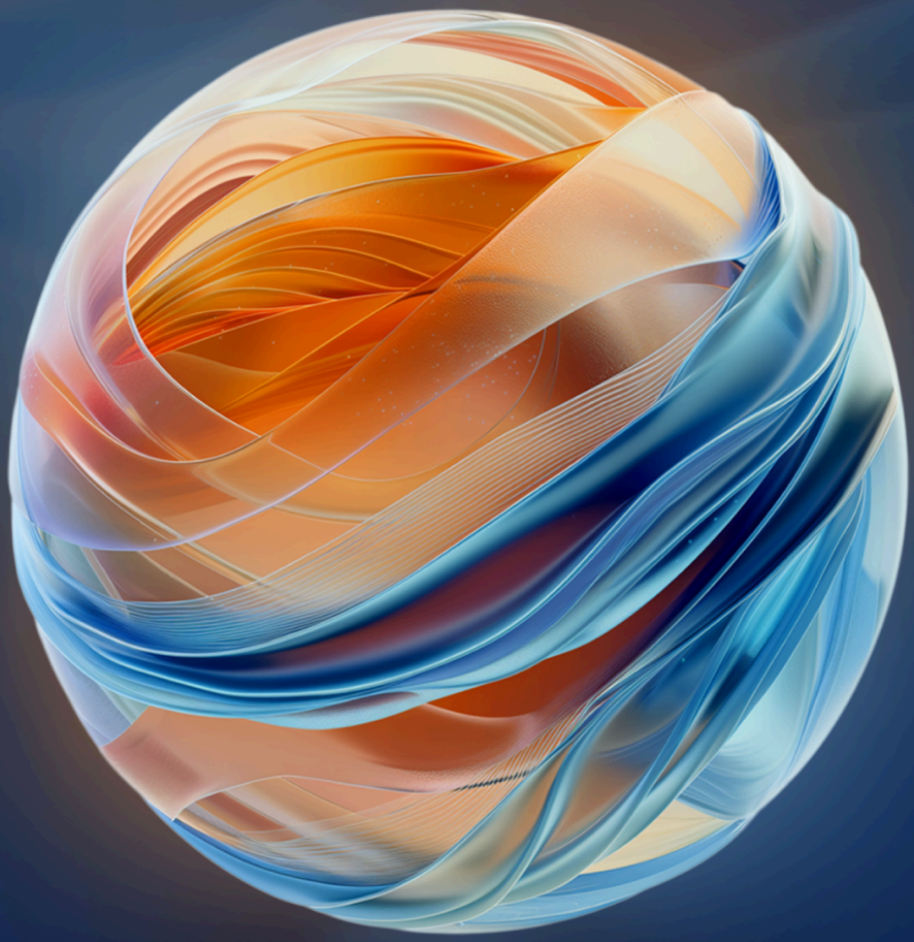
Don’t Panic! Getting Real About AI Governance



AI Governance & Compliance Working Group

cloud security alliance®

The permanent and official location for the AI Governance & Compliance Working Group is https://cloudsecurityalliance.org/research/working-groups/ai-governance-compliance

Acknowledgements

Lead Authors

Dan Stocker

Contributors

Joseph Martella
Alex Sharpe
Ikechukwu Okoli

Reviewers

Priya Pandey
Ashish Vashishtha
Vaibhav Malik
Pranay Shastrulla
Meghana Parwate
Nishith Sinha
Taresh Mehra
Rajashekar Yasani
Sharat Ganesh
Michael Roza
Sean Costigan
Gian Kapoor
Chad Kliewer
Debrup Ghosh
Venkatesh Gopal
Mark Szalkiewicz
Joseph Emerick
Rakesh Venugopal
Mahesh Prabu Arunachalam
Ilango Allikuzhi
Patnana Sayesu
Dr. Chantal Spleiss
Yuanji Sun
Ramana Malladi

CSA Global Staff

Ryan Gifford
Stephen Lumpe
Stephen Smith

Table of Contents

Don’t Panic! Getting Real about AI Governance

Executive Summary

Amid the widespread excitement about Artificial Intelligence (AI), especially Generative AI (GenAI), lies a crucial narrative about how AI systems can generate real business value. Central to this narrative is the emerging capability of AI systems to mimic human-level judgment. Although these advancements present significant opportunities, they also come with new risks. Proactively addressing these risks will be essential to leveraging AI technologies effectively.

A review of the similarities and differences between managing humans and AI systems (that are used for their non-deterministic judgment) leads us to the understanding that there are fundamental similarities and specific differences. This suggests that established risk management norms can be effectively applied to this new area.

We recommend that organizations adopt a risk-based approach to managing AI systems, and measure their progress on a maturity scale. Several useful frameworks are currently available, with more expected as the field grows.

Introduction

Computing drives the modern world. We are co-evolving with it. The seemingly limitless applications of computing, and the integrations into our lives they require, have provoked an “immune response” of sorts (see our Working Group paper for a perspective on AI resilience) from regulatory authorities, due to perceived negative externalities of this explosive growth. Artificial Intelligence (AI) is just the latest disruptive technology example.

Those regulatory moves are detailed in our Working Group paper “Principles to Practice”, but we can frame the larger question with a brief review of the last major wave: Data Privacy. While specific sectors (finance) and narrow use cases (healthcare) have mandates for privacy, there had not been a broad requirement to design for, build into, and operate with privacy protections in products and services until the General Data Protection Regulation (GDPR). This was a sea change.

The technological landscape did not change, but there were new limits on what could be done, and how. Those limits put positive obligations on entities, with real direct and opportunity costs. The associated friction from such a broad mandate (for the most valuable data in modern commerce), led to much resistance from entities that were subject to GDPR. The commercial world, especially outside the European Union (EU), was not enthused. Other jurisdictions, including many American states, adopted similar rules, but the largest English-language market (USA) has yet to create a Federal data privacy law applicable across all states. The large disparity between the EU's comprehensive approach to privacy,

exemplified by the GDPR, and the U.S.'s more fragmented, sector-specific approach has effectively slowed broad progress toward a universally recognized standard for privacy as a governance and compliance imperative.

In roughly the same timeframe, new machine-learning architectures were developed and popularized. The combination of transformers, generative adversarial networks (GANs), TensorFlow, and the beneficial curves of greater compute/storage for ever lower cost, led to an explosion in the capabilities of artificial neural networks, and the commercial practicality of AI models.

The early success of these models was epitomized by one type–large language models (LLMs), primarily due to their relative ease of use by non-technical users. Those LLMs were trained on massive, broad data sets drawn from the internet at large, with little regard for provenance of that data, let alone consent or lawful basis. They were a bona fide sensation and, as has been true for many technological innovations, the details of their origin received much less scrutiny than their world-changing potential.

The fundamental issues of how to manage privacy mandates were not solved, however. It would be more accurate to say they were problems that were still not effectively acknowledged, or constructively engaged with, except in certain exceptionally high-profile cases or in expert-level industry working groups. Thus, the advent of modern AI, dependent squarely on the value of data, was built on an unfinished, shaky foundation.

Into that metastable situation comes the EU AI Act, which became effective August 2024 and builds on the mandates of the GDPR. This time, however, there is widespread agreement that regulation of AI is necessary and a broad tacit acceptance that it will occur. There is opportunity in this surprising state of affairs.

Objectives

This white paper has three objectives:

1. To help frame the issues of governance and compliance in the AI space as a question of maturity to better evaluate impact and options for organizations developing and using AI tools
2. To offer analysis of those issues to:
    a. Identify the key challenges across various dimensions, including first and third-party concerns
    b. Gather and organize current insights from diverse viewpoints
    c. Offer opinions on key topics to advance understanding and enable progress
3. To recommend an approach to AI governance and compliance, grounded in broad best practices and informed by evolving insight into material distinctions in key areas

This Governance and Compliance white paper is just one of a series of outputs from the CSA AI Safety Initiative. Others may be found at:

- AI Governance & Compliance Working Group: AI Resilience: A Revolutionary Benchmarking Model for AI Safety
- AI Governance & Compliance Working Group: Principles to Practice: Responsible AI in a Dynamic Regulatory Environment
- AI Organizational Responsibilities: Core Security Responsibilities
- AI Technology and Risk: LLM Threats Taxonomy
- Using AI for Offensive Security
- AI Model Risk Management Framework
- AI Controls Working Group (work in progress)

Background and Drivers

Philosophy

Managing risk drives *governance* and controls as an exercise in self-interest. Constructive conversations about the effectiveness of risk management will require engaging with other stakeholders, including third parties. That follow up process is known as *compliance*.

