Don’t Panic! Getting Real about AI Governance

AI Governance and Compliance Working Group

Release Date: 09/19/2024

Agenda

1. Introduction
2. Background and Drivers
3. Challenges
4. Key Factors for AI Adoption
5. Recommendations
6. NIST Artificial Intelligence Risk Management Framework
7. AI Maturity Models
8. Conclusion

Introduction

- Computing drives the modern world, provoking an "immune response" from regulatory authorities due to perceived negative externalities. AI is the latest disruptive technology example.

- Data Privacy: The General Data Protection Regulation (GDPR) marked a significant change, placing limits on what could be done with data and adding direct and opportunity costs for entities. Resistance was notable outside the EU, especially in the U.S., which has no federal privacy law.

- Around the same time, machine learning architectures like transformers and LLMs led to an explosion in AI capabilities, trained on broad datasets with little regard for provenance or consent.

- Privacy mandates remain unsolved, and modern AI was built on a shaky foundation of unresolved issues.

- The EU AI Act, effective August 2024, builds on GDPR mandates, with broad agreement that AI regulation is necessary, presenting new opportunities.

Background and Drivers

- **Philosophy**: Managing risk drives governance and controls as an exercise in self-interest. Effective risk management requires engaging with stakeholders, including third parties. This process is known as compliance.

- **Governance**: All organizations building or using AI need to develop a situated understanding of their risk appetite and tolerance. Governance is necessary for effective risk management, whether or not compliance is a goal.

- **Compliance Frameworks**: Compliance frameworks are tools for addressing key questions and mitigating risk. Examples include PCI DSS, NERC CIP, and GDPR. AI compliance is still immature in 2024, with efforts like the EU AI Act establishing broad mandates.

Background and Drivers

- **Why Treat AI Differently?** AI disrupts the traditional model of analyzing people, processes, and technologies. Technologies are deterministic, measurable, and an extension of people, but AI, with its non-determinism, challenges this understanding.

- **AI as a Tool:** AI implements human judgment, which is non-linear and non-deterministic. This disrupts governance assumptions and raises the question of how to govern AI as compared to human behavior.

- **Key Comparisons**:
  - **Reliability**: Both humans and AI require fact-checking.
  - **Ethics**: Humans develop ethics through socialization; AI requires guardrails.
  - **Governance Model**: The governance model for humans has strong analogues in AI but requires new tailored controls.

Inherent Challenges

- **Tension:** AI brings tension between performance, robustness, and safety.
- **Abstraction Expansion:** Progress in computer science expands abstractions, challenging trust in established technology stacks, leading to unsecured surface areas.
- **Non-Determinism:** The non-deterministic nature of AI models makes testing for correctness difficult.
- **Data Needs:** AI models need lots of data, which can test the limits of propriety. Synthetic data can mitigate some risks but introduce others (e.g., bias or inaccuracies).
- **Rapid Change:** The rate of change in AI, with no prior analogue, complicates issues like explainability and ethics.

Topical and Near-Term Challenges

- **Privacy:** AI compounds the challenges of data governance and compliance with laws.
- **Explainability:** One of the hardest and most important challenges is explainability, which is essential for substantive compliance but remains frustratingly out of reach.
- **TEVV:** Testing, Evaluation, Validation, and Verification (TEVV) processes are needed, but consensus on them is still developing.
- **Model Evolution:** Self-training models and multi-agent systems present risks of reinforcing negative characteristics and emergent behaviors, challenging threat modeling and third-party risk management.

Challenges: Technology Trends and Risk

- **Zero Trust:** Extending the principle of least privilege through Zero Trust Architectures provides assurance in access limitations, mitigating AI model vulnerabilities.

- **Confidential Computing:** Innovation in confidential computing helps secure data in distributed training and inference, reducing risks of data spillage.

- **Quantum Computing:** The adoption of quantum computing will further disrupt AI, leading to advancements but also magnifying explainability challenges.

- **Third-Party Risk:** AI development spans multiple organizations, heightening the importance of third-party risk management, especially regarding data provenance and model lineage.

Key Factors for AI Adoption: SMBs and Growing Companies

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

- SMBs have fewer resources, leading to short-term planning and a tactical approach to AI adoption.
- Key factors:
  - Organizational size (cycles/expertise, budget constraints for AI talent).
  - Higher tactical priority for operations, accelerating existing debt.
  - Lack of governance sophistication echoes the tactical approach.

