Risk Management & AI Governance: Comments on the NIST GAI-PWG Concept Note

By Dennis D. McDonald, Ph.D.[1] Sept. 2, 2023

Introduction

The NIST (National Institute of Standards and Technology) is part of the U.S. Department of Commerce. The GAI Profile Concept Note, distributed to members of the NIST Generative AI Public Working Group (NIST GAI-PWG), is a short document that present's NIST's

> *"…initial perspective for a cross-sectoral AI RMF [Risk Management Framework] profile for generative AI models or systems. It is intended for broad feedback from working group members. NIST expects and encourages direct critique and challenging of its content."*

[1] *Dennis D. McDonald, Ph.D., Alexandria, Virginia. Interests: writing, business development, data governance, project management, technology adoption, digital strategy. Industry experience: civilian and military contracting, higher ed, consulting, nonprofits, manufacturing, software, and international development. Memberships: NIST Generative AI Public Working Group (NIST GAI-PWG); Alexandria Virginia Public Records Advisory Commission; American Association for the Advancement of Science. Volunteer, Alexandria Film Festival. Professional web site: www.ddmcd.com. Book reviews: http://www.ddmcd.com/books. Movie & media reviews: http://www.ddmcd.com/movies.*

Article Focus

My experience as a consultant, project manager, analyst, or writer on IT- and data-related research, planning, or system development projects colors my review of the Concept Note. My interests related to AI risk management include both AI governance as well as how information sources and intellectual property ownership are managed. In this article my primary focus is on the "governance" section of the Concept Note as it impacts GAI (Generative AI) related risk management. Future articles may address other areas including GAI provenance and intellectual property management.

Risk Management

Managing risk is always something to consider when planning or managing any project involving changes to technology and/or business processes. Planning the use of a specialized LLM (Large Language Model) application (such as ChatGPT) to support a specific business process within an organization must consider both familiar and unfamiliar types of risk. Risk types will vary and may include data ownership issues, problems with data accuracy, unplanned or surprise outcomes, ethical and privacy issues, and cybersecurity breaches.

Generative AI's Use of Intellectual Property

Regarding the question of Generative AI's incorporation (potentially unacknowledged) of someone else's intellectual property in its training, or in its generation of prompted output that incorporates local context information, there may be some similarity to how we govern contracts with consultants in tech related work.

It is not unusual when contracting for work for a consultant to sign both a nondisclosure agreement as well an agreement regarding intellectual property ownership. A typical clause in these agreements is that information if generated previously or already publicly known is excluded from such agreements. Consultants, after all are hired because of what they know or what they can do, so their state of prior knowledge and skill is expected to influence what they will generally be asked to perform under contract.

Are there analogs between this and how we might govern GAI applications to manage risk? As noted above it is normal when contracting with a management consultant to specify in writing what the job requirements are and what "rules" will be followed in carrying out a job. When building a GAI based application it is also normal to specify rules for how tasks will be carried out—including tasks that need to be flagged for management (human) review. For an example, see OpenAI's Using GPT-4 for content moderation which describes a method for automating the implementation of "content moderation policies" to reduce workload on human content moderators. Another example of using AI based rules in te review of content is in the incorporation of AI-based "plagiarism detection" features in content editing products such as Grammarly.

Concept Note Line 13: "Lifecycle" Reference

It is not clear what "lifecycle" refers to in the Concept Note. GAI applications can be built and queried directly or can be used as the basis for developing a variety of different purpose-specific applications through intermediary systems or API toolsets. Risk management will require different actions depending on where one is in the application development "lifecycle." Understanding the application's lifecycle context will be important in deciding on important actions. How—and when—"training" is performed for will also be an important consideration especially if an externally-maintained system or database of content is used under license as part of the risk management process.

Concept Note Text Box at Line 60: Governance and Actor Role Changes

Text box at line 60 of the Concept Note refers to:

1. Identifying changes to governance practices forced by managing the risk of GAI applications and
2. Defining "actor role changes" and the types of professions and disciplines involved.

Understanding how these two relate to each other will at least partly depend on the scope and complexity of the organization's intended GAI application. The more complex and far-reaching the application, the more complex and far reaching may be the risks and the actions needed on the part of concerned staff and stakeholders in the adopting organization. In other words, understanding or even limiting the scope of a Generative AI application within on organization will be an important element in risk management.