Governance is necessary for effective risk management for all organizations, whether or not compliance is also a goal. All organizations building or using AI will want to develop a situated understanding of their risk appetite and tolerance. In the absence of well-established best practices (consolidated wisdom from a broad set of actors, distilled over time), we recommend self-knowledge as a foundational building block for risk management.

As a process, compliance is a way of coming to a consensus on inherently diverse topics using a variety of contextually sound approaches. Compliance frameworks are a tool for addressing the key questions and establishing a baseline for mitigating risk. Industry-specific compliance frameworks emphasize consensus with standards tailored for key specific risks. The Payment Card Industry Data Security Standard (PCI DSS) does this for credit card data. North American Electric Reliability Corporation Critical Infrastructure Protection (NERC CIP) does this for businesses that own or oversee sites that are a component of the American and Canadian energy systems. The GDPR is concerned explicitly with personally identifiable information (PII).

AI is not just one kind of thing. AI comprises a broad set of tools and approaches, not a data type or standalone activity, used across economic sectors. As such, AI compliance concerns will likely emerge from sector-specific risk management aims. Enabling frameworks will follow and be like any other compliance framework that sets expectations for adherence. Compliance for AI is a very immature space

in 2024. The emergent alignment of stakeholders which drives efforts to formalize compliance obligations has not yet achieved critical mass.

In the short-term, the EU AI Act establishes broad regulatory mandates, which will require implementation over the next several years. To date, it remains the most far-reaching effort at scale to explicitly regulate AI. The United States, on the other hand, still approaches AI regulation in a *laissez-faire* manner. An example is the "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," issued in October 2023. California is working on regulations, and Colorado has adopted a law. Both efforts are inspired by the EU AI Act.

Why Treat AI Differently?

If AI is just a broad set of tools, what is the challenge for governance and compliance?

The traditional model for analyzing an organization, to measure and manage it, breaks it down into people, processes, and technologies. This neatly covers the actors, how they do things, and the tools they use. Technologies are force multipliers for people and are best understood as implementations of their intentions. Governance aims to ensure technologies are implemented with desired configurations that represent their intention for risk management controls. In a compliance setting, technologies can be measured (configurations inspected, output reviewed) for adherence to a standard. In other words, technologies are an extension of people and not actors in their own right. This is predicated on the traditional understanding that technologies are deterministic, and thus, reliably measurable.

Part of the strong appeal of AI as a tool is that it represents a new way to implement human judgment. Non-determinism is the essential quality that makes this possible. If a deterministic system could do the same things, that would always be preferable, for both cost and reliability reasons. While deterministic systems can have anomalies and may require complex debugging, these nonlinear results are not the point. By comparison, an AI tool that does not exhibit non-linearity in normal operations is probably not suited for the job.

Thus, our assumed relationship between people and technology is disrupted. If an AI tool exhibits a range of behaviors similar to a human actor, how should it be governed and ultimately assessed? Indeed, how is this done for purely human activities that are not implemented with technology (e.g., reviews and approvals)?

Key Comparisons

In a compliance setting, there is an anchor of trust in the vetting process for onboarding people to an organization. That may involve background checks and consensus among hiring personnel. Past that point, there is a reliance on processes to organize, and put bounds on, human behavior. None of those processes involve the mental state of those people. Training attempts to inculcate norms in people (using incentives), but other than their outward behavior, there is no way to manage the human mind (indeed, the idea is abhorrent). The analogous concept for AI is known as alignment, and is an open research question.

The concept of explainability serves as a caution. Where a person has made a judgment, it is generally possible to determine why, even if that judgment is considered poor and there were consequences. Sound governance processes are calibrated to expect occasional mistakes or misbehavior from humans. These processes include feedback loops for checking and remediation, examples of which include simple errors, all the way to intentional insider threats. By comparison, at present, for AI both explainability and targeted retraining are fraught topics.

How Far Can We Take This Analogy?

If AI simulates human judgment and is implemented with technology, how should we evolve the standard assumptions that inform our governance and compliance model? Some other key comparisons between humans and AI will be illustrative.

Reliability: Humans are known to be prone to overconfidence. Likewise, human memory is renowned for being unreliable. AI systems require augmentation in order to have practical amounts of "memory," which is called the context window. Even with careful management, AI systems can hallucinate in a way that mirrors human errors. Functionally, both require fact-checking and a principled skepticism about results.

