Harvard Business Review
ANALYTIC SERVICES

White Paper

PROTECTING BRANDS FROM MALICIOUS MISINFORMATION

Sponsored by

Teleperformance

DANIEL JULIEN

CHAIRMAN AND CHIEF EXECUTIVE OFFICER

TELEPERFORMANCE

Almost 3.5 billion people are active social media users, spending more than two hours a day generating massive digital content. Our current love affair with social media is accelerating at an unprecedented pace. According to Domo's Data Never Sleeps report, these social media metrics are generated every minute of the day:

- 390,030 apps downloaded
- 4,500,000 YouTube videos viewed
- 511,200 tweets
- 277,777 Instagram stories

Beyond the numbers is the fact that many of these videos, photos, and tweets contain content that is fluid, polymorphic, and erratic. It's very difficult to separate the wheat from the chaff.

Social media is an integrated part of human life today. It keeps us connected, enlarges our communities, makes us more aware, and is a formidable source of information and education. But like any powerful invention in the digital era, social media is also used to spread fake news, defame individuals, diffuse hate speech, promote terrorism, and, on more than one occasion, brainwash the impressionable to become religious fanatics or commit hate crimes.

The amount of user-generated data shared is both constructive and destructive in nature, meaning social media needs to have a reliable and efficient gatekeeping system to filter out the latter in order to protect the users.

But how can we control 2.5 quintillion bytes of data shared by 3.5 billion users across the world?

By merging cutting-edge technology with the human touch, we can isolate online risks through content moderation.

When it comes to easily recognizable illegal, criminal, and vicious content, artificial intelligence (AI) is the perfect choice. But as I always say, no matter how advanced the technology is, it lacks humanness. AI does not understand human emotions, culture, and context. As a result, after the initial triage made by AI-led content moderation programs, social media gatekeepers have to decide whether content is accurate, legal, and acceptable.

The role of content moderators is a difficult one. These social media gatekeepers have to review large quantities of data and are exposed to disturbing, shocking, and violent content. They are like the social web's field cops and emergency services professionals. They are deeply needed but hard to find. A high sense of ethics is required to value service for the community, as well as resilience to handle the stress that goes with the mission.

Fortunately, social media companies and content moderation firms are extremely conscious of the importance of protecting users, and have put policies in place to better guide their social media gatekeepers, including:

- Selecting content moderators based on stable and resilient psychological profiles
- Reducing effective working hours versus current labor practices
- Providing permanent psychological support
- Offering a better work environment
- Giving higher compensation than traditional customer service jobs

A content moderator's job is definitely not a sinecure. Content moderators need to be taken care of. I am sure that many cops, emergency rescue personnel, veterans, or security officers would appreciate getting the same working conditions and care provided to their counterparts in the virtual world.

To all the social media gatekeeper teams around the world, a great thank you for your service!

PROTECTING BRANDS FROM MALICIOUS MISINFORMATION

INTRODUCTION

In 2017, a prankster in the U.K. created a fake London restaurant called The Shed. Pictures of its cuisine were made from round toilet bowl deodorizers and cropped shots of people's heels. Reservations were by appointment only, and the restaurant always seemed booked. It served a meal only once to see how far the ruse would go. The main dish was frozen lasagna. But through a carefully orchestrated campaign of phony user-generated content such as reviews, the fake restaurant soared some 15,000 spots on a review site and ended up as London's top-ranked eatery.[1]

The comical nature of the prank notwithstanding, the tale of The Shed demonstrates how easy it is to promulgate fake news, reviews, and offensive content to manipulate beliefs about a business. Many of the incidents aren't funny, however. In 2019, a phony video showed a supposedly self-driving Tesla car catching on fire after hitting a robot. Although Tesla had no self-driving cars, its stock price took a hit nonetheless.[2] Another misinformation campaign struck at Coca-Cola's Dasani bottled water. A social media campaign claimed that Coca-Cola was recalling Dasani because it contained dangerous parasites. Another malicious attack alleged that an Xbox video console exploded and killed its teenage user.[3]

Although user-generated content has been a boon for business, it has a dark side that is increasingly rearing its head. Fake news, fake reviews, and offensive content are on the rise and can severely damage a brand's reputation among customers, investors, and other critical constituencies.