An analog to managing risk when implementing an organization's AI governance strategy is the work involved in developing and implementing a corporate data governance strategy to improve an organization's access to and management of its own data. Attempting too much or employing a too comprehensive "boil the ocean" strategy, risks failure due to a variety of factors including the widespread problem of project "scope creep," the likelihood that conditions will change the longer and more complex a project is configured, political and organizational resistance, and the "unknowns" associated with the implementation of any innovative technology.

This latter issue of "unknowns" may be a special problem associated with implementation of generative AI applications, given what is stated in the NIST Concept Note about "capabilities and adaptability" (lines 40-43): "GAI systems - and LLMs in particular — have been shown to exhibit capabilities that were not part of model training. Deployed models might therefore perform in an unpredicted or undesired manner." (This appears to be a recommendation to "expect the unexpected"!)

Concept Note Page 3: "Governance" Table

The Concept Note's Governance table lists seven points in column 1 along with three accompanying blank columns: "About," "Suggested Actions," and "Transparency and Documentation." To these three I've added "Role of AI" in the discussion below.

I interpret the "Govern" items listed in the table's first column as proposed requirements for effectively governing how risks are managed in a planned generative AI project.

Imagine, for example, an RFP issued by a public or private sector organization listing such requirements along with the requirement that the bidder's proposal describe how, for a particular generative AI application, each requirement will be satisfied.

The following are the Concept Note's listed requirements, along with my own comments and questions. The numbering is the original numbering supplied in the Concept Note.

Govern 1.1: Legal and regulatory requirements involving AI are understood, managed, and documented.

About. Responding to a government or private sector RFP requires knowledge of relevant legal and regulatory requirements and showing proof that these requirements will be adhered to.

Suggested Actions. Compliance with relevant regulations must be demonstrated. Such demonstration can take several forms including:

- A statement of compliance by an organization's legal representative based on self-assessment.
- Evidence of formal certification based on a third party assessment conducted within a specific time period.
- Provision of specific detailed evidence of how relevant regulations will be complied with during implementation and operation of the proposed system.

Transparency & documentation. How such compliance is documented will be key. As with cybersecurity management there is a limit to how public one must make explicit risk detection and mitigation measures, especially since cybersecurity threats are constantly evolving.

Role of AI. Government issued regulations, regardless of how they are enforced, tend to lag technology. This will be likely be true with generative AI applications given not just rapid upgrades to LLM tools and toolsets but also to the unpredictability the Concept Note points out, i.e.

GAI systems - and LLMs in particular — have been shown to exhibit capabilities that were not part of model training. Deployed models might therefore perform in an unpredicted or undesired manner.

Perhaps, then, it makes sense to consider development of new and creative approaches to regulation, possibly through development of regulatory monitoring and reporting tools based on generative AI technology.

Govern 1.6: Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.

About. I interpret the phrase "mechanisms in place to inventory AI systems" to mean that we know what AI system we will be using as well as relevant details concerning how they are installed, supported, trained, deployed, updated, and integrated with local and contextual systems and data. "Organizational risk priorities" may refer to an organization's exposure to loss of intellectual property or the personal information of its employees, customers, or clients or to liability associated with the generation or dissemination of inaccurate or false information inside or outside the organization.

Suggested Actions. Ideally, we want to identify and track how GAI based systems are being used in the organization. Some mix of tools providing digital asset management, network monitoring, data governance, cybersecurity attack detection, and the ability to monitor input and output (text, graphics structured & unstructured data, etc.) will be required. Training such systems to securely identify and track sensitive data will be essential, as will the efficient incorporation of human involvement to provide judgement, prioritization, evaluation, and decision-making.

Transparency & documentation. "Transparency" needs to be defined, i.e., "transparency of what to whom"? Consider the recent controversy surrounding Zoom's retraction of a Terms of Service revision that appeared to give Zoom permission to use contents of Zoom session conversations to train AI systems. Some users feared this revision would allow for the release of proprietary or confidential information. While Zoom has provided some clarification and retraction of its TOS changes, this episode does point out some of the potential transparency pitfalls associated with ensuring that people understand how contextual, prompting, and tuning information might somehow be "shared" without their permission.

Role of AI. AI tools are already being used in programming, coding, and prompt development. How might related tools support the "inventorying" and "resourcing" of tasks related to risk management?

Govern 1.7: Processes and procedures in place for decommissioning and phasing out of AI systems safely and in a manner that does not increase risks or decrease the organization's trustworthiness.

