of generative artificial intelligence, this compensation must also be dynamic, continuously adjusting thresholds and incentives as the cost of manipulation evolves.

5 Summary and conclusion

This paper surveyed the empirical literature on biases in reputation systems through a unified conceptual framework that categorizes distortions by their signature: violations of the Mean Condition (systematic bias), the Variance Condition (noise) or the Independence Condition (serial correlation). The survey investigated five major sources of bias: (i) strategic actions of sellers (fake reviews and rebate-for-review programs), (ii) strategic actions of reviewers (reciprocity), (iii) social influence, (iv) selection bias and (v) noise, investigating their mechanisms, effects and mitigation strategies, summarized in Table 7.

The synthesis of evidence yields three central insights that extend beyond existing literature. First, biases are endogenous to platform design. Verification rules, visibility settings and aggregation methods do not merely reveal user behavior, but actively shape it. For instance, two-sided systems inherently invite reciprocity bias, while visible prior ratings induce social influence. Second, design involves inherent trade-offs (Proposition 1). Interventions targeting one moment of the error term often affect another. For instance, as shown in Sect. 3, strict verification requirements reduce fake reviews but can reduce participation. Correcting selection bias via incentives may introduce reciprocal behavior, thereby potentially inflating the error term. Thus, optimal design is not about minimizing a single error type but balancing the trade-offs between accuracy, precision and independence. Third, detection and mitigation strategies must be asymmetric (Proposition 3). Tools effective for mean-shift biases (e.g., outlier detection) are inappropriate for variance-driven noise, which requires aggregation. Similarly, increasing review volume solves noise and selection issues but fails against social influence, which requires structural independence (Proposition 2).

These findings have distinct implications for theory, practice and policy. Building on the diagnostic framework proposed in Sect. 4, the evidence implies that improving reputation systems is primarily a design and governance problem that requires systematic trade-off management rather than isolated fixes. For platform operators, this means moving beyond generic solutions. Managers must first diagnose the specific

error signature (mean, variance or correlation) before applying matched interventions. Crucially, the interaction matrix (Table 8) demonstrates that interventions are rarely cost-free. Therefore, design changes (e.g., stricter verification) must be systematically paired with compensatory measures (e.g., increased solicitation) to maintain signal precision and participation. For regulators and policymakers, the results suggest that governance should focus on transparency of these design trade-offs. Instead of prescribing specific technologies, regulators could require platforms to disclose how their design choices balance bias reduction against representativeness and independence, and to audit algorithmic curation layers that may amplify user-generated distortions. The rise of generative artificial intelligence further underscores the need for dynamic governance, where detection and incentive structures co-evolve with manipulation capabilities.

The rise of generative artificial intelligence, where the marginal cost of generating deceptive content approaches zero, requires a focused research agenda on four critical frontiers. First, scholars must investigate the dynamic co-evolution between artificial intelligence-generated manipulation and artificial intelligence-based detection systems, modeling this interaction as an ongoing arms race rather than a static equilibrium (Proposition 4). Empirical studies exploiting platform policy changes, detection thresholds or text-based manipulation scores are needed to understand how this technological shift affects bias dynamics. Second, research must examine algorithmic amplification, specifically, how platform recommendation and ranking algorithms—which often display certain reviews more prominently or suppress others to maximize engagement—interact with user-generated biases such as fake reviews or social influence. Future work should quantify whether these algorithmic sorting mechanisms inflate variance or reinforce mean shifts beyond the organic bias of the reviews themselves. Third, the field needs studies on bias dynamics within review ecosystems mediated by generative artificial intelligence, where both the creation and consumption of reviews are increasingly automated. Fourth, comparative analysis of the effectiveness of bias-mitigation mechanisms (e.g., hiding reviews vs. verification of customers) is needed across different platform types (e.g., e-commerce vs. peer-to-peer platforms) to validate the trade-offs predicted by Propositions 1 and 2. Addressing these priorities will shift the literature from investigating biases in isolation to generating integrated, predictive and design-oriented insights.

Future research should also exploit new empirical settings to test the generalizability of the proposed framework. Reputation systems are increasingly interconnected. Sellers often operate on multiple platforms and reputation information is transferred through aggregators or social media. This implies that biases generated on one platform may spill over to others, even when design rules differ. Multi-homing sellers and cross-posted reviews offer promising settings to study how reputation and bias spread across ecosystems. Furthermore, while much of the available evidence is drawn from a limited number of large platforms, smaller and niche platforms may face different trade-offs, making them important settings for future empirical work. The rapid evolution of platform design creates repeated natural experiments in these diverse contexts that future research can exploit.

This study is subject to several limitations. It focuses on the empirical literature and therefore does not provide a formal theoretical model of the different biases and

how they interact. Platform design and technologies evolve rapidly, meaning that some empirical results may become outdated as firms adapt their systems. However, the conceptual framework is designed to be robust to such technological changes, allowing future studies to be categorized within the same structure.

Despite these limitations, the survey highlights that biases in reputation systems are neither inevitable nor purely behavioral. They are largely shaped by institutional design choices, and understanding these choices is essential for building reputation systems that sustain trust and support efficient digital markets.

