Reference
https://www.regulations.gov/document/NIST-2021-0004-0001

Questions

1. The greatest challenges in improving how AI actors manage AI-related risks—where "manage" means identify, assess, prioritize, respond to, or communicate those risks;

Over the course of the last decade, the adoption of Machine Learning within organizations has seen a rapid rise. Although the promise and value of Machine Learning are great, and ought to be encouraged, a myriad of pitfalls currently accompany this modern practice of data science.

In our experience, some of the greatest challenges or impediments to managing AI-risks include, amongst other things, (a) a lack of organizational maturity in implementing and facilitating Machine Learning and adequately supporting the Machine Learning professionals, (b) the immediate bulky and burdensome process of operationalizing Machine Learning for the benefit of organizational process(es) and decision-making, (c) the lack of transparency (and possible bias) that Machine Learning can foster, and (d) the scarcity of experienced and skilled Machine Learning professionals to implement best practices that ensure appropriate risk management. The above shortcomings have generated acknowledged operational, ethical, legal and governance risks.

There are two key components to this question, namely the definition of "AI actors" and "challenges in managing these risks".

(a) AI actors

It would be useful to define a taxonomy of AI actors and their corresponding responsibilities in creating the right framework and culture for managing these risks. Some examples (non-exhaustive) of such roles could include: Board of Directors, Regulators, Committees (e.g. Audit Committee, Ethics Review Board, CEO, Chief AI Officer (CAIO), Business Owner, Data Engineer, Quality Assurance, CISO, etc.)

While we may never capture the full spectrum of possible roles, defining it would be important as it sets the necessary frame of references and responsibilities in defining and managing AI-related risks.

(b) Challenges

It would be useful to classify the challenges into the oft-understood *People*, *Process*, *Technology* dimensions.

(i) People

There must first be a book of knowledge to close this general knowledge gap amongst the stakeholders identified (the AI-actors). Without this appreciation of why such risks not only pose an issue to the organization but the society at large, there will not be the right traction and movement to develop the right measures.

(ii) Process
There must also be a set of organizational best practices that help to define the necessary managerial oversight, industry-consistent metrics (to prevent the abuse of statistics to less technical audience), Product and Model Management, Resource Management, Incident Management, and Ethics, Public Interest and Legal policies.

The Foundation of Best Practices for Machine Learning (FBPML) - which the respondents are submitting this RFI on behalf, has also launched an Organizational Best Practices covering the main points above (https://www.fbpml.org/the-best-practices/the-best-practices).

(iii) Technology
This is an ever-evolving field, and the work can leverage the existing NISTIR 8269 Draft that defines the adversarial attacks on Machine Learning. The Berryville Institute of Machine Learning has also produced a good paper on the taxonomy of known attacks and how to perform an architectural risk analysis by referencing a typical 9-stage ML workflow.

The Foundation of Best Practices for Machine Learning (FBPML) - which the respondents are submitting this RFI on behalf of, has also launched a Technical Best Practices which comprises 20 sections cover two broad categories of (1) Product Management, and (2) Model Design, Development and Production (https://www.fbpml.org/the-best-practices/the-best-practices).

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

In addition to the dimensions highlighted in the RFI, the other oft-referenced areas would include:

- ***Fairness and impartiality***. While this topic is broadly discussed in subjective terms, there is a body of knowledge around fairness metrics (e.g. https://ai-fairness-360.org/) which provides a more objective view to this quality and businesses make informed decisions on the impact they are creating).

  While Fairness is not an attribute highlighted in the various international standards such as ISO 24028: Artificial intelligence — Overview of trustworthiness in artificial intelligence, this is an important attribute to be considered to achieve trust in the algorithms.

- **Responsibility and accountability.** In continuation to point (1) above, it would also be useful to define the responsibilities and accountability of the various "AI Actors" / stakeholders in the overall ecosystem. This helps to ensure the right oversight is achieved and serves to drive the "tone-at-the-top" to encourage adoption. The challenge today is that there is no real push for this. Building trust into the algorithmic system often gives way to faster time-to-market (similar to the journey of how Cybersecurity was before it became a more matured topic today). This is covered in Part A of the Foundation's Organizational Best Practice guide.

- **Data quality.** Often the unintended consequences occur as a result of training data (in the case of supervised algorithms) which is not necessarily representative of the environment (e.g. training images used for autonomous vehicles not including possibilities of edge cases, which could be a cause of a chain of accidents triggering further investigations by the NHTSA in 2021).

  Also, biases are often embedded in the data as the historical labels may not be representative of the ideal state that society needs. In such situations, algorithms used on historical data blindly will serve to amplify the inequality and biases we have observed in the past. This can then be made worse if humans are not in the loop.

- **Human Oversight and Control.** The involvement of humans in the decision processes of AI systems is an important component of trustworthiness (especially in high-risk applications) and moreover contributes to the prevention and mitigation of harms. The extent to which the human is involved should be determined following a risk-based approach (various extents and formats being known under different names, e.g. 'human-in-the-loop', 'human-on-the-loop', 'active learning')

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: Transparency, fairness, and accountability;

In addition to the aforementioned principles of transparency, fairness, and accountability, we recommend consideration of the following additional principles that we believe are not currently encompassed:

- **Stability:** To prevent (in)direct adverse social and environmental effects as a consequence of interactions amongst Products, Models, the Organization, and the Public due to unexpected volatility in AI predictions or outcomes.

- **Safety and Security**: To (a) prevent adversarial actions against, and encourage graceful failures for, Products and/or Models; (b) avert malicious extraction of Models, data and/or intellectual property; (c) prevent Model based physical or irreparable harms; and (d) prevent erosion of trust in outputs or methods due to security breaches.

- **Provenance/Traceability:** To ensure the clear and complete Traceability of Products, Models and their assets (inclusive of, *inter alia*, data, code, artifacts, output, and documentation) for as long as is reasonably practical; To promote the documentation and recording of Product, Product Team and employees tasks, deliverables and progress.

- **Monitoring and Maintenance**: To build sustained trust beyond the initial deployment, it is important to also have continual assurance that the models are operating as intended over time. This is to cater for either gradual drifts in the inference due to changing trends, preferences, or disruptions to the operating environment. Section 15 of the Foundation's Technical Best Practice includes our views around continual monitoring and maintenance to ensure ongoing trust.

4. The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management—including, but not limited to, the management of risks related to cybersecurity, privacy, and safety;

NIL

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

We, at The Foundation for Best Practices in Machine Learning, want to help data scientists, governance experts, managers, and other machine learning professionals implement ethical and responsible machine learning. We are a team of seasoned data scientists, machine learning engineers, AI ethicists and governance experts, who are enthusiastic about lowering the barriers for pragmatic ethical and responsible machine learning. Our team is passionate about ethical and responsible machine learning and believes that, in order to best promote real change, grassroots, democratic movement is key.

We do this through championing our Technical and Organizational Best Practices for machine learning, specifically via our free, open-source guidelines. Our Best Practices are structured around the following core subjects:

- Product Management
- Fairness and Non-Discrimination
- Data Quality
- Representativeness and Specification
- Performance Robustness
- Monitoring and Maintenance

- Explainability
- Security
- Safety
- Human-Centric Design
- Systematic Stability
- Product Traceability

Within these subjects, the Technical Best Practices (TBPs) are scoped for a single product (which includes the ML models) and are aimed at helping your team best develop and maintain this product in an ethical and responsible way. The subjects within the Best Practices are approached through Product Lifecycle Phases including Product definition, Data Exploration, Model Development, and Production.

The TBPs inherently incorporate the natural iteration that comes with the design, development, and deployment of ML systems.

The Organization Best Practices (OBPs) are scoped for the entire organization. It advises how to effectively support product teams within an organization. This support is clustered around the core subjects mentioned above. These are approached through Policies, Management, and Governance aspects.

6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;

NIL

7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts;

Please see our response to Question 5.

8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation—and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society. Start Printed Page 40813

To (a) identify possible risks[1] for protected classes of persons, animals and the natural environment; and (b) minimize the unequal distribution of Products and Models errors[2] to prevent reinforcing and/or deriving social inequalities and/or ills.

To (a) ensure that Product data and Models are representative of, and accurately specified for, target environments as far as is reasonably practical; and (b) objectively assess and mitigate against[3] unintentional Products and Models behaviours and outputs as far as is reasonably practical.

Please refer to Response (2) above on the discussion on biases in training data.

9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, “AI RMF Development and Attributes”);

NIL

10. Effective ways to structure the Framework to achieve the desired goals, including, but not limited to, integrating AI risk management processes with organizational processes for developing products and services for better outcomes in terms of trustworthiness and management of AI risks. Respondents are asked to identify any current models which would be effective. These could include—but are not limited to—the NIST Cybersecurity Framework or Privacy Framework, which focus on outcomes, functions, categories and subcategories and also offer options for developing profiles reflecting current and desired approaches as well as tiers to describe degree of framework implementation; and

NIL

11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.

Section 2 of the Technical Best Practices highlight the various Team Compositions that can be found in the machine learning team. As part of the proposed NIST Framework, it would help to define the stakeholders (“AI Actors”) and consequently the roles and responsibilities of the different groups.

With an objective framework to develop, assess, monitor, and rectify trust issues identified

[1] Reference: Sections 6.4 and 6.5 of the Technical Best Practices by The Foundation for Best Practices in Machine Learning.

[2] Reference: Sections 5.4, 5.6, 5.7 of the Technical Best Practices by The Foundation for Best Practices in Machine Learning.

[3] Reference: Section 5 of the Technical Best Practices by The Foundation for Best Practices in Machine Learning.

in the ML Operations, this allows industries and nations to start building up a training and skills matrix needed for fresh graduates and mid-career transitions.

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Refer to the Foundations' Organizational Best Practices Sections 3, 4 and 5, and the Technical Best Practices Section 15 for monitoring and evaluation.

These are essential components of any risk management framework and are important to establish the tone-at-the-top and help to drive adoption in the organization.

Similar to Cybersecurity years ago, this needs to be seen as a business issue and responsibility, and not just a data team's problem to deal with.