About. I am reminded of government RFPs that require the bidder to include a description of how the work will be turned over to a subsequent contractor — how documentation will be turned over, how new contract employees will be trained, and in some cases, how existing employees might be evaluated in terms of being transferred to new contract management. The goal is to provide a seamless transition of work so that the essential work continues. Applying this concept to GAI based systems that may eventually be replaced or upgraded, one can anticipate some challenges in transitioning to a new system. While it is possible that a new system might simply replace an old system with a few changes to its associated business processes, that is highly unlikely. "Decommissioning" an old GAI application will require changes to how it is managed and used; planning will be essential.

Suggested Actions. Data generated over time by the GAI based system must be evaluated as part of the decommissioning process. Requirements for privacy, security, and access may have changed since data were generated. If such data are stored, appropriate and up to date security and access control measures may be required for older legacy data that are not destroyed.

Transparency & documentation. The GAI system decommissioning process may be streamlined if adequate *and regularly updated* documentation exists concerning system design, operation, and maintenance. Such documentation can provide a baseline against which proposed decommissioning and retirement processes can be measured.

Role of AI. Decommissioning and phasing out systems require planning and use of GAI based tools can accelerate project planning; see Using ChatGPT to Accelerate Creation of Business Case and Project Definition Documents.

Govern 3.2: Policies and procedures are in place to define and differentiate roles and responsibilities for human AI configurations and oversight of AI systems.

About. Given uncertainty of how GAI systems may evolve and operate, both manual and automated monitoring and review of operations will be needed. This should include regular review and updating of risk definitions and trigger conditions driving risk mitigation or other action including changes in prompting.

Suggested Actions. A key assumption is that the scope, management, and usage of the GAI application are well defined along with business and technical roles and responsibilities that vary according to where the application is regarding phased deployment including:

- Pre-deployment requirements, planning, design, and development.
- Training and deployment including both technology and business process changes.
- Post-deployment operation, monitoring, and support.
- Monitoring based updates, corrections, and modification.

Transparency & documentation. Roles and responsibilities must be documented and communicated to stakeholders so that management and operations staff know who is responsible for what. One challenge will be determining how much technical and operational knowledge will be needed by management and system users who are not engaged in programming or software development tasks.

Role of AI. Another challenge will be clearly defining when task responsibilities shift between human and machine. (For a simple example of the latter in terms of defining the role of AI in developing project management documentation, see Figure 1 in Orchestrating AI Large Language Model Tools to Enhance Project Management Document Creation.)

Govern 5.1: Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks.

About. An analog here is how traditional system and user support operations are organized and managed. System performance is monitored and tracked with problems being recorded and addressed by dedicated staff according to priority and severity. Feedback from users and those charged with user support is also a source of potential issues to be addressed through dedicated maintenance and support operations. Those responsible for managing the technology are also usually responsible for tracking industry technical developments especially in areas touched on by the product or technology being supported.

Suggested Actions. With generative AI systems in the loop, both similarities and differences emerge with respect to traditional system and user support operations. Technical and user support operations may be similar in that technical and usage related issues will need to be captured, prioritized and addressed, both for the GAI-specific system components as well as any other systems with which they are integrated. Differences from traditional support operations may however be significant. Generative AI systems may incorporate natural language input and output. This may be different from what many traditional systems depend upon. As a result, human-machine interaction may be more conversational with both content and context evolving over time. How to comparably capture and monitor such evolving meaning across sessions may require both natural language processing as well as AI based analysis focused on how user sessions are conducted.

Transparency & documentation. Potential impacts will be internal and external to the organization adopting a GAI based system. Internally, the possible need to modify system operation based on outright “errors” or planned or unplanned AI system behaviors may drive changes to technology or business processes. Externally, industry developments as GAI technologies and services rapidly evolve must be tracked for relevance.

Role of AI. AI can play a role throughout the system lifecycle all the way from initial design to the monitoring and support of system operation. This raises an interesting issue; if it is appropriate to

separate system development from system operation, as suggested by this Govern 5.1 requirement, should it also not be required to separate AI based performance roles in the same manner?

Govern 6.1 Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third party's intellectual property or other rights.

About. Regarding the question of Generative AI's incorporation (potentially unacknowledged) of someone else's intellectual property in its training, or in its generation of prompted output that incorporates local information, there may be some similarity to how we govern contracts with consultants in tech related work.

Suggested action. It is not unusual when contracting for work for a consultant to sign both a nondisclosure agreement as well an agreement regarding intellectual property ownership. A typical clause in these agreements is that information if generated previously or already publicly known is excluded from such agreements. Consultants are usually hired because of what they know or what they can do, so their state of prior knowledge and skill is expected to influence what they will generally be asked to perform under contract.

