

CALYPSO AI

The Way Forward: What It Takes to Adopt AI and Win



This paper is adapted from *The Role of the CISO Shaping Trust in an AI World* by Jim Routh, published by the [Institute for Critical Infrastructure Technology](#) on July 25, 2023.

Introduction

The widespread adoption of large language models (LLMs) is bringing artificial intelligence (AI) and machine learning (ML) capabilities into everyday business practices and products. It's also expanding attack surfaces. Recent history reminds us that an organization's adoption of new technology typically begins with a focus on utility and time to market, with consideration of controls and governance lagging a distant second. The evolution of governance for emerging AI/generative models is following the same pattern.

Establishing an executable plan for AI governance at enterprise scale is not difficult, but it requires diversity of thought from different functions, and a commitment to follow through with implementation. Organizational structure will dictate which role has the most skin in this game, in that AI/ML and generative AI (GenAI) tools are gaining traction across the enterprise, from Operations and Sales to Legal and Human Resources. It's the CISO who must bring together a cross-functional team with participants from every organizational function, who must balance the increasing demand for AI/ML tools against the expanding attack surfaces they present, and who must get buy-in from stakeholders and the C-suite about the need for a robust, enterprise-specific and enterprise-wide governance framework. It's no small task.

This paper details the roadmap for this endeavor, including current options and opportunities, and the challenges ahead.



Where We Are Now

Amid growing GenAI markets, enterprise is the one to watch, with its near-limitless use cases. The surge in available AI/LLM products and services provides enterprise users many options to update or redefine business processes, improve the digital consumer experience, and enhance professional productivity. Increasingly, businesses are opting for LLMs designed for industry-specific utility, with many developing in-house models that use proprietary data. If this trend continues and expands, jobs will likely consolidate, with some roles becoming obsolete and new roles emerging.

This shift will demand new skills both in AI model creation and integration with evolving business processes. Given the rapidly growing applications of LLMs in contemporary enterprises, it is impractical and unrealistic for decision makers to restrict their usage until all risks are known and assessed. Such a move might harm the organization's credibility, especially given LLMs' potential for business innovation and optimization. And should another stakeholder propose a ban on LLM usage, I advise dissuasion followed by suggesting alternative approaches, including educating them regarding the relevant historical context, which is discussed in the following section.



**Amid growing GenAI markets,
enterprise is the one to watch.**

¹ The GenAI market is growing at a compound annual growth rate (CAGR) of 35.6%. "Large Language Models (LLM): An Ultimate guide for 2023," Shopdev (June 16, 2023) [ShopDev.co/blog/what-are-large-language-models](https://shopdev.co/blog/what-are-large-language-models)

Adapting to Disruptive Technology Innovation

Following the Y2K panic, companies, including the one I worked for, focused their attention on e-commerce and enhancing online customers' digital experience, with an emphasis on shrinking their "brick and mortar" footprint. We first-generation dot-commers considered time-to-market the key and drove new ways to accelerate the software development lifecycle and deliver functionality sooner. The proliferation of attack surfaces created by that intense code development went unnoticed, as did the risk created by linking web applications to back-end databases not designed for Internet-facing exposure.

Looking back, I feel pretty guilty about not using the means available then to build more resilient consumer-facing software. Typical web-application security included little more than basic authentication, and the focus on lowering operating costs caused [the cost of security breaches](#) to be overlooked for more than a decade. Cybercriminals, however, specialize in adapting and, thus, constantly discover new ways to exploit advancements.

I often wonder why I ignored the need for software security back then, when I should have incorporated all available tools and methods. Now a cybersecurity practitioner and recovering CISO, I realize I was a victim of "Ostrich Syndrome" and counsel other security leaders not to follow suit. Recently, several CISOs told me they believe most enterprises already have the necessary cybersecurity guardrails to deal with the risks presented by widespread LLM usage and don't need to consider additive or different controls for enterprise-wide AI governance. In my opinion, this is a classic example of "Ostrich Syndrome." If technological controls—AI governance, in this case—don't evolve at close to the same pace the technology advances, then, as a field, we have learned nothing from our history. Every Cybersecurity leader who has not already done so must make AI governance an enterprise priority—now.

Every leader who has not already done so must make AI governance an enterprise priority—now.

² For example, *Java Security* by Dr. Gary McGraw (John Wiley & Sons: New York) 1996 or *Building Secure Software: How to Avoid Security Problems the Right Way* by John Viega and Gary McGraw (Addison-Wesley: Boston) 2001.

³ ChatGPT3.5's response to the author's query on July 7, 2023: "Cybersecurity leaders who exhibit the ostrich syndrome may neglect to acknowledge the severity or existence of cyber threats, adopt a reactive rather than proactive approach to security, or underestimate the impact of potential breaches. The ostrich syndrome can be detrimental to an organization's cybersecurity posture. By ignoring or downplaying risks, leaders may fail to allocate sufficient resources, implement necessary security measures, or develop proactive incident response strategies." "Contrary to popular belief, ostriches do not actually stick their heads in the sand. This is a common misconception."

Toward an AI Governance Framework

The subsections below detail the concepts and actions I recommend incorporating when establishing an AI governance framework.

People

It falls to the person bringing the governance steering up together to educate leaders of other functions, such as Legal, Privacy, Compliance, Sales, Operations, IT, and HR, about the importance of establishing consensus-driven principles to guide development of a framework applicable across the enterprise. Limiting the team to six to eight participants will lower decision-making and other hurdles. The leader, as the facilitator bringing this process to consensus, must avoid being a subject matter expert—even though that’s exactly what they are—and must instead demonstrate neutrality and a lack of bias toward the result.



Principles

Such a group must focus on the changes that will result from incorporating new AI solutions into the workflow and identify ownership for those changes. Aligning the AI governance principles to organizational values and brand reputation affords a clear foundation for the principles’ business value. Establishing principles early on also enables the committee to review and amend existing company policies. Examples of principles used by enterprises include:

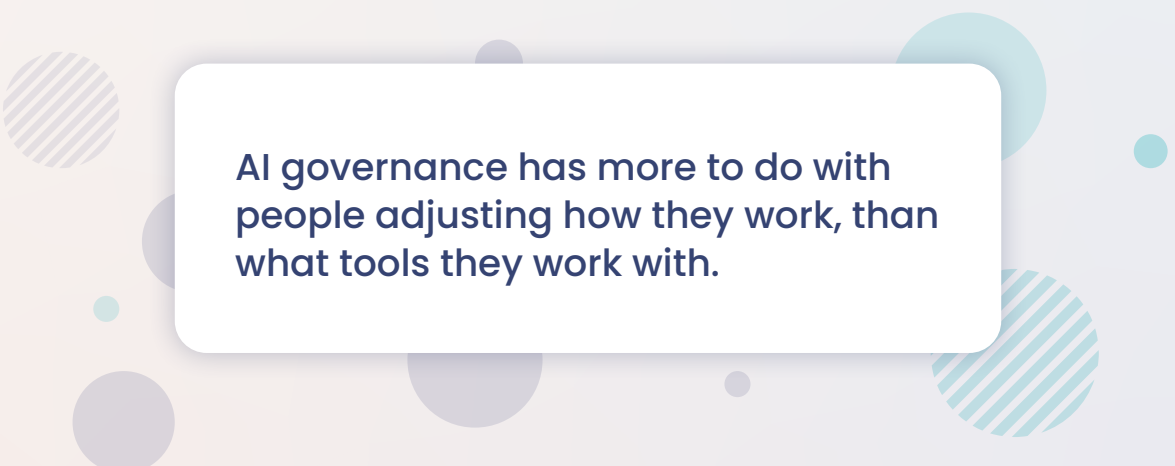
- All LLM output will be reviewed and owned by a person.
- LLM-generated software code must be checked by a human in a standard or automated review process prior to being entered into the code repository or software pipeline.
- Personnel writing internal correspondence must identify any LLM used and the date.

Regulation

National and global regulations addressing AI governance are influenced by existing privacy laws and practices and, as such, executives responsible for privacy, compliance, and related issues must be involved in developing AI governance principles. Most GenAI models, including [multi-modal models](#), which utilize text, images, and audio, [do not meet General Data Protection Regulation \(GDPR\) standards](#) about consent, data anonymization, and data deletion. Regulators have not focused on these omissions—yet—perhaps hoping companies will self-govern. But the risks are real; consider a VP of Corporate Communications shown digital content that could be genuine or could be a deepfake, who requests leadership’s guidance about how to determine what they’re dealing with so they don’t make a brand-damaging mistake.

Ethics

AI tools are labor-saving in many ways, but AI governance is labor-dependent and labor-intensive. The reason is AI governance has more to do with people adjusting how they work, than what tools they work with. This begs the question of the need for incorporating ethics into AI development, particularly for certain [AI use cases](#). And since we haven’t created a technology that makes decisions based on ethical judgment, humans must be the arbiters of ethics when it comes to technology design.



AI governance has more to do with people adjusting how they work, than what tools they work with.

Controls

AI governance models identify three tiers of human input in AI/ML technology design/operation:

- Human-in-the-loop (HITL): Human experts operate or are actively involved in the decision-making process.
- Human-on-the-loop (HOTL): Humans are involved in the decision-making process, and play a role in continuously monitoring results and evaluating AI/ML model performance.
- Human-in-control: Humans maintain ultimate control and responsibility for AI/ML systems regarding ethical, legal, and societal boundaries.

The following subsections detail more common issues that must be addressed by behavioral protocols, as well as technological solutions and security controls.

Data Leakage

While data leakage has always been near the top of a Cybersecurity professional's list of nightmares, the inclusion of natural language processing (NLP) models, including LLMs and other GenAI models, has added a new layer of concern. Without adequate safeguards, prompts sent to an LLM could easily include personally identifiable information (PII) or proprietary information. Consider a customer service representative in a call center for a financial services or healthcare company who uses their personal phone to send a query containing PII to ChatGPT or BERT to help them help the customer/patient. However innocent the intent behind that query, the customer or patient's information is now in the model's knowledge base and could be used as training data for the next iteration. Existing data leakage prevention (DLP) policies and controls that address only cybersecurity and not AI security could not prevent this classic data-leakage event.

Malicious Code

A developer queries a GenAI model to find and fix defects in their code and includes the company's code library in the prompt. Within minutes, the model provides ostensibly better code. The developer loads the new code into the firm's code repository and begins the automated build process without realizing the generated code contains an easily exploited vulnerability.

The benefit of the developer's increased productivity is offset by several very serious concerns:

- The potential cost—financial and otherwise—of a system breach by bad actors who could find and exploit the vulnerability before it can be patched
- The lack of controls to prevent externally-sourced code from entering the company's system
- The lack of controls to prevent proprietary content from being shared externally

LLM Robustness

While NLP and GenAI models are accessible, seemingly authoritative sources of mostly sound information, each model's robustness—including the data sets it relies on, the model itself, and the software running the model (taken together, the AI or ML system)—[must be verified and validated](#).

“Model robustness” refers to three characteristics of a model's output:

- Integrity (Is it factual and accurate?)
- The presence of bias (Is it ethical?)
- The ease with which it could be used maliciously

Early-stage companies are designing tools to protect models from specific vulnerabilities. Cybersecurity leaders should invest time learning about platform capabilities that improve robustness:

- Meet with vendors to learn about technologies that can support AI governance at scale.
- Identify capabilities that support all the AI models in use, as well as LLM-specific tools.

Threat Vectors

Cyber criminals invest serious time and money devising ways to compromise AI/ML systems. Shrinking models' attack surfaces, including plug-ins, is a critical part of an AI governance program, as is understanding how adversaries exploit these new technologies. Authorities whose research on AI/ML attacks varies in perspective, but is equally rigorous and insightful, include [Scott Alfeld](#) from Amherst College and [Alexey Rubtsov of the Global Risk Institute](#), as well as [MITRE](#) and the [Berryville Institute of Machine Learning](#). Become familiar with all of them.

Traceability

Effective governance of AI tools across an enterprise requires that users' interactions be logged and traceable, although the level of detail will vary by organizational needs and policies, and regulatory compliance. If one framework principle is human verification of all LLM results, for example, the company must be able to track the user, the prompt, the response, and the verifier. Even the amount of information generated in an hour is instantly too much for humans to handle. While the models themselves offer no such capability, tools exist that offer traceability at varying levels and for a wide array of purposes for both internal and external LLMs.

Maintenance

For years, technology solutions from industrial robots to doorbell cameras have included AI/ML features and elements. Mirroring the web app development history discussed earlier, enterprises actively using such solutions have not hired personnel to verify, validate, or account for the results, or even track the purchase, use, and maintenance/obsolescence of the tools. IT and security leaders familiar with the challenges of IT hygiene (configuration and vulnerability management, asset inventory and management) at enterprise scale must now integrate management practices for AI governance into the definition of IT hygiene.

Resources

Use existing research and expertise to ground the foundations of your AI governance program. The following list provides current resources; the first two are particularly useful:

- [Artificial Intelligence Risk Management Framework](#) (NIST)
- [Model AI Governance Framework](#) (Government of Singapore)
- [Principles on Artificial Intelligence](#) (OECD)
- [Ethics Guidelines for Trustworthy AI](#) (European Commission)
- [Global Initiative on Ethics of Autonomous and Intelligent Systems](#) (IEEE)
- [Ethical Guidelines and Principles in the Context of Artificial Intelligence](#) (ACM)
- [Recommendation on the Ethics of Artificial Intelligence](#) (UNESCO)
- [MITRE ATT&CK® Framework](#) (MITRE)

These frameworks are heavily influenced by current and pending privacy-related regulations and legitimate concerns about expanding AI development, such as [Artificial General Intelligence](#) (AGI). This topic has been discussed in academia for years as an alternative way to describe computers and software that perform complex tasks as a human might or could.

Business Case For AI Governance

The business case for AI governance has two critical cost pillars: technology and labor. While using LLMs securely demands new technology that is costly, it is less so than the HITL needed for all AI applications to meet regulatory and cyber risk management goals.

Software use and license costs for the AI governance platforms are typically based on usage and users, and dependent on the size of the enterprise. Presuming Year 1's cost is \$500k with \$1.0M for Years 2 and 3 and an additional \$200k for SaaS plug-in costs, the annualized Y1 cost is \$1.5M, and \$2.0M for Y2 and Y3.

Implementing an AI governance program enterprise-wide requires an accountability model that identifies the AI system, its owner, and the individual providing traceability confirmation. The committee must ensure identified employees understand they are responsible for the output of a given model or models, and are stewards of those models on behalf of the enterprise. If the accountability model is enforced effectively, the need to add support personnel should be insignificant. For example, an in-house data scientist who creates and deploys a customized LLM must own the verification and validation of that AI/ML system. Their labor cost is already in the budget, with the time spent acting as the AI/ML system owner incorporated as part of that data scientist's responsibility. This mirrors a "[champion](#)" structure in software security, in which developers collaborate to improve software resilience and the enterprise benefits because the need for support staff is minimal.

Using LLMs in core business workflows can lead to significant productivity gains and labor cost savings in the immediately subsequent years—as much as 20% in key areas—which is critical to justifying both the investment and the sensitive topic of staff reductions. Take, for example, a company with 50,000 employees, half of whom could be more productive using AI models. Multiplying that half by the average cost (e.g., \$150/hour) and then by 10% shows an annual benefit of \$750K that offsets the investment in Y1.

The business case focuses on counterbalancing Y1 implementation and purchase expenses with productivity benefits. If these gains don't meet expectations, adjust the resource time to manage costs. Factor in the potential reduction in privacy/compliance fines to further strengthen the business case.

Conclusion

LLMs and GenAI models are enjoying very high visibility right now and their deployment across the enterprise is increasing steadily, which means the technology has assimilated into the business landscape. Now is the time for AI governance to be integrated as a new component of the enterprise cybersecurity ecosystem.

Creating a business case for this investment of time, technology, and resources to support AI governance has an inherently strong foundation for a solid return on investment: minimal costs to achieve high productivity gains. Incorporating lessons from the past by bringing automation controls into the software build and/or implementation process strengthens the case.

In combination with the AI accountability model, the approach outlined in this paper establishes a comprehensive, operative AI governance model that enables enterprises to manage extended risk and experience measurable business benefits aligned with company values, principles, and policies. But, in the final analysis, successful enterprise-wide AI governance frameworks depend on the commitment of the people who design, develop, deploy, and use the AI tools, the Cybersecurity leadership, and the steering committee to implement and adhere to the program principles and metrics.





About the Author

Renowned cyber security industry expert and thought leader [Jim Routh](#) has served as the CISO/CSO for CVS Health, Aetna, KPMG, DTCC, American Express, and MassMutual. He is a member of the Board of Advisors for CalypsoAI.



About Us

CalypsoAI is the leader in developing and delivering AI security solutions. The company's vision is to be the trusted partner and global leader in the AI security domain, empowering enterprises and governments to leverage the immense potential of generative AI solutions and Large Language Models (LLMs) safely and securely.

CalypsoAI is driving the field in shaping a future in which technology and security coalesce to transform how businesses operate while contributing to a better world. Founded in Silicon Valley in 2018 by top minds in the fields of data science, machine learning, and defense, the company has secured investments from Lightspeed Venture Partners, Lockheed Martin Ventures, Paladin Capital Group, Hakluyt Capital and Expeditions Fund, and strategic angels, including Auren Hoffman and Anne and Susan Wojcicki.

To learn more, visit our [website](#) or follow CalypsoAI on [Twitter](#) and [LinkedIn](#).

CALYPSOAI