Despite the danger of fake content, combating it is still low on executive agendas. Rapid advances in technology are lulling many into believing that artificial intelligence (AI) and other software will be able to rid the world of fake and offensive content before it spreads. However, putting all the eggs in the technology basket can put organizations in harm's way. Humans are the only ones who can really understand important nuances such as dialect and cultural context. Only people can teach AI what to learn as bad actors come up with ever more devious ways to fool the technology. Technology alone can let fake content slip through the cracks.

HIGHLIGHTS

- Fake news and other malicious content targeting businesses are on the rise and wreaking havoc on everything, from sales to stock prices.
- Business leaders put too much stock in technology to screen malicious content and expect too much from online media and platform companies to control the problem. Technology can't catch everything, and there are practical limitations to what platform companies can do.
- Corporations need to moderate user-generated content on their own and implement potent defensive and offensive strategies to protect their brands.

"Corporations spend hundreds of millions—even billions—to develop their brands, but they **often devote an almost infinitesimal percentage of that amount** to protect them," says Mike Paul, president of public affairs at Reputation Doctor.

In politics, fake news and offensive content are running rampant. It is still the early days for corporations, and many are yet to feel the heat. But governments, media businesses, and platform companies will be able to do only so much to protect companies. Corporations need to be at the ready for these attacks, which are clearly on the rise. Organizations need a fine-tuned machine combining technology and human intervention to fight the increasing incidence of phony content designed to wreak havoc on a company's reputation.

Ignoring the Storm on the Horizon

Mike Paul, public relations specialist and president of the public affairs consultancy Reputation Doctor, rarely sees executives express strong concerns about fake content—until it's too late. As a result, many of the challenges become full-blown crises. "It is very similar to the early days of cyber attacks," he says. "Even when the number and seriousness of these attacks were on the uptick, many business leaders had their heads in the sand and hoped the problem would just go away. So when the number of big cyber attacks started growing, corporations were caught completely off guard."

Paul also points to a huge schism between what major corporations invest to build brands versus how much they spend to protect them. "Corporations spend hundreds of millions—even billions—to develop their brands," he says. "But they often devote an almost infinitesimal percentage of that amount to protect them. Most businesses aren't even aware of their vulnerabilities to fake content and might not even know how to go about identifying them."

Most companies already have their hands full trying to protect personal data and fending off cyber attacks, according to Harlan Loeb, global chair of risk and reputation management at the public relations agency Edelman. "Retail is an excellent example," he says. "They have thousands of data points about each customer that they have to protect. Hackers have already seized mountains of that data."

Platform companies such as Facebook, YouTube, and Reddit are doubling down on eliminating hate speech and harmful content. They aren't as focused on fake news targeting businesses. Thus executives may be putting too much stock in their ability to moderate misinformation targeting companies. "Online media and platform companies are more concerned about content that incites violence or harms elections," says Aviv Ovadya, founder of the Thoughtful Technology Project and a fake news researcher. "They rarely have public policies to address phony content targeted at businesses."

Moreover, Vince Sollitto, senior vice president of global communications and public policy at Yelp, says media and platform companies are not always comfortable determining what should be banned. "The content may be untrue but might also be fiction or satire," he says. "It can be difficult to

make that call with billions of pieces of user-generated content flowing online every day."

Platform companies also don't have the incentives needed for them to play a much bigger role. "Their business, unlike Yelp's, is to engage users for as long as they can," says Sollitto. "Thus, they are only going to take action when users object and engagement declines."

The Onslaught of Fake Content About Businesses

Although fake news in politics captures the most attention, attacks on corporations are mounting and bad actors have both consumer and financial markets in their sights. In 2015, a website designed to look like Bloomberg.com posted a letter announcing a $31 billion takeover bid for Twitter. The stock price soared, and the motive seems to have been unscrupulous investors planning to sell their Twitter shares as more legitimate investors drove the price up and lost their shirts.

Devious short sellers are also hiding online. Broadcom's stock price began to fall after a fake memo circulated on the internet purporting to be from the U.S. Department of Defense. The memorandum claimed that the agency was investigating national security risks posed by Broadcom's actual $19 billion bid for CA Technologies. CA Technologies' stock sank along with Broadcom's until the short-selling scheme was finally exposed.[4]

Products and services are also on the hit list. In 2017, an anonymous man launched a particularly malicious campaign against Starbucks in order to damage a "liberal" company. The perpetrator created a Twitter campaign claiming that Starbucks was promoting a "Dreamer Day" event and giving free coffee to illegal immigrants.

Businesses themselves are starting to target their competitors with fake news. Sollitto says that competitors are the biggest source of fake reviews. They are trying to manipulate Yelp to disparage competitors or inflate their own reputations. A major electronics company was caught hiring customers to write bad reviews of its competitor's products and services. The company was socked with a multimillion-dollar fine. The phony Tesla video was apparently the creation of agents in Russia acting on behalf of Russian interests trying to derail the U.S. electric vehicle market.

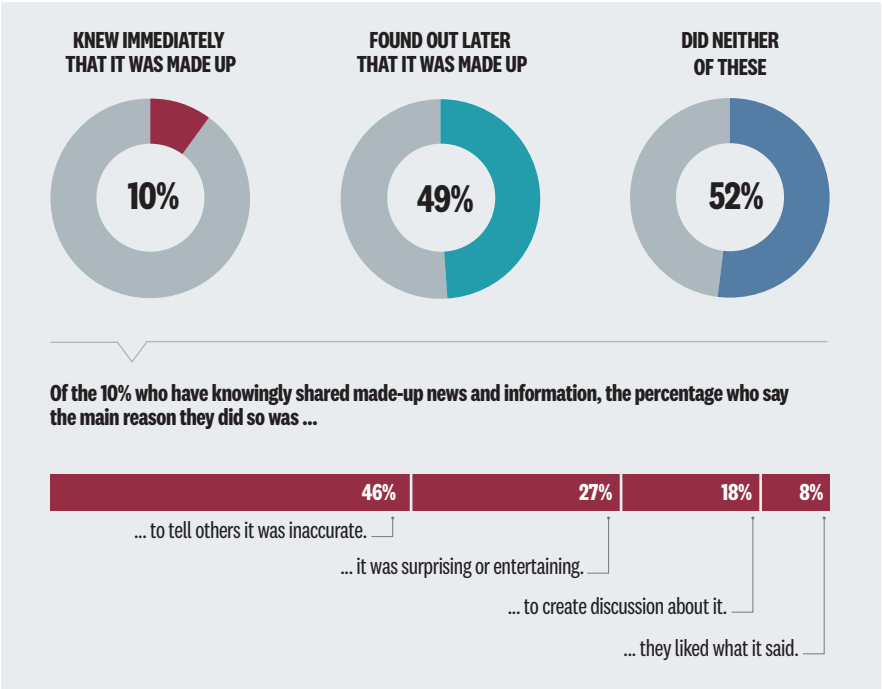
People Can Be Duped

Phony content works because people can be manipulated. According to a recent study by the Pew Research Center, more than 50% of individuals have unknowingly shared fake news and information. FIGURE 1 The 2018 annual Edelman Trust Barometer found that two-thirds of individuals

FIGURE 1

FALLING PREY TO FAKE CONTENT

About half of U.S. households have unknowingly shared phony content.