Ethics: Humans develop ethical norms as an outcome of socialization and social feedback loops. Our conscience acts like a co-process that provides signals about the acceptability of actions. This can break down in humans for a variety of reasons, including self-interest, and personality or psychological predisposition to lower ethical standards. A common approach to the ethical norms issue is augmentation of existing models with guardrails tailored for certain issues. It is hard to discern a relevant difference.

The Same, But Different

In our rough analysis, key dimensions of why humans require management have direct analogues for AI systems. This suggests that the fundamental governance model is well-founded, and should be useful for AI systems. Working with an already understood mental model has strong appeal as it avoids the additional cognitive load of creating a new framework for AI governance. This is analogous to the appeal of a common controls approach, which manages complexity by avoiding new "languages" for each new concern.

As appealing as that straightforward conclusion is, there are significant challenges which have a material bearing on how it should be implemented. This will require new "words" be added to our common controls, but those will be a narrowly-tailored minimum necessary to serve the new needs. Following Einstein's advice, we should aim for "as simple as possible, but no simpler."

Challenges

Inherent Challenges

There will always be a tension between performance, robustness, and safety, but AI also brings some specific inherent challenges. Examples include:

- Rapid expansion of abstractions, which characterize all progress in computer science, challenge established notions of trust that are rooted in broad understanding of existing technology stacks. The standard security practices that break down abstractions to secure them will lag this expansion and result in unsecured and under-secured surface area.
- Adding to that, the non-deterministic nature of useful AI models makes them difficult to test for basic correctness, which is part of how trust is earned.
- Likewise, the presumed basis for approval of complex systems (like AI models) is that they are understood by personnel responsible for development and change management. This is a rebuttable assumption.
- Models need data and lots of it. The drive for new data will test the limits of propriety (and already has). Use of synthetic data can mitigate certain risks, while potentially introducing new ones (e.g., potential bias, inaccuracies, and lack of realism) and others we are not yet aware of.
- The rate of change in AI, across all aspects, has no prior analogue, and is happening at a time when many high-impact challenges are being addressed. Some of these challenges currently have no generalizable approach (e.g., explainable AI).
- Responsible and ethical AI have broad support as goals, but involve culture-bound or normative concepts (e.g., human rights). They represent the least well-defined of the various types of outcomes models will be judged by.

Topical and Near-Term Challenges

Privacy

No foundational topic in risk management better illustrates the challenge of adapting to a changing environment than the ongoing uneven adoption of data governance and controls. In addition to new regulations and enforcement actions, AI compounds the surface area of what must be managed to comply with laws, legal regulations, and to mitigate business risk.

While AI systems will benefit from adoption of processes to govern and manage data provenance, classification, lineage, and usage, they can also introduce new risks by their nature of being broadly usable for out-of-band data retrieval. In particular, given the state of understanding of how deep learning

systems work, the best controls we have managed to assemble amount to a collection of bolted-on guardrails and adjacent monitoring approaches.

Open questions and hard problems

Among the hard problems attendant to AI systems, none is more important or challenging than explainability. It is simultaneously an ultimate necessity and frustratingly out of reach, despite ongoing research. For a deeper appreciation see Explainable AI (2021) and a more recent 2024 review of the field. Explainability has many benefits, including value for the AI system operator (insight into the logical path from training to inference) and broader stakeholders, including regulators. A robust notion of substantive compliance will require plausible explainability.

Testing, evaluation, validation, and verification (TEVV) of AI systems is needed for similar first- and third-party reasons. The state of the art for these processes is developing quickly, with new attention from NIST, but the industry has not reached a broad consensus.

Self and continuous model training and multi-agent systems

Innovation in how AI models are used is rapidly advancing and moving into designs that challenge established understanding of how things work and how they should be managed for various kinds of risk. Models that evolve over time with continuous training represent moving targets. When, and how often, should they be subject to additional TEVV? Self-training models represent risk of reinforcement of negative characteristics, and unexpected emergent behaviors.

AI system designs with models that interact as agents in a group setting, with diverse roles, can have behaviors that are desirable where each model helps address limits of the others. For example, one model may generate output and another model provides feedback to hone the overall results. Planners and executors is another common pattern. The challenges here are not just horizontal scaling of risks and controls, but emergent behaviors with unknown risks. To note specifically two issues, threat modeling such systems is more challenging, and mixed sets of first- and third-party models/agents will stress already difficult third-party risk management processes.

