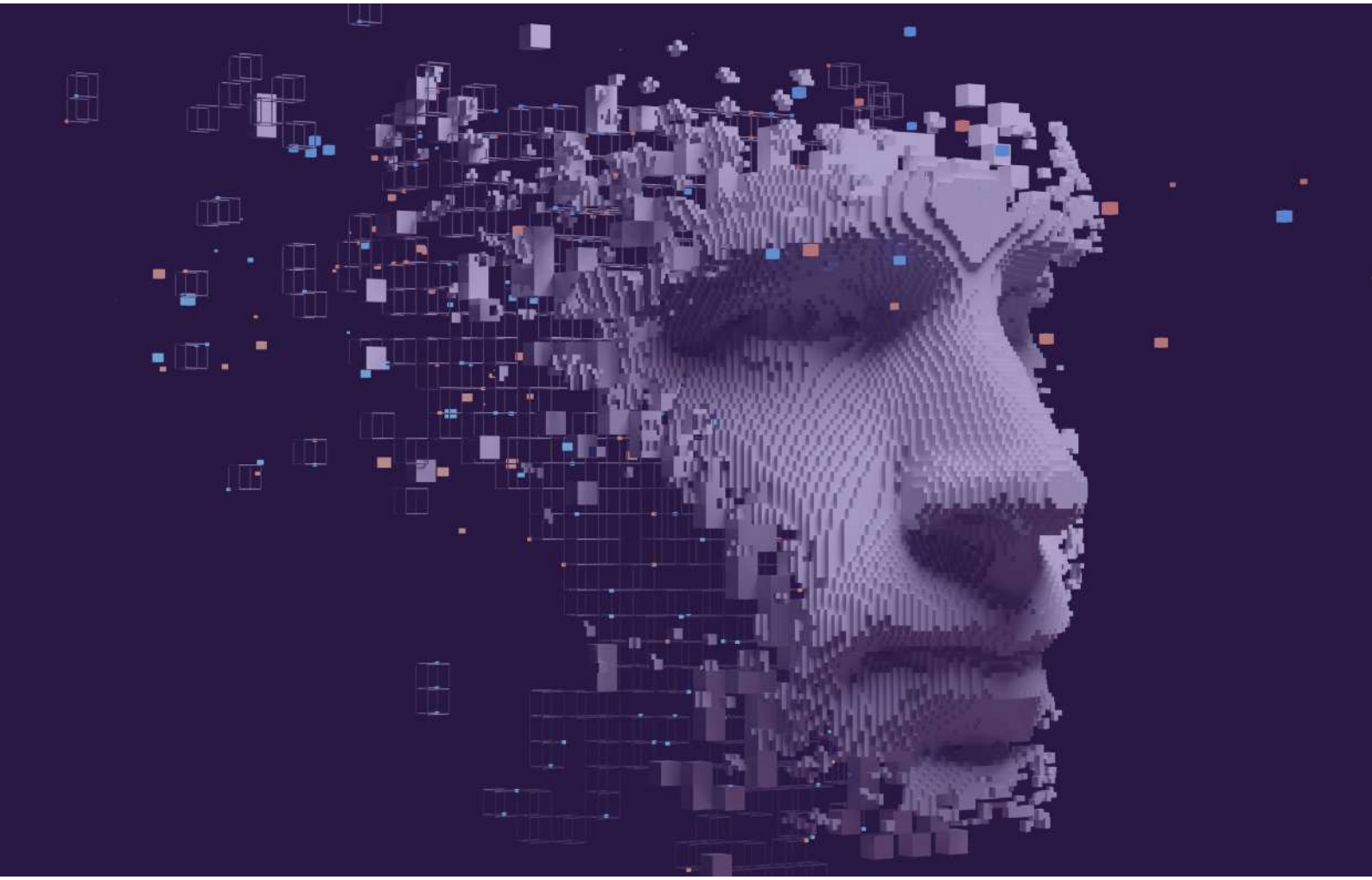
Luxoft
A DXC Technology Company



How to prevent AI-solution abuse

Mitigate the risks of AI-solution abuse thorough human-centered strategies to ensure its ethical and responsible use

Mitigating the risk of AI-solution abuse:

Human-centered strategies

In brief:

- In his latest blog, Jeroen Bet, Director of Strategy at Luxoft's Smashing Ideas, explores the growing risk of AI-solution abuse and highlights the importance of proactive measures in safeguarding ethical AI use
- The article presents four human-centered strategies: Creating bad actor personas, using the Six Thinking Hats technique, establishing guiding principles, and emphasizing continuous improvement and monitoring
- Jeroen underscores the need for organizations to anticipate potential abuse scenarios, foster a culture of accountability, and maintain an ongoing dialogue with users to ensure responsible AI adoption

Introduction

Artificial intelligence (AI) has revolutionized various industries, offering tremendous potential for innovation and transformation. However, as businesses adopt AI solutions, it is crucial to learn from past mistakes and take proactive measures to prevent abuse. This article offers four human-centered strategies that can effectively mitigate the risk of AI-solution abuse, ensuring ethical and responsible use of AI technologies.

