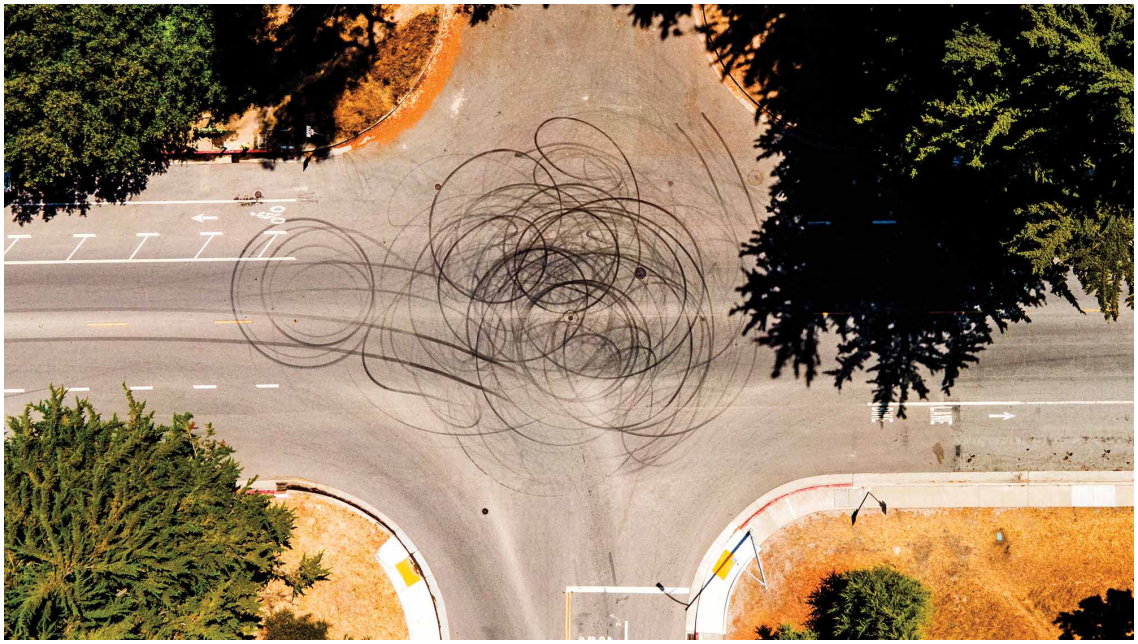
HARVARD BUSINESS REVIEW

Marketing

Don’t Let an AI Failure Harm Your Brand

by Julian De Freitas

From the Magazine (July–August 2025)



Winni Wintermeyer

Summary. AI is being adopted in everything from automobiles to chatbots, but AI products—like all products—will eventually fail. And *how* a company markets its AI systems affects the repercussions it will face when a failure... **more**

In October 2023 an automated vehicle (AV) operated by Cruise, a robotaxi subsidiary of General Motors, was involved in a serious accident in San Francisco. A Nissan, driven by a human, struck a pedestrian, who was thrown into the AV’s path. According to an independent engineering consultant’s investigation of the accident, no prudent human driver under

those circumstances would have been able to steer the AV to avert the crash. But Cruise’s initial report to regulators omitted the fact that the pedestrian was then dragged underneath the AV for 20 feet. She suffered severe injuries but survived.

Even though Cruise was not at fault for the collision, the accident set off a crisis at the company. Its failure to be transparent led to a $1.5 million fine by the National Highway Traffic Safety Administration. It also sparked a criminal investigation by the Department of Justice; the firm ultimately settled the lawsuit and paid a $500,000 fine. But that’s not all. Its permit to operate in San Francisco was revoked. Half of Cruise’s workforce lost their positions, the CEO stepped down, and the company’s valuation dropped by more than 50%. The broader AV sector also felt tremors. Within months, a driverless Waymo taxi (owned by Alphabet) was attacked and set on fire in San Francisco by a crowd, and the NHTSA opened investigations into multiple AV developers, including Waymo and Zoox (owned by Amazon). By the end of 2024, GM announced that it would end development of its robotaxi business altogether.

AI is being rapidly adopted in everything from automobiles to chatbots, but the Cruise example highlights a stark reality: Eventually, AI fails. When it does, many organizations—whether they build their own AI systems or integrate others’—find themselves in the crosshairs of public scrutiny. Although much has been written on how to market AI to boost its adoption, less attention has been paid to how to market AI in a way that prepares for its inevitable failures.