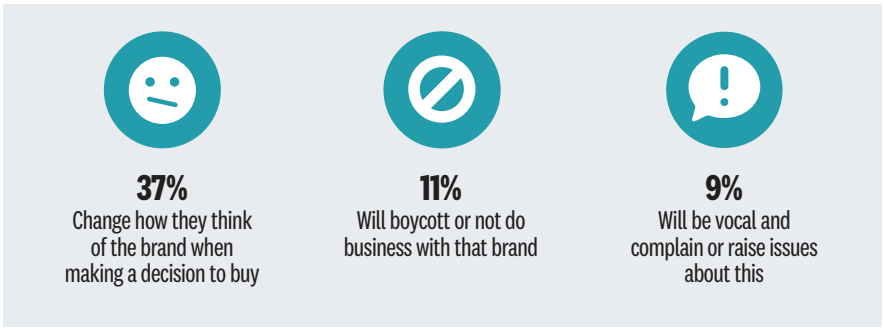


SOURCE: PEW RESEARCH CENTER 2019

FIGURE 2

BRAND TRUST IS AT RISK

Consumers lose trust in brands that advertise near offensive content.



SOURCE: 2018 CMO COUNCIL STUDY

don't care if the facts in a piece of content are correct or not.

Equally daunting, most individuals believe that advertisers know and approve of where their content ends up.[5] According to a 2018 study by the CMO Council, when individuals see the ads placed on offensive or dangerous websites, their opinions of the company will decline. In addition, a significant majority will change buying decisions, boycott a product, or voice their concerns to all who will listen. FIGURE 2

Protecting the Brand

Given the limitations of platform and media companies, businesses would be well-advised to pick up the mantle to protect their brands from malicious misinformation. Although technology isn't the only solution, it can play a significant role.

Most advanced content moderation technologies use image recognition to spot pornography and violent crime. These are among the most prevalent forms of offensive content and a major driver of public fear and rancor. However, technologies are advancing that can scrutinize more varied forms of content, including the written word.

For example, while she was a professor at the Yale School of Management, Marina Niessner was interested in how fake news could be used to influence financial markets. She focused on news stories covering small-cap companies listed on the New York Stock Exchange. To determine which articles might be fake on crowd-sourcing sites such as Seeking Alpha, Niessner and her team turned to the latest in AI. "We had access to very sophisticated algorithms that were able to pretty accurately distinguish phony articles from true ones," she says. "It wasn't 100% accurate, but much better than random guessing."

The software didn't check facts or statements or look for obvious signs such as grammatical errors. Instead, the algorithms identified phony content by identifying the intent to deceive in the writing. "The software works much like a lie detector," she says. "It can detect news that is likely fake by the emotional nature of how it is written."

Patrick Hillmann, executive vice president of crisis and risk management at Edelman, describes another innovative use of AI. Banning users from a website can be difficult. So instead of trying to keep bad actors off sites, some companies are letting them continue to post content and using AI software to learn how they are trying to mislead and why people are responding. That information becomes a gold mine of knowledge in the quest to find and intervene in further instances of fake or offensive content.

During the 2016 U.S. presidential election campaigns, purveyors of fake news used social media analytics to pinpoint exactly who was most likely to believe the misinformation. Social media and data about lifestyles and personal interests can paint a very detailed portrait of target audiences. Reputable organizations can use these same data sets to identify who are the most likely to believe fake news written about the company and why.

In a 2019 episode of the CBS news magazine *60 Minutes*, Hany Farid, a digital forensics expert and professor at the University of California, Berkeley, says that the sophistication

"THE CONTENT MAY BE UNTRUE BUT MIGHT ALSO BE FICTION OR SATIRE. IT CAN BE DIFFICULT TO MAKE THAT CALL WITH BILLIONS OF PIECES OF USER-GENERATED CONTENT FLOWING ONLINE EVERY DAY." VINCE SOLLITTO, SVP, YELP

"... human moderators **determine if the content is really fake or offensive** and flags it to be investigated," says Harlan Loeb at public relations agency Edelman.

of so-called deep fakes—videos that use AI to make them look like someone is saying something they aren't—has reached an alarming stage: many are now undetectable to the human eye. Currently, there are AI applications coming on the market that analyze thousands of facial images and videos to detect slight errors in deep fakes. Subtle hints such as eyebrows not moving when they usually do can give away the fake.

The Human Factor

New technologies are providing eye-opening capabilities and will no doubt continue to do so. Nonetheless, even artificial intelligence can't supplant human moderators and the knowledge they bring to the table.

AI deep learning software can learn a great deal on its own, which goes a long way in spotting fake content. But these applications discover only what people have taught them to learn. Sarah Roberts, an assistant professor at the University of California, Los Angeles, and the author of *Behind the Screen: Content Moderation in the Shadows of Social Media*, stresses that deep learning algorithms must be "trained" with sample data sets in order to understand what they are looking for. For example, AI applications can use image recognition to identify pornography. But pornographers keep coming up with new ways to conceal their content, such as labeling it in Arabic. AI can learn Arabic, but someone has to let it know to look for Arabic and how to learn about how pornographers are using it.

Humans also play a crucial role in identifying nuances that software may never be able to detect. Understanding these subtleties is essential to focus on actual fake content and not chase and challenge content that is simply misunderstood. For example, people deeply familiar with nuance such as sarcasm can understand the actual intent of the content.

The same holds true for cultural contexts. In the U.S., for example, colloquial expressions such as "sick"—meaning "great"—are constantly entering the language. In the U.K., the phrase "do away with" is a figure of speech used in an ironic sense. In other countries, it could constitute a deadly threat. Dialects make the challenge even greater. Virtually every language has dozens, if not hundreds, of dialects, and native speakers often don't understand them all. Lack of indigenous language and cultural knowledge can be a large chink in the armor.

Industry and technical knowledge are also best left to people. Fake content attacks on corporations can target industries in which specific knowledge is needed that AI may not be able to see through. When Niessner was conducting her research into stock market manipulation, biotechnology was one of the industries the team zeroed in on. "Much of the content was highly specialized and technical," she recalls. "Companies may need experts to determine what is actually fake."

A hybrid model of technology and people is the best strategy, according to Loeb. "Technology is the first line of defense that flags potentially problematic content," he says. "Then human moderators determine if the content is really fake or offensive and flags it to be investigated."

Playing a Defensive and Offensive Game

The adage that a good offense is the best defense holds true for content moderation. In addition to being ready to fight an attack, companies can and should take strong preventive measures that minimize the impact.

As the Reputation Doctor's Paul pointed out above, business leaders should treat content strategy much like they approach cybersecurity. The first step is to understand where the company's vulnerabilities lie. In the case of fake and offensive content, the biggest vulnerability may be the company's content—or lack of it. "The best offense is to have a critical mass of positive, authentic, and ethical content in the market," says Paul. "Convincingly positive material can serve as a shield against anyone trying to attack the brand."

As a hypothetical example, Paul often asks executives and their boards to imagine they are heads of a breakfast cereal company. A fake video suddenly appears claiming that insects such as flies were found in the cereal during manufacturing. If the company has a strong following, many are probably likely to have been customers since childhood. But for those customers who aren't, the business needs to have engaged them about its values around nutrition and food safety. If the business has become an exemplar of these values, customers are far less likely to believe the fake story from the get-go.

A critical mass of convincing material is only the first step, however. Businesses need to know who their fans are—customers, investors, and other stakeholders—and what makes them tick. If an individual has negative feelings about a brand or even the category, companies should know and work to change their opinions.

Hillmann says that media companies have become experts at identifying fans and reaching out to them. They segment markets down to the level of individual customers and develop deep relationships. They know which content customers consume and how broadly they share it. This information can be combined with behavioral data to identify and predict people's reactions and communicate directly with fans and their communities. The value of the work can't be underestimated: Edelman's Loeb says that the firm's proprietary research discovered that product loyalists don't take the fake news bait. It doesn't affect their buying decisions.

Loeb also stresses that businesses can leverage the faith and interest of their fans. Ovadya echoes the sentiment and points out that it was consumers who discovered the offensive websites where company commercials ended up and reported it to the brands. Paul adds that an organization's sales force or call center can be good listening posts. "These employees are speaking directly with customers and prospects," he says. "Customers will almost always report what they are hearing about a business and its offerings that could breach their trust. Trust is the ultimate currency of brand and reputation."

