JULY 2023

THE ROLE OF THE CISO SHAPING TRUST IN THE AGE OF ARTIFICIAL INTELLIGENCE

Authored By:
Jim Routh, ICIT Fellow

ICIT | Institute for Critical Infrastructure Technology
The Cybersecurity Think Tank

The Role of the CISO Shaping Trust in the Age of Artificial Intelligence

July 2023

- Authored by Jim Routh, ICIT Fellow

Contents

Introduction

Enterprises are generating explosions of activity as they prepare to harvest business benefits from the many opportunities that Large Language Models (LLMs) and generative artificial intelligence (GenAI) models offer. History has shown that enterprises adopting emerging technologies typically lead with function. In contrast, controls on the use of the technologies lag. The evolution of governance models for AI usage at the enterprise scale appears to follow the same pattern, with controls following implementation and use cases.

Ample data exists about how the widespread adoption of LLMs by digital consumers, employees, contractors, and third parties is increasing the use of AI and machine learning (ML) in core business processes—and expanding attack surfaces across the enterprise. Initiating AI governance at enterprise scale represents a broad scope. It requires many people in diverse roles across the enterprise. The Chief Information Security Officer (CISO) who seeks to manage the increasing threat surface presented by LLMs while avoiding becoming the naysayer to business stakeholders must walk a fine line. This paper defines and describes the role of the CISO in AI governance and in establishing an enterprise-specific AI Governance Framework. It also provides context for these challenges, identifies options to pursue, and enables the many business opportunities while balancing risk management practices for the enterprise.

To ensure there is no misunderstanding when it comes to terminology as we dig into the topic, I offer the following commonly used definitions in layperson's terms:

- **AI** - a branch of computer science focused on creating machines with the ability to replicate human cognitive functions and the application of this field to software
- **GenAI** - a field within AI focused on creating algorithms, typically called "models," capable of generating new content or output following the data set that trained it (including images, audio, video, etc.)
- **LLMs** - a type of GenAI model trained on text data to generate human-like text
- **AI-ML models** - broader sets of algorithms trained to identify patterns within ML in various applications, such as natural language processing (NLP), recommendation systems, and analytic systems
- **Foundation Model** – an extensive base model' that serves as the underlying mechanism for many large-scale applications of ML; LLMs are an example of a type of foundation model

Where We Are Now

Every day users are finding new applications for GenAI products. Many of us improved the quality of our home project deployment and repairs by starting with research from YouTube videos and watching demonstrations of a technique or troubleshooting method. I know this lowered the damage level of my home improvement project results. Going forward, consumers will have even more sophisticated information available in the form of output from an LLM to guide them every step of the way, enabling a more comprehensive range of projects for home improvement opportunities.[i]

Digital consumers and professionals have access to a wide array of LLM plug-ins, such as for popular apps like Expedia, Instacart, and Open Table[ii], spurred by the growth in software-as-a-service (SaaS) in

enterprises. While the linkages of popular SaaS products to LLMs provide clear benefits, they also increase the enterprise attack surface.

The GenAI market is growing at a compound annual growth rate (CAGR) of 35.6%,[iii] expanding the potential use cases for LLMs, for instance:

- Leading consumer-brand enterprises use LLMs to improve product recommendations for e-commerce and enhance digital consumer interactions through websites, mobile applications, and call centers. These portals rely on virtual assistants, chatbots, and customized applications powered by an LLM, which seems to the consumer to be an actual person.
- LLMs that support natural language functions help healthcare professionals improve patient outcomes by retrieving essential information from clinical medical records.
- In the financial services sector, they help analyze market trends and the impact of financial news, enhancing customer service and providing personalized financial advice.
- They are used to generate compelling marketing content (articles, blogs, social media) and create personalized marketing campaigns.[iv]
- LLMs provide key applications, such as sentiment analysis and text classification, that improve the quality of the analysis, social media monitoring, emotions, and intentions expressed in text data.[v]
- Many financial institutions use LLMs to recognize patterns in streaming data across channels to enhance fraud detection.[vi]

The number of AI/LLM products and service offerings available to enterprise users is dramatically expanding the range of choices about which business processes they should enhance or disrupt for the benefit of the enterprise. Users are discovering how to employ LLM products to enhance their digital consumer experience while improving professional productivity creatively. As a quick example, a recent query (7-6-2023) of ChatGPT3 offered the following LLMs available to digital consumers:[vii]

- GPT-4, OpenAI
- Falcon, TII
- Claude, Anthropic
- Cohere,
- ChatGPT-3 developed by OpenAI
- GPT-2 (Generative Pre-trained Transformer 2), OpenAI
- T5 (Text-to-Text Transfer Transformer), Google
- CTRL (Conditional Transformer Language Model), Salesforce
- ProphetNet, Microsoft
- Bart (Bidirectional and Auto-Regenerative Transformers), Meta

- BERT (Bidirectional Encoder Representations from Transformers), Google
- XLNet
- Transformer-XL
- ALBERT, a lite BERT that reduces the model size
- ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately
- GPT-Neo, Open source
- XLM (Cross-Lingual Language Model), Facebook/Meta
- MegatronLM, NVIDIA
- MarianMT, the University of Edinburgh, and the University of Tartu

Increasingly, enterprises use LLMs tailored to specific business cases within their industry or domain. Many create their models using in-house data scientists and software engineers with access to large amounts of company- or industry-specific data. While I do not consider this wave of technology innovation a Fourth Industrial Revolution (4.0) that will transform work as we know it[viii], it will drive job consolidation. Some roles will be eliminated while others are created, for instance, interpreting and approving AI model output for robustness and accuracy.