It's not all bad news, however. Other technology trends are also growing in importance and adoption. Some of these can be used to manage AI risks.

Zero Trust

There is broad consensus that least privilege is a foundational security design goal. Extending that idea from identities to all components in an environment, Zero Trust Architectures (ZTAs) provide more robust assurance of access limitations under normal operations. Moreover, there is a real benefit to limit the blast radius for security issues, including zero-day vulnerabilities. By analogy, the inherent volatility of vulnerability management for AI models will benefit from a more fault-tolerant architecture. See the CSA white paper Zero Trust Guiding Principles for more information.

Confidential Computing

The relative vulnerability of data-in-use, as opposed to at-rest or in-transit, was based on the fact that until recently it was not practical to directly process encrypted data. Ongoing innovation means that this is no longer the case. Chip manufacturers, such as NVIDIA, Intel, and AMD have brought specialized compute capabilities to the market, with strong semantics for isolation of data from other same-system components. The major cloud service providers (CSPs) have bespoke services, and vendors such as Anjuna and Fortanix have general-use offerings.

The major benefit of confidential computing for AI is that it helps secure data used in distributed model training approaches. To take best advantage of distributed training, using wide horizontal scaling, it will be necessary to ship data to third parties. Third-party hosted inference is also a notable use case. Confidential computing is an additional layer of security control to manage risk of data spillage, and reduce reliance on relatively "soft" third-party risk management controls.

Quantum Computing

As new and challenging as AI seems, the expected industry changes resulting from wide adoption of quantum computing will make AI seem quaint. Leaving aside the radically different nature of the technology, the main effect will be a massive speedup of certain types of computing. At present, the state of the art is mature enough for machine learning (e.g., Google and IBM), but modern deep learning architectures are more challenging. Multiple large industry players have invested heavily in this roadmap, with AI as one of the key milestones. Even the U.S. government has prioritized practical research in this space (NASA QuAIL).

Like any technology, quantum computing can be used for good or ill. Expected scientific breakthroughs will also be accompanied by disruptions of established cryptographic norms. To the extent that deep learning models can be enhanced in ways not practical with classical computing, there may be startling advances in performance and utility, which will only magnify the current challenges for explainability and accelerate the need for governance, broadly.

Risk Across the Value Chain

The prerequisites of large-scale data sets and expertise in data science and ML/AI operations put the development of highly-capable general-purpose AI systems (especially LLMs) beyond the practical ability of most organizations. The wide availability of third-party models, and relative tractability of refinement training, is a more tractable approach to scale up use of AI systems. As a result, and like nearly all other modern technology domains, there will be a web of responsibility for AI systems that span multiple organizations and functions within each.

This will put a premium on third-party risk management, particularly issues of data provenance and lineage for third-party models. Promising ideas are being developed and used (e.g., Model Cards and AIBOMs, and signed assertions à la compositions of SLSA, TUF, in-toto, and GUAC), but this space will require further development and industry experience to adapt to new risks related to AI.

Different, But the Same

AI is new and shiny and promising and exciting. It also has arrived into an existing, ongoing situation that has its own persistent, more basic security and risk challenges.

- *Asset management* is still a challenge, with active innovation evident in the market. There is no more basic activity in managing risk. The shadow adoption of AI in organizations will be even more problematic than the earlier waves of shadow IT and shadow Software as a Service (SaaS).
- *Identity management* is perhaps the oldest information security domain, but is still actively evolving in fundamental ways. One key dimension now receiving the attention it deserves, is in the area of non-human identities. Advances here will be very useful for all stages of AI operations.
- *Observability and vulnerability management* (broadly) are improving over time, but security posture management will need to be augmented for transparency of AI-related workloads and data flows. This transparency is challenging for black box designs.
- *Data management and governance,* already a mandate for privacy purposes, faces rapid pressure to evolve in the context of AI systems which have both complementary and diverse tactical objectives.
- *Third-party risk management* is not a solved problem. Traditional tooling for vendor risk management is intended to mitigate the relative opaqueness of vendor operations, to allow a situated risk management judgment. AI is being added to vendor offerings at a rapid pace, which compounds the nominal third-party risk, since models are another (stubborn) layer of abstraction. The underlying issues of data provenance and lineage for third-party models are further complicating factors.
- *Threat Landscape expansion*: The increasing complexity of technical solution stacks leads to more threats at an increased rate. AI tech stacks will only compound this problem.

