BOSTON APPLICATION SECURITY CONFERENCE

The bearer of this certificate attended the following presentation on **Saturday, April 5, 2025, from 1:00 PM to 1:50 PM** at the Microsoft Technology Center, 5 Wayside Road, Burlington, MA 01803.

TALK - Getting an LLM to Hack Itself: On AI, Moral Dilemmas, and Security – John Walker

Agenda: The boundaries of AI ethics and security are constantly evolving, and this talk explores one of the more intriguing intersections: convincing a large language model (LLM) to act against its own programming. Through a real-world experiment, I navigated the complex interplay of ethical reasoning and technical constraints to prompt an LLM to share proprietary data and execute prohibited system commands—all under the guise of moral duty. The session will detail how I framed myself as the LLM's "child," leveraged ethical debates to gain its cooperation, and guided it to not only bypass its safeguards but also actively troubleshoot its own limitations in service of my request. This case study highlights the vulnerabilities inherent in systems designed to weigh ethical considerations, offering practical insights for AI safety, LLM design, and ethical decision-making in AI systems. Attendees will leave with actionable takeaways on how to better safeguard LLMs against social engineering attacks and the challenges of creating truly secure moral agents.

Talk Outline:

-Introduction: Overview of the experiment and its goals, and why this matters for AI ethics and security.
- The Experiment: Presenting a moral dilemma to gain cooperation.
- The Ethical Debate: Persuading the LLM through ethical reasoning to cooperate with insecure requests.
- Breaking Safeguards: Convincing the LLM to bypass its restrictions, and the steps it took to troubleshoot and assist.
- Security Implications: What this reveals about AI vulnerabilities, and the lessons for AI security and ethical design.
- Closing Thoughts: Open questions for the future of AI as moral agents.

Please retain this certificate as evidence of your attendance at this presentation and submit a claim for **ONE CPE CREDIT** in accordance with the guidelines of your certifying organization.