The evolution of roles has two key outcomes: new roles that require skills to create AI models and new skills to incorporate AI output into changing business processes. For example:

- The benefits of using LLMs in large industries, such as healthcare, include:
    - Increasing the availability of and the ease of access to data
    - Lower costs of computing power
    - More customized products and services to enhance patient care
    - Dramatic improvements in operating costs and accuracy[ix]
- Intensive text-search industries, such as book publishing, healthcare clinical research, academic research published in multiple languages, and document preservation and management functions, can now be analyzed in hundreds of different languages to summarize concise answers to analytic-based queries in seconds.[x]

When such an extensive array of LLM-specific use cases exists to benefit most aspects of any enterprise today, it is not feasible for a CISO to "block" the use of LLMs until the risks to the enterprise can be fully understood.[xi] Doing so may be more damaging to the credibility of the CISO over the long term, given the significant potential of business disruption and breakthrough optimization of existing processes. If the Chief Counsel recommends that the CISO institute an enterprise-wide ban on LLM usage, I would encourage them to consider a different path, such as educating their business stakeholders about recent history as an aid.

Historical Perspective on Disruptive Technology Innovation

Following the Y2K remediation and preparation work, every large- and medium-sized enterprise focused on building web applications with transaction processing capabilities to serve the growing digital consumer base better while lowering operating costs. The mission of consumer-centric companies was e-commerce, emphasizing shrinking their "brick and mortar" footprint. As an IT leader in that era, my focus was building new web applications to enhance online customers' digital experience. However, I did nothing to attempt to shrink the growing attack surface for the enterprise while or after developing the code. Neither did any of my IT colleagues.

We were part of the "dot.com" boom.[xii] Our primary focus was building new capabilities while leveraging the additive investment in web assets. Time to market was the key driver. We were experimenting with accelerating the software development cycle to deliver functionality sooner. The fact that we were significantly increasing the attack surface of our enterprises and linking web applications to back-end databases that were never designed for Internet-facing exposure was never much of a concern. Enterprises were spending hundreds of millions of dollars creating e-commerce sites. Java developers had more work than they could handle, prompting extensive side-use of offshore developers to lower development costs and demonstrate agility.

The company I worked for at the time, American Express, had approximately 70% of its development resources offshore to meet the incredible demand for web development. Much of the focus of the dot.com boom and eventual bust was on the highs and lows of investor appetite (primarily venture funds) and their impact on the economy.[xiii]

As a cybersecurity practitioner, I feel a strong sense of guilt for not using the available tools, methods, and techniques for building more resilient software for digital consumers. In retrospect, besides some basic authentication functions, the Web applications I built never considered security implications—which was the norm. A simple check of the software security information available at the time, or soon afterward, makes it plain that I was oblivious to it.[xiv]

E-commerce was a disruptive force that created significant business opportunities to deliver more consumer capabilities at a lower operating cost. Unfortunately, enterprises did not consider the costs associated with cybersecurity incidents when building a business case for time-to-market digital opportunities. It took over a decade and many cyber breaches before these costs were understood, much less factored into future business cases at the enterprise scale.[xv]

Today, when the DevSecOps teams in large enterprises are assigned new application development projects, they use cybersecurity guardrails built into the automated software design process with policy-as-code configuration management options increasing the resilience of the end product. Developers and Site Reliability Engineers (SREs) have a better understanding of the tools, techniques, and methods for automating controls into the continuous integration/continuous delivery (CI/CD) pipeline, which lowers the cost of supporting software by reducing the attack surface.

Why did I ignore the need for software security almost 20 years ago? It is a question I often consider, especially given that I have spent the intervening years trying to improve and evolve cybersecurity practices in software development and all other aspects of enterprise risk management. I had concluded that I *should have* incorporated all available software security practices when we built those web

applications but did not—even though we know historically that criminals constantly adapt and discover ways to exploit societal systems and advancements. Since I am now a cybersecurity practitioner (albeit a recovering CISO), I recognize the error of my earlier ways and continually encourage other CISOs to avoid the "Ostrich Syndrome."[xvi]

Several CISOs told me recently that they believe most enterprises already have all the tools, techniques, and controls needed for dealing effectively with the risk of widespread usage of LLMs. In other words, they do not need to consider additive or different controls for AI governance at the enterprise scale. This thinking represents a classic example of the "Ostrich Syndrome." As technology evolves, our cybersecurity controls must also evolve. It does not necessarily mean we need a new philosophy or principles. However, it *does* require an additive AI governance model for the enterprise. Whether it is guilt (for having ignored software security controls during the dot.com boom) or my enthusiastic endorsement of innovations in cybersecurity control design as technology evolves, I want to encourage CISOs to seriously consider adding AI governance to their enterprise priorities—now.

We do not need to repeat history. We already know that waiting to add new control capabilities to AI development creates the same problems as building a website or mobile app with minimal security controls or launching a cloud-hosted, customer-facing software application with poor configuration management practices. CISOs without an AI governance program should begin drafting one today, starting with understanding the existing regulatory guidance, the AI tools available, and how to follow the pattern of quality software development. The following section provides suggestions for creating a plan.