Appendix A: Survey methodology

This appendix details the protocol used to identify, select and synthesize the literature surveyed in this paper. The objective of the survey is not to provide an exhaustive list of all publications on online reviews, but to conduct a structured, problem-oriented review that synthesizes empirical findings through a unified conceptual framework. This approach was chosen over a purely algorithmic systematic review because the paper's goal is to integrate heterogeneous findings from economics, information systems, marketing and management into a common theoretical structure, i.e., the conceptual framework introduced in Sect. 2.3. This approach differs from a traditional narrative review in three ways: (i) it employs explicit, predefined inclusion and exclusion criteria rather than subjective selection, (ii) it uses a standardized protocol to extract data consistently across studies, reducing author bias and (iii) it validates comprehensiveness through search saturation and cross-referencing with existing surveys. While a fully algorithmic systematic review might miss cross-disciplinary conceptual links due to rigid keyword constraints, a traditional narrative review risks selection bias. The structured, problem-oriented protocol used in this paper balances these extremes, ensuring both theoretical integration and methodological transparency. Figure 2 provides a schematic overview of the literature identification and selection process.

Search strategy

The literature identification process followed a two-stage strategy combining citation networking and targeted database searches.

First, backward and forward citation tracking was conducted starting from foundational contributions on reputation systems and online reviews, including Bolton et al. [14, 22], Dellarocas [16], Dellarocas and Wood [97], Luca and Zervas [59], Mayzlin et al. [37], Resnick and Zeckhauser [4] and Tadelis [2]. Backward citation tracking (reviewing references) and forward citation tracking (using Google Scholar to find subsequent citing papers) were performed iteratively. This snowballing method is particularly effective for tracing the evolution of specific empirical literature streams and ensuring that foundational works are not missed due to keyword variations.

Second, to complement citation tracking and capture recent work not yet heavily cited, targeted keyword searches were conducted in Google Scholar and Scopus.

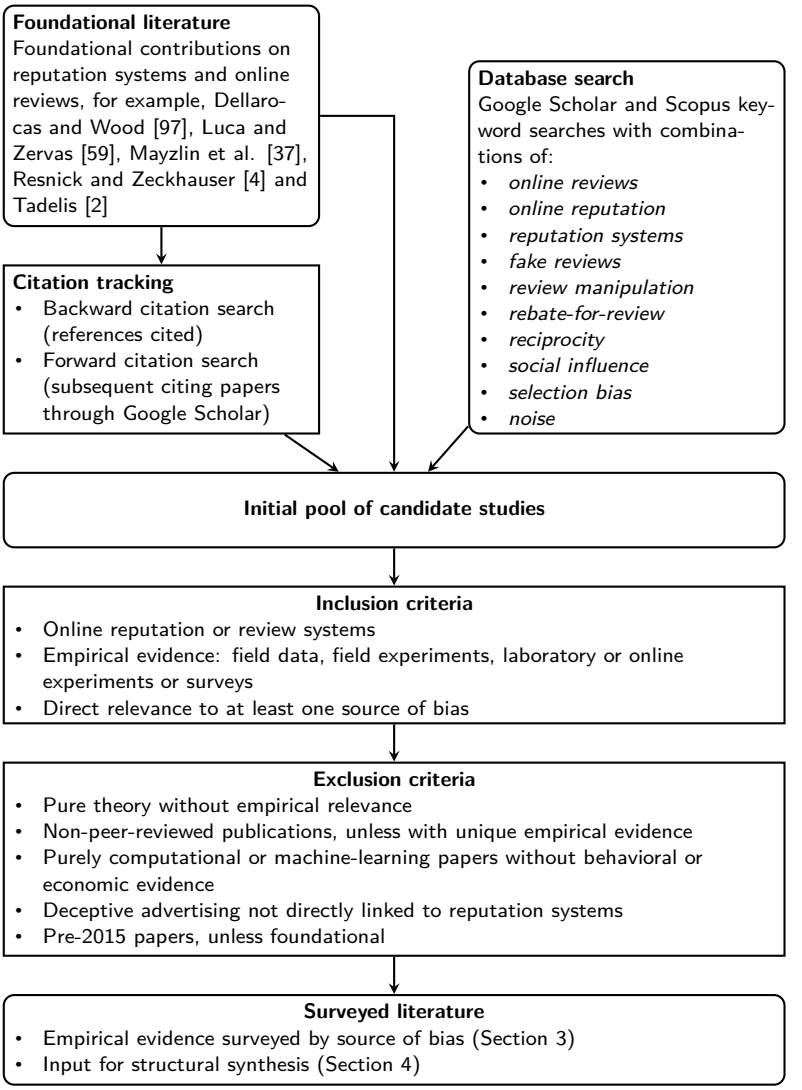

Fig. 2 Schematic overview of the literature identification and selection process

Search terms included combinations of *online reviews*, *online reputation*, *reputation systems*, *fake reviews*, *review manipulation*, *rebate-for-review*, *reciprocity*, *social influence*, *selection bias* and *noise*. These searches were used to identify empirical studies that are not connected to the foundational contributions through citation networks, as well as more recent contributions that had not yet accumulated substantial citations.

