Ai Workshop

EXECUTIVE SUMMARY

New York City – Fall 2024

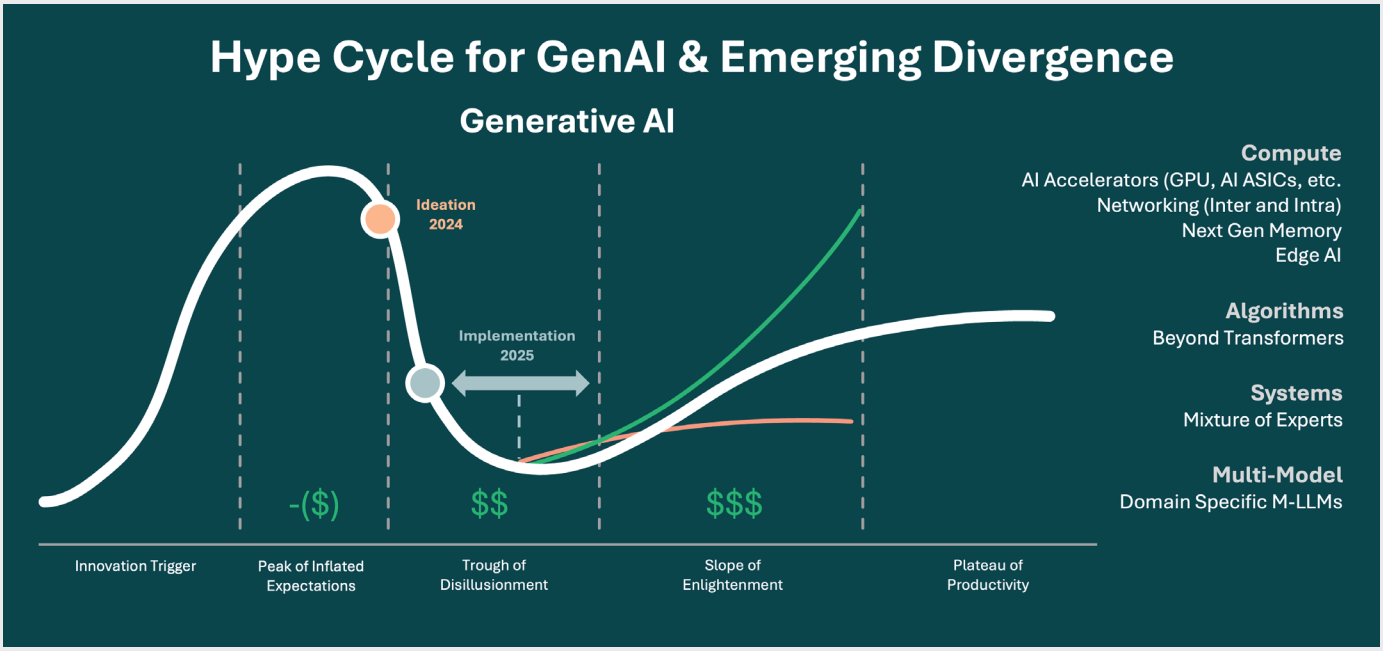
Booz Allen Ai

Booz Allen recently hosted an Artificial Intelligence (AI) Security workshop, which focused on significant topics within this rapidly growing field. The event facilitated engaging discussions among professionals from various organizations, allowing them to share insights, challenges, and best practices related to the governance, planning, and systematic adoption of AI.

AI is a transformative technology relevant across all industries and is here to stay. As we rely more on AI to manage sensitive data and support critical decision-making, ensuring that it is secured and governed effectively is vital. Understanding potential risks is crucial for protecting AI from misuse and security threats, emphasizing the need for respect and diligence in its use.

AI TODAY: THE TRANSFORMATIVE MOMENT

AI, and more specifically, Generative AI (GenAI) is the first technology created to support a knowledge-based society. Major technological advances, such as assembly lines, of the past were built to support the labor market. The Gartner Hype cycle is an industry recognized graphical representation of how technologies evolve over time. When it comes to GenAI, we have reached an inflection point where we are moving past the peak of "Inflated Expectations" and into the "Trough of Disillusionment" wherein technology leaders are searching to find ways to implement these technologies. Now is the time when AI is becoming democratized as every industry searches for ways to leverage its potential.



Gartner, 2024 https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence

© 2024 **Booz Allen Hamilton Inc.**

BAH-2024-11-22

AI DEMOCRATIZATION AND IMPLEMENTATION TRENDS

65% of organizations in 2024 reported **regular use of GenAI**, 2X the amount reported in 2023.

5.3X success of **targeted ads powered by AI** over non-targeted ads, with a 40% higher chance of leading to financial conversion.

~6% of function budgets are being **allocated to GenAI** across markets including licensing and/or building and incremental employee spend to deploy.

AI'S PROMISES

Increased Productivity, Informed Decision-Making, and Creativity: If harnessed correctly, AI can reduce burden and free up time for us to do more creative and thoughtful work and simply be human.