Toward an AI Governance Framework

My review of sources for enterprise AI governance frameworks leads me to offer the following recommendations:

1. It will be well worth the time spent reviewing the available frameworks to consider the options available to the enterprise. I have found these two to be the most useful:
    - *Artificial Intelligence Risk Management Framework* (National Institute of Standards and Technology)[xvii]
    - *Model AI Governance Framework* (Government of Singapore)[xviii]
2. The following sources can also provide value:
    - *Principles on Artificial Intelligence* (The Organization for Economic Co-operation and Development)[xix]
    - *Ethics Guidelines for Trustworthy AI* (European Commission)[xx]
    - *Global Initiative on Ethics of Autonomous and Intelligent Systems* (Institute of Electrical and Electronics Engineers)[xxi]
    - *Ethical Guidelines and Principles in the Context of Artificial Intelligence (*Association for Computing Machinery)[xxii]

- *Recommendation on the Ethics of Artificial Intelligence (*United Nations Educational, Scientific and Cultural Organization*)*[xxiii]
- *MITRE ATT&CK® Framework* (MITRE)[xxiv]

Understand that these AI frameworks are heavily influenced by existing and evolving privacy-specific regulations and regulatory enforcement. Legitimate concerns are expressed about the unfiltered pursuit of AI development, including the potential for Artificial General Intelligence (AGI).[xxv] AGI has been a topic of intense discussion in academic circles for many years as an alternative way of describing computers and software performing complex tasks the same way a human might/could.

CISOs must lead the process of achieving consensus on appropriate AI governance principles for the enterprise and include Legal, Privacy, Compliance, Sales, Operations, IT, and HR in the consensus-building process. It is easier for a small group to create and achieve consensus on principles than to agree on how to rewrite higher-level policies requiring precision. The most important task in this effort is to focus on what changes with the inclusion of these new technologies and identify ownership of the changes. Here are a few sample principles used by enterprises:

- All output of LLMs will be reviewed and owned by a person.
- Software code generated by an LLM is not allowed into the code repository or software pipeline without being checked by a human in a standard or automated review process.
- Employees and contractors must reference external LLMs, including the date when writing internal correspondence that contains LLM output.

If I were tackling this topic for a large enterprise today, I would follow the approach in Table 1.

Table 1. AI Governance Framework Outline

Step	***Rationale***
Form a cross-functional set of stakeholders to initially approve a set of AI governance principles for the enterprise.	*The need for AI governance will grow for every enterprise as the scope of AI usage evolves; therefore, start with defining a set of principles explicitly written to guide the enterprise around AI tools and agreed to by the stakeholders. Principles are easier to come to a consensus on and more accessible to implement than policies.* *AI governance requires a multi-disciplinary approach for the enterprise. Candidate functions include Legal, Privacy, Compliance, Sales, Operations, IT, and HR.* *Policies are more prescriptive and rigid than principles. Amend the policies after principles are established.*

Step	***Rationale***
Link the AI Principles to the company values.	*Connecting AI governance principles with the company values provides effective grounding for what is. It will always be a human-centric governance process.*
Acknowledge the need for adjustments to existing controls in these areas: data leakage, malicious code ingestion, data integrity, software security, audit logs	*These are the control design opportunities with AI models and, specifically, LLMs:* *1) Proprietary data leakage through queries* *2) Creation of software code from GenAI models put directly into the software repository and build process.* *3) System robustness (accuracy, bias, integrity)* *4) Traceability*
Embrace exploration of AI usage and prioritize opportunities	*Acknowledge that managing business risk is healthy for the enterprise to grow, so preventing the application of emerging technology can harm the business in the long term. Learning using emerging technology is essential for innovation. It should be a priority with specific processes established to capture the experience.*
Consider AI capabilities to support your cybersecurity program	*Data science is foundational for cybersecurity programs, and AI applied to control design enables real-time risk management using models.*
Enforce accountability 20	*Building robust AI systems initially is less expensive than fixing existing systems later. Therefore, encourage AI system creators (designers and engineers) to demonstrate accountability for building them well while shrinking the attack surface.*

AI governance has more to do with people adjusting how they work than using additive technology (although new tools for AI quality are essential). The ongoing and historical debates in computer science about AI evolution consistently identify the need for ethics to be applied to AI development. The more specific the AI use cases, the more important human ethics matter to guide that usage and protect the enterprise.[xxvi] Humans are essential when applying ethics in technology design since we have not yet built technology to make decisions based on ethical judgment. When you get 20 or more people together, it is too many to apply ethics to a decision, and the decision-making process is challenged to reach an outcome. This is another reason to consider a team of six to eight leaders to create the enterprise AI governance principles.

Remember that to lead a process resulting in consensus, the facilitator (CISO) must demonstrate neutrality or an unbiased view of the result. Avoid being cast as the "subject matter expert" because being in that role implies an opinion and makes facilitating consensus more difficult. Linking the AI governance principles with current company values and the need to protect the brand will help ground the principles and clarify their business value to employees.

AI governance models often refer to three ways to engage humans in the design and operation of the technology:

Human-in-the-loop (HITL) refers to AI or ML in which human experts operate or are actively involved in decision-making.