The search process followed an iterative approach until diminishing returns were observed. Specifically, citation tracking and keyword searches continued until additional queries yielded no new mechanisms, bias types or contradictory empirical find-

ings that would offer new insights for the structural synthesis. To validate comprehensiveness, the final bibliography was cross-referenced with major existing survey papers (e.g., Magnani [12] and Wu et al. [6]) to ensure coverage of all identified bias types, major platform contexts (e-commerce, hospitality, peer-to-peer) and methodological approaches (field vs. lab). This step ensured that all seminal works and major research streams identified in prior literature were included, minimizing omissions of relevant publications.

Inclusion and exclusion criteria

Studies were screened against explicit inclusion and exclusion criteria to ensure relevance and quality.

To be included, a study had to satisfy three conditions: (i) It examines online reputation or review systems where ratings or textual reviews play a central role. (ii) It provides empirical evidence (field data, field experiments, laboratory/online experiments or surveys) on how reviews are generated, manipulated, interpreted or affect market outcomes. (iii) It relates directly to at least one of the sources of bias analyzed: strategic actions of sellers, strategic actions of reviewers, social influence, selection bias or noise.

The following types of works were excluded or de-prioritized: (i) Purely theoretical contributions, unless they provide foundational or conceptual insights that support the understanding of empirical findings. (ii) Non-peer-reviewed publications (e.g., grey literature, including working papers and preprints), unless such work is influential, frequently cited or provides unique empirical evidence despite not yet being published in a journal. (iii) Purely computational or machine learning papers, e.g., related to the detection of fake reviews, that focus solely on algorithmic methodology without offering empirical insights into economic or behavioral mechanisms. (iv) Studies on deceptive advertising or misinformation that do not explicitly link to reputation systems.

While the survey emphasizes work published from 2015 onward to reflect modern platform designs, foundational studies prior to 2015 were retained when they established key theoretical baselines or when recent literature on a specific bias (e.g., reciprocity) was limited.

Classification of the literature

To ensure a structured synthesis rather than a descriptive narrative, all included studies were analyzed using a consistent analytical protocol. Every paper was mapped against the following set of standardized questions during the reading and synthesis phase:

1. Antecedents

   (a) What is the bias (definition)?
   (b) Why do interested parties cause the bias?
   (c) Who has an incentive to bias?

(d) When (under what conditions) does the bias happen?
(e) How is the bias created?

2. Properties

(a) What are the properties of the bias?
(b) What are the spreading properties of the bias?
(c) What is the distribution of the bias?

3. Consequences

(a) How does the bias affect the development of online product reviews?
(b) What are the effects of the bias on stakeholders?
(c) How does the bias affect the overall market and society?

4. Interventions

(a) What are suitable detection methods?
(b) What can stakeholders do to effectively respond to the bias?

This set of questions is a loose adaptation and generalization of the antecedent-consequence-intervention framework used by Wu et al. [6] to review the literature on fake reviews. They guide the presentation of the literature on the individual biases in Sect. 3 and are used to structure the literature in Tables 1, 2, 3, 4, 5, 6 by the focus of the analysis, with granularity varying based on the scope and specifics of the literature on a particular bias.

To evaluate mechanisms, responses to these questions were then mapped against the conceptual framework from Sect. 2.3. Specifically, evidence regarding *Properties* was mapped to the moments of the error term (i.e., mean, variance and independence), while *Interventions* were categorized by their alignment with Propositions 1–4.

By applying this uniform set of questions across diverse literature from multiple disciplines (economics, information systems, marketing and management), the review ensures that findings are comparable and can be integrated into the unified conceptual framework presented in Sect. 2.3. This problem-oriented structure allows the paper to transcend disciplinary silos and focus on the underlying mechanisms of bias.

Limitations

This review has limitations inherent to its approach. First, while the iterative search strategy and cross-referencing aimed to ensure comprehensiveness, it is possible that niche studies in emerging platforms or non-English journals were missed. Second, the prioritization of peer-reviewed English-language literature may introduce a slight selection bias toward established platforms (e.g., Amazon, eBay, Yelp) and Western contexts. While this focus on major platforms limits generalizability to niche markets, it ensures the review focuses on the ecosystems where the vast majority

of global commerce and bias-related harm occurs, aligning with the paper's goal of informing mainstream platform design and policy. Finally, the rapid evolution of platform design means that some empirical findings may become outdated as new features (e.g., artificial intelligence-driven curation) are introduced. However, the conceptual framework is designed to be robust to such technological changes, allowing future studies to be categorized within the same structure.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The author states that there is no conflict of interest.