Growing Companies and Larger Efforts

- These companies appreciate AI risks and complexity, integrating them into overall risk management.
- Key factors:
  - The strategic importance of AI safety for leveraging technology and managing reputational/legal risks.
  - Cross-functional collaboration between IT, data science, legal, and business units.
  - Continuous monitoring of AI models to meet desired outcomes and adapt to changing conditions.

Key Factors for AI Adoption: Mature Enterprises and Common Challenges

Mature Enterprises

- Characterized by long-term thinking and engagement with responsible and ethical AI.
- Key factor:
  - Preparation for regulatory mandates through evolving risk management processes that support compliance.

Common Challenges

- Cost and complexity of first-party AI models vs. third-party risk.
- Specialized knowledge is required, leading to talent challenges.
- Harmonizing AI risk management with existing risk management programs requires broad stakeholder agreement.
- Data privacy challenges remain, putting pressure on systems and increasing compliance debt.

Recommendations

- Adopt a risk-based approach
- 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 Overview

- NIST AI RMF is the gold standard for vendor and product-neutral risk and security governance tools.
- Composed of four functions: Govern, Map, Measure, and Manage. Applied across 19 categories and 72 subcategories.
- Tools:
  - AI RMF Playbook: Detailed guidance for each subcategory.
  - Profile for Generative AI and guidance for Adversarial Machine Learning.
- Central focus on Govern function, which is surrounded by Map, Measure, and Manage in the framework's operational cycle.

NIST AI RMF: Data and Training

Data

- Valid & Reliable: Data acquired on a lawful basis (e.g., GDPR, CCPA).
- Safe: No risks to human life, health, or property.
- Privacy-enhanced: Use of PETs to mitigate privacy risks (e.g., differential privacy).
- Fair with bias managed: Bias mitigation is a goal in data processing.

Training

- Valid & Reliable: Training is lawful and ethical.
- Safe: Models do not pose risks to health, property, or critical infrastructure.
- Explainable & Interpretable: Models are explainable to inform risk management (safety, privacy, bias).
- TEVV: Testing, Evaluation, Validation, and Verification are crucial to mitigate risks.

NIST AI RMF: Inference

Inference

- Valid & Reliable: Inference requests are served for lawful purposes only.
- Safe: Responses do not risk endangering life, health, or property.
- Explainable & Interpretable: Inference results can be explained.
- Privacy-enhanced: Inference requests mitigate privacy risks.
- Fair with bias managed: Bias mitigation is ensured during inference.
- More generally: Secure & Resilient systems are in place, ensuring confidentiality, integrity, and availability. All activities are accountable and transparent.

AI Maturity Models

- Maturity models in 2024 focus on business value, not risk or security. The popular Gartner framework consists of levels (1-5): Awareness, Active, Operational, Systemic, and Transformational, which are not very useful for risk management.

- CMMI Breakdown is better suited for AI governance:
  - Incomplete – ad hoc/unknown
  - Initial – unpredictable/reactive
  - Managed – in silos
  - Defined – proactive, rather than reactive
  - Quantitatively managed
  - Optimizing – with feedback process

MITRE AI Maturity Model & NIST Integration

- MITRE AI Maturity Model adapts CMMI levels across six categories (pillars):
  - Ethical, Equitable, and Responsible Use
  - Strategy and Resources
  - Organization
  - Technology Enablers
  - Data
  - Performance and Application

- NIST AI RMF Integration: For organizations aligning to NIST AI RMF, a calibrated maturity model is available, which emphasizes “adaptive” as the ultimate tier. The scoring rubric is tied to AI RMF categories and sub-categories, facilitating reporting and feedback on maturity.

Conclusion

- AI systems are new to many organizations, but the underlying technologies and processes are fundamentally similar.
- AI's non-deterministic quality raises the bar for governance, requiring greater attention to managing specific risks, especially those magnified in this new context (e.g., data privacy).
- The rapid pace of AI technology will accelerate regulatory mandates, urging organizations to adopt a risk-based approach and measure progress on a maturity scale.
- Viewing progress as a journey toward mature risk management aligns better with the reality of AI adoption.

Document Acknowledgements

Lead Author

Dan Stocker

Contributors

Joseph Martella
Alex Sharpe
Ikechukwu Okoli

CSA Global Staff

Ryan Gifford

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