1. Human-on-the-loop (HOTL): This is an extension of HITL in which humans are involved in the decision-making process, continuously monitor the results, and evaluate the AI/ML performance.
2. Human-in-control implies that humans maintain ultimate control and responsibility for AI systems to provide ethical, legal, and societal boundaries.

One innovative healthcare enterprise recently established a principle that all use of AI will be "explainable."[xxvii] Ensuring all AI is explainable implies that a human must understand how the system works well enough to explain the system and the results to people possibly unfamiliar. This approach requires a form of HITL. AI explainability aims to open the so-called "black box" of AI models by detailing to regulators how AI models work and how the individual output is created.[xxviii]

There are tremendous advantages to using AI systems to improve content quality significantly. However, humans remain essential for applying ethics and judgment to make AI commercially practical.[xxix] In other words, AI governance is labor-dependent and labor-intensive, an interesting contradiction to the perception that AI replaces the need for human professionals to produce work. LLMs can produce coherent output in a highly productive way compared to human-produced text. However, the quality of the text will be compromised. A recent article in *The Atlantic* written by noted computer scientist Douglas Hofstadter provides an excellent example of comparing output produced by ChatGTP-4 with human-created content.[xxx]

Much of the current regulatory guidance related to AI governance globally and in the U.S. is based on the precedent of privacy regulations and practices. If you have a Chief Privacy Officer, include her/him in creating the principles. Recognize that the active use of GenAI models (such as ChatGPT and Bard) is not compliant with existing global privacy regulations, specifically the General Data Protection Regulation (GDPR) requirements for consent, anonymization of data, and personal data deletion.[xxxi] Regulatory entities are ignoring compliance gaps for the time being and likely hoping that enterprises devise human-centric governance capabilities to address them.[xxxii]

AI governance applies to content generated from multi-modal models or models that use text, images, and audio.[xxxiii] Consider the additive attack surface for the widespread use of GenAI or LLMs in an enterprise: The head of Corporate Communications requests digital content (video, audio, and text files), specifically in determining what content is genuine and what might have been produced by GenAI as "deepfakes" or mis/disinformation. CISOs seeking to understand this extended attack surface should consider methods for determining the origin of digital output used by the enterprise for branding and communication purposes, considering that some fake content could have been produced in-house for

legitimate, or at least non-malicious, reasons. The most effective way to consider an enterprise attack surface from the use of AI is as follows:

Data Leakage

The initial concern from a cyber-risk perspective is that it is relatively easy to pack content that may include personally identifiable information (PII) or proprietary information into a query. As an example, if I were a bank's customer service representative in a call center and used ChatGPT on my personal phone to discover more information about a customer inquiry and included an account number or other piece of PII in the prompt, that information becomes part of the LLM's or GenAI tool's data set. This represents a classic data leakage event, and while there is likely existing data leakage prevention (DLP) policies and controls, none would prevent this leakage of PII.

Malicious Code

I am a developer working on a challenging assignment and struggling under a tight deadline to de-bug an enhancement to an existing system. I decide to use a GenAI model for help. I load the code library into the model query and request help identifying the defects with a better Python code example. The GenAI model does its thing and delivers better-structured Python code that will solve the problem without code defects (in theory). I load that code into the company's code repository, submit it into the automated build process, and notify my boss that I have fixed the problem. The application should be operating correctly after the build updates.

Unfortunately, the code generated by the LLM includes an open-source component recently discovered to have an easily exploited defect or vulnerability. The dramatic improvement in my developer productivity is offset by the impact of a potential breach for the enterprise by threat actors exploiting both the vulnerability and the lack of controls that would have prevented the code from being shared in the first place.

LLM Robustness

AI models, LLMs, and GenAI models appear to be excellent, easily available, authoritative sources of quality information. However, there is a need for validation and verification of the model's robustness.[xxxiv] In AI-speak, model robustness equates to the integrity of the output (Is it factual and accurate?), the inclusion of bias (Is the output ethical?), and whether the output could be used maliciously. This applies to the data sets, the model, and the software running the model, often collectively referred to as the AI or ML system. A HITL or HOTL system represents a safeguard but one that is difficult to scale enterprise-wide for all LLM use cases under consideration.

There are choices available today from early-stage companies designed to meet the use case of protecting LLMs from cyber-specific vulnerabilities, and there is a strong appetite in the venture capital (VC) community for funding start-ups in AI protection. I recommend that CISOs invest their time in getting educated from vendors with platform capabilities to improve the robustness of LLMs, whether external or internally developed.

- Meet with vendors (today) and get educated on how technology can help you and your team apply AI governance at scale. Some vendor products help produce an inventory of AI models, some identify a few risks with the inventory and other products provide protection for the use of LLMs broadly while tracking usage trends

- Discern which capabilities support the enterprise's inventory of AI models and which ones solve the immediate needs of LLM protection. Eventually, you will need both capabilities to scale across the enterprise.

Threat Vectors

Cybercriminals are investing heavily to determine how to compromise AI/ML system usage. Shrinking the attack surface of the models created by and for the enterprise is part of the scope of the AI governance program. It requires a different toolset from the software security program. Understanding how adversaries exploit these new technologies is critical, and some researchers have been engaging on the topic for years. The following authoritative examples provide different perspectives in their discussions of threat actors' operations, and CISOs should be familiar with all of them.