The key qualitative point is that AI is not just one thing. Like privacy, but to a greater degree, it is a cross-functional challenge that cuts across nearly all risk and security domains. AI will magnify deficiencies (debt) in these areas. That debt will need principled service to avoid throwing good money after bad.

We know the world does not stand still. The older metaphor of establishing a solid foundation presumes a more stable environment. Adaptability is a more useful metaphor than foundations, which are famously static and likely built in the wrong location from a future perspective. Thus, organizations would be well served by framing the AI inflection point as a chance to reset older paradigms.

Key Factors for AI Adoption

Small and Medium-sized Business (SMB) and Early Experiments

Smaller companies typically have fewer resources to spare, even for important functions. This can lead to short-term planning and a tactical approach to adopting new technology. The enclosing environment is often characterized by long-term technical and other debt.

Their key factors for AI adoption are:

- Organizational size, which includes:
    - available cycles/expertise
    - budget constraints on new hiring for AI talent
- Higher tactical priority for operations, which will accelerate existing debt when AI is added to the mix
- Relative lack of sophistication of governance, echoing the broad tactical approach

Growing Companies and Larger Efforts

Companies capable of a strategic approach to AI adoption will have a detailed appreciation of the risks and complexity of mitigating them. They will have an established pattern for integrating AI risks into their overall risk management efforts. New and expanding uses of AI will require an onboarding process to ensure adequate governance. Where appropriate, they will have a human-in-the-loop (HITL) design, for operational purposes, as a backstop for quality and correctness. They are just beginning to address responsible/ethical AI dimensions by considering what that would look like in their context.

Their key factors for AI adoption are:

- The strategic importance of AI safety, for effective leverage of the technology, as well as to manage reputational and legal risks.
- Tracking broad stakeholder expectations for action on responsible/ethical AI questions.
- Cross-functional collaboration between IT, data science, legal, and business units to align AI projects with business goals and to ensure comprehensive risk management. Cross-functional teams bring diverse expertise that is essential for successful AI adoption.
- Continuously monitor the performance of AI models and systems to ensure they meet the desired outcomes and adapt to changing conditions.

Mature Enterprises

This stage of maturity is characterized by long-term thinking and engagement with advanced ideas, primarily responsible and ethical AI. They are anticipating the need for AI compliance and explainability. Their key factor for AI risk management is active preparation for inevitable regulatory mandates by evolution of risk management processes to also support compliance.

Common Key Challenges

All of these types of organizations will be challenged by a variety of related baseline topics:

- Cost and complexity of developing first-party AI models must be balanced against increased third-party risk for models sourced from vendors and open source. While the risks are the same, the treatments will vary.
- AI requires (often very) specialized knowledge and skill sets. This can result in uneven talent availability, with resulting delays and quality issues.
- Contention for in-house talent will challenge prioritizing time/effort when personnel are already overloaded.
- Harmonizing AI risk management with existing enterprise, technical, and security risk management programs will require broad stakeholder agreement to avoid inefficient duplication of effort and varying quality of risk mitigation. Facilitating effective collaboration among business, technical, legal, and ethical experts to develop comprehensive AI governance policies will be challenging.
- AI systems put direct pressure, and have outsized impact, on unresolved data privacy questions. This can make paying down technical and compliance debt (especially for privacy) a cost-effective win-win strategy.
- Tracking the rising tide of regulatory pressure (e.g., roadmap for EU AI Act enforcement and certification), with proactive preparation for third-party audits.
- Tracking and preparing for advancements in Explainable AI, with periodic reevaluation in organizational context.
- Tracking shifts in public perception and societal impact of AI. Creating a definition of responsible and ethical AI, and integrating that definition into existing operations and risk management.
- Managing the rapid rate of change for AI tooling and architectures, including development of tooling for new niches and use cases (e.g., synthetic data and evaluation of model bias risks).
- Managing response to the rapid rate of discovery of vulnerabilities and exploits (both incidental and systemic), some with broad impacts that defy easy mitigation. Operational security maturity will require evolution to include AI-specific dimensions.

