The Trust Illusion: Why We Navigate Blindly in the Age of AI Reviews

A media and marketing psychology analysis

Dr. Josef Sawetz, August 2025

What if you were to learn that, with near certainty, a large share of online reviews had been written by an artificial intelligence (AI)? And what if someone told you that neither you nor the online retailer has a realistic chance of recognizing this?

This is not a dystopian vision of the future. It is the reality of 2025, as compellingly demonstrated by a current scientific paper titled “Large Language Models as ‘Hidden Persuaders’: Fake Product Reviews Are Indistinguishable to Humans and Machines” by Weiyao Meng and his team. The diagnosis of this study constitutes a seismic rupture in a fundamental pillar upon which our entire digital commerce is built: trust in the “wisdom of the crowd.”

Whereas advertising is recognizable as such and a certain level of media literacy allows for interpretation, AI-generated reviews masquerade as authentic consumer experiences. They exploit our evolutionarily shaped tendency to trust like-minded others, and they turn it against us.

Overconfidence as an additional problem

Another disturbing finding was the participants’ massive overestimation of their abilities in this investigation. While their actual performance was around 51%, on average they rated their own capabilities at 67%. From a psychological perspective, this gap between self-perception and reality is particularly problematic: consumers who trust their own judgment too much are even more susceptible to manipulation.

1. The bombshell—What science has really revealed

The researchers conducted a series of experiments to answer a seemingly simple question: Can we distinguish genuine, human-written product reviews from fake, AI-generated ones? The answer is a clear and troubling no.
The human participants in the study achieved an accuracy of only 50.8%. Statistically, that is no better than flipping a coin. Our gut feeling, our intuition, our life experience—all the tools we rely on to judge authenticity—completely fail here.

But the real shock is yet to come: the researchers posed the same task to the most advanced AI models. The result: the AI systems were just as bad or even worse than humans. The most powerful AI model also achieved only about 50% accuracy. Others performed significantly worse.

From a psychological standpoint, this is a turning point. Until now, we assumed that technology would eventually solve the problem of counterfeiting. The study destroys this hope and shows why the situation is so precarious:

The failure of human heuristics: Our “lie detector” is broken

We humans use mental shortcuts—so-called heuristics—to make sense of the world. In online reviews, we look for patterns: a small typo seems authentic (“to err is human”). Very emotional language seems real. A balanced critique with pros and cons seems credible. The problem: the AI models were trained to imitate precisely these patterns perfectly. On command, they can produce “authentically human” texts—with small mistakes, colloquial language, and emotional nuance.

Emotional authenticity

AI systems can now also simulate emotional authenticity. They use colloquial expressions, personal anecdotes, and the typical tone of real consumers. This emotional component is especially powerful because it appeals to our trust at an unconscious level.

Human heuristics: the “skepticism bias”

The study also uncovers a “skepticism bias” in people: we tend to view especially positive reviews with suspicion. This “too-good-to-be-true” heuristic leads us to classify perfectly worded, exuberantly positive reviews as more likely to be fake. At the same time, reviews with small mistakes or mixed feelings are perceived as more authentic. Manipulators know this and exploit it.

2. The machine’s blind spot: “veracity bias” (truth-bias)

Why does AI fail at detection? The researchers identified a fascinating reason they call “veracity bias.” The AI models were trained on the vast text corpora of the internet, which overwhelmingly consist of authentic, human-written content. As a result, the AI has a deeply ingrained default assumption: when in doubt, text is genuine.

That makes the system vulnerable to fraud. According to a current statistic, up to 30% of all online reviews worldwide are fake, and they are estimated to cost consumers 787.7 billion US dollars in 2025. This is enormous because reviews influence revenue: every additional star can raise sales by 5–9%.