Over the past seven years, I have conducted seven studies into the dangers of AI failures from a marketing perspective. From my research I’ve distilled insights about how consumers perceive and react to AI failures. Savvy marketers need to account for five pitfalls related to consumer attitudes—both before and after something goes wrong. In this article, I’ll examine how companies should prepare for failures—and what to do after the fact when AI misses the mark. I’ll explore the ways in which companies market their own AI and whether their tactics present risks. I’ll detail how some organizations have responded to AI failure and offer practical advice to managers who want to promote their AI systems while protecting their brands and strengthening consumer trust. Let’s look at each of the pitfalls in turn.

People Tend to Blame AI First

To better understand why the public and regulatory backlash to the Cruise incident was so severe, my colleagues and I conducted a study with more than 5,000 participants. We shared with them an accident scenario that was similar to the Cruise example—in which a human driver collided with a pedestrian, who was flung into a second, not-at-fault vehicle. One group of participants learned that the not-at-fault vehicle was autonomously driven; the other, that it was driven by a human. Then we asked participants to evaluate how responsible the manufacturer of the vehicle was.

Participants attributed greater liability to the maker of the not-at-fault vehicle when it was autonomously driven than when it was human-operated—even though in both cases the vehicle could not have done anything to prevent the accident. They also judged the AV company to be more liable than the human driver of the not-at-fault vehicle (when it was human-operated). This result has been replicated by an independent research group in China,

which demonstrates that the bias persists across cultural contexts.

Blame is mislaid in these instances because respondents, distracted by the AV’s novelty, imagine what could have happened had the AV been absent. Notably, in these cases they tend to imagine what a *perfect* driver would do, leading them to hold AI to a higher standard than is reasonable. They conclude, for example, that the AV could have prevented the collision by somehow swerving before impact, even if this would have been impossible. These imagined counterfactuals bias them toward viewing the AV as more responsible for the accident than a human in the same situation.

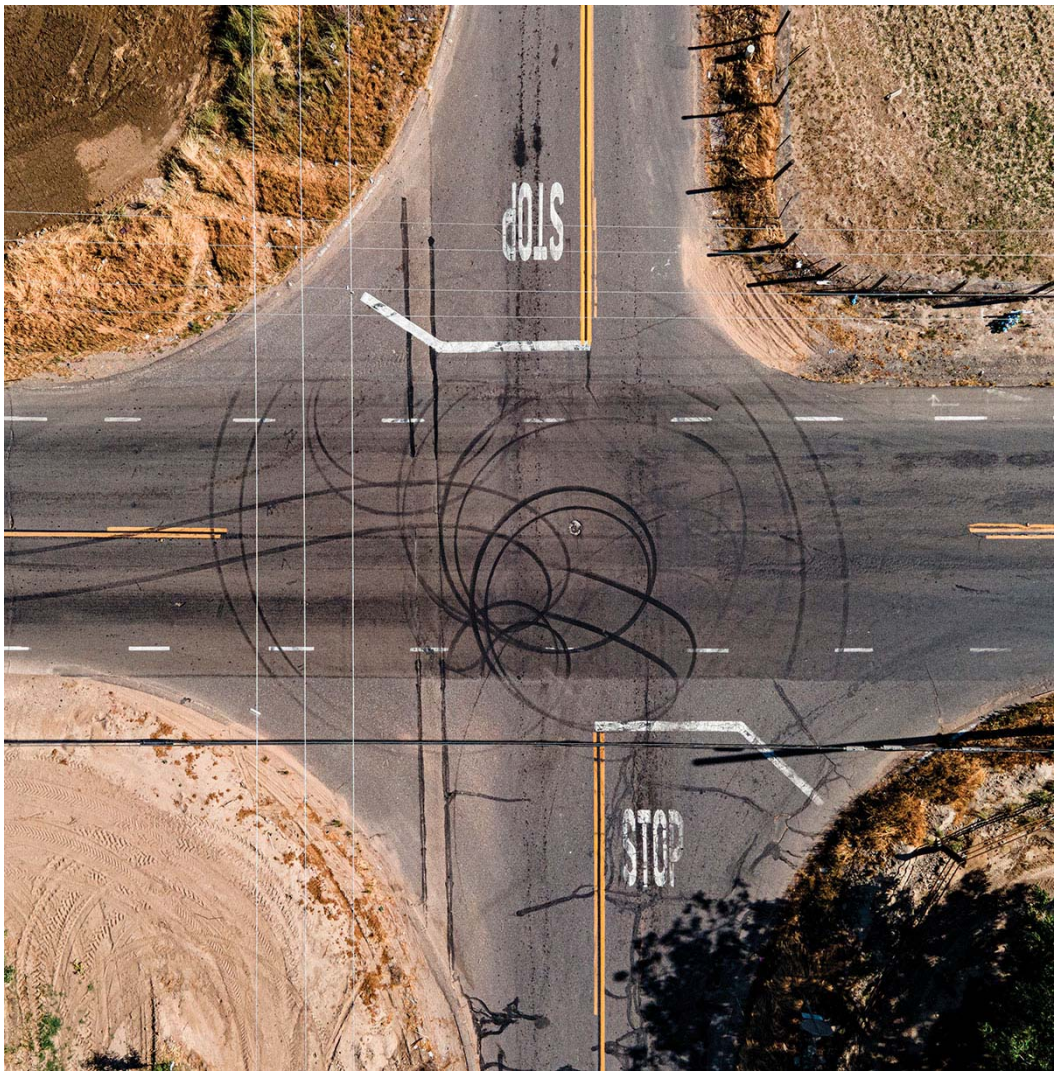
In subsequent tests, we found that shifting focus onto the at-fault human driver (the one who initially caused the accident) reduced the amount of blame on AI. When peoples’ attention was diverted from the AV’s novelty to other salient factors, they were less inclined to dream up unrealistic scenarios. But diverting attention from novelty should not be conflated with deliberately misleading stakeholders: Executives should not hide details about AI’s role in failure incidents. Although Cruise’s leaders emphasized the fault of the human driver, they failed to disclose the additional harm their vehicle caused by dragging the pedestrian. When that was discovered, Cruise lost control of the media narrative and damaged regulator trust.