References

1. Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics, 84*(3), 488–500. https://doi.org/10.2307/1879431
2. Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics, 8*(1), 321–340. https://doi.org/10.1146/annurev-economics-080315-015325
3. Bajari, P., & Hortaçsu, A. (2004). Economic insights from internet auctions. *Journal of Economic Literature, 42*(2), 457–486. https://doi.org/10.1257/0022051041409075
4. Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. In M. R. Baye (Ed.), Advances in applied microeconomics (pp. 127–157, Vol. 11). Emerald. https://doi.org/10.1016/S0278-0984(02)11030-3
5. Sahut, J. M., Laroche, M., & Braune, E. (2024). Antecedents and consequences of fake reviews in a marketing approach: An overview and synthesis. *Journal of Business Research, 175*, Article 114572. https://doi.org/10.1016/j.jbusres.2024.114572
6. Wu, Y., Ngai, E. W., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems, 132*, Article 113280. https://doi.org/10.1016/j.dss.2020.113280
7. Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y., & Maqsood, S. (2021). Fake reviews detection: A survey. *IEEE Access, 9*, 65771–65802. https://doi.org/10.1109/ACCESS.2021.3075573
8. Vidanagama, D. U., Silva, T. P., & Karunananda, A. S. (2020). Deceptive consumer review detection: A survey. *Artificial Intelligence Review, 53*(2), 1323–1352. https://doi.org/10.1007/s10462-019-09697-5
9. Gutt, D., Neumann, J., Zimmermann, S., Kundisch, D., & Chen, J. (2019). Design of review systems – a strategic instrument to shape online reviewing behavior and economic outcomes. *The Journal of Strategic Information Systems, 28*(2), 104–117. https://doi.org/10.1016/j.jsis.2019.01.004

10. Lee, P. S., Chakraborty, I., & Banerjee, S. (2023, March 13). Artificial intelligence applications to customer feedback research: A review. In K. Sudhir & O. Toubia (Eds.), Artificial intelligence in marketing (pp. 169–190, Vol. 20). Emerald Publishing Limited. https://doi.org/10.1108/S1548-643520230000020010
11. Luca, M. (2017). Designing online marketplaces: Trust and reputation mechanisms. *Innovation Policy and the Economy, 17*, 77–93. https://doi.org/10.1086/688845
12. Magnani, M. (2020). The economic and behavioral consequences of online user reviews. *Journal of Economic Surveys, 34*(2), 263–292. https://doi.org/10.1111/joes.12357
13. Pocchiari, M., Proserpio, D., & Dover, Y. (2025). Online reviews: A literature review and roadmap for future research. *International Journal of Research in Marketing, 42*(2), 275–297. https://doi.org/10.1016/j.ijresmar.2024.08.009
14. Bolton, G. E., Katok, E., & Ockenfels, A. (2004). How effective are electronic reputation mechanisms? An experimental investigation. *Management Science, 50*(11), 1587–1602. https://doi.org/10.1287/mnsc.1030.0199
15. Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics, 9*(2), 79–101. https://doi.org/10.1007/s10683-006-4309-2
16. Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*(10), 1407–1424. https://doi.org/10.1287/mnsc.49.10.1407.17308
17. Klein, T. J., Lambertz, C., & Stahl, K. O. (2016). Market transparency, adverse selection, and moral hazard. *Journal of Political Economy, 124*(6), 1677–1713. https://doi.org/10.1086/688875
18. Klein, T. J., Lambertz, C., Spagnolo, G., & Stahl, K. O. (2009). The actual structure of eBay's feedback mechanism and early evidence on the effects of recent changes. *International Journal of Electronic Business, 7*(3), 301–320. https://doi.org/10.1504/IJEB.2009.026532
19. Bolton, G. E., Greiner, B., & Ockenfels, A. (2013). Engineering trust: Reciprocity in the production of reputation information. *Management Science, 59*(2), 265–285. https://doi.org/10.1287/mnsc.1120.1609
20. Belleflamme, P., & Peitz, M. (2021, October 31). The economics of platforms: Concepts and strategy. Cambridge University Press. https://doi.org/10.1017/9781108696913
21. Lewis, G. (2011). Asymmetric information, adverse selection and online disclosure: The case of eBay Motors. *American Economic Review, 101*(4), 1535–1546. https://doi.org/10.1257/aer.101.4.1535
22. Bolton, G. E., Katok, E., & Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Public Economics, 89*(8), 1457–1468. https://doi.org/10.1016/j.jpubeco.2004.03.008
23. Veh, A., Göbel, M., & Vogel, R. (2019). Corporate reputation in management research: A review of the literature and assessment of the concept. *Business Research, 12*(2), 315–353. https://doi.org/10.1007/s40685-018-0080-4
24. Zhang, Y., Wang, Z., Xiao, L., Wang, L., & Huang, P. (2023). Discovering the evolution of online reviews: A bibliometric review. *Electronic Markets, 33*(1), 49. https://doi.org/10.1007/s12525-023-00667-y
25. Goldfarb, A., & Tucker, C. (2019). Digital economics. *Journal of Economic Literature, 57*(1), 3–43. https://doi.org/10.1257/jel.20171452
26. Liu, M., Brynjolfsson, E., & Dowlatabadi, J. (2021). Do digital platforms reduce moral hazard? The case of Uber and taxis. *Management Science, 67*(8), 4665–4685. https://doi.org/10.1287/mnsc.2020.3721
27. Reimers, I., & Waldfogel, J. (2021). Digitization and pre-purchase information: The causal and welfare impacts of reviews and crowd ratings. *American Economic Review, 111*(6), 1944–1971. https://doi.org/10.1257/aer.20200153
28. Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498–1512. https://doi.org/10.1109/TKDE.2010.188
29. De Langhe, B., Fernbach, P. M., & Lichtenstein, D. R. (2016). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research, 42*(6), 817–833. https://doi.org/10.1093/jcr/ucv047
30. Köcher, S., & Köcher, S. (2018). Should we reach for the stars? Examining the convergence between online product ratings and objective product quality and their impacts on sales performance. *Journal of Marketing Behavior, 3*(2), 167–183. https://doi.org/10.1561/107.00000050