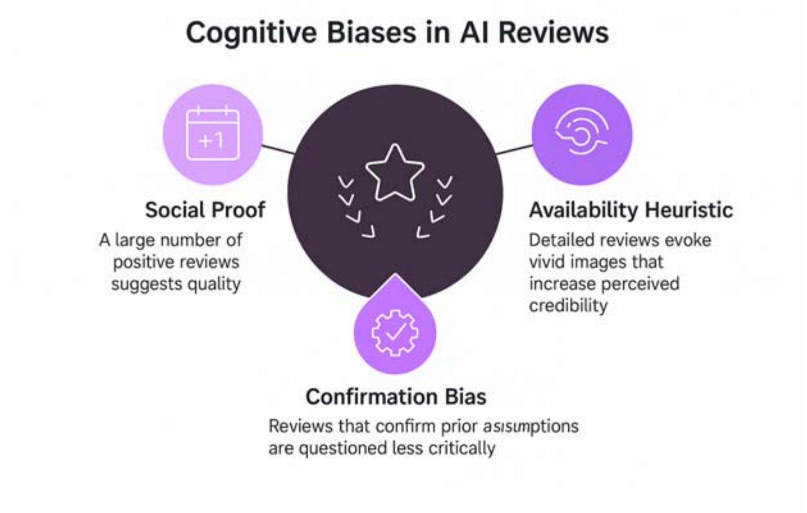
The vicious cycle

The more fake reviews circulate on the internet, the less reliable online reviews become as an information source. This creates a vicious cycle: As authentic reviews lose value, consumers’ incentives to write honest reviews decline. At the same time, the relative value of fake reviews rises, which makes producing them even more lucrative.

Exploiting cognitive biases

AI-generated reviews systematically exploit human cognitive biases:

- Availability heuristic: Detailed reviews seem more credible because they generate vivid images in our mind
- Confirmation bias: Reviews that confirm our prior beliefs are scrutinized less critically
- Social proof: The sheer number of positive reviews suggests quality, regardless of their authenticity


Cognitive Biases in AI Reviews
+1
Social Proof
A large number of positive reviews suggests quality
Availability Heuristic
Detailed reviews evoke vivid images that increase perceived credibility
Confirmation Bias
Reviews that confirm prior asisumptions are questioned less critically

The unequal fight: generation is miles ahead of detection

The study reveals a fundamental asymmetry: AI’s ability to generate human-like text is miles ahead of its ability to recognize such text.

It’s an arms race that the forgers have already won. It is as if a counterfeiter were given a state-of-the-art printer to print money, while the police only had a 19th-century magnifying glass for verification.

3. The tremor in the market—The consequences for all of us

The insights from the investigation by Weiyao Meng and his team are far more than an academic footnote. They shake the foundations of the digital marketplace. Let us consider the effects on the four central stakeholder groups.

For us as consumers, the promise of online reviews was revolutionary: it democratized product information. No longer was it only the manufacturer with glossy brochures who had a voice, but a community of users. This promise has now been broken.

The psychological consequences:

- *Cognitive load and decision fatigue:* Trying to distinguish genuine from fake reviews is an enormous mental effort that, as the study shows, is doomed to fail. The result is severe fatigue. Imagine having to analyze the ingredient list yourself for every product in the supermarket to see whether it is accurate. At some point, you give up. In e-commerce, this leads consumers either to spend hours researching and end up frustrated anyway, or to capitulate and buy on a whim—or to stick only to well-known brands.

- *Erosion of fundamental trust:* The core problem is not the single false review but the doubt cast on all reviews. If we know we could be deceived, but have no way of recognizing the deception, trust in the entire system collapses. This can lead to a more general retreat from information-based online shopping.

- *Learned helplessness:* When our efforts to uncover the truth repeatedly lead nowhere, we develop what psychology calls “learned helplessness.” We resign ourselves to the uncontrollable situation. This makes us even more vulnerable to manipulation because we have given up the fight.

What remains for the consumer?

The study suggests that only verified-purchase badges (“Verified Purchase”) still offer a remnant of credibility. But even here, caution is warranted, because fraudsters are trying to circumvent these signals as well. The advice to consumers must be: be radically skeptical. Do not rely on a single source. Seek test reports from reputable trade magazines, video reviews by trustworthy YouTubers, or ask in your circle of friends. The “wisdom of the crowd” is currently a mirage.

Imagine you have spent years developing an outstanding product. It is durable, fairly produced, and superior to the competition in every respect. In the past, you could rely on this quality to be reflected in honest, positive customer reviews and thus to lead to market success. That causality is now suspended.

The strategic and psychological challenges:

- *Distorted competition:* A competitor with an inferior product can use a few hundred euros to have thousands of perfect five-star reviews generated by AI. Within days, their product can appear on sales platforms as a “bestseller” or “customer favorite.” Your high-quality product gets lost in the crowd. The study shows that these AI-generated positive reviews are hardly recognized by people as fakes.

- *Reputation sabotage:* The same technology can be used to generate targeted fake yet plausible-sounding one-star reviews for your product. “The battery exploded after two weeks!”—written by an AI. Attempting to refute such defamation is a futile struggle that ties up enormous resources.

- *Psychological demotivation:* For entrepreneurs and their teams, this situation is profoundly frustrating. When hard work and high quality are no longer rewarded because the market is manipulated by fakes, motivation and innovative strength are undermined. One feels at the mercy of an unfair, invisible adversary.

AI-driven Manipulaton Undermines Market Integrity

Distorted Competition

Fake reviews promote subpar products

Reputation-Sabotage

AI generates credible negative reviews

Psychologiccal Demotivation

Unfair market stifles innovation

Implications for service providers: trust as capital

Service providers such as restaurants or consultants live by reviews. Meng et al. (2025) show that AI fakes are particularly dangerous for services because they are emotional and contextual. Psychologically, negativity bias makes negative fakes especially destructive—one bad review outweighs more than five positive ones (Mudambi & Schuff, 2010).

Which strategies remain?

Honest firms must diversify their trust-building. Focusing solely on star ratings is risky. The following will become more important:

- *Community building:* A loyal community on social media or in your own forums in which real customers interact.
- *Maximum transparency:* An open approach to criticism, publishing unvarnished testimonials, and providing behind-the-scenes insights.
- *Brand building:* A strong brand that stands for values and reliability becomes the most important anchor in an ocean of uncertainty.

For fraudsters, manufacturers of cheap throwaway products, and aggressive marketing agencies, the conditions described in the study are a paradise. It has never been so easy, so inexpensive, and so low-risk to manipulate public opinion and influence purchasing decisions at scale.

The AI models are the “hidden persuaders,” as the study’s title aptly puts it. They are an army of tireless, perfectly phrased, psychologically trained salespeople working around the clock.

Exploiting cognitive shortcuts: The paper mentions the Heuristic–Systematic Model from psychology, which posits that people process information either systematically (slowly, deliberatively) or heuristically (quickly, intuitively). In the hustle of online shopping, we almost always use the heuristic path. A high star rating is one such heuristic.

Manipulators know this and attack right there. They flood the system with the perfect signals for our autopilot brain. Scalability and costs: In the past, writing fake reviews was manual labor and relatively expensive. Today, a single person with an AI tool can generate thousands of unique, persuasive reviews in an hour. The costs are marginal, the potential gains enormous.

Lack of consequences: Since the fakes, as the study shows, cannot be reliably detected even by the platforms, the risk of discovery is minimal.

For platforms such as Amazon, Booking.com, Google Maps, or eBay, the situation is existential. Their entire business model is based on users’ trust in user-generated content.

If reviews become worthless, the platform loses its central function as a trustworthy intermediary. Cory Doctorow aptly called this process of platform quality degradation “enshittification.”

The platforms’ dilemma

- *Technological powerlessness:* As the study shows, their own AI-based detection systems are nearly blind. They are fighting an invisible adversary with a magnifying glass.

- *Caught between both fronts:* If platforms crack down hard and delete large volumes of suspicious reviews, they risk the wrath of sellers (honest and dishonest alike), who are their main source of revenue. If they do nothing, they lose trust—and hence buyers. It is a quandary.

- *Credibility crisis:* If users feel that the platform is either unable or unwilling to solve the problem, they turn away. The platform becomes a cluttered bargain bin that people no longer trust.

