Cyber Risks & Liabilities

Protecting Your Business From AI Chatbot Errors

Operational efficiency and customer satisfaction are top priorities for organizations, and many are turning to artificial intelligence (AI) chatbots to support these goals. AI chatbots can perform a range of tasks, including handling customer service inquiries and troubleshooting technical issues, while delivering key benefits such as speed, scalability and around-the-clock availability. In addition, AI chatbots can enhance sales efforts by engaging potential buyers, recommending products and identifying high-quality leads.

However, as organizations increasingly adopt AI chatbots to enhance efficiency, they may expose themselves to new vulnerabilities. These tools can generate inaccurate or biased responses, especially when pulling from outdated or poor-quality datasets. In some cases, AI chatbots have fabricated medical and financial advice, posing serious risks to users and organizations alike. Moreover, chatbots may collect and process sensitive data without informed consent or full transparency, raising ethical and privacy concerns.

This article explores the key risks posed by AI chatbots and outlines practical strategies organizations can implement to mitigate them.

Understanding Misrepresentations or Hallucinations in AI Chatbots

As organizations integrate AI chatbots into operations, it's vital that they understand the risks of inaccurate outputs, particularly misrepresentations and hallucinations.

Misrepresentations refer to false or misleading statements made by chatbots, such as incorrect product details, policy terms, or service information. These errors may stem from outdated data, poor logic in interpreting user intent or weak system design—meaning flaws in the chatbot's underlying architecture, such as its conversation flow, escalation protocols or response-handling mechanisms.

Hallucinations occur when an AI chatbot generates responses that sound plausible but are factually incorrect or entirely fabricated. They are common in generative AI models, which produce fluent and contextually relevant text by predicting likely word sequences based on statistical patterns in their training data. However, such models typically don't have access to real-time external knowledge, so they may confidently generate false information, especially if fed ambiguous prompts.

Fundamentally, AI chatbots can produce inaccurate outputs that users can mistake for authoritative guidance. In one recent high-profile incident, an AI-powered chatbot launched by New York City inadvertently gave advice that violated state and federal laws. The risks of misrepresentations and hallucinations remain a persistent challenge that organizations must actively manage to avoid legal liability, reputational damage and loss of public trust.

Key Risks for Businesses Due to Chatbot Misrepresentations and Hallucinations

Deploying AI chatbots without proper safeguards can expose businesses to a range of risks, including the following:

- **Customer trust erosion**—When chatbots provide inaccurate or misleading information, customers may lose confidence in the brand, leading to reduced engagement, reputational harm and diminished long-term customer loyalty.
- **Legal liability**—Businesses may be held accountable for the statements made by their chatbots. Misrepresentations about products, services or policies can lead to breach-of-contract claims, consumer protection violations or even class-action lawsuits. The risk may be higher in regulated sectors like health care, finance and legal services.
- **Financial consequences**—When chatbots make errors, businesses may face direct costs like refunds or compensation, as well as indirect losses from legal fees, regulatory fines and lost customers.
- **Regulatory scrutiny**—Chatbots that violate privacy, enable fraud or deceive users in harmful ways may attract enforcement from regulators. Agencies like the Federal Trade Commission have warned that AI tools are subject to existing consumer protection laws, and misuse can lead to investigations and regulatory penalties.
- **Security and privacy risks**—Chatbots often handle sensitive customer data such as personal identifiers, payment information or health records. If improperly secured, they can become vectors for data breaches, identity theft or unauthorized access to internal systems.
- **Disinformation and reputational attacks**—Bad actors can manipulate chatbots through prompt injection, data poisoning or jailbreaking techniques to spread false information, impersonate individuals or generate harmful content. These tactics can damage brand reputation, mislead customers and undermine public confidence.

Preventive Measures to Reduce Risks

To reduce the risks associated with AI chatbots, organizations should implement the following proactive measures:

- **Regular monitoring and testing**—Organizations should continuously evaluate chatbot performance through automated checks and manual audits to detect misrepresentations, hallucinations and inappropriate responses. They should also conduct scenario-based testing with realistic customer interactions before deployment and throughout the chatbot's lifecycle. Moreover, organizations should consider real-time monitoring tools to actively track outputs, flag anomalies and trigger alerts for human review, particularly in high-risk sectors such as health care, finance and law.
- **Human oversight**—Organizations should establish clear protocols for human involvement in chatbot workflows, particularly for sensitive, complex or high-impact interactions. AI systems should be designed to recognize when to defer to human agents, using predefined escalation criteria such as regulatory triggers or ethical risk indicators. Organizations must train employees to review flagged outputs and make final decisions in cases where bias, legal exposure or reputational harm could arise.
- **Clear disclaimers and transparency**— Organizations should clearly inform users that chatbot responses are generated by AI and may not constitute professional or authoritative advice. Disclaimers should be prominently displayed within chatbot interfaces and reinforced through user agreements that define the scope and limitations of the service. Regular reviews of disclaimer language, in line with evolving regulations and industry standards, are essential to maintain clarity and mitigate legal liability.
- **Restrict chatbot authority**—Organizations should tightly control what actions chatbots can perform, especially in high-risk sectors. For example, chatbots should not be permitted to initiate financial transactions, modify account settings or access sensitive data without explicit user consent and secure authentication. Organizations should implement robust safeguards (e.g., authentication checks and permission filters) to keep chatbots within approved boundaries.

- **Training with high-quality, diverse data—** Organizations must ensure AI chatbots are trained using accurate, current and context-specific data relevant to their products and services. Data must include representation across demographics and geographies and account for factors such as socioeconomic diversity, disability status and age range to reduce bias and improve reliability. Organizations must regularly audit and refresh datasets to reflect evolving user needs, market conditions, and regulatory changes.
- **Robust data privacy and security measures—** Organizations must implement strong safeguards to protect user data throughout chatbot interactions. Chatbots should support data minimization by collecting only the information necessary to fulfill their operational purpose. Businesses must obtain explicit user consent (e.g., through consent prompts in chats and linked privacy policies), offer opt-in and opt-out mechanisms, and provide users with control over their data, including access, correction and deletion upon request. Organizations should conduct regular security audits and penetration testing to spot and address vulnerabilities.
- **Incident response plans—**Organizations must develop AI-specific incident response plans to manage the risks associated with chatbot failures, including data breaches, algorithmic bias, misinformation, hallucinated content and other unexpected AI capabilities. Organizations should include clear protocols for detection, containment, investigation and recovery in their response plans.
- **Monitor and counter disinformation campaigns—** Organizations must implement proactive strategies to detect and mitigate AI-driven disinformation, including chatbot-generated falsehoods and the spread of hallucinated or misleading content. Organizations could conduct frequent bias audits and deploy threat intelligence detection tools to monitor and respond to disinformation.

Why Proactive Risk Management Matters

Proactive risk management is essential for organizations deploying AI chatbots. It helps protect customers and build trust by minimizing harmful errors, such as hallucinated responses, misinformation or biased outputs. By anticipating and addressing these risks early, organizations can avoid costly legal and regulatory repercussions and support sustainable AI adoption. In doing so, organizations can position themselves to operate more efficiently and maintain a competitive edge.

Conclusion

As AI chatbots continue to reshape business operations, their deployment must be done responsibly. While these tools offer significant advantages, they also introduce complex risks. As such, organizations must implement robust safeguards and maintain human oversight to harness the power of AI chatbots while minimizing their vulnerabilities.

Contact us today for additional risk mitigation guidance.