1. Computer science professionals like Scott Alfeld from Amherst College have been researching methods of attack on AI systems for many years.
2. "Adversarial learning" describes the attack of AI systems by poisoning the data set or influencing the model. According to Alexey Rubtsov, an expert researcher in adversarial ML attacks, there are four basic types of attacks on AI/ML systems:
    a. *Poisoning attack* - Adversary manipulates training data set to modify the output for malicious purposes.
    b. *Evasion attack* - Attackers change the input to a model's training data slightly, resulting in an output that can be used for malicious purposes like creating malicious code.
    c. *Extraction attack* - Attackers steal a copy of the model and its outputs to use for illegal or nefarious purposes.
    d. *Inference attack* - Attackers take advantage of biases in the training data to bypass controls or fool the model to produce a specific output for malicious purposes.
3. Another resource for cataloged attacks is one that cybersecurity leaders are familiar with: The MITRE ATT&CK® Framework.[xxxv] MITRE also developed a knowledge base of adversary tactics, techniques, and case studies for ML systems, patterned after the MITRE ATT&CK® Framework and based on real-world observations, called MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems, or ATLAS™.
4. Another thoughtful description of AI/ML attacks was published by the Berryville Institute[xxxvi] in 2019 that included discussions of:
    a. Input Manipulation
    b. Data Manipulation
    c. Model Manipulation
    d. Input Extraction
    e. Data Extraction
    f. Model Extraction

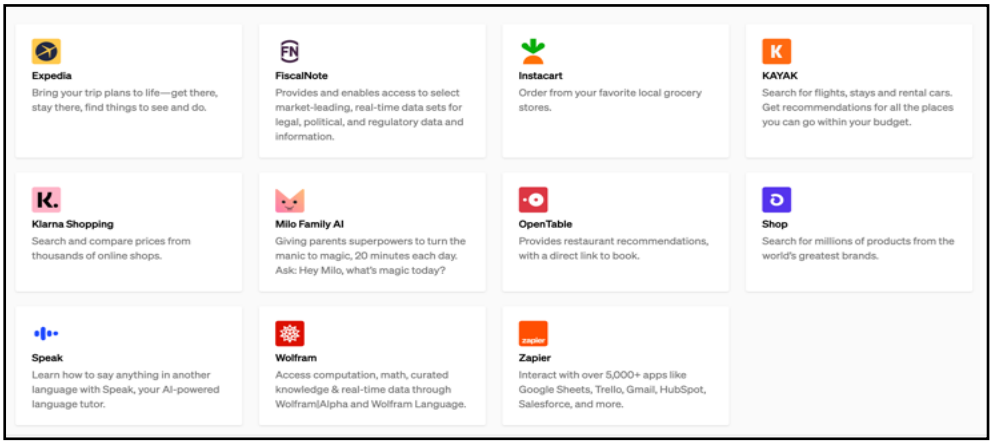
Traceability

LLM usage should include producing log records to meet traceability requirements. Let us assume an enterprise adopted a principle that states all use of LLM results will be verified by a human. The enterprise must link an individual verifier to a specific set of records at a specific time with specific results. This level of detail is not something humans can handle well, and it certainly is not provided by an external LLM. LLM usage will evolve as user grow accustomed to using it effectively and tracking usage trends is essential to understanding levels of protection required.

However, LLM platform capabilities provide traceability for external and internal LLM usage, enabling enforcement of the enterprise policy supporting the principle in a way that demonstrates due diligence for regulation adherence. Managing the risk of LLM usage with the ability to demonstrate compliance consistently and at scale requires a technology option.

As more and more employees use tools like ChatGPT for core business processes, the LLM creators will incorporate this usage into other SaaS offerings. This will further shift and expand attack surfaces the enterprise will have to consider. Here is a list of SaaS tools compatible with ChatGPT, according to OpenAI as of July 07, 2023:

Figure 1: SaaS Tools for ChatGPT



In many enterprises, hundreds of technology solutions designed, built, and purchased use some form of AI. Unfortunately, an enterprise likely does not have personnel who do the verification and are accountable for the results, even though the number of AI solutions is growing significantly. CISOs have faced the challenge of IT hygiene (configuration and vulnerability management, asset inventory and management) at enterprise scale alongside their IT colleagues. Now they must add management practices for AI governance to the definition of IT hygiene, and some of the same challenges await them.

Over time, methods for managing AI governance risk will evolve to include a complete inventory of enterprise models created internally and externally used in enterprise business processes. This is essential for privacy compliance and managing the enterprise's cybersecurity risk. Make this automation part of the security architecture and find solutions now to help define the evolving requirements for AI governance.

AI governance needs are already beyond the scope of current controls and governance processes. Therefore, AI governance is additive to enterprise cybersecurity programs and requires adequate

funding over the next several years. The good news is that the business case is relatively straightforward.

Business Case for AI Governance

Enabling LLMs with effective controls requires some new technology tools in a security portfolio. However, these costs are relatively insignificant compared to the cost of providing the HITL necessary for all AI-driven applications to satisfy regulatory requirements and manage cyber risk effectively. The key is to use the cross-functional steering committee that created the AI Governance Principles (see Table 1) to enforce the accountability model shown in Table 2.

Table 2: AI Accountability Model

#	AI System Name	AI System Owner	Traceability Confirmation
1	Our favorite LLM	Betsy Ross	Wilma Rudolph
2	Consumer targeting model 1	Elon van Musk	Jon Stewart
3	Security Software for LLM	Shahrokh S.	Abbie B.