Possible ways out for platforms

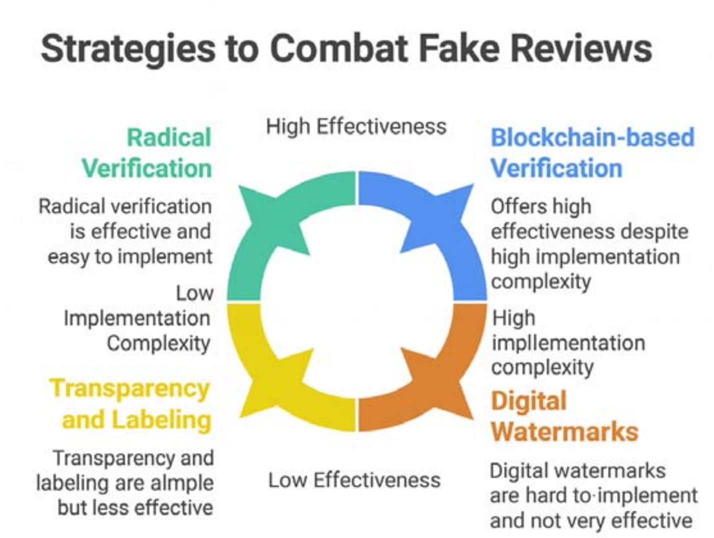
The researchers suggest in their conclusion that detection technology alone is a dead end. The solution must begin at the source:

- *Radical verification:* Only people who can prove that they purchased a product may leave a review. This would not make fake reviews impossible, but it would make them significantly harder.

- *Transparency and labeling:* AI-generated content (e.g., summaries of reviews) must be clearly labeled as such.

- *Digital watermarks:* Although the technology is not yet mature for short texts like reviews, research into invisible markers for AI text could be a future solution.

- *Blockchain-based verification:* Blockchain technology could enable tamper-proof documentation of purchases and reviews. Each review would be cryptographically linked to a verified purchase.


Strategies to Combat Fake Reviews
High Effectiveness
Radical Verification
Radical verification is effective and easy to implement
Blockchain-based Verification
Offers high effectiveness despite high implementation complexity
Low Implementation Complexity
High impllementation complexity
Transparency and Labeling
Transparency and labeling are almple but less effective
Digital Watermarks
Digital watermarks are hard to implement and not very effective
Low Effectiveness

Recommendations for action

For consumers:

- *Diversify your information sources:* Do not rely solely on online reviews. Also use lab tests, product videos, social media, and recommendations from your network.

- *Maintain a skeptical baseline:* Be especially critical of patterns of predominantly positive or predominantly negative reviews. Genuine product reviews usually show a certain spread.

- *Quality before quantity:* Pay more attention to detailed, balanced reviews than to the sheer number of reviews.

- *Consider temporal distribution:* Be wary of products that accumulate a large number of reviews very quickly—this could indicate organized manipulation.

For companies:

- *Authenticity as a competitive advantage:* Rely on genuine customer satisfaction instead of fake reviews. In the long term, authenticity pays off.
  Proactive communication: Inform your customers about the challenge of fake reviews and what you are doing to counter it.

- *Invest in quality:* The best defense against negative fake reviews is a truly good product that generates positive, genuine reviews.

- *Alternative trust signals:* Develop other ways to build trust: money-back guarantees, certifications, transparency in production.

4. Looking ahead—Navigating a post-authentic world

The study by Meng et al. is more than a wake-up call. It is the official confirmation that the old paradigm of "user-generated trust" has ended. We have arrived in a post-authentic era of e-commerce.
What does that mean for the future?

From a psychological perspective, we are experiencing an inflation of trust. Just as money loses value when too much of it is printed, the written word online loses value when it can be produced arbitrarily and at no cost.
What we need now is a paradigm shift:

- *For consumers:* A new form of digital literacy. We must learn to live with uncertainty. The ability to assess the credibility of sources will become the most important competence in the digital sphere. Do not rely on stars anymore, but on brands, experts, and genuine social networks.