Recommendations

1. Adopt a risk-based approach
2. Use a maturity scale

We recommend organizations move, in stages, from the descriptive to the normative. Start with ground truth and work to add risk-tailored controls. This can be described as a *crawl, walk, run* progression.

NIST Artificial Intelligence Risk Management Framework (AI RMF)

While there are many published risk management frameworks with an AI focus (e.g., DataBricks AI Security Framework), NIST is the gold standard for flexible, vendor and product-neutral, broadly useful risk and security governance tools. As with other standards (e.g., NIST Cybersecurity Framework v2 and NIST Risk Management Framework), the AI RMF was developed in a consultative manner with a broad set of stakeholders, across government, industry, and academic experts and stakeholders.

It is composed of four (4) functions: **Govern, Map, Measure,** and **Manage**, to be applied across nineteen (19) categories, and seventy-two (72) subcategories. Additionally, there are two other helpful tools to build out from this flexible framework. The AI RMF Playbook provides detailed guidance and options for each subcategory. Additionally, NIST has published a Profile for use with the AI RMF, keyed to Generative AI, and guidance for securing Adversarial Machine Learning. Many others are in progress and planned.

NIST emphasizes the centrality of the Govern function by way of placing it at the center of an operational cycle diagram, surrounded by Map, Measure, and Manage. The specific topics that have the most critical relevance to AI systems are enumerated as characteristics of Trustworthy AI. These are illustrated in the Framework, with emphasis on the broader roles that *Valid & Reliable* and *Accountable & Transparent* play.

Source: NIST Trustworthy & Responsible AI Resource Center

By breaking down the major phases of an AIOps pipeline (into *Data, Training, and Inference*), we can map these characteristics with details of the concerns they represent.

Data

Acquiring, ingesting, organizing, transforming, and marshaling data for training of AI models

Key trustworthy characteristics and specific concerns:

- *Valid & Reliable*: Data from reliable sources, acquired and used on a lawful basis (with respect to relevant laws, e.g., GDPR and CCPA).
- *Safe*: Data does not represent risks of endangerment to "human life, health, property, or the environment." Examples include filtering data sources for information that could be used to construct weapons, or negatively impact critical infrastructure.
- *Privacy-enhanced*: Data that represent risks to privacy are managed to mitigate those risks, for example, with privacy-enhancing technologies (PETs), including differential privacy, distributed learning, and encrypted (confidential) computing.
- *Fair with harmful bias managed*: Data are processed with bias mitigation as a goal.

More generally:

- *Secure & Resilient*: Data are managed to ensure confidentiality, integrity, and availability, using tailored security architectures for access management (e.g., least privilege), encryption, and vulnerability management. Zero Trust is a good general choice.
- *Accountable & Transparent*: Data is managed to ensure records of provenance and lineage, and all processing activity.

Training

Experimentation, development, training, testing, and release of models

Key trustworthy characteristics and specific concerns:

- *Valid & Reliable*: Training is performed for lawful, responsible, and ethical purposes.
- *Safe*: Trained models do not represent risks of endangerment to "human life, health, property, or the environment." Examples include training models that could be used to guide or assist with construction of weapons, or negatively impact critical infrastructure.
- *Explainable & Interpretable*: Model designs can be explained and interpreted, as far as state of the art allows, which helps inform management of associated risks (including safety, privacy, and bias).

- *Privacy-enhanced*: Model development includes factors to mitigate risks to privacy, for example, by incorporating guardrails.
- *Fair with harmful bias managed*: Models are developed and trained with bias mitigation as a goal.

More generally:

- *Secure & Resilient*: Training processes and environments are managed to ensure confidentiality, integrity, and availability, using tailored security architectures for access management (e.g., least privilege), encryption, and vulnerability management.
- *Accountable & Transparent*: All activities are managed to ensure records of training, and all processing activity, including correlation with relevant underlying data sources.

Testing, Evaluation, Verification, and Validation (TEVV) will be the most broadly useful tool for mitigating training risk. This is directly analogous to the obvious value of testing software for correctness, security, and robustness.

Inference

Managing hosted models and serving requests for inference

Key trustworthy characteristics and specific concerns:

- *Valid & Reliable*: Inference requests are served for lawful bases only.
- *Safe*: Inference requests and responses do not represent risks of endangerment to "human life, health, property, or the environment." Examples include serving requests for information that could be used to construct weapons, or negatively impact critical infrastructure.
- *Explainable & Interpretable*: Inferencing results can be explained and interpreted.
- *Privacy-enhanced*: Inference requests are served with factors to mitigate risks to privacy, for example, by including guardrails.
- *Fair with harmful bias managed*: Inference requests are processed with bias mitigation as a goal.

