Response to NIST Request for Comment regarding an Artificial Intelligence (AI) Risk Management Framework

University of Illinois at Urbana-Champaign

The management of risk as it relates to algorithms, machine learning, and artificial intelligence capabilities presents unique challenges. At the same time, foundationally, AI risk management is rooted in the same principles as many other disciplines.

Organizations implement many strategies to mitigate, transfer, accept risk. Organizations create structures to distribute functions in a risk management framework. Governance models support the evaluation of policy and controls frameworks, as well as the evaluation of the benefit of taking some action measured against the potential risk of taking said action, across a series of metrics , often including a human or political metric. An organization often establishes a risk register, evaluates the cultural tolerance and risk appetite for engaging in activities, and may prioritize investment in and around managing the highest risks across the enterprise.

A Model for Calculating AI Risk

Because AI Risk is unique, NIST should include a framework or model for calculating risk. The final model should be agnostic as to the enterprise risk management framework which AI risk management fits under at a particular organization. It may support other NIST frameworks such as NIST SP 800-39 (NIST RMF) but should also be flexible enough to facilitate application to other risk management frameworks (e.g. COSO).

Fundamentally, the risk calculation remains the same as any other calculation of risk as a measure of the impact and likelihood of some adverse event occurring, potentially mitigated by some compensating factors or controls.

AI risk management must be fundamentally differentiated from other forms of risk management in two manners: 1) AI lacks “empathy” and any independent form of ethical or emotional intelligence and 2) AI suffers from the “black box problem” by which it is difficult or impossible to, as the human observer, understand the nature of how the algorithm generated its output due to complexity or design, and as such, risk-based oversight of the decisions made and rationales for those decisions may prove challenging. Transparency, trust, fairness, accountability, mitigation of bias, reliability, privacy, security, among other noteworthy goals articulated by NIST in the RFC, must be carefully considered in the context of these two present-day limitations of AI.

We propose the consideration of these metrics as part of the NIST AI Risk Management Framework

1. **AI Maturity and Acceptance** - An algorithm processing data and generating new knowledge or actionable data should be evaluated against historical outputs
    a. In models that have a basis of comparison to prior results, these algorithms should then be measured on a maturity model which characterizes the reliability and repeatability of outcomes
    b. For algorithms/AI generating knowledge which has no basis for comparison, a metric should be assigned based on how robust the AI model and data set are anticipated to be, and the outcomes must be assessed by an appropriate expert, such as a professional in the relevant field, a data scientist, and/or others capable of evaluating the likelihood of AI maturity.
        i. Nevertheless, for the purposes of a risk calculation, we may wish to assign this AI model a "0" value despite the potential for accurately generating new knowledge
        ii. The maturity of an AI model should be a factor in performing a risk calculation. Early decisions using the algorithm, as the model is trained could affect the model's ability to recover from early outlier data affecting decisions.
2. **Context –** The context of how the output of an AI will be used should be ranked on a scale indicating high, medium, or low risk or analogous measure. For example, AI used to inform treatment or care of patients or to inform policy decisions may be of considerably higher risk than an AI used to inform an individual of a low-stakes pattern identification ("your power bill is higher than usual")
3. **Assistive/Facilitative or Authoritative–** Will the AI be used to support decision making by an expert or professional and evaluated against other data points? Or will the AI be used to provide an expert decision to a layperson or audience with limited experience or limited reference points to compare against?
4. **Expectation of User** - To what degree will the recipient of the AI output have an expectation of accuracy of the output? What is the extent of trust the user places on the AI as an authoritative source for the output?
5. **Cultural/Political** - An outcome generated through the processing of data/inputs by artificial intelligence must be evaluated in the context of the cultural norms within which the benefit or harm exists.
    a. Risk related to culture may relate to either the desired level of alignment to culture or the desired challenge to current cultural norms, depending on the context. Will success of the algorithm be measured in terms of how well cultural norms are preserved, or rather how effectively cultural expectations are evolved over time to support the organizational mission?

   b. This measure may require an additional factor or inversion to accommodate the desired outcome.
6. **Data Security & Privacy** - The extent to which data input to train an AI model or to generate outputs is personally identifiable as compared to mitigating factors to secure the algorithm, data, and outputs, as well as to de-identify, reduce re-identification risk, and/or obfuscate the identity of individuals if data is published must be measured and input to the AI risk calculation. What are user expectations of security, privacy? Does the use to which data is put AI is put comport with the consents provided by the persons described by the data? The Genetic Information Nondiscrimination Act (GINA) exists in part to prevent data given for one purpose from being appropriated and used, perhaps in an AI algorithm, to generate discriminatory decisions that affect individuals, using their data in a way not consented to.

Some of these concepts are based on concepts listed in a recent memo regarding AI and that memo may also add to inform some aspects to consider related to AI Risk Management.[i]

Use of the Privacy Framework as a Model

The RFC suggests the use of existing frameworks to align the AI Risk Management Framework. The AI Risk Management Framework could benefit from the context of the NIST Privacy Framework:[ii]

1. Section 1.2 - Privacy Risk Management & 1.2.1 broadly and Cybersecurity and Privacy Risk Management specifically.
   a. Data Actions – the AI Risk Management Framework must consider data at the core of the framework in terms of input, processing, and output, as well as considerations around the physical security of the systems and data as well as privacy considerations for any/all identifiable data inputs / outputs.
   b. Perception – AI risk suffers a similar issue as referenced in the Privacy Framework in terms of perception and real application. Where algorithms, data mining, and AI models generate new insights and knowledge about an individual, perhaps with less personal bias and awareness than the individual knows, and possibly without their permission or expectation, the individual has real and legitimate concerns regarding their own right to privacy and/or perception of surveillance. It is critical an AI risk framework meaningfully anticipates and uses both training and awareness, as well as informed consent, when a person/persons engage in sharing personally identifiable data within an AI.

2. Section 3.4 - System Development Lifecycle Considerations ---AI planning should include data end-of-life disposition provisions

3. AI and Privacy Risk both benefit from Risk Assessment, and the AI Risk Management model could align and/or use a very similar framework as outlined in section 1.2.2 of the NIST Privacy Framework.
    a. Response Approaches of mitigation, transferring/sharing risk, avoiding risk, and accepting risk are valid in both a Privacy and AI Risk Framework
    b. Evaluating the benefit: risk across multiple values and often competing values appears to have analogs in both the Privacy Framework as well as AI risk assessment.
    c. The distinction between AI risk and compliance risk has similar parallels given the maturity of the fields as well as ethical, moral and legal/compliance obligations.
4. "Strengthening Accountability" - The AI Risk framework might benefit from adapting significant portions of the accountability model for individuals and organizations as outlined in the Privacy Framework in Section 3.2. The AI Risk framework might supplement this model by inclusion of ethics boards similar to an IRB or operational data ethics board and/or other data governance models which focus on data privacy and security, compliance, and ethical use of AI in practice.

Use of the NIST SP-800-39 Risk Management Framework (RMF) as a Model

AI risk can also fit neatly into the NIST RMF and thus integrates AI risk management into the enterprise risk management efforts rather than needing to develop a separate AI Risk Management program

The NIST RMF allows organizations to develop a robust risk management program, while leaving the organization free to tailor the program to its mission and operational needs.

It also examines risks at three levels:

- Senior Leadership tracks treats to the overall mission (financial, reputational, operational, and "duty of care to others" and oversees Operational risk management at the business process level
- Operational Leadership ("upper middle management") analyzes risks to specific business processes and oversees risk management at the level of the individual systems that deliver those business processes.
- System level risk management is operationalized by the staff responsible for the systems that deliver those business functions

AI risk, like other business risks would be treated by following a repeatable, predictable cycle:

a. Framing the risk—Senior leaders in consultation with risk professionals define the parameters of risk decisions and the definition of acceptable risk. Operational Leadership creates a risk catalog listing the specific threats that AI might present to the organization's mission.
b. Assessing the Risk— Operational Leadership, working with the System managers assess the systems that provide AI services to see if gaps exists that might to enable those threats to come to pass. The assessment report is provided to Senior Leadership.
c. Responding to risk--- Operational Leadership, using the risk response guidance provided by Senior Leadership, will decide whether specific gaps will be **avoided** (by not using AI in that process), by taking steps to **mitigate** risk (using administrative or technical controls over AI systems and processes), **transferring** risk with third-party contracts or insurance policies), or **accepting** residual risk in accordance with risk management instructions from Senior Leadership.
d. Monitoring risk by looking for changes in the legal/regulatory environment around the use of AI, and considering how changes in the technical environment or new cyber attack methods alter the AI risk landscape. The results of such monitoring will be used to begin the risk management cycle again by framing the now-current risk.

Artificial Intelligence then becomes another risk factor to be managed along with other business risks, and folded into the overall risk management program without necessitating a separate process specifically for managing AI risk.

---

[i] "Guidance for Regulation of Artificial Intelligence Applications." White House Memorandum for the Heads of Executive Departments and Agencies. 2019. https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf Last Accessed Sept 15, 2021.

[ii] NIST Privacy Framework: National Institute of Standards and Technology. January 16, 2020 https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf . Last Accessed September 15, 2021.