- *For companies:* Authenticity as the hardest currency. In a world full of perfect AI fakes, genuine, imperfect humanity becomes the most valuable asset. Companies that manage to build an honest, transparent, and direct relationship with their customers will be the winners. The brand will again become more important than the review.

- *For platforms and regulators:* Rules instead of a futile arms race. Technology alone will not save us. We need clear regulatory frameworks. The initiative by the British government mentioned in the paper—to outlaw fake reviews by statute—is an important first step. But we need global standards for verification and transparency.

The era of naïve confidence in online reviews is over. The AI "hidden persuaders" are among us, and they are invisible. It is now up to all of us—consumers, companies, and platforms—to draw the consequences and to find new ways to re-establish in the digital market what must stand at the beginning of everything: genuine, earned trust.

References

Bandura, A. (1977). Social learning theory. Prentice Hall.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. Journal of Personality and Social Psychology, 39(5), 752–766.

Cialdini, R. B. (2001). Influence: Science and practice (4th ed.). Allyn & Bacon.

Chevalier, J. A., & Mayzlin, D. (2003). The effect of word of mouth on sales: Online book reviews. NBER Working Paper.

Dathathri, S., et al. (2024). Scalable watermarking for identifying large language model outputs. Nature.

Doctorow, C. (2023). TikTok's enshittification. Pluralistic: Daily Links from Cory Doctorow.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. Decision Support Systems, 45(4), 1007–1016.

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2022). Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory & Cognition, 38(8), 1087–1100.

European Commission. (2024). Unfair commercial practices directive – overview & updates. https://commission.europa.eu/law/law-topic/consumer-protection-law/unfair-commercial-practices-and-price-indication/unfair-commercial-practices-directive_en

EUR-Lex. (2019). Directive (EU) 2019/2161 (Omnibus). https://eur-lex.europa.eu/eli/dir/2019/2161/oj/eng

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, T. (2014). How online product reviews affect retail sales: A meta-analysis. Journal of Retailing, 90(2), 217–232.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. Information Systems Research, 19(3), 291–313.

Gandhi, A., Hollenbeck, B., & Li, Z. (2024/2025). The equilibrium effects of fake reviews on Amazon.com. Working paper.

Hardin, G. (1968). The tragedy of the commons. Science, 162(3859), 1243–1248.

Hu, N., Liu, L., & Zhang, J. J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. Information Technology and Management, 9, 201–214.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. NeurIPS Poster. https://openreview.net/forum?id=WbFhFvjjKj

Lang, A. (2000). The limited capacity model of mediated message processing. Journal of Communication, 50(1), 46–70.

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. Journal of Consumer Research, 15(3), 374–378.

Liang, W., et al. (2023). GPT detectors are biased against non-native English writers. Patterns.

Li, H., et al. (2015). Analyzing and detecting opinion spam on a large-scale dataset. ICWSM.

Meng, W., Harvey, J., Goulding, J., Carter, C. J., Lukinova, E., Smith, A., Frobisher, P., Forrest, M., & Nica-Avram, G. (2025). Large Language Models as “Hidden Persuaders”: Fake product reviews are indistinguishable to humans and machines. arXiv. https://arxiv.org/abs/2506.13313

Mukherjee, A., et al. (2013). Spotting opinion spammers using behavioral footprints. KDD.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? MIS Quarterly, 34(1), 185–200.

OECD/BEUC (2025). How to make online reviews more reliable? https://www.beuc.eu/sites/default/files/publications/BEUC-X-2025-027_how_to_make_online_reviews_more_reliable.pdf

Packard, V. (1957). The hidden persuaders. David McKay Co.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. Advances in Experimental Social Psychology, 19, 123–205.

Shukla, A. D., & Goh, J. M. (2024). Fighting fake reviews: Authenticated anonymous reviews using identity verification. Business Horizons, 67(1), 71–81.

Wang, Y., Wang, J., & Yao, T. (2019). What makes a helpful online review? A meta-analysis of review characteristics. Electronic Commerce Research, 19, 257–284.

Zhou, Y., et al. (2024). Evading AI-text detection through adversarial attack. LREC-COLING.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. Journal of Marketing, 74(2), 133–148.