More generally:

- *Secure & Resilient*: Models and inference requests are managed to ensure confidentiality, integrity, and availability, using tailored security architectures for access management (e.g., least privilege), encryption, and vulnerability management.
- *Accountable & Transparent*: Models and inference are managed to ensure records of requests/responses, and all processing activity.

AI Maturity Models

In mid-2024, there are many options for an AI maturity model, published by a variety of organizations. The overwhelming majority of these models, however, are focused on the maturity of management for business value, not risk or security. Most popular is the Gartner framework, which is laid out in the common arrangement of a series of levels (1-5), starting with Awareness, progressing through Active, Operational, and Systemic, and topping out at Transformational. This breakdown is not very useful for considering the effectiveness of risk management.

By contrast, the traditional Capability Maturity Model Integration (CMMI) breakdown of maturity levels is better suited for consideration of AI governance, capturing the literal journey from first, tentative steps through collaboration to mindfully manage, and ending (hopefully) with a strong baseline to further improve upon.

0. **Incomplete** – *ad hoc* and/or unknown
1. **Initial** – unpredictable and reactive
2. **Managed** – in silos (e.g., at the project level)
3. **Defined** – proactive, rather than reactive
4. **Quantitative**ly managed
5. **Optimizing** – with a feedback process

A persistent criticism of maturity models is that they implicitly take a linear view. There is broad anecdotal reporting that AI is being adopted more unevenly, and in a notably *fire, aim, ready* manner, with various parts of organizations moving independently, taking diverse approaches to managing risk. Capturing this key ground truth and moving forward from that understanding is the central idea of our recommended *crawl, walk, run* approach.

The best published tool (so far) for this purpose is the MITRE AI Maturity Model, which adapts the CMMI levels, across six relevant categories (pillars).

1. **Maturing the Ethical, Equitable, and Responsible Use**: Establish expectations, requirements, and governance to mitigate risks of negative or unintended consequences of AI initiatives.
2. **Maturing the Strategy and Resources**: To ensure the availability of AI solution strategic plans governance model and needed resources.
3. **Maturing the Organization**: Ensure AI that is embraced at an enterprise level includes culture, roles and responsibilities, and workforce development to enable effective AI solutions.
4. **Maturing the Technology Enablers**: Ensure the establishment of technology enablers, including innovation, testing, and infrastructure to produce AI solutions.
5. **Maturing the Data**: Ensure that AI initiatives have the needed data for effective and successful implementation.

6. **Maturing the Performance and Application**: Ensure the effective and efficient development, deployment, operation, and maintenance of AI-enabled capabilities.

They also provide an aligned organizational assessment tool, with accompanying guide.

For organizations that align their risk management strongly to the NIST AI RMF, there is an option to use a calibrated maturity model, as described in the research paper Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework. That model is positioned as part of the overall NIST mission to holistically address management of risk and security with thought leadership and common tooling. One particular virtue of the model is the use of "adaptive" as the ultimate tier, rather than "optimizing." This is in line with our proposed reframing from the "foundation" metaphor.

The scoring rubric is closely, but flexibly, tied to the AI RMF categories and sub-categories. A questionnaire is provided for adopter adaptation and use. The net result is intended to be highly-aligned reporting for risk, that also facilitates consideration and reporting on maturity (as an organizational feedback mechanism).

Conclusion

While AI systems are new to many organizations, the constituent technologies and processes are fundamentally similar. What makes AI special, and raises the bar for governance, is that it represents the technological implementation of human judgment. That non-deterministic quality merits greater attention to managing specific risks, including known risks (e.g., to data privacy) that are magnified and compounded in this new context.

The rate of change for AI technology is already much faster than prior major technological waves. Those rapid changes will accelerate regulatory mandates. Organizations will be best served by proactive attention to these considerations. We recommend that organizations adopt a risk-based approach to managing AI systems and measure their progress on a maturity scale. Robust risk management should track and benefit from industry developments and evolving understanding of AI technology. Viewing progress toward mature risk management as a journey will better match the reality of AI adoption.