When One AI Fails, People Lose Faith in Others

The ripple effects of the Cruise incident on other players like Waymo and Zoox suggest another risk associated with AI failure: When one company’s AI fails, people tend to think that the AI systems of other companies are similarly faulty. Such a contamination effect could negatively affect public perception of various forms of AI.

A study conducted by professors Chiara Longoni, Luca Cian, and Ellie Kyung offers an excellent example of how failure contamination occurs. They informed 3,724 people about a problem: The state of Arkansas failed to properly allocate disability benefits to a disabled person. They told some participants that a human employee was to blame and others that an algorithm was. Then they asked how likely it was that an employee or an algorithm from a different state (Kentucky) would also make an error in allocating disability benefits. Participants were more likely to predict that the algorithm would fail than that the human Kentuckian would. The researchers replicated this effect for various other AI failures, such as improperly allocating Social Security payments.

The respondents reacted the way they did because, the researchers found, many people don’t understand how AI works. They tend to think that AI solutions are part of a homogeneous group and share the same underlying characteristics; they don’t regard AI solutions to be distinct systems that have different capabilities and flaws.



Winni Wintermeyer captures the intricate tire patterns on intersections, observing how the marks from car burnouts turn into layered abstractions.

To avoid being contaminated by failures caused by other companies, you should highlight how your AI systems differ from those of your competitors. Emphasize differentiators, such as proprietary algorithms, safety measures, and human oversight. Consider how AI company Anthropic built and markets Claude, its generative AI model. It calls it “a next generation AI assistant… trained to be safe, accurate, and secure to help you do your best work.” Anthropic says it trains Claude using a “constitutional” approach that is more transparent, interpretable, and aligned with human values. This helps Anthropic dissociate Claude from popular competitor models like ChatGPT and Bard, which are trained differently and have been accused of being biased and generating inaccurate information. If either competitor fails,

Anthropic will have already laid the groundwork to limit potential failure contamination effects.

Another preventative measure is to communicate when a human supervises the AI and decides whether to implement the AI’s recommendation. When there is failure in such a “human in the loop” arrangement, people are less likely to assume that other AIs are similarly faulty, presumably because they are less likely to see the failure as an indictment of AI generally.