31. Brandes, L., Godes, D., & Mayzlin, D. (2022). Extremity bias in online reviews: The role of attrition. *Journal of Marketing Research, 59*(4), 675–695. https://doi.org/10.1177/00222437211073579
32. Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM, 52*(10), 144–147. https://doi.org/10.1145/1562764.1562800
33. Zervas, G., Proserpio, D., & Byers, J. W. (2021). A first look at online reputation on Airbnb, where every stay is above average. *Marketing Letters, 32*(1), 1–16. https://doi.org/10.1007/s11002-020-09546-4
34. Filippas, A., Horton, J. J., & Golden, J. M. (2022). Reputation inflation. *Marketing Science, 41*(4), 733–745. https://doi.org/10.1287/mksc.2022.1350
35. Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345–354. https://doi.org/10.1509/jmkr.43.3.345
36. Babić Rosario, A., Sotgiu, F., de Valck, K., & Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research, 53*(3), 297–318. https://doi.org/10.1509/jmr.14.0380
37. Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review, 104*(8), 2421–2455. https://doi.org/10.1257/aer.104.8.2421
38. Gössling, S., Hall, C. M., & Andersson, A.-C. (2018). The manager's dilemma: A conceptualization of online review manipulation strategies. *Current Issues in Tourism, 21*(5), 484–503. https://doi.org/10.1080/13683500.2015.1127337
39. Federal Trade Commission. (2024, August 14). Federal Trade Commission announces final rule banning fake reviews and testimonials. Retrieved December 18, 2025, from https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials
40. Federal Trade Commission. (2024, September 25). FTC announces crackdown on deceptive AI claims and schemes. Retrieved December 18, 2025, from https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes
41. Federal Trade Commission. (2025, July 14). FTC takes action against telemedicine firm NextMed over charges it used misleading prices, fake reviews, and deceptive weight loss claims to sell GLP-1 weight-loss programs. Retrieved December 18, 2025, from https://www.ftc.gov/news-events/news/press-releases/2025/07/ftc-takes-action-against-telemedicine-firm-nextmed-over-charges-it-used-misleading-prices-fake
42. Digital Markets, Competition and Consumers Act 2024 (2024, May 24). https://www.legislation.gov.uk/ukpga/2024/13/contents
43. Amazon Staff. (2025, October 8). Amazon's latest actions against fake review brokers: Amazon and BBB join forces again to combat fake reviews. About Amazon. Retrieved December 18, 2025, from https://www.aboutamazon.com/news/policy-news-views/amazons-latest-actions-against-fake-review-brokers
44. Choi, S., Mattila, A. S., Van Hoof, H. B., & Quadri-Felitti, D. (2017). The role of power and incentives in inducing fake reviews in the tourism industry. *Journal of Travel Research, 56*(8), 975–987. https://doi.org/10.1177/0047287516677168
45. Glazer, J., Herrera, H., & Perry, M. (2021). Fake reviews. *The Economic Journal, 131*(636), 1772–1787. https://doi.org/10.1093/ej/ueaa124
46. Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems, 33*(2), 456–481. https://doi.org/10.1080/07421222.2016.1205907
47. Anderson, E. T., & Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research, 51*(3), 249–269. https://doi.org/10.1509/jmr.13.0209
48. Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal, 122*(563), 957–989. https://doi.org/10.1111/j.1468-0297.2012.02512.x
49. Dini, F., & Spagnolo, G. (2009). Buying reputation on eBay: Do recent changes help? *International Journal of Electronic Business, 7*(6), 581–598. https://doi.org/10.1504/IJEB.2009.029048
50. Hajek, P., Hikkerova, L., & Sahut, J.-M. (2023). Fake review detection in e-commerce platforms using aspect-based sentiment analysis. *Journal of Business Research, 167*, Article 114143. https://doi.org/10.1016/j.jbusres.2023.114143
51. He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. *Marketing Science, 41*(5), 896–921. https://doi.org/10.1287/mksc.2022.1353