Transparency and documentation. As noted above it is normal when contracting with a management consultant to specify in writing what the job requirements are and what "rules" will be followed in carrying out a job. Such documentation should be created when specifying responsibilities of the AI application.

Role of AI. When building a GAI based application it is also normal to specify rules for how tasks will be carried out—including tasks that need to be flagged for management (human) review. For an example, see OpenAI's Using GPT-4 for content moderation which describes a method for automating the implementation of "content moderation policies" to reduce workload on human content moderators. Another example of using AI based rules in the review of content is incorporation of AI-based "plagiarism detection" features in content editing products such as Grammarly.

Govern 6.2 Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.

About. Contingency processes must be developed in advance to address what must be done to maintain operations and recover in the event of an AI failure related to third party data and systems.

Suggested action. Actions may involve:

1. Creation of a dedicated response team (in advance of a failure).

2. Clear definition of human and automated actions to take in case issues are detected regarding both internal and external data and systems.
3. Development and dissemination of documentation as well as training describing "who does what" in the event of a detected failure.
4. Continuous updating of 1 through 3 as internal and third party data and systems continue to evolve.

Transparency & documentation. One transparency issue concerns the resources and methods required to continually monitor third party data and systems to detect and then respond to failures. Some monitoring can be automated and rule based alongside likely involvement of human judges to resolve ambiguous or borderline situations.

Role of AI. Given how AI systems can evolve and "learn," continued testing and monitoring of both internal and external systems will be required; this will require both AI and human involvement in both detection and resolution.

Conclusions

Governing a moving target such as risk management and Generative AI applications is not without precedent given how often policy and law lag technological innovation. At least two unique generative AI challenges exist:

1. Generative AI systems themselves may have the ability to evolve. We need ways to monitor such changes.
2. We need to understand how best to incorporate Generative AI systems into how we monitor and control other Generative AI systems.

How we define an "assistant" role for AI in risk management will be critical given how AI systems are already permeating a wide range of management and research processes. AI applications are already used to generate computer code and to synthesize test data for software application testing. Also, AI systems have been proposed for use in generating synthetic data as a (partial) replacement for humans in certain types of behavioral science research.

At a minimum, we may need to separately maintain (a) how we use Generative AI systems to do work and (b) how we use Generative AI systems to monitor and control how that work is done.

About the Graphic

The graphic at the head of this article was generated via the Microsoft Bing and Microsoft's Edge browser via a vintage Apple iMac running Ubuntu Linux. Bing uses OpenAI's DALL-E neural network to create images from text prompts. The prompts used here were variations of the statement,

"Generate a graphic showing teams of diverse programmers navigating a risky passage between Scylla and Charybdis." (Why these programmers are wearing ties escapes me.)

References

- "NIST Generative AI Public Working Group Virtual Tour," **NIST**, https://www.nist.gov/video/nist-generative-ai-public-working-group-virtual-tour
- "All in 'AI Governance'," **Dennis D. McDonald's Web Site**, http://www.ddmcd.com/managing-technology/category/AI+Governance
- "Using GPT-4 for content moderation," **OpenAI**, https://openai.com/blog/using-gpt-4-for-content-moderation
- "Orchestrating AI Large Language Model Tools to Enhance Project Management Document Creation," **Dennis D. McDonald's Web Site**, http://www.ddmcd.com/managing-technology/orchestrating
- "AI Detection and Grammarly," **Turnitin educator network, AI Writing & Pedagogy**, https://turnitin.forumbee.com/t/g9hcsxv/ai-detection-and-grammarly
- "A Framework for Defining the Scope of Data Governance Strategy Projects (Part 2)," **Dennis D. McDonald's Web Site**, http://www.ddmcd.com/managing-technology/strategy2
- "Zoom Contradicts Its Own Policy About Training AI on Your Data," **Gizmodo**, https://gizmodo.com/zoom-ai-privacy-policy-train-on-your-data-1850712655
- "Using ChatGPT to Accelerate Creation of Business Case and Project Definition Documents," **Dennis D. McDonald's Web Site**, http://www.ddmcd.com/managing-technology/strategy2http://www.ddmcd.com/managing-technology/accelerate
- "How to generate synthetic data from real data - zero to hero," **MOSTLY AI**, https://mostly.ai/blog/how-to-generate-synthetic-data
- "Guinea Pigbots," **Science**, https://www.science.org/doi/epdf/10.1126/science.adj6791

Acknowledgement

I would like to thank my colleague Michael Kaplan, PMP, for involving me in his work on how to incorporate Generative AI systems with project and program management processes. Michael can be reached via email at kaplan.usa@gmail.com.