# AI safety is now critical to protecting brand reputation in marketing



The marketing and advertising industry is undergoing a profound transformation powered by artificial intelligence (AI). From automated media buying and personalised content delivery to generative creative tools and conversational brand agents, AI has become an integral part of the modern marketing stack. As these systems take on increasingly visible and autonomous roles, a critical issue demands attention at the highest levels of leadership: AI safety.

Traditionally viewed as a technical concern relegated to data science or engineering teams, AI safety must now be acknowledged as a strategic brand priority. As AI generates content, makes real-time decisions, or directly engages with consumers, any failures in alignment, accuracy, or fairness reflect not just on the technology but significantly on the brand itself.

In this new landscape, ensuring AI safety is fundamentally intertwined with brand safety. Marketers have historically acted as custodians of brand reputation, but with AI systems now significantly shaping consumer experiences and brand perceptions, new complexities arise. Generative AI can inadvertently produce factually incorrect or misleading information, posing a risk of reputational damage, consumer backlash, or even regulatory scrutiny. Automated targeting models might exhibit biases, leading to potential exclusions or unfair categorisations that erode trust and breach compliance standards. Furthermore, AI-driven chatbots and agents may malfunction or be subject to manipulation, compromising customer experience. The potential for brand content to appear alongside harmful or inappropriate material further underscores the urgent need for brands to safeguard their image.

These issues are not mere hypothetical scenarios; they are emerging realities faced by organisations employing AI tools on a large scale. A survey from July 2023 indicated that nearly 30% of marketing professionals perceive generative AI as a significant risk to brand safety, while 55% view it as a moderate threat to reputation. This rising concern highlights a crucial shift in the industry's understanding of the impacts of AI on brand integrity and the dissemination of accurate information.

Defining AI safety within the marketing context hinges on establishing governance frameworks that ensure AI-operated tools align with brand values and legal obligations. Practically, this encompasses maintaining factual integrity to prevent the spread of misinformation, ensuring fairness and inclusivity through algorithm audits, and guaranteeing robustness and reliability to minimise adverse outputs. Essential to this governance is the principle of explainability and accountability, ensuring that stakeholders understand how AI systems arrive at conclusions.

To manage these risks, leading organisations are proactively integrating AI governance into their broader marketing strategies. They are developing policies that outline responsible AI use in various functions, establishing cross-functional oversight committees comprising experts from marketing, legal, compliance, and data science, and providing training sessions for marketing teams to deepen their understanding of AI's capabilities and limitations. Pre-deployment testing procedures such as red teaming and bias detection are increasingly viewed as vital steps before launching AI-driven campaigns.

Regulatory and public scrutiny of AI is intensifying, as evidenced by initiatives like the European Union's AI Act and the US Executive Order on Artificial Intelligence, which underscore a shift towards mandatory transparency and accountability. Concurrently, consumer expectations around data usage and automated engagement are evolving rapidly; trust is no longer an ancillary concern but a cornerstone of consumer loyalty and advocacy.

In light of these insights, marketing leaders are urged to conduct comprehensive audits of AI use within their organisations, establish specific guidelines for generative systems, and foster cultures that champion responsible experimentation. Notably, collaborations with compliance and legal teams can ensure that AI practices align with emerging regulatory frameworks, while layered safeguards, including human oversight and grounding mechanisms for generative content, can protect brand integrity.

This imperative extends beyond mere adherence to ethical standards; it targets the safeguarding of trust and long-term brand success in an increasingly sceptical environment. As AI continues to revolutionise marketing, it is clear: AI safety is not a luxury but a necessity—a vital component of brand safety itself.

Lionel Sim, founder of AI agency Capitol and former global Chief Commercial Officer for Livewire, asserts, “In today’s world, AI safety isn’t just a nice-to-have. It is brand safety.” Companies that effectively adopt AI-driven strategies stand to benefit considerably, enhancing consumer trust and bolstering brand reputation, ultimately steering themselves towards a more responsible and equitable future in marketing.