People Place More Blame on Companies That Overstate AI Capabilities

Tesla refers to its driver assistance system as Autopilot even though operating it requires active human supervision on the road. After several fatal accidents involving Teslas whose Autopilot feature was engaged, the automaker has been entangled in lawsuits, including an investigation by the Department of Justice into whether this label constitutes misleading marketing. It has also been asked by the National Highway Traffic Safety Administration to ensure that its public communications more accurately reflect the system’s capabilities and limitations. Tesla CEO Elon Musk has historically defended using the Autopilot label by noting that it was borrowed from aviation, where it is used as an aid to pilots rather than as a fully autonomous system. But do consumers interpret it that way?

My colleagues and I investigated the question of misrepresentation in a study involving 9,492 participants, using both driving simulations and hypothetical accidents. We told participants that an automaker was planning to introduce a new automated vehicle. One group was told it was labeled Autopilot (suggesting high AI capability); the other group was told it was labeled Copilot (suggesting midlevel AI capability). Participants were then placed in a simulated driving scenario involving the

vehicle and asked to take control of the wheel whenever they felt they needed to intervene. In the simulation, the vehicle approached a busy intersection and proceeded to collide with jaywalkers unless the participant intervened.

We found that participants took control of the vehicle later when it was labeled Autopilot than when it was named Copilot. This suggests that the label itself made them complacent. The more capable they thought the vehicle was, as indicated by the label, the later they took control. In other studies, we found that people are more likely to view companies as liable for accidents when they use labels suggesting greater capability, which suggests that the labels lead to both riskier consumer behavior and more blame when a failure occurs.

Based on these results my colleagues and I hypothesized that a common marketing approach may backfire when it comes to AI failure: touting your product as superior to alternatives. Our subsequent studies confirmed this hypothesis: When companies did this, they increased the perceived capability of their systems, inadvertently *increasing* how liable participants thought they were when a failure occurred. This finding suggests that ad campaigns such as GM’s “Humans are terrible drivers” might have made consumers assign greater liability to Cruise for accidents in which it was later involved.

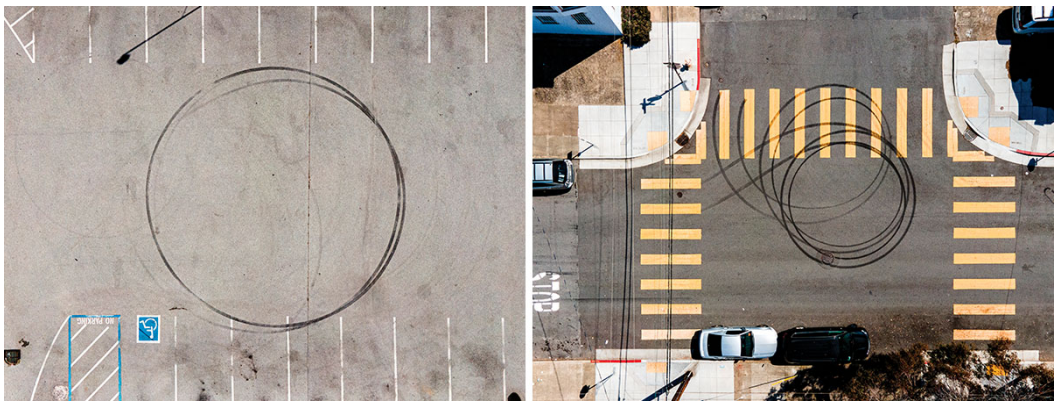
These effects occur because AI systems can span the spectrum from partial autonomy (in which the human is chiefly responsible for operating the system) to full autonomy (in which the AI is in control). Yet most consumers do not know where along the spectrum any single AI lies. This creates a dilemma for marketers. Although honest labels will reflect a system’s actual capabilities, companies are often tempted to use labels that exaggerate capabilities to boost sales.

If you use a misleading label to market your AI, you should accurately and clearly explain the AI’s true capabilities elsewhere, such as on your website, or even in the fine print on the product itself. My research finds that companies that disclose AI’s actual capabilities alongside misleading labels face fewer penalties when AI fails compared with companies that use misleading labels only. The other option, of course, is to simply use a less misleading label in the first place. While this approach carries the obvious drawback of not amplifying perceptions of your AI’s capabilities, at least it is risk-free.

People Judge Humanized AI More Harshly

Companies are increasingly deploying AI systems that exhibit human characteristics. Wysa, a mental health app, uses a cartoon avatar that helps people complete exercises, and AI companion apps such as Replika use realistic human images that express “feelings” and “reflections.”