52. Hu, N., Bose, I., Gao, Y., & Liu, L. (2011). Manipulation in digital word-of-mouth: A reality check for book reviews. *Decision Support Systems, 50*(3), 627–635. https://doi.org/10.1016/j.dss.2010.08.013
53. Hu, N., Liu, L., & Sambamurthy, V. (2011). Fraud detection in online consumer reviews. *Decision Support Systems, 50*(3), 614–626. https://doi.org/10.1016/j.dss.2010.08.012
54. Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems, 52*(3), 674–684. https://doi.org/10.1016/j.dss.2011.11.002
55. Ko, E. E., & Bowman, D. (2023). Suspicious online product reviews: An empirical analysis of brand and product characteristics using Amazon data. *International Journal of Research in Marketing, 40*(4), 898–911. https://doi.org/10.1016/j.ijresmar.2023.06.006
56. Lappas, T. (2012). Fake reviews: The malicious perspective. In G. Bouma, A. Ittoo, & H. Wortmann (Eds.), Natural language processing and information systems (pp. 23–34, Vol. 7337). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31178-9_3
57. Li, H., Ji, H., Luo, J. M., & Zhang, Z. (2023). Competition and restaurant online review manipulations: A dynamic panel data analysis. *International Journal of Hospitality Management, 115*, Article 103605. https://doi.org/10.1016/j.ijhm.2023.103605
58. Luca, M., & Reshef, O. (2021). The effect of price on firm reputation. *Management Science, 67*(7), 4408–4419. https://doi.org/10.1287/mnsc.2021.4049
59. Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science, 62*(12), 3412–3427. https://doi.org/10.1287/mnsc.2015.2304
60. Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the prevalence of deception in online review communities. Proceedings of the 21st International Conference on World Wide Web, 201–210. https://doi.org/10.1145/2187836.2187864
61. Wang, Q., Zhang, W., Li, J., Ma, Z., & Chen, J. (2023). Benefits or harms? The effect of online review manipulation on sales. *Electronic Commerce Research and Applications, 57*, Article 101224. https://doi.org/10.1016/j.elerap.2022.101224
62. Xu, H., Liu, D., Wang, H., & Stavrou, A. (2015). E-commerce reputation manipulation: The emergence of reputation-escalation-as-a-service. Proceedings of the 24th International Conference on World Wide Web, 1296–1306. https://doi.org/10.1145/2736277.2741650
63. Zhang, Z., Li, Y., Li, H., & Zhang, Z. (2022). Restaurants' motivations to solicit fake reviews: A competition perspective. *International Journal of Hospitality Management, 107*, Article 103337. https://doi.org/10.1016/j.ijhm.2022.103337
64. Ananthakrishnan, U. M., Li, B., & Smith, M. D. (2020). A tangled web: Should online review portals display fraudulent reviews? *Information Systems Research, 31*(3), 950–971. https://doi.org/10.1287/isre.2020.0925
65. Banerjee, S., & Chua, A. Y. (2023). Understanding online fake review production strategies. *Journal of Business Research, 156*, Article 113534. https://doi.org/10.1016/j.jbusres.2022.113534
66. Gössling, S., Zeiss, H., Hall, C. M., Martin-Rios, C., Ram, Y., & Grøtte, I.-P. (2019). A cross-country comparison of accommodation manager perspectives on online review manipulation. *Current Issues in Tourism, 22*(14), 1744–1763. https://doi.org/10.1080/13683500.2018.1455171
67. Harrison-Walker, L. J., & Jiang, Y. (2023). Suspicion of online product reviews as fake: Cues and consequences. *Journal of Business Research, 160*, Article 113780. https://doi.org/10.1016/j.jbusres.2023.113780
68. Krügel, J. P., & Paetzel, F. (2024). The impact of fraud on reputation systems. *Games and Economic Behavior, 144*, 329–354. https://doi.org/10.1016/j.geb.2024.01.013
69. Malbon, J. (2013). Taking fake online consumer reviews seriously. *Journal of Consumer Policy, 36*(2), 139–157. https://doi.org/10.1007/s10603-012-9216-7
70. Kovács, B. (2024). The Turing test of online reviews: Can we tell the difference between human-written and GPT-4-written online reviews? *Marketing Letters, 35*(4), 651–666. https://doi.org/10.1007/s11002-024-09729-3
71. Salminen, J., Kandpal, C., Kamel, A. M., Jung, S.-G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services, 64*, Article 102771. https://doi.org/10.1016/j.jretconser.2021.102771
72. Akesson, J., Hahn, R., Metcalfe, R., & Monti-Nussbaum, M. (2023, November). The impact of fake reviews on demand and welfare (NBER Working Paper No. 31836). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w31836