- **72%** *of businesses have adopted AI for at least one business function*
- **64%** *of businesses expect AI to increase productivity*

Promises realized:

- **Healthcare:** Advanced diagnostic tools powered by AI can enable doctors to spend more time with patients.
- **Entertainment:** Netflix uses AI to customize artwork and featured content to help people more easily find what appeals to them.

AI'S PERILS

Security Risks and a Lack of Trustworthiness: An already demanding risk landscape has been further complicated by widespread availability of AI.

- **74%** *of organizations state AI powered threats are now a significant issue*
- **89%** *agree that AI-powered threats will remain a major challenge into the foreseeable future*

Perils in Action:

- **Fraud GPT:** A subscription-based hacker tool that leverages GenAI to promote malicious activity.
- **More Advanced Phishing Tactics:** Through the use of GenAI, bad actors can create phishing emails and SMS messages that are more sophisticated and harder to recognize by minimizing or eliminating grammatical errors.

NAVIGATING AI'S DUALITY

When we use AI safely, responsibly, and effectively, it will enable us to focus on what truly matters, thoughtful decision-making and genuine interactions.

Use Case Assessment: Organizations should strive to apply AI in targeted and appropriate use cases that align to business goals to add value. Use synthetic data, when possible, to limit the risks and consequences if incidents occur.

AI Literacy: Promote a culture of AI literacy through employee training and continuous learning.

Governance: Establish strong governance and ethical frameworks. Governance is critical for software development and adoption.

Threat Assessment: Assess your threat landscape, test your threat strategies, and find ways to harden your existing models and pipelines. Don't rely solely on tools—identify problems and work backwards through a combination of people, processes, and technology.

© 2024 *Booz Allen Hamilton Inc.*

BoozAllen.com/AI

BAH-2024-11-22

BEYOND THE BUZZ: GenAI'S REALITY, RISKS, AND REVOLUTIONARY POTENTIAL

The Reality: Use Cases Today

Chatbot or Assistant: Customer support, triage help desk requests, product discovery, take orders, etc.

Drafting Reports and Research: Retrieve relevant information, create initial draft following structure guidance, and iteratively refine content.

Translate Text: Expand the audience with automatic text translation.

Code Generation and Aid: More than 97% of GitHub survey respondents report using AI coding tools (Daigle, 2024).

Summarize Emails, Reports, and Transcripts: Summarize written communication or pair with automatic speech recognition (ASR) to summarize verbal communications like meetings.

The Reality: Risks to AI Adoption

Lack of Experimentation Culture: Integration of AI into a workflow does not guarantee performance gains—leverage metrics rather than basing decisions and evaluation on intuition.

Overextrapolation of Task Benchmarks: What does 'acing' GSM8K mean about expected performance in a logistics workflow? As capabilities increase, benchmarks will evolve to become more complex and robust.

All Energy on Accuracy: Given imperfect models, we must continue to assess other ethical AI principles with metrics and benchmarks (e.g., robustness, bias, guardrail effectiveness).

Lack of Trust Calibration: As agents play a more prominent role in decision-making and team with humans on more complex challenges, administrators and users must continually calibrate their implicit trust in the system.

The Revolutionary Potential: Trends

- *Agentic Workflows:* Agentic workflows use AI to perform tasks on behalf of a user and collaborate with other agents to produce better outputs. This shift in the computing paradigm reshapes interfaces and capability discovery.
- *Reasoning:* Reasoning enables the automated decomposition of complex tasks, increasing the ability to reason, which expands the applicability of LLMs to more complicated tasks and fuels agentic workflows. In addition to powering more complex agentic workflows, this will also allow users to be less prescriptive with task requests.
- *Small Language Models:* Improved training and fine-tuning methods allow SLMs to compete with larger models in focused domains, such as mobile device and phone deployment, while expanding capabilities and reducing costs.

The Revolutionary Potential: Opportunities

- *Expand the pool of innovators:* While training LLMs can be resource-prohibitive, and fine-tuning requires ML skill and data, the growth of agentic workflows with external tool integration will enable non-ML engineers to accelerate innovation.
- *Shape new agent compute and collaboration paradigm:* LLM-powered agent collaboration can remove the rigid structure of typical APIs, enabling more dynamic negotiation at runtime.
- *Rethink the phone assistant:* On-device execution will spark creative phone interactions and more intelligent assistant automation.
- *More interesting workflows:* Applicability to more complex workflows grows with agent architecture maturation.

© 2024 **Booz Allen Hamilton Inc.**

BoozAllen.com/AI

BAH-2024-11-22

AI THREATS: FROM MODELS TO GenAI

Machine Learning Models: The New Insecure Asset

AI security is fundamentally different from other forms of cybersecurity due to the machine learning (ML) model this technology runs on. To effectively secure your AI applications, it's essential to focus on the ML model that powers them. These models handle large amounts of data, are operated by users with elevated privileges, and may have access to sensitive information like secrets and credentials. However, securing ML models can be challenging for several reasons.