## Reference Map:

* Paragraph 1 – [[1]](https://www.campaignlive.com/article/brand-safe-ai/1919868), [[4]](https://www.weareliquid.com/latest/exploring-the-impact-of-ai-safety-summit-2023-on-marketing/)
* Paragraph 2 – [[1]](https://www.campaignlive.com/article/brand-safe-ai/1919868), [[2]](https://www.marketingdive.com/news/meta-ai-brand-safety-social-media-marketing/646639/), [[3]](https://www.statista.com/statistics/1405095/ai-risk-level-brand-safety/)
* Paragraph 3 – [[5]](https://www.decisionmarketing.co.uk/news/ai-safety-summit-isba-and-ipa-outline-ad-industry-code), [[6]](https://aibyhumans.com/blog/the-role-of-ai-in-marketing-compliance-and-ad-safety)
* Paragraph 4 – [[7]](https://magsmarketing.com/whats-new/ai-driven-brand-safety-in-digital-advertising-spaces/)
* Paragraph 5 – [[5]](https://www.decisionmarketing.co.uk/news/ai-safety-summit-isba-and-ipa-outline-ad-industry-code), [[3]](https://www.statista.com/statistics/1405095/ai-risk-level-brand-safety/)

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.campaignlive.com/article/brand-safe-ai/1919868> - Please view link - unable to able to access data
2. <https://www.marketingdive.com/news/meta-ai-brand-safety-social-media-marketing/646639/> - Meta has introduced AI-driven brand safety and verification tools for Facebook and Instagram advertisers in English- and Spanish-speaking markets. These tools employ a multi-stage AI review system to classify text, photo, and video content, offering advertisers control over ad proximity to potentially objectionable material. The controls categorize content into high, medium, and low-risk levels, allowing advertisers to select settings that align with their brand safety preferences. Additionally, Meta has partnered with Zefr to develop an AI-powered tool that assesses the context of ads appearing in Facebook Feeds, enhancing content suitability for advertisers.
3. <https://www.statista.com/statistics/1405095/ai-risk-level-brand-safety/> - A survey conducted in July 2023 among marketing and advertising professionals across North and South America and Europe revealed that nearly 30% considered generative AI to pose significant risks to brand safety and misinformation. Additionally, 55% viewed the technology as a moderate risk to maintaining a brand's reputation. This underscores the growing concerns within the industry regarding the potential challenges AI presents to brand integrity and the dissemination of accurate information.
4. <https://www.weareliquid.com/latest/exploring-the-impact-of-ai-safety-summit-2023-on-marketing/> - The AI Safety Summit 2023, held at Bletchley Park, brought together international governments, leading AI companies, civil society groups, and research experts to address AI risks and opportunities. A key outcome was the Bletchley Declaration, which outlines shared principles for AI safety, emphasizing responsible and ethical AI development. For marketers, this means adhering to emerging AI safety standards and aligning with customer expectations for ethical AI use, ensuring that AI-powered marketing solutions are trustworthy, fair, and transparent.
5. <https://www.decisionmarketing.co.uk/news/ai-safety-summit-isba-and-ipa-outline-ad-industry-code> - In conjunction with the UK's AI Safety Summit, the Institute of Practitioners in Advertising (IPA) and the Incorporated Society of British Advertisers (ISBA) have published 12 guiding principles for the use of generative AI in advertising. These principles aim to ensure ethical AI adoption in the industry, covering areas such as transparency, intellectual property rights, and human oversight. The guidelines are designed to protect consumers and professionals in the creative sector, promoting responsible AI usage in advertising practices.
6. <https://aibyhumans.com/blog/the-role-of-ai-in-marketing-compliance-and-ad-safety> - The integration of Artificial Intelligence (AI) in marketing has revolutionized compliance and ad safety by automating content reviews and identifying potential risks early. AI systems analyze vast amounts of data to proactively manage compliance, ensuring that marketing content aligns with legal and regulatory standards. Additionally, AI enhances decision-making by providing real-time analytics, allowing marketing teams to make informed choices. Continuous learning and adaptation of AI models ensure that compliance assessments remain up-to-date with evolving regulations, safeguarding brands from reputational damage and regulatory penalties.
7. <https://magsmarketing.com/whats-new/ai-driven-brand-safety-in-digital-advertising-spaces/> - To effectively implement AI-driven brand safety measures, advertisers should adopt multi-layered strategies that combine AI-powered content analysis with human review. Continuous monitoring and optimization of these solutions are essential to detect and mitigate brand safety risks promptly. Transparency and collaboration with AI vendors are crucial to ensure that AI systems align with brand values and objectives. Real-world examples demonstrate that successful implementation of AI-driven brand safety leads to increased consumer trust, improved brand reputation, and higher engagement with target audiences.