To guide the creation of an effective offense, Paul recommends that business leaders engage with professionals who understand the dark web and what is looming on the horizon. These professionals can range from SEO experts to former government online intelligence investigators. "You need to think deeper and darker than you are now," says Paul. "What you think is the worst-case scenario right now is probably much rosier than what could

BUSINESS LEADERS SHOULD TREAT CONTENT STRATEGY MUCH LIKE THEY APPROACH CYBERSECURITY. THE FIRST STEP IS TO UNDERSTAND WHERE THE COMPANY'S VULNERABILITIES LIE.

A CORNERSTONE OF MANAGING THE RISK OF FAKE CONTENT IS A DEEP UNDERSTANDING OF CUSTOMERS AND MEANINGFUL CONVERSATIONS WITH THEM.

happen. Those working as criminals on the dark web are ruthless and will commit their crimes by any means necessary."

A strong offense will go a long way in protecting a brand from false and damaging information. But companies can't stop all fake content from appearing and spreading. They need a good defensive strategy. There are three main components to launching a counterattack.

The first step, according to Loeb, is to gather all the facts. "You need to go to the beginning of the chain of how the content was created," he says. "Businesses need to uncover all the misinformation used and counter each point with facts that prove unequivocally that the content is false." To add heft to the counteroffensive, business leaders can turn to acknowledged experts and information sources to challenge the veracity of misinformation.

Next, Loeb recommends that organizations try to find out who is behind the content and discredit them. "I used to be a trial lawyer," he says. "A critical strategy was to impeach the witness. You need to show why they are not credible."

As humans moderate content with the help of machines, they need to implement measures that categorize the urgency needed to deal with the fake content found. The categorization can be as simple as yellow and red. What is important is understanding the potential impact of the content on fans and other important constituencies.

Before launching an attack, business leaders should choose their battles wisely. "You need to do no harm," says Hillmann. "That is why it is so important to understand who your brand supporters are and what they are likely to believe. If the fake news won't affect them, it may be better to leave it alone and stop the news cycle."

It's Customer Intimacy in the End

Defending an organization against fake news and offensive content is no small task. However, the investment pays off in more than one way. A cornerstone of managing the risk of fake content is a deep understanding of customers and meaningful conversations with them. To build their support requires content and an idea-rich dialogue—the essential drivers of brand loyalty.

The ability to reach fans is also paramount. Bad actors in politics avail themselves of sophisticated data and analytics to target exactly those who they can influence. Many brands, on the other hand, don't necessarily know who their fans are nor how susceptible they are to different types of misinformation. The ensuing customer knowledge can be enormously valuable in bolstering loyalty. At the same time, it is also a powerful offense against the mounting threat of fake and harmful content.

"The ultimate irony of all this is that connections with customers are exactly what brands are seeking as a core part of their business," says Loeb. "Deep relationships foster trust while immunizing audiences against malicious misinformation."

Endnotes

1 https://www.washingtonpost.com/news/food/wp/2017/12/08/it-was-londons-top-rated-restaurant-just-one-problem-it-didnt-exist/

2 https://www.nbcnews.com/business/business-news/fake-news-can-cause-irreversible-damage-companies-sink-their-stock-n995436

3 https://www.washingtonpost.com/outlook/fake-news-threatens-our-businesses-not-just-our-politics/2019/02/08/f669b62c-2b1f-11e9-984d-9b8fba003e81_story.html?noredirect=on

4 https://www.washingtonpost.com/outlook/fake-news-threatens-our-businesses-not-just-our-politics/2019/02/08/f669b62c-2b1f-11e9-984d-9b8fba003e81_story.html

5 https://www.cnbc.com/2017/03/24/google-ad-scandal-how-companies-buy-youtube-and-google-display-ads.html

Harvard Business Review
ANALYTIC SERVICES

hbr.org/hbr-analytic-services

CONTACT US

hbranalyticservices@hbr.org