COSMOS
Navigate the Compliance Universe

CEP Magazine - January 2024

Hilary Wandall (complianceofficer@dnb.com) is Chief Ethics and Compliance Officer at Dun & Bradstreet in Jacksonville, Florida, USA.

The digital frontier: Mitigating risk in the face of AI and emerging cyber threats

By Hilary Wandall

In today's rapidly changing digital landscape, artificial intelligence (AI) has emerged as a disruptive force with the potential to reshape entire industries and professions. Its ability to augment human capabilities, introduce intelligent automation, and analyze vast data sets has attracted many companies seeking to unlock new levels of innovation and efficiency.

Despite enormous opportunities, AI's emergence has brought a slew of new risks and regulations to the private sector.[1],[2] Executives now face added complexity as they integrate new AI tools into existing business models and seek to balance potential benefits and threats. More than ever, an updated framework is needed to address risk management and ensure companies are prepared for the new era of digital innovation.

To combat threats in a compliant way, executives must embrace the enhanced and accelerated sharing of information across the organization, build awareness of new opportunities, and educate employees on cyber and related AI risks. Only by integrating transparency across the entire business function can leaders prepare their firms for what's ahead.

While AI and other digital innovations offer immense promise for businesses, many firms have rushed to get involved without developing a comprehensive risk framework. Instead of acting immediately, an optimal strategy involves planning to address new and evolving risks thoughtfully and with agility while protecting employees and customers for the long term. Bringing together siloed teams into a broader compliance and ethics program is key to getting ahead of hidden risks and ensuring leaders are equipped with the necessary information to make informed decisions. Additionally, cybersecurity readiness must be a priority across the C-suite as cyber threats grow more sophisticated.

Rethinking the AI approach

For business leaders, the fear of missing out on potential AI benefits can lead to an accelerated adoption strategy that prioritizes speed at the expense of security, safety, and resilience. This failed AI approach often begins at the outset when executives neglect to examine potential opportunities through a risk management lens.

Decision-makers should begin their evaluation process by first asking, "What's the real opportunity, and how do we define it?" Only when an organization has identified an opportunity or set of options can it identify and prioritize risks in terms of their short-term and long-term impact.

While unknowns abound in the AI space, most companies are not attuned to the novel perils of generative AI—a

type of AI that can create new content or data in the form of text, images, audio, or code. Due to the intricate nature of prompt engineering skills required to avoid hallucinations and generate accurate results, most employees currently lack the skills to use these tools effectively.

To harness the full power of generative AI, executives should implement proper awareness and education, especially for consumer-facing employees. Training employees to write prompts that yield relevant, trustworthy results—which can generate quality data and new insights—should be top of mind for executives.

Despite the power of these tools, companies must also make their employees aware that generative AI can produce inaccurate responses, even if used correctly. Users often don't take the time to carefully examine what's being generated. This can lead to substantial risks around the quality and accuracy of the outputs and an additional risk for the person who receives misleading results.

However, the risks of generative AI extend beyond inaccurate results. Prompt injections, for example, are a type of attack that tricks a language model into producing unwanted or malicious output. This involves injecting malware or untrusted text into the prompt, resulting in the system delivering hate speech, false information, or other unexpected behaviors.

Currently, there are not many good strategies for mitigating prompt injection risks. This means that companies must devote additional scrutiny when choosing and testing opportunities and framing the utilization of the tools in the context of those opportunities, with a recognition that some risk will always be part of a cyber strategy.