The challenge of AI-solution abuse

A couple of years ago, a client aimed to enhance diversity within their predominantly homogeneous workforce by implementing an AI engine. The concept was to utilize the AI engine for sorting and filtering resumes, generating merit-based hiring recommendations, and ultimately eliminating bias associated with traditional hiring methods. However, the AI model's training data set consisted solely of profiles from employees who had historically excelled within the company. Consequently, the training data set introduced inherent bias into the algorithm. As a direct consequence, the AI engine's recommendations unsurprisingly favored new hires who closely resembled the existing workforce, perpetuating the lack of diversity.

In the aforementioned scenario, mitigating unintentional bias in the system might have proven challenging without manipulation of the learning data set. From an AI perspective, the system itself did not commit any wrongdoing as it merely acted based on the defined success criteria of hiring candidates similar to the existing workforce. However, a comprehensive approach encompassing bias detection, auditing, and human evaluation of the data could have resulted in broader recommendations and helped steer the program away from biases. Although initial human intervention would be crucial, over time, the AI would progressively learn to act more responsibly and reduce bias tendencies.

AI-solution abuse can take various forms, ranging from unintentional biases in decision-making, as described above, to intentional manipulation for malicious purposes. These incidents highlight the need for strategies to address edge cases (extreme or unusual scenarios) and prevent abuse scenarios. Organizations must consider the potential risks associated with AI solutions and take steps to proactively mitigate them. The following are four human-centered strategies to get started:

1. Creating bad actor personas

To proactively identify potential abuse scenarios, organizations can create virtual bad actor personas. These personas represent users who intentionally misuse or exploit the AI solution. By understanding their motivations and behaviors, organizations can gain valuable insights that expose vulnerabilities in the system. This, in turn, enables them to develop specific edge cases that simulate scenarios and analyze how the personas interact with the system. Bad actor personas provide valuable insights into the deceptive practices, biases, security risks, privacy concerns, and system performance issues that could be exploited. This approach empowers teams to uncover potential areas of concern, anticipate abuse scenarios, and implement preventive measures.

2. Utilizing the Six Thinking Hats

Generating a comprehensive set of edge cases is crucial in addressing abuse scenarios. The Six Thinking Hats technique provides a structured approach to decision-making and problem-solving, enabling team members to consider multiple angles and perspectives. Each "hat" represents a distinct viewpoint, encouraging team members to consider various angles of abuse, such as the perspectives of the abuser, the system designer, the user, the ethical implications, the legal considerations, and the potential consequences. By systematically analyzing aspects related to AI-solution abuse, organizations can identify potential vulnerabilities and develop robust designs. Integrating the creation of bad actor personas with the Six Thinking Hats technique can foster a more holistic approach to addressing abuse scenarios.

3. Defining guiding principles

Guiding principles serve as a moral compass for the design and development of AI solutions. They establish a framework that guides decision-making and behavior, ensuring ethical and responsible practices. By defining a set of guiding principles, organizations proactively address potential abuse scenarios. These principles can outline what constitutes acceptable and unacceptable use of the AI solution. Edge cases can be built around these guidelines, helping teams navigate complex ethical considerations and minimizing the risk of abuse. Guiding principles foster accountability, monitoring, and promote a culture of responsible AI use within the organization.

4. Monitoring and improving with a human-centered focus

Engaging in continuous monitoring and improvement with a human-centered focus is crucial to mitigating the risk of AI-solution abuse. Organizations must establish mechanisms for regular audits, ethical evaluations, and soliciting user feedback to identify emerging abuse scenarios promptly. By cultivating a culture of transparency, accountability, and open communication, individuals are encouraged to report concerns regarding potential abuse, ensuring their experiences and well-being are prioritized. Staying updated with industry standards and ethical best practices ensures that AI solutions remain aligned with human-centric principles, safeguarding the interests and needs of users.

Conclusion

Mitigating the risk of AI-solution abuse is of paramount importance in today's rapidly evolving technological landscape. By adopting human-centered strategies, such as creating bad actor personas, utilizing the Six Thinking Hats technique, defining guiding principles, and continuously improving, organizations can proactively address abuse scenarios. These strategies empower teams to identify vulnerabilities, generate edge cases, and develop responsible AI solutions. As businesses continue to innovate with AI, the ethical and responsible use of these technologies must remain at the forefront. By integrating these strategies and fostering a culture of accountability and continuous improvement, organizations can ensure that their AI solutions align with their values and objectives, promoting trust and responsible AI adoption.

About **the author**

Jeroen Bet
Director of Strategy,
Luxoft's Smashing Ideas

Jeroen is a strategy director with a solid background in customer experience. Over the course of his 25-year career, Jeroen has worked with companies such as Chempoint, Expedia, Costco, Amazon, and Microsoft. He has collaborated with a variety of stakeholders, including scientists, engineers, plant managers, marketers, conservationists, end-users, and business leaders. Jeroen has successfully led project teams in developing human experiences that incorporate AI and other cutting-edge technologies.

Ready to take a human-centered approach to your AI strategies?

At Luxoft, our strategy team can help you ask better questions to get to the right solutions.

Visit **luxoft.com** or **contact us** today to learn more about how we can help your organization innovate faster towards business objectives.

About Luxoft

Luxoft, a DXC Technology Company delivers digital advantage for software-defined organizations, leveraging domain knowledge and software engineering capabilities. We use our industry-specific expertise and extensive partnership network to engineer innovative products and services that generate value and shape the future of industries.

For more information, please visit **luxoft.com**