WAVESTONE

NIST Request for Information – AI Executive Order

Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Response of Wavestone

February 2, 2024

Table of contents

Introduction and scope of Wavestone's answer

Interest and development of Artificial Intelligence (AI) capabilities have increased over the past years in all industries and geographies, raising unique challenges and risks. Several frameworks provide first guidance to companies developing AI solutions. However, given the complexity of this new technology, its broad impact within each organization and in its usage, further support and guidelines need to be defined and standardized across industries to foster a safe, secure, and trustworthy development and use of AI.

In this context, Wavestone welcomes the opportunity to contribute to the development of guidelines and best practices for AI safety and security by bringing its perspective as an international management consulting firm, highlighting the main challenges faced by our clients, and proposing initial mitigating actions.

With 5,500 consultants across 4 continents, the firm[1] provides consulting services to various industries in the United States, specializing in areas such as Cybersecurity & Operational Resilience, IT Strategy, Data Governance. Our teams rely on several frameworks (either available on the market or developed internally) to improve the cybersecurity and privacy maturity of organizations, with transformations impacting the board, management, and operational levels. Over the last few months, built in-house tools, methodologies have been built to speed up the AI journey of our clients. While supporting 15+ companies, their main challenges are around the following:

- Organizing the governance for AI and generative AI (GenAI) use case evaluation
- Evaluating security risks and requirements of AI use cases
- Testing and checking the level of trust in AI products, specifically through AI red teaming
- Identifying and ensuring the security level of new AI features in software (e.g., AI Copilot/companion initiatives from Microsoft, GitHub, SAP, Salesforce, Zoom, etc.)
- Evaluating compliance against the current version of the EU AI act or US AI executive order
- Identifying efficient cybersecurity-related use cases to ease daily activities.

With this perspective, this document will focus on securing AI development mainly from a cybersecurity and privacy aspects, by covering selected points from sections 1 and 3 of the NIST RFI[2], respectively "*Developing Guidelines, Standards, and Best*

[1] https://www.wavestone.us/
[2] https://www.federalregister.gov/documents/2023/12/21/2023-28232/request-for-information-rfi-related-to-nists-assignments-under-sections-41-45-and-11-of-the

Practices for AI Safety and Security", "Advance responsible global technical standards for AI development".

Wavestone is eager to pursue its contribution to industry developments regarding AI safety and security and would be pleased to participate in any future definition of guidelines, standards, best practices. Our experts are available to answer any questions the RFI reviewers will have.

Matthieu Garin
Partner

M +1 917 415 4535
matthieu.garin@wavestone.com

Matthieu is leading Wavestone US office and the Cybersecurity Global practice. He is promoting cyber excellence as a speaker and contributor in various forums

Gérôme Billois
Partner

M + +33 (0)6 10 99 00 60
gerome.billois@wavestone.com

Gérôme is leading Wavestone Cybersecurity development as well as AI capability. He works with authorities & legislators, shares his expertise through multiple media

1. Governance and actors in AI security

This section will address the below points from the NIST RFI:

- Recommended changes for AI actors to make to their current governance practices to manage the risks of generative AI *(Section 1.a.1, p.4)*
- The types of professions, skills, and disciplinary expertise organizations need to effectively govern generative AI and what roles individuals bringing such knowledge could serve *(Section 1.a.1, p.5)*
- Roles that can or should be played by different AI actors for managing risks and harms of generative AI (e.g., the role of AI developers vs. deployers vs. end users) *(Section 1.a.1, p.5)*

The integration of generative AI into various facets of operations has become a priority, driven by top management directives to keep a competitive edge and leverage the best technology capabilities through innovative use cases. This impetus results in a proliferation of AI use case proposals (~ 150+ use cases to evaluate for our clients), each vying for consideration based on business impact, return on investment, and related risks. However, the pivotal juncture at which these use cases necessitate thorough risk analysis often becomes a stumbling block for many organizations. While cognizant of the many risks associated with AI, companies frequently find themselves grappling with the challenge of effectively addressing these risks.

The NIST AI RMF emerges as a promising initial step in delineating the multifaceted dimensions of AI risks. It serves as a cornerstone, offering a structured approach to identify and assess risks.

However, from a governance standpoint, there exists a crucial need to further develop clear criteria for evaluating these dimensions of risk. A fundamental shift in existing governance structures involves the early integration of cybersecurity awareness and practices into the AI development lifecycle. This entails equipping the data team (data scientists, data analysts, database managers, etc.) with knowledge, tools, and requirements pertinent to securing various aspects of the AI system throughout its lifecycle.

Presently, the governance of AI projects often lacks well-defined parameters in the business landscape. Two prevalent approaches—centralized and decentralized— vie for adoption, with centralized models demonstrating more success in governing AI strategies. For instance, the establishment of a dedicated Design Authority meeting, encompassing cybersecurity, infrastructure, privacy teams, and the AI project team, has demonstrated efficacy in ensuring a unified and holistic assessment of each AI project.

In our experience, a centralized approach thrives with the implementation of an "AI Hub," overseeing security, privacy, legal, purchasing and ethics. A data/development manager operates under this body, responsible for the model's transparency,

explainability, and interpretability, ensuring reliability and validity aligned with business use cases. This model promotes collaboration between disparate teams, fostering a cohesive and comprehensive approach to AI governance.

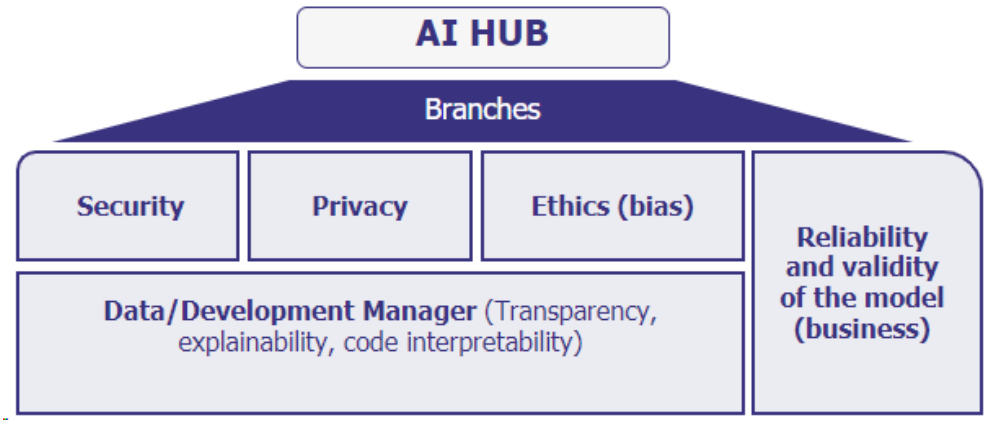

Figure 1 - Example of AI Hub structure

As organizations structure AI governance frameworks, it becomes imperative to delineate and address the intertwined yet distinct aspects of AI trustworthiness and AI cybersecurity. AI cybersecurity involves safeguarding AI systems from both general and AI-specific cyber threats, encompassing elements such as network security and data poisoning. On the other hand, AI trustworthiness aims for responsible, reliable, safe, secure, and unbiased AI systems. The proposed model recognizes the need to differentiate security countermeasures applicable at various levels—embedded in the AI model, applied at the AI level, or enforced at the infrastructure or governance layer.

Moreover, the model emphasizes the need to clearly assign responsibility for implementing these security measures across different teams involved in the project. Data scientists or engineers with technical expertise and access to data and models, business use case owners with domain knowledge, and IT teams managing operational aspects all play distinct roles in ensuring comprehensive security measures at different levels of abstraction. This holistic approach aligns security measures with each team's specific expertise and responsibilities.

In response to the market need for guidance on governance, role definition, our team defined a guide to limit AI risks within project development and deployment, highlighting key cybersecurity criteria of trustworthy AI and main roles and responsibilities throughout the cycle (based on NIST AI RMF and the guidelines from UK National Cyber Security Centre (NCSC), the US Cybersecurity and Infrastructure Security Agency (CISA), and their international partners[3]). Our team can provide experience feedback from diverse clients, not consolidated yet, but are willing to share with respect to confidentiality agreements and clients' privacy.

[3] https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

2. Tools & methodology for identifying and mitigating negative impacts

This section will address the below points from the NIST RFI:

- Risks and harms of generative AI, including challenges in mapping, measuring, and managing trustworthiness characteristics as defined in the AI RMF *(Section 1.a.1, p.4)*
- Tools for identifying impacts of generative AI systems and mitigations for negative impacts *(Section 1.a.1, p.5)*

One of the main challenges faced by our clients is the definition of the appropriate security measures for the AI components. A risk analysis is conducted in the first phase, but there is still a need for alignment with best practices and standards to follow. Existing cyber frameworks, such as the NIST CSF, can be leveraged to secure the infrastructure and the underlying systems, such as the network, the nodes, and the edge; however, the security practitioner community will need more specific guidance on the AI security controls that are relevant for various use cases.

Some initiatives provide concrete recommendations on how to secure AI systems. For example, the European Union Agency for Cybersecurity (ENISA) has published a guide on securing machine learning across the entire life cycle[4]. Our team collaborated on this to identify and develop specific AI uses-cases in the areas of biometric identification, critical infrastructure, healthcare, and personal connected devices, considering their core functionalities, cyber and privacy threats, vulnerabilities as well as security and privacy controls. Another useful resource is the OWASP Top 10 LLM, which lists the most common threats and risks for machine learning models[5].

A similar effort with NIST can help identifying and implementing the security countermeasures per business use case, and indicating what should be embedded in the AI model, around the AI model, at the infrastructure level, etc.

For instance, our team developed a methodology and related tool used as pre-framing stage for an overall risk analysis of system using AI, which could be leveraged more broadly by NIST. This tool focuses on four main categories based on the life cycle of AI and its constructive element: context, input data, model and output data. Several criteria are considered for the risk rating, such as:

- The geographic context and related regulations

[4] https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms
[5] https://owasp.org/www-project-top-10-for-large-language-model-applications/

/ The estimated impact on individuals & on the firm of decisions made by the system
/ The confidence levels associated with the dataset at various stages
/ The training methodologies, hosting environment, and availability
/ The usage of expected output (i.e. automatic decision, support for decision)

Based on the answers to multiple criteria, a risk scoring is generated in addition to tailored recommendations, directing attention to specific security focus areas (e.g. reliability, transparency, bias mitigation).

In addition to this first risk analysis, Wavestone has undertaken an extensive effort to develop a cutting-edge AI threat and counter-measures framework, i.e. the Global AI Risks and Mitigation Radar (see figure 2). This graphical representation organizes mitigating measures into six major categories, classified by feasibility/complexity of implementation: performance & resilience, adversarial attack mitigation, bias & fairness, explainability, data protection & privacy, governance & compliance.

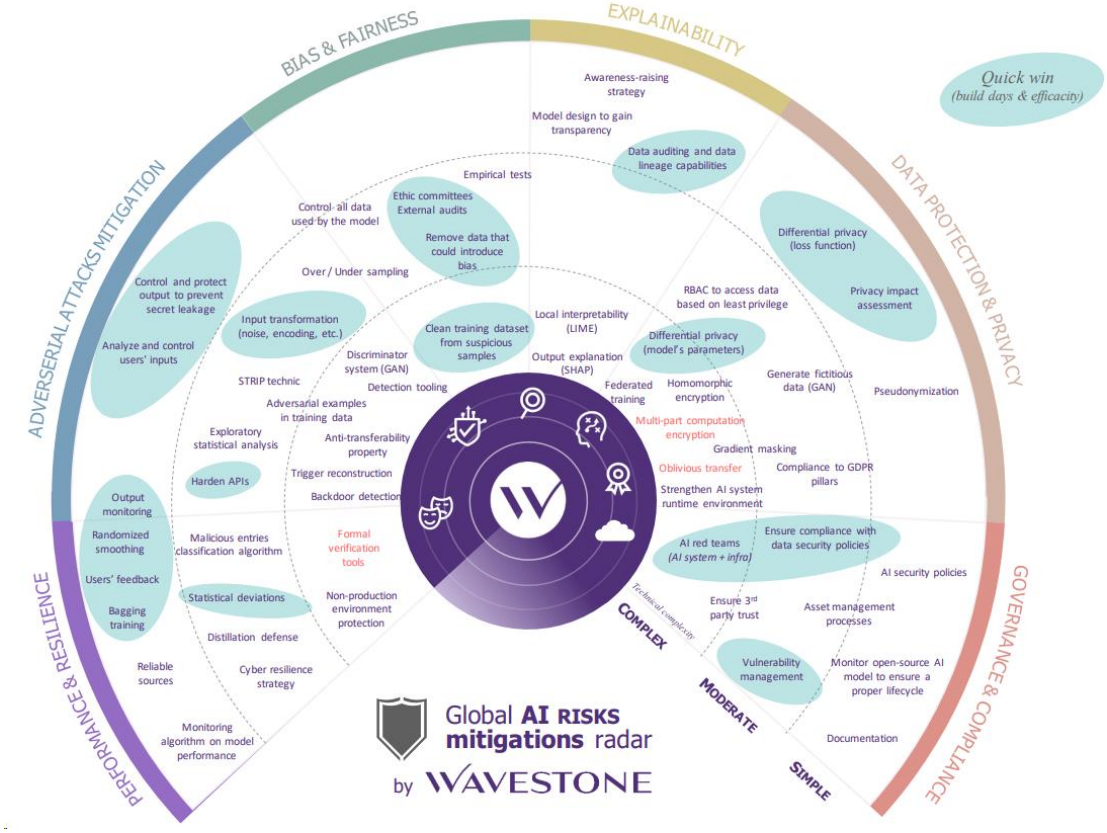

Figure 2 - Global AI risks mitigations radar by Wavestone

Beyond simplification, the radar serves as a practical roadmap to prioritize mitigating actions to implement based on main areas to secure and the maturity of the organization. This comprehensive model, aimed at supporting the AI security community, is poised for widespread publication and could be shared more broadly as a first guidance for organizations.

3. Checks, controls and best practices before go-live

This section will address the below points from the NIST RFI:

- The need for greater controls when data are aggregated *(Section 1.a.1, p.6)*
- The possibility for checks and controls before applications are presented forward for public consumption *(Section 1.a.1, p.6)*
- Best practices regarding data capture, processing, protection, quality, privacy, transparency, confidentiality, handling, and analysis […] *(Section 3.a., p.11)*

Before deploying an AI solution, the decision makers need to get a certain level of confidence that the tool is secure, safe, trustworthy, based on the inputs from the build phase, testing, certification at the release time, as well as monitoring over time. The ISO/IEC 42001[6] provides standard requirements for the development and maintenance of AI systems. The larger tech providers can deploy AI red teaming to assess the related risks. However, only a handful of firms can invest that level of effort, others will need to rely on external providers, such as startups offering support to assess if the level of risk is acceptable or existing cybersecurity actors with pen testing teams. This raises an additional challenge to ensure the tests proposed by external provider are reliable.

Nonetheless, developing a certification standard will support firms in developing reliable solutions in an industrialized manner. This standard should cover the different steps of development of the AI (e.g. inputs gathering, learning, processing) and be tailored to sectors and use cases.

In addition, specific measures around privacy needs to be included throughout the lifecycle and by different stakeholders. Firstly, the usage of personal data should be limited as much as possible to ensure privacy by design. If this is needed, our recommendation is to rely on a legitimate purpose, given the inherent complexity to collect and manage consent, especially the opt-out of those models.

Throughout the development lifecycle, some key points need to be addressed to limit impact on privacy for the inputs, their processing, as well as the users, such as:

- Is the personal and/or sensitive data used as input relevant and necessary?
- Can the model make automatic decisions using this data?

[6] https://www.iso.org/standard/81230.html

- Are the users well informed that personal and/or sensitive information are used by the model?
- Are the users accordingly trained?
- How long is the personal and/or sensitive data stored during the project phase?
- What are the options to delete personal and sensitive data if needed?
- Is this usage of personal and/or sensitive data reflected in the data processing record (for states with enforced privacy laws)?
- Will the company be able to answer a request to exercise rights (for citizens from states with enforced privacy laws)?

Some best practices can be prioritized and implemented to respect privacy and usage of personal and/or sensitive data (see table below). Given our projects, we have extended knowledge of privacy laws throughout the US as well as Europe that we can share during workshops as well.

Priority	Best practices
P0	Apply usual data protection measures (e.g environment segregation, role-based access control on database, encryption)
	Limit the model learning on personal or sensitive data through anonymization, pseudonymization
	Implement data auditing & data lineage capabilities to ensure transparency on data processing along the lifecycle
	Comply with data privacy laws (depending on geographical scope / data handled)
P1	Filter model outputs to limit the answer with personal or sensitive data
	Deploy privacy preserving / reinforcement learning for the system to recognize personal / sensitive data and process them accordingly
P2	Leverage machine unlearning techniques
	Use synthetic data for the model to use a realistic and anonymized dataset
	Implement differential privacy
P3	Deploy counter measures to limit oracle attacks and related consequences

Furthermore, the reliability of AI depends on the trust from citizens. If individuals are not confident in AI solutions, this could lead to non-use or even misusage: the public will provide false information as input, decreasing the data quality, thus the reliability of the model. Such behaviors have already been observed, raising concerns among our clients. The development of such standards will foster transparency, and confidence.