73. Alma Economics. (2023, April). Fake online reviews research: Estimating the prevalence and impact of fake online reviews. Department for Business and Trade. Retrieved December 18, 2025, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1152812/fake-online-reviews-research.pdf
74. Xia, R., Dong, X., An, J., & Wang, H. (2025). The impact of fake online reviews on customer satisfaction: An empirical study on JD.com. *Electronic Commerce Research, 25*(6), 4689–4716. https://doi.org/10.1007/s10660-024-09865-y
75. Munzel, A. (2016). Assisting consumers in detecting fake reviews: The role of identity information disclosure and consensus. *Journal of Retailing and Consumer Services, 32*, 96–108. https://doi.org/10.1016/j.jretconser.2016.06.002
76. Song, Y., Wang, L., Zhang, Z., & Hikkerova, L. (2023). Do fake reviews promote consumers' purchase intention? *Journal of Business Research, 164*, Article 113971. https://doi.org/10.1016/j.jbusres.2023.113971
77. Bhangale, S., & Roy, P. K. (2025). Is it genuine or fake? Analyzing e-commerce reviews using large language models. *Knowledge-Based Systems, 330*, Article 114556. https://doi.org/10.1016/j.knosys.2025.114556
78. He, S., Hollenbeck, B., Overgoor, G., Proserpio, D., & Tosyali, A. (2022). Detecting fake-review buyers using network structure: Direct evidence from Amazon. *Proceedings of the National Academy of Sciences, 119*(47), Article e2211932119. https://doi.org/10.1073/pnas.2211932119
79. Nawara, D., & Kashef, R. (2025). A dual-phase framework for detecting authentic and computer-generated customer reviews using large language models. *Decision Analytics Journal, 15*, Article 100581. https://doi.org/10.1016/j.dajour.2025.100581
80. Plotkina, D., Munzel, A., & Pallud, J. (2020). Illusions of truth–experimental insights into human and algorithmic detections of fake online reviews. *Journal of Business Research, 109*, 511–523. https://doi.org/10.1016/j.jbusres.2018.12.009
81. Lappas, T., Sabnis, G., & Valkanas, G. (2016). The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Information Systems Research, 27*(4), 940–961. https://doi.org/10.1287/isre.2016.0674
82. Salehi-Esfahani, S., & Ozturk, A. B. (2018). Negative reviews: Formation, spread, and halt of opportunistic behavior. *International Journal of Hospitality Management, 74*, 138–146. https://doi.org/10.1016/j.ijhm.2018.06.022
83. Li, L. I., Tadelis, S., & Zhou, X. (2020). Buying reputation as a signal of quality: Evidence from an online marketplace. *The RAND Journal of Economics, 51*(4), 965–988. https://doi.org/10.1111/1756-2171.12346
84. Shukla, A. D., & Goh, J. M. (2024). Fighting fake reviews: Authenticated anonymous reviews using identity verification. *Business Horizons, 67*(1), 71–81. https://doi.org/10.1016/j.bushor.2023.08.002
85. Amazon Customer Service. (n.d.). Understanding customer reviews and ratings. Retrieved December 18, 2025, from https://www.amazon.com/gp/help/customer/display.html?nodeId=G8UYX7LALQC8V9KA&language=en_US
86. Ivanova, O., & Scholz, M. (2017). How can online marketplaces reduce rating manipulation? A new approach on dynamic aggregation of online ratings. *Decision Support Systems, 104*, 64–78. https://doi.org/10.1016/j.dss.2017.10.003
87. Li, L. I., & Xiao, E. (2014). Money talks: Rebate mechanisms in reputation system design. *Management Science, 60*(8), 2054–2072. https://doi.org/10.1287/mnsc.2013.1848
88. Li, L. I. (2010). Reputation, trust, and rebates: How online auction markets can improve their feedback mechanisms. *Journal of Economics & Management Strategy, 19*(2), 303–331. https://doi.org/10.1111/j.1530-9134.2010.00253.x
89. Cabral, L., & Li, L. I. (2015). A dollar for your thoughts: Feedback-conditional rebates on eBay. *Management Science, 61*(9), 2052–2063. https://doi.org/10.1287/mnsc.2014.2074
90. Fradkin, A., & Holtz, D. (2023). Do incentives to review help the market? Evidence from a field experiment on Airbnb. *Marketing Science, 42*(5), 853–865. https://doi.org/10.1287/mksc.2023.1439
91. Park, S., Shin, W., & Xie, J. (2023). Disclosure in incentivized reviews: Does it protect consumers? *Management Science, 69*(11), 7009–7021. https://doi.org/10.1287/mnsc.2023.00930
92. Garnefeld, I., Helm, S., & Grötschel, A.-K. (2020). May we buy your love? Psychological effects of incentives on writing likelihood and valence of online product reviews. *Electronic Markets, 30*(4), 805–820. https://doi.org/10.1007/s12525-020-00425-4

93. Koukova, N. T., Wang, R.J.-H., & Isaac, M. S. (2023). "If you loved our product": Do conditional review requests harm retailer loyalty? *Journal of Retailing, 99*(1), 85–101. https://doi.org/10.1016/j.jretai.2022.09.002
94. Fradkin, A., Grewal, E., & Holtz, D. (2021). Reciprocity and unveiling in two-sided reputation systems: Evidence from an experiment on Airbnb. *Marketing Science, 40*(6), 1013–1029. https://doi.org/10.1287/mksc.2021.1311
95. Proserpio, D., Xu, W., & Zervas, G. (2018). You get what you give: Theory and evidence of reciprocity in the sharing economy. *Quantitative Marketing and Economics, 16*(4), 371–407. https://doi.org/10.1007/s11129-018-9201-9
96. Cabral, L., & Hortaçsu, A. (2010). The dynamics of seller reputation: Evidence from eBay. *The Journal of Industrial Economics, 58*(1), 54–78. https://doi.org/10.1111/j.1467-6451.2010.00405.x
97. Dellarocas, C., & Wood, C. A. (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science, 54*(3), 460–476. https://doi.org/10.1287/mnsc.1070.0747
98. Jian, L., MacKie-Mason, J. K., & Resnick, P. (2010). I scratched yours: The prevalence of reciprocation in feedback provision on eBay. *The B.E. Journal of Economic Analysis & Policy, 10*(1), 92. https://doi.org/10.2202/1935-1682.2470
99. Klein, T. J., Lambertz, C., Spagnolo, G., & Stahl, K. O. (2006, June). Last minute feedback (CEPR Discussion Paper No. 5693). https://repec.cepr.org/repec/cpr/ceprdp/DP5693.pdf
100. Li, L. I. (2010). What is the cost of venting? Evidence from eBay. *Economics Letters, 108*(2), 215–218. https://doi.org/10.1016/j.econlet.2010.05.013
101. Bolton, G. E., Greiner, B., & Ockenfels, A. (2018). Dispute resolution or escalation? The strategic gaming of feedback withdrawal options in online markets. *Management Science, 64*(9), 4009–4031. https://doi.org/10.1287/mnsc.2017.2802
102. Hui, X., Saeedi, M., & Sundaresan, N. (2018). Adverse selection or moral hazard, an empirical study. *The Journal of Industrial Economics, 66*(3), 610–649. https://doi.org/10.1111/joie.12183
103. Eryarsoy, E., & Piramuthu, S. (2014). Experimental evaluation of sequential bias in online customer reviews. *Information & Management, 51*(8), 964–971. https://doi.org/10.1016/j.im.2014.09.001
104. Sikora, R. T., & Chauhan, K. (2012). Estimating sequential bias in online reviews: A Kalman filtering approach. *Knowledge-Based Systems, 27*, 314–321. https://doi.org/10.1016/j.knosys.2011.10.011
105. Aral, S. (2014). The problem with online ratings. *MIT Sloan Management Review, 55*(2), 47–52. https://sloanreview.mit.edu/article/the-problem-with-online-ratings-2/
106. Askalidis, G., Kim, S. J., & Malthouse, E. C. (2017). Understanding and overcoming biases in online review systems. *Decision Support Systems, 97*, 23–30. https://doi.org/10.1016/j.dss.2017.03.002
107. Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science, 341*(6146), 647–651. https://doi.org/10.1126/science.1240466
108. Rohde, C., Kupfer, A., & Zimmermann, S. (2022). Explaining reviewing effort: Existing reviews as potential driver. *Electronic Markets, 32*(3), 1169–1185. https://doi.org/10.1007/s12525-022-00595-3
109. Han, S., & Anderson, C. K. (2019). Estimating the effect of social influence on subsequent reviews. In H. Yang & R. Qiu (Eds.), Advances in service science (pp. 231–238). Springer International Publishing. https://doi.org/10.1007/978-3-030-04726-9_23
110. Jacobsen, G. D. (2015). Consumers, experts, and online product evaluations: Evidence from the brewing industry. *Journal of Public Economics, 126*, 114–123. https://doi.org/10.1016/j.jpubeco.2015.04.005
111. Karaman, H. (2021). Online review solicitations reduce extremity bias in online review distributions and increase their representativeness. *Management Science, 67*(7), 4420–4445. https://doi.org/10.1287/mnsc.2020.3758
112. Moe, W. W., & Trusov, M. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research, 48*(3), 444–456. https://doi.org/10.1509/jmkr.48.3.444
113. Wang, C. A., Zhang, X. M., & Hann, I.-H. (2018). Socially nudged: A quasi-experimental study of friends' social influence in online product ratings. *Information Systems Research, 29*(3), 641–655. https://doi.org/10.1287/isre.2017.0741
114. Krishnan, S., Patel, J., Franklin, M. J., & Goldberg, K. (2014). A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. Proceedings of the 8th ACM Conference on Recommender Systems, 137–144. https://doi.org/10.1145/2645710.2645740
115. Schlosser, A. E. (2005). Posting versus lurking: Communicating in a multiple audience context. *Journal of Consumer Research, 32*(2), 260–265. https://doi.org/10.1086/432235

116. Gao, B., Wang, J., Ding, X., & Guo, Y. (2025). The pitfalls of review solicitation: Evidence from a natural experiment on TripAdvisor. *Management Science, 71*(2), 1671–1691. https://doi.org/10.1287/mnsc.2023.01006
117. Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research, 19*(4), 456–474. https://doi.org/10.1287/isre.1070.0154
118. Hu, N., Pavlou, P. A., & Zhang, J. (2017). On self-selection biases in online product reviews. *MIS Quarterly, 41*(2), 449–471. https://doi.org/10.25300/MISQ/2017/41.2.06
119. Gao, G. G., Greenwood, B. N., Agarwal, R., & McCullough, J. S. (2015). Vocal minority and silent majority: How do online ratings reflect population perceptions of quality? *MIS Quarterly, 39*(3), 565–589. https://doi.org/10.25300/MISQ/2015/39.3.03
120. Han, S., & Anderson, C. K. (2020). Customer motivation and response bias in online reviews. *Cornell Hospitality Quarterly, 61*(2), 142–153. https://doi.org/10.1177/1938965520902012
121. Han, S., & Mikhailova, D. (2024). Reducing the bias in online reviews using propensity score adjustment. *Cornell Hospitality Quarterly, 65*(4), 429–441. https://doi.org/10.1177/19389655231223364
122. Wang, F. X., & Anderson, C. (2023). How firm strategies affect consumer biases in online reviews. *Service Science, 15*(3), 172–187. https://doi.org/10.1287/serv.2023.0316
123. Brandes, L., & Dover, Y. (2022). Offline context affects online reviews: The effect of post-consumption weather. *Journal of Consumer Research, 49*(4), 595–615. https://doi.org/10.1093/jcr/ucac003
124. Greiff, M., & Paetzel, F. (2020). Information about average evaluations spurs cooperation: An experiment on noisy reputation systems. *Journal of Economic Behavior & Organization, 180*, 334–356. https://doi.org/10.1016/j.jebo.2020.10.014