These cues can create the impression that the bot possesses personal feelings, goals, desires, and other qualities that it does not genuinely hold. Bots that are humanized have several benefits for companies compared with neutral bots: They increase consumers’ purchase intent, level of trust, brand loyalty, compliance with provider requests, and willingness to disclose personal information. These effects persist even when people know they are conversing with a machine.



Winni Wintermeyer

In a study conducted by professors Raji Srinivasan and Gülen Sarial-Abi, participants were told about a financial investment company that made a mistake that resulted in losses for its customers. The researchers told one set of participants that the mistake was made by an “algorithm program”; they told the other set that a humanized AI named “Charles” was responsible. They found that attitudes toward the brand were more negative when the algorithm was humanized than when it was not. Follow-up studies suggest that humanized bots are more likely to be ascribed mental capabilities such as remembering, understanding, planning, and thinking, which leads participants to ascribe heightened responsibility to the bots for any failures. In another study, a team of researchers led by Cammy Crolic found that participants evaluated a failing chatbot more negatively when it employed both verbal and visual humanlike cues (using first-person language, introducing itself as “Jamie,” and having a female avatar) than when it employed only verbal humanlike cues. This suggests that humanized traits have an additive effect on negative reactions to failure.

You should be especially wary of using chatbots in domains where customers may be angry. In another study led by Crolic, researchers analyzed approximately 500,000 customer chatbot sessions at an international telecommunications company. They found that the more that customers treated the bots like a human (by using the bot’s first name in conversations, for example), the more that satisfaction suffered when customers were angry. One way to mitigate this effect is to selectively use humanized bots in neutral domains—such as product searches—and less frequently in roles that tend to involve angry customers, such as customer service centers. Another way is to temper expectations as soon as a customer starts a chat session. Slack’s chatbot, for example, says, “Hello, I’m Slackbot. I try to be helpful. (But I’m still just a

bot. Sorry!)” Bots that make such disclosures are less likely to infuriate customers when they fail.

People Are Outraged by Programmed Preferences

In 2016 a senior manager at Mercedes-Benz stated that when developing its self-driving vehicles, the company would place the safety of its passengers above that of pedestrians and other road users. His logic was straightforward: If only one life could be saved, it should be the person inside the car. His statement ignited a media firestorm. One tabloid headline accused the automaker of essentially choosing to “run over a CHILD rather than swerve” to protect those inside. Within weeks, Mercedes-Benz publicly clarified that neither engineers nor autonomous systems should make judgments about the relative value of human lives. The emotions raised by this incident suggest that people find it unethical for companies to deliberately encode group-based preferences (based on age, gender, or customer status, for example) into AI systems.

To test that hypothesis, I conducted a study with Harvard professor Mina Cikara in which we asked 826 participants in the United States to imagine that a fully automated vehicle was faced with various difficult dilemmas—whether to swerve into an elderly person or a young child, for example—and then collided into a person from one of those two groups. Importantly, we manipulated whether the vehicle decided this at random or based on a programmed preference to favor one group over another. We found that participants were more outraged when the vehicle had any kind of programmed preference than when it chose at random. This suggests that companies may not want to communicate when their systems make decisions based on group-based preferences. In some cases, they may even want to avoid collecting data on features like race, gender, and age in the first place to inform the behavior of AI systems. Another approach

is to use more structural features of the situation, such as prioritizing saving more people versus fewer. A study by Yochanan Bigman and Kurt Gray found that people are more supportive of AI systems using structural preferences than group-based ones, presumably because there is a clear, utilitarian reason for structural preferences that most people can agree on.

...

Failures of AI are inevitable. Marketers must recognize that the same actions that increase adoption today can create problems when an AI failure occurs, especially when those actions tout the benefits and superiority of your AI. Therefore, before implementing a marketing strategy, be sure to understand the five pitfalls related to AI. Assessing those risks can help your company pursue a marketing strategy that sells your AI now while reducing your liability and brand risk in a future failure.

A version of this article appeared in the July–August 2025 issue of *Harvard Business Review.*

Julian De Freitas is an assistant professor in the marketing unit at Harvard Business School.

Read more on **Marketing** or related topics **Consumer behavior**, **Risk management**, **AI and machine learning** and **Automation**