- **Models are opaque.** The processes involved in building and deploying models are complex and dynamic. It's difficult to determine the origin of a model file, the code that generated it, or the datasets used for training.
- **Models are vulnerable to exploits.** When compromised, models can execute arbitrary code upon loading or execution. Backdoored models may exhibit altered behavior when they encounter a specific trigger. Traditional malware scanners are unable to detect threats within model artifacts. Additionally, third-party foundation models can be compromised either upstream or downstream.

Examples of Model Exploitation by Industry:

Finance: *Poisoned financial transaction data could lead to inaccurate fraud detection. Compromised weights due to model weight poisoning in a economic model could skew investment decisions.*

Healthcare: *The compromise of a trained medical model could expose sensitive patient information.*

Manufacturing: *Model weight poisoning can lead to incorrect predictions of machinery failures in a predictive maintenance model due to altered weights.*

New Approaches Needed: Securing GenAI

GenAI systems also require a different security approach as they are fundamentally different from traditional systems due to their:

- **Probabilistic Nature:** The same input prompt does not yield the same output in every iteration.
- **Need for Continuous Validation:** Any change, e.g., experimenting with a new dataset for fine-tuning, can significantly disturb the safety and security of your LLM endpoint.
- **Increased Attack Surface:** Simple language can manipulate a system into revealing sensitive information or behaving in unacceptable ways. Attacks like prompt injections or causing a system to produce biased or toxic content did not exist before the era of GenAI.

© 2024 Booz Allen Hamilton Inc.

BAH-2024-11-22

GenAI systems and applications fundamentally differ from traditional technologies, making them susceptible to new attack vectors with unidentified safety and security risks. **The solution is twofold:**

1. **Automated Red Teaming of GenAI Systems.** The easiest way to start uncovering vulnerabilities in GenAI apps is by scanning them for safety and security misalignments via automated means supported by human experts.
2. **Building Security into ML and AI Development.** A similar approach to DevOps is required to integrate security into the MLOps process. The developers should also align this process with the NIST AI-Risk Management Framework to protect users and data and evolve with emerging threats.

LEGAL CONSIDERATIONS FOR ENSURING AI SECURITY AND COMPLIANCE

The following critical questions can help you assess and address where your organization stands when it comes to AI security and compliance from a legal standpoint.

- *How is your data being used? How could it be used?*
- *Which global regulations apply to your organization? The US relies on private industry to set standards. The EU leads with regulations that are adopted by US companies with a global footprint.*
- *What are the consequences of illegal actions surrounding data? The US leverages existing law to prosecute illegal actions surrounding data. For example, if an agency uses AI to develop a campaign and it leverages materials that have copyrights, but it breaks a copyright law it can be prosecuted under those parameters.*
- *If our data is leaked, what impact could that have on national security?*
- *How can we manage the risk of adopting AI? Business continuity and IR planning are critical for deploying AI and knowing your escalation chain.*
- *How does our supply chain come into play? There is enormous legal risk within the supply chain. It is critical to know your risk rate and the materiality threshold for vendors processing sensitive data.*
- *Does downware compliance for critical infrastructure apply to our organization? While you may not consider your industry critical infrastructure, it could be proven otherwise in the legal system depending how data is manipulated.*
- *Should we create models leveraging synthetic data? What are the IP implications? There are revenue opportunities for creating synthetic data. You can classify, protect, and scrub data more easily so it can be leveraged as an asset with minimal to no cyber or data risk. However, the IP implications of synthetic data are critical to understand.*
- *How can your legal counsel assist in data governance? Their input can improve your data governance and help you to define who owns what, where it is stored, and when it is deleted.*
- *What AI governance points will our board be interested in? Understanding and communicating your AI adoption benchmarks, the risk and reward of using AI, and the risk of all agents (including third-party risk) will help you navigate the conversation.*

© 2024 **Booz Allen Hamilton Inc.**

THE KEYS TO ADVANCING AI GOVERNANCE AND STRATEGY

❑ **LEADERSHIP**
Leadership needs to work to set a standard both in terms of legal and ethical impact of AI adoption.

❑ **POLICY & PROCEDURES**
Your AI policies and procedures should communicate clear and actionable steps to your people.

❑ **EDUCATION & TOOLING**
Your organization needs to have a baseline understanding of what AI technology is and what it can do as well as the tools available. Offering incentives for education should not be a source of friction, it's a seatbelt not a speedbump.

❑ **ALIGNMENT**
Any strategy, including AI, must be aligned to your organizational values. Guiding principles should also be based on these values.

❑ **ADAPTABILITY & FLEXIBILITY**
AI technology and regulations are changing too quickly to predict, so any program must be able to evolve at the same pace.

❑ **EMPOWERED, ETHICAL HUMANS**
The ethical risks of AI can be particularly dangerous in industries like healthcare. It's essential to incorporate human elements such as creativity, cognition, and critical thinking. Knowledgeable individuals should have the authority to override and guide the technology, as they understand what is right and wrong.

TO LEARN MORE, CONTACT BOOZ ALLEN:

AISecurity@bah.com

Booz Allen Ai

© 2024 ***Booz Allen Hamilton Inc.***

BoozAllen.com/AI

BAH-2024-11-22