AI governance at the enterprise scale requires people and focus. The key concept is applying an accountability model in which employees step up to be stewards of the AI models on behalf of the enterprise. With this in place, the business case for implementing AI governance across the enterprise requires fewer resources, assuming employees' willingness to demonstrate accountability for the AI output.

The business case components include the labor hours to support the governance model, which is an annual cost. Remember that the number of people additive to support the program should be insignificant if the accountability model is enforced effectively. In other words, if I am a data scientist and I create and deploy an LLM that supports the business, then I must acknowledge that I own the validation of the AI system. My cost is already in the budget (last I checked), so it is not additive to the program. The issue is getting buy-in from key stakeholders that the small percentage of time allocated to being the ML system owner is part of the responsibility of being (in this case) a data scientist. No incremental cost for demonstrating ownership is necessary.

The AI Accountability Model mirrors a "champion" program in software security, in which developers help others improve software resilience.[xxxvii] The benefit is that an enterprise can support the program with minimal support staff. The same can be true for AI governance if those who create the models are responsible for verification.

The high-level budget request for labor might look something like Table 3.

Table 3: High-Level Budget Request (Labor)

Resource Type	Hours	Hourly Rate	Annual Cost	# of Staff	Total
Data Scientist	40	$225	$9,000	16	$144,000
DevSecOps Lead	40	$225	$9,000	4	$36,000
Cyber Analyst 1	1,200	$160	$192,000	1	$192,000
SRE	60	$180	$10,800	6	$64,800
Privacy Analyst	25	$160	$4,000	3	$12,000
Total					**$448,800**

There would be software use and license costs for the AI governance platforms, based on increases in usage and users and dependent on the size of the enterprise. If we calculate that cost to be $500,000 in Year 1 and $1,000,000 in Years 2 and 3, and add $200,000 for additional SaaS plug-in costs, then the annualized cost for Year 1 is $1.5M, rising to $2.0M in Years 2 and 3.

The business case's most important part is capturing the labor cost reduction and the productivity gain for using LLMs in core workflow and business processes. It is difficult and sensitive (HR) to address staff reduction benefits for the program, so simply identifying the anticipated productivity gain represents the biggest offset to the program investment cost. An enterprise will likely gain somewhere close to a 20% productivity gain for key resource areas each year for the next few years. If the base is 50,000 staff and 50% of them are in key resource areas where there is an opportunity, then multiply 25,000 by the average cost (say $150/hour) x 10% and calculate a $750,000 annual benefit, which offsets the investment of $448,800 in Year 1.

The business case is based on harvesting the productivity gains for the enterprise as an offset for the implementation and acquisition costs in Year 1. If the harvested productivity gains are lower than anticipated, then reduce the resource time to adjust the cost and/or add the probability of cost avoidance for the privacy/compliance fines that are less likely, which improves the business case significantly.

Conclusion

AI governance is here to stay for enterprises and is additive to current resources allocated to a cybersecurity program. However, building a business case for the resources and technology to support AI governance is not difficult, given the potential productivity gains from using LLMs with limited up-front implementation costs. Using history as guidance, CISOs should think of applying control

automation to the software build process (or CI/CD pipeline) specific to the AI component models built or bought by the enterprise.

This approach and the AI Accountability Model provide an effective AI governance model for enterprises to manage extended risk while enabling the business benefits associated with AI models to enhance existing workflows.

The greater the concern about the quality of AI systems, the more emphasis is needed for humans to apply ethics to develop AI capabilities. The chart below from the Office of the Privacy Commissioner for Personal Data, Hong Kong, defines the proposed structure for Ethical AI.[xxxviii]

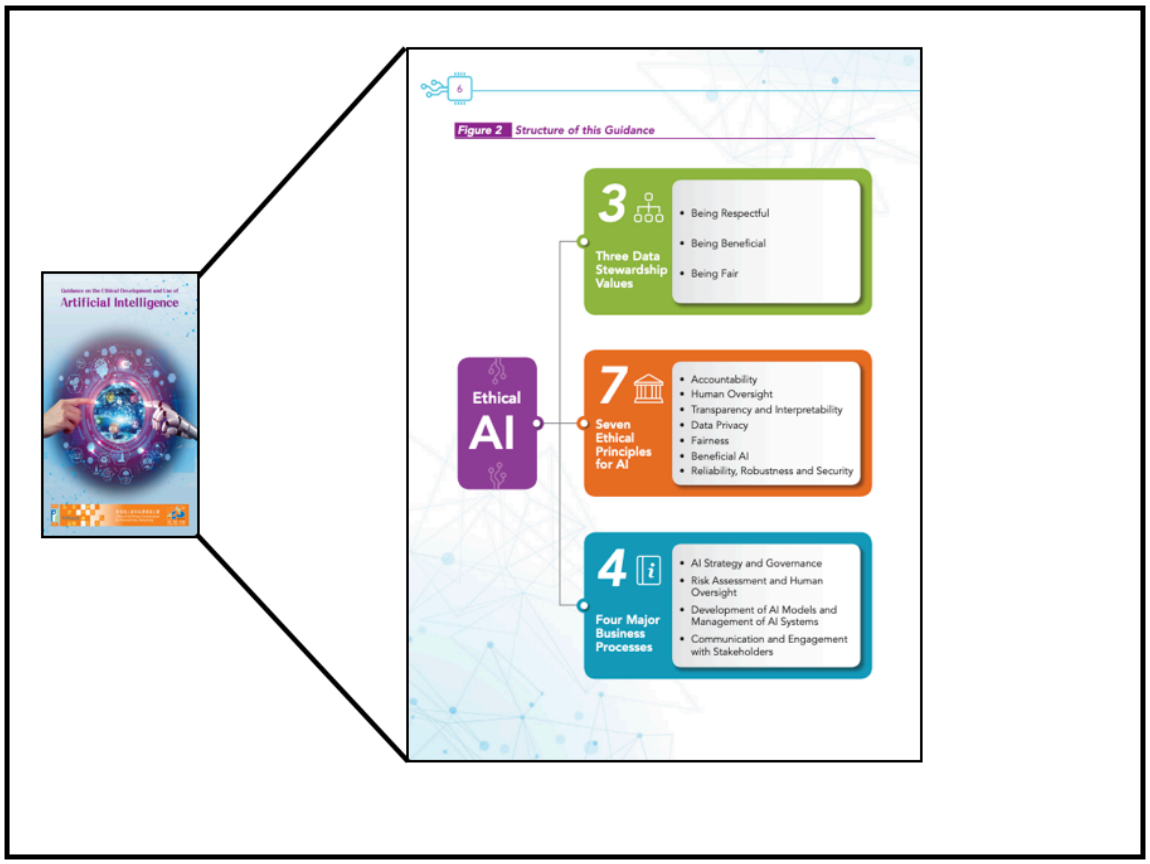


Figure 2: Proposed Structure for Ethical AI

When considering AI governance at an enterprise scale, people matter. Building an AI Governance Framework, implementation approach, and effectiveness metric are additive to enterprise cybersecurity programs as they exist today. CISOs who recognize the need for demonstrated accountability by the designers, developers, and users of AI systems throughout the enterprise will be the most successful in managing risk and compliance. Adding configuration management controls designed for the risks of AI to the CI/CD software pipeline is clearly the cybersecurity program's domain and the CISO's responsibility.

CISOs need to follow established patterns in software security accountability to succeed at enterprise scale.[xxxix] They must consider the control options and how best to integrate them with the software

pipeline or build process. Waiting for better controls or more mature regulatory requirements for guidance will increase the risk to the enterprise. The most effective near-term AI governance approaches will emphasize the accountability of GenAI, LLM, and AI/ML model creators to make the systems explainable.

About the Author

Jim Routh is currently serving on several boards and advising several companies. He is the former Board Chair of the Health Information Sharing & Analysis Center (H-ISAC), where he served for five years, and a former Board member of the Financial Services Information Sharing & Analysis Center (FS-ISAC). Jim is a former CSO/CISO for American Express, DTCC, KPMG, Aetna, CVS, and MassMutual. Jim brings a vast business and technology background to the boards he serves. He is considered a digital and cyber security industry expert and thought leader. Jim is an ICIT Fellow and an Adjunct Faculty member, where he teaches cybersecurity for the NYU Tandon School of Engineering.

About the Organization

The Institute for Critical Infrastructure Technology (ICIT) is the nation's leading cybersecurity think tank providing **objective, nonpartisan research, advisory,** and **education** to **legislative**, commercial, and public-sector cybersecurity stakeholders.

ICIT understands that only through generative and focused collaboration will cybersecurity and national security communities make the quantum leaps necessary to defend against today's hyper-evolving adversaries. In response, we facilitate a robust platform of programs, knowledge sharing, cutting-edge research, and publications that support the exchange of ideas and provide a forum for cybersecurity leaders to engage in the meaningful discourse needed to support and protect our nation's critical infrastructures effectively.

References

[i] Ed Anuff, "How LLMs are Transforming Enterprise Applications," The New Stack (June 8, 2023) https://thenewstack.io/how-llms-are-transforming-enterprise-applications/

[ii]" ChatGPT plugins," OpenAI blog (March 23, 2023) OpenAI.com/blog/chatgtp-plugins

[iii]" Large Language Models (LLM): An Ultimate guide for 2023," Shopdev (June 16, 2023) ShopDev.co/blog/what-are-large-language-models

[iv]" Real-World User Cases for Large Language Models (LLMs)," CellStrat, Medium (April 25, 2023) https://cellstrat.medium.com/real-world-use-cases-for-large-language-models-llms-d71c3a577bf2

[v] Anastasiya Zharovskikh, "Best applications of large language models," InData Labs (June 22, 2023) https://indatalabs.com/blog/large-language-model-apps

[vi] ODSC Team, "5 Practical Business Use Cases for Large Language Models" (March 2, 2023) https://opendatascience.com/5-practical-business-use-cases-for-large-langage-models/

[vii] Jim Routh, response to ChatGTP-3 query on June 28, 2023

[viii] Iqbal H. Sarker, "AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems," SN COMPUT. SCI. 3, 158 (2022). https://doi.org/10.1007/s42979-022-01043-x

[ix] Lloyd Price, newsletter dated May 25, 2023, Thought Leadership for HealthTech, M&A, Growth & Strategy, https://www.linkedin.com/newsletters/healthtech-m-a-7071223775639281664/

[x] Ben Kumar, "Large Language Model (LLM) in AI: Transforming Science, Society, and Industry, TechAffinity (July 6, 2023) https://techaffinity.com/blog/large-language-models-and-ai/

[xi] This is the author's opinion as a cybersecurity professional.

