SPECIAL COMPETITIVE STUDIES PROJECT

National Security Addition to the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework Playbook

(NIST AI RMF)

APRIL 2023

Rama Elluru
Chuck Howell
Michael Garris

The Special Competitive Studies Project (SCSP) “National Security Addition to the NIST AI RMF Playbook” addresses the national security risks posed by AI. The NIST AI RMF is a voluntary use tool to manage AI associated risks to individuals, organizations and society.

Risk Management Playbook Framework

The NIST AI RMF calls on developers, deployers, and users of AI systems to assess and address risks including to national security. National security threats may be posed by AI systems even if they were not created for the national security domain. While AI applications offer a variety of benefits to society and the economy, they may also introduce risks to national security through for example, intentional or unintentional misuse, extreme scalability, generative capabilities, and/or corrupted data or software. Sidestepping national security considerations could lead to unintended and harmful consequences. This document adds considerations and questions that are meant to assist with identifying and assessing risks specific to national security.

National security risks in AI applications include revealing sensitive data about strategic infrastructure, populations, or other subjects through geospatial mapping software and publicly posted fitness tracking data to identify U.S. military facilities. Unprecedented scaling in using AI analytics to handle the massive volume of cell phone locations (e.g., from data aggregators) enables the identification of cell phones associated with regular visits to sensitive facilities and other geographic locations, including individuals' homes, putting individuals and locations at risk of exposure and or targeting.[1] Generative AI illustrates another novel risk of deep fake technology as evidenced by the U.S. ambassador to Russia announcing that he was being impersonated by deepfake technology that was sufficiently convincing to fool some Ukrainian officials on video calls.[2]

The NIST AI RMF Playbook is a living document that will evolve. The AI RMF's cross-sectoral profiles, such as for national security, cover risks of models or applications that can be used across use cases or sectors. Cross-sectoral profiles can also cover how to Govern, Map, Measure, and Manage risks for activities or business processes common across sectors such as the use of large language models, cloud-based services or acquisition. The Map function establishes the context to frame risks related to an AI system. This SCSP crafted resource is intended to aid users of the RMF Playbook functions, particularly the Map function, to navigate the framework from a national security perspective. Without contextual knowledge, and awareness of risks within the identified contexts, risk management is difficult to perform. Map is intended to enhance an organization's ability to identify risks and broader contributing factors. As each section of the Map function is considered, a corresponding section of this document can be checked to see if additional guidance is provided (e.g., in Suggested Actions) to help address national security risks. This resource follows the Map function order, however, considerations in the following sections are of priority (e.g., 1.6 & 5.2).

[1] Christopher Burgess, OPSEC Nightmare: Tracking Cell Phone Data in the U.S. and Abroad, Clearance Jobs (2022).
[2] David Sadler, Rogue Russian Duo Targeting High-Ranking Western Officials With Video Calls, Globe Echo (2023).

Map 1.1

Intended purpose, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes; uses and risks across the development or product AI lifecycle; TEVV and system metrics.

About

AI actors can work collaboratively, and with external parties such as community groups, to help delineate the bounds of acceptable deployment, consider preferable alternatives, and identify principles and strategies to manage likely risks. Context mapping is the first step in this effort, and may include examination of the following (added text in **bold**):

- Potential negative impacts to individuals, groups, communities, organizations, and society, **including national security** – or context-specific impacts such as legal requirements or impacts to the environment.

Suggested Actions

- Consider intended AI system design tasks along with unanticipated purposes **and uses** in collaboration with **national security**, human factors and socio-technical domain experts.

Transparency and Documentation

Organizations can document the following:

- **Who is the person(s) accountable for identifying, assessing, and mitigating the national security considerations across the AI lifecycle?**

References (new)

- Under “Identification of harms”,
    - Miles Brundage, et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, Future of Humanity Institute, et al. (2018).
    - John Villasenor, Artificial Intelligence, Geopolitics, and Information Integrity, Brookings at 131-142 (2019) (paper within The Global Race for Technological Superiority: Discover the Security Implication).
    - Greg Allen & Taniel Chan, Artificial Intelligence and National Security, Belfer Center for Science and International Affairs (2017).

Map 1.2

Inter-disciplinary AI actors, competencies, skills and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.

Suggested Actions

- Create and empower interdisciplinary expert teams to capture, learn, and engage the interdependencies of deