2021

NIST AI Risk Management Framework RFI Responses

Artificial Intelligence Risk Management Framework
National Institute of Standards and Technology
Request for Information

VA National Artificial Intelligence Institute
VA Cybersecurity Innovation Program
September 2021

The National Artificial Intelligence Institute at the Department of Veterans Affairs is submitting the following responses to the NIST AI RMF RFI, in concert with the VA's Office of Information Technology's Office of Information Security Cybersecurity Innovations Program (CIP)

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;*

- Understanding all forms of bias
- Understand what the terms 'explainability' and 'interpretability' fully mean and when they should be addressed in the AI service lifecycle
- Identifying AI-related risks without an organizational AI inventory;
- Identifying, assessing and, prioritizing AI-related risks without standardized and organization-wide AI threat modeling and attack frameworks;
- Identifying, responding to, and communicating AI-related risks without proactive collaboration between security teams, privacy teams and business stakeholders during design, development, and testing; and
- Identifying, assessing, prioritizing, and responding to risks with limited organizational AI / cybersecurity talent.

Obvious forms of **bias** include racial bias that has been in the news. Yet there are other forms of bias that can limit the effectiveness of AI models, including but not limited to;

- Religious
- Ethnic
- Gender
- Age
- Geographic (i.e., a model created during on a pilot in Los Angeles may not achieve the same success metrics when used on populations in the Midwest, New England, Florida, etc)

- Theaters of US Service Personnel active duty deployment (different theaters can have different challenges, stresses, injuries, etc for active duty personnel that can then be reflected in variations of their needs for and approaches to benefits at the VA)

When does an AI development team know when they have fully addressed **explainability and interpretability**? Is it when the end user agrees that the explanations and interpretation out are sufficient, or that "they can't be improved anymore so just take it as-is"? Another risk in this area is waiting until late in the prototype/pilot lifecycle to add explainability and interpretability, as the techniques for them are frequently a limiting factor in the choices a AI modeler must consider at the start of the model design.

Various stakeholders and offices may sponsor AI deployments or pilots at large organizations without a centralized **inventory** or consistent process for tracking organizational AI-related development and deployment efforts. The lack of an inventory or consistent tracking complicates organizational efforts to identify AI-related risks because the total scope of the organization's AI deployments is unknown. Risks cannot be properly identified or otherwise managed without first identifying and tracking organizational AI systems.

The lack of a common and authoritative **AI attack framework** complicates organizational efforts to identify, assess, and prioritize AI risks. Securing AI is still in its infancy, making it challenging for organizations to conduct comprehensive threat modeling. Common attack techniques include backdoor attacks, data contamination, denial of service, and oracle attacks (i.e., an adversary using an API to present the model with inputs and to observe the model's outputs for reconnaissance purposes). However, the security field is still cataloging and analyzing adversarial tactics, techniques, and procedures (TTPs). Additionally, even if organizations accurately identify AI risks, the lack of known best practices, countermeasures, and solutions makes it difficult for organizations to respond effectively and cost-efficiently.

The lack of **security team** involvement early and often in the design, development, testing, and deployment of AI-related systems complicates efforts to identify and respond to AI-related risks and effectively communicate risks amongst siloed departments. As a result, organizations face challenges, including security, during initial development, presenting undesirable alternatives such as the need to redesign or patch the models retroactively. For example, model training data may be deliberately contaminated, or "poisoned," by attackers or otherwise biased and cascade throughout the AI model, negatively impacting effectiveness and reliability. Such a consequence will be more difficult to resolve, more damaging to

organizations, and more expensive to mitigate if not addressed by proper channels during the planning phase.

Finally, there is generally **a shortage of skilled workers** in both AI and cybersecurity. Skilled talent shortages negatively impact an organization's ability to manage AI-related risks. The intersection of these talent shortages contributes to the sentiment that organizations often feel underprepared and poorly equipped to secure their AI and machine learning (ML) systems.

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;*

- Both data drift and concept drift are risks can affect the accuracy of AI model outputs and should be proactively monitored.
- Auditability - AI models should be externally auditable to allow organizations to demonstrate compliance with regulations, standards, and best practices. Additionally, Audibility provides organizations and third parties the ability to check AI capabilities to promote desired outcomes and prevent consequences that may compromise the best interest of citizens, customers, and previously established characteristics.
- Informed by Science and Technology - Organizations should continuously monitor and employ advances in research and best practices from the public and private sectors that evolve with changing security requirements, techniques, and tools.
- Privacy protection - Organizations should prioritize processes for securely and efficiently removing personal data from the model (i.e., unlearning). Building in processes to support unlearning provides organizations the ability to remove sensitive personal information or biased data that may negatively impact users without completely restarting the AI model.
- Decommissioning - Organizations should create processes for securely decommissioning and disentangling AI models from other organizational systems when no longer in use. Compromised or untrustworthy AI systems may negatively impact other organizational systems that they interact with, and risk management or decommissioning activities should extend to include related systems.

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;*

Implementing a trustworthy AI framework can help organizations address ethical and compliance obligations, meet security and privacy requirements, identify AI-related risks, and assign accountability and responsibility to areas of AI development, testing, and operations.

Noteworthy frameworks in the public sector include those developed by the Intelligence Community and Department of Defense. Frameworks commonly include the following principles:

- **Safety / Security** – AI systems can be protected from risks (including cybersecurity risks) that may cause physical and/or digital harm;
- **Privacy** – Data privacy is respected and is not used beyond its intended and stated use, and subjects can opt-in and out of sharing their data;
- **Responsibility / Accountability** – Policies are in place to determine who is held responsible and/or liable for the output of AI system decisions;
- **Robustness / Reliability** – AI systems can learn from humans and other systems, and produce consistent and reliable outputs without excessive failures or anomalies; and
- **Fairness / Impartiality** – AI applications include internal and external checks to help enable equitable application across all kinds of participant types

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;

Government decision-makers should refine existing or develop new policies, processes, and procedures for mitigating AI-related risks, identify their organization's overall AI risk appetite, and incorporate inputs from different offices, including cybersecurity, privacy, and DevSecOps.

Beyond traditional enterprise risk management best practices, AI deployments should include cybersecurity, privacy, and safety assessments that evaluate architecture, data ingestion and management, model development and deployment, and technical monitoring solutions. Assessments should document proper security control implementations, include plans for managing residual risk, and be centralized in an organization-wide risk registrar.

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;

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;

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;

NA

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.

AI design and development should also include security teams in the AI design and development processes. Security teams can help manage cybersecurity and privacy risks that may alter the AI model or make it vulnerable to tampering and malicious activity. For example, security teams can help establish an authorized secure development platform for AI-related products and services to mitigate risks to identify, protect, and detect security functions.

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

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;

Effectively addressing organizational AI trustworthiness and risk management may benefit from two complementary approaches:

1. A secure AI development platform; and
2. An AI risk management governance structure.

Establishing a secure development platform for AI-related products and services can help lessen the burden on organizations to identify, protect against, and detect AI-related cybersecurity incidents. Organizations can proactively harden and secure their AI instances by incorporating trustworthy AI framework principles (and their associated controls) into DevSecOps and system authorization processes. AI software factories can standardize security with AI-related penetration testing tools and processes (e.g., Microsoft's open-source tool, Counterfit, which tests AI systems during red team operations and pre-production development), secure code repositories, and model robustness checks. The platform can facilitate model training using privacy-preserving data repositories based on concepts such as synthetic data or homomorphic encryption.

At the organization-wide governance level, AI products and services are not radically distinct from other types of systems. Existing frameworks (e.g., NIST CSF and Privacy Framework) should still apply. For example, given an AI model's heavy reliance on data, existing guidance on data encryption (at rest, in transit, and in use) are applicable and would be covered by outcomes expressed in the NIST CSF. Similarly, Transparency / Explainability are related to the privacy objective of predictability in the NIST Privacy risk management framework. Where necessary, frameworks can be augmented to include AI-specific outcomes, and the AI risk management framework could be dedicated to principles that are unique to this discipline. Other frameworks such as MITRE's ATT&CK framework could also be enhanced to identify AI-specific TTPs.

Regardless of the framework or approach, organizations should create a maturity model to assess the organization's overall level of maturity and progress across different AI components. Organizations should also designate champions to supervise key areas that can impact AI trustworthiness, including:

- **Architecture** – Requires AI champions to remain informed on architecture and set policies so stakeholders can implement and maintain best practices;
- **Data Ingest and Management** – Requires AI champions to stay informed and develop policies about data practices such as who can use the data, acceptable sources, and necessary prior authorizations;
- **Model Development and Deployment** – Requires AI champions to develop policies and monitor development, training, and deployment best practices; and
- **Technical Monitoring Solutions** – Requires AI champions to develop policies and monitor model drifts, unauthorized scope creep, and other maintenance best practices.

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.

1. The framework should prioritize establishing and supporting education focused programs (EFP), especially those focused on AI and emerging technologies, for use by and benefit to the federal workforce.

    a. Research[1] indicates that the contemporary workforce is more motivated by the ability to gain further knowledge and skills from a career rather than mere pay levels.

        i. Given this, these EFPs would be a massive recruitment and retention asset as well as their inherent development applications.

    b. Also, any gains for an individual employee in terms of knowledge or skill necessarily adds to Governments human capital.

i. While the benefit of one, or even many, EFPs on a single individual would be like a drop in a lake in terms of Government's capacity, but even undertaking only one EFP per person on average across the workforce would be an unfathomable gain.

c. Moreover, through the government work better done by these enriched personnel, the benefits back to society from having higher capacity civil servants will multiply the aggregate ROI.

2. These EFPs should themselves exemplify foundational Governmental and Ethical AI tenets in their design and execution. These tenets include, but are not limited to:

a. The pursuit of global leadership in AI
b. Trustworthy AI
c. Highly secured systems
   i. Including domestic production
d. Assessment and accountability
e. High Reliability systems and organizations
f. Adaptability and updatability
g. Personalization of content and learning modality
h. Interagency and Private-Public cooperation

3. The Framework should direct government to interact with higher- and secondary-education institutions to provide guidance and materials from this framework and any EFPs to them to help better the education system.

a. This will in turn later provide the federal workforce and wider US labor market with AI, data science, and more generally technologically capable recruits.
b. The Framework should also direct government to consult with higher education institutions on the development of any EFPs.

4. The Framework should suggest incentivization tied to employees' undertaking and succeeding in these EFPs and/or success on any related assessments to encourage their retention and thus the retention of access to returns on Governments educational investment.

5. Also, to further appeal to those who are pay focused, The Framework should direct Agencies to use their Critical Position Pay Authority under 5 USC 5377 and other special hiring authorities to make near-market-rate or at-market-rate offers to expert level job candidates.

6. Similarly, the Framework should direct agencies to expand Pathways hiring programs to accelerate the finding, vetting, and onboarding of already skilled and educated personnel

The Framework should advise the inclusion of careful recordkeeping and personnel profile tracking for any personnel who receive training over knowledge or skills that might be an asset to our adversaries, or that are otherwise classified, to help maintain the security of this knowledge/skills/information and the nation.

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.

Human Centered Design (HCD) aspects focus on how the end users can best receive and interpret the information and capabilities provided to them from applications. A team's HCD practice should include explainability and interpretability, at a minimum, for efforts that include AI.

Continuous monitoring and evaluation of model metrics is essential to avoiding out-of-tolerance data draft and concept drift.

Page Break

References

1. Makridis, Christos, March 2021, "(Why) Is There a Public/Private Pay Gap?", Journal of Government and Economics, https://doi.org/10.1016/j.jge.2021.100002 (also attached to input submission

email)

Journal of Government and Economics 1 (2021) 100002

Contents lists available at ScienceDirect

Journal of Government and Economics

journal homepage: www.elsevier.com/locate/jge

(Why) Is There a Public/Private Pay Gap?☆

Christos A. Makridis[a,*]

Arizona State University and MIT Sloan School of Management, 245 First St, Room E94-1521 Cambridge, MA 02142-1347

ARTICLE INFO

Keywords:
Careers
Development and training
Earnings
Management
Public service motivation
Public-private pay gap

ABSTRACT

The government is facing a severe shortage of skilled workers. The conventional wisdom in branches of policy and public administration is that the shortage is driven by low salaries that are not competitive for attracting top talent. Using longitudinal data on high skilled workers between 1993 and 2013, this paper shows that, if anything, government employees earn more than their private sector counterparts. Although government workers tend to earn less in the raw data, these differences are driven by the correlation between unobserved productivity and selection into private sector jobs. Instead, this paper provides empirical evidence that low non-pecuniary amenities, such as development opportunities and management, can explain earnings differences between the public and private sectors.

1. Introduction

The public sector, especially the Federal government, faces a significant shortage of skilled workers (Goldenkoff, 2015), including information technology and cyber security jobs (Libicki et al., 2014). While the acquisition and retention of skilled workers has been a challenge in the public sector since at least the 1970s (Lewis, 1991a; 1991b), it has intensified in recent years, including across other countries, such as Britain (Murphy et al., 2019) and France (Bargain et al., 2016). One proposed solution to the skills gap, advanced by a combination of researchers, the popular press, and think-tanks, involves an increase in compensation for those serving in the public sector (Donahue, 2008).

Identifying genuine earnings differences between public and private sector employees is challenging because selection into public service is not a random decision. In particular, individuals sort into jobs based on their preferences and productivity, meaning that simple comparisons of means between public and private sector jobs could prompt spurious implications for public policy. Using longitudinal data on high skilled workers between 1993 and 2013, my primary contribution is to examine the earnings differences between public and private sector workers and their source. Together with data on job satisfaction and work-place practices, I show that bureaucracy and poor management practices are more plausible reasons for the skills shortage in the public sector. These results suggest that public sector organizations may find that focusing on non-pecuniary amenities, such as development opportunities and social impact, are more effective vehicles for raising retention and attraction of skilled workers, relative to increasing pay.

The first part of the paper introduces the data and estimates differences in pay between government and private sector workers. Although government employees earn 4.1% less compared to their private sector counterparts after controlling for demographic characteristics, these estimates of the public-private earnings difference may be downwards biased if higher productivity workers sort into the private sector. The preferred specification, instead, exploits longitudinal variation among individuals who switch between public and private sector jobs, thereby comparing the same individual before versus after a switch from the private to public sector, or vice versa. These results suggest that government workers earn 3.9% more than their counterparts. Moreover, I show that 3.9% is a lower bound for the overall compensation premium since government employees receive greater non-wage benefits, such as healthcare and pensions, on top of their salary income. The compensation premium among public sector workers is not driven by differences in labor supply; the data suggests that public sector workers allocated less time towards work activities.

The second part of the paper examines an alternative explanation behind the articulated skills gap in the public sector. Using additional information on work-place practices and job satisfaction for a subset of the sample, I show that government workers report significantly fewer opportunities for advancement in their careers, less intellectually stimulating and challenging work, less independence and autonomy, and less responsibility and ownership, relative to their private sector counterparts. Moreover, after controlling for these differences in work-place practices, the earnings gap between government and private sector workers becomes statistically indistinguishable from zero. These results suggest

☆ The paper reflects my views only, rather than any affiliated individuals / organizations or the United States.

* Corresponding author.
E-mail address: makridis@mit.edu
URL: https://www.christosmakridis.com

https://doi.org/10.1016/j.jge.2021.100002
Received 2 February 2021; Received in revised form 3 March 2021; Accepted 5 March 2021
Available online 26 March 2021
2667-3193/Published by Elsevier B.V. on behalf of Academic Center for Chinese Economic Practice and Thinking, Tsinghua UniversitySociety for the Analysis of Government and Economics. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

NA