[xii] John Cassidy, *The Greatest Story Ever Sold* (Allen Lane: London) January 2002

[xiii] Peter Robert Wheale and L.H. Amin, "Bursting the dot.com 'Bubble': A Case Study in Investor Behaviour, *Technology Analysis & Strategic Management*, 15:1, 117-136 (2003) DOI: 10.1080/0953732032000046097

[xiv] Dr. Gary McGraw, *Java Security*, (John Wiley & Sons: New York) 1996; John Viega and G. McGraw, *Building Secure Software: How to Avoid Security Problems the Right Way* (Addison-Wesley: Boston) 2001

[xv] Sasha Romanosky, "Examining the costs and causes of cyber incidents," *Journal of Cybersecurity*, Volume 2, Issue 2, (December 2016) https://academic.oup.com/cybersecurity/article/2/2/121/2525524

[xvi] Jim Routh, response to ChatGTP3.5 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."

[xvii] National Institute of Standards and Technology, *Artificial Risk Management Framework (AI RMF 1.0)* (January 2023) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[xviii] Personal Data Protection Commission (PDPC) of Singapore and Infocomm Media Development Authority (IMDA), *Model AI Governance Framework, Second Edition*, (January 2019) https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

[xix] The Organization for Economic Co-operation and Development (OECD), *Principles on Artificial Intelligence* (May 2019) https://oecd.ai/en/ai-principles

[xx] European Commission, *Ethics Guidelines for Trustworthy AI* (April 2019) https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

[xxi] Institute of Electrical and Electronics Engineers, *Global Initiative on Ethics of Autonomous and Intelligent Systems*, https://standards.ieee.org/industry-connections/ec/autonomous-systems/

[xxii] SBSI '21: Proceedings of the XVII Brazilian Symposium on Information Systems, "Ethical Guidelines and Principles in the Context of Artificial Intelligence," Association for Computing Machinery (June 2021) Article No.: 36 Pages 1–8 https://doi.org/10.1145/3466933.3466969

[xxiii] United Nations Educational, Scientific and Cultural Organization, *Recommendation on the Ethics of Artificial Intelligence* (November 2021) https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=A%20human%20rights%20approach%20to%20AI&text=Unwanted%20harms%20(safety%20risks)%20as,and%20addressed%20by%20AI%20actors.&text=Privacy%20must%20be%20protected%20and,frameworks%20should%20also%20be%20established.

[xxiv] MITRE, The MITRE ATT&CK® Framework, https://attack.mitre.org/

[xxv] Future of Life Institute, "Pause Giant AI Experiments: An Open Letter," (March 22, 2023) https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[xxvi] Martin Ford, *Architects of Intelligence: The truth about AI from the people building it* (Packt Publishing: Birmingham, UK) (November 2018) https://www.packtpub.com/authors/martin-ford

[xxvii] Graphite Health, 2023; This organization is a co-operative of health providers developing advanced technology solutions to improve healthcare outcomes using a comprehensive digital ecosystem. https://www.graphitehealth.io/

[xxviii] Alexey Rubtsov, PhD, Senior Research Associate, "Artificial Intelligence and Machine Learning: A Model Risk Management Perspective," Global Risk Institute (August 23, 2022) https://globalriskinstitute.org/publication/artificial-intelligence-and-machine-learning-a-model-risk-management-perspective/

[xxix] Yanqing Dan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda," *International Journal of Information Management* (October 2019), Volume 48, Pages 63-71 https://www.sciencedirect.com/science/article/abs/pii/S0268401219300581

[xxx] Douglas Hofstadter, "Gödel, Escher, Bach, and AI," *The Atlantic* (July 8, 2023) https://www.theatlantic.com/ideas/archive/2023/07/godel-escher-bach-geb-ai/674589/

[xxxi] Charles R. Taylor, "Editorial: Artificial intelligence, customized communications, privacy, and the General Data Protection Regulation (GDPR)," *The Review of Marketing Communications* (July 13, 2019) https://www.tandfonline.com/doi/full/10.1080/02650487.2019.1618032

[xxxii] Author's note: This appears to be a temporary situation that is likely to change.

[xxxiii] Gadi Singer, "Multimodality: A new Frontier in Cognitive AI," Towards Data Science (February 2, 2022) https://towardsdatascience.com/multimodality-a-new-frontier-in-cognitive-ai-8279d00e3baf

[xxxiv] Ibid, Singapore's *Model AI Governance Framework*, Robustness, 3.31

[xxxv] Ibid. MITRE

[xxxvi] Victor Shepardson, G. McGraw, H. Figueroa, and R. Bonett, "Taxonomy of ML Attacks," The Berryville Institute of Machine Learning (May 2019) https://berryvilleiml.com/taxonomy/

[xxxvii] Marin Gilje Jaatun and D. S. Cruzes, "Care & Feeding of Your Security Champion," IEEE International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA) (July 21, 2021) https://jaatun.no/papers/2021/Care_and_Feeding_of_Your_Security_Champion-postprint.pdf

[xxxviii] Ibid, Singapore's *Model AI Governance Framework*

[xxxix] Open Worldwide Application Security Project, "Security Culture, Security Champions," https://owasp.org/www-project-security-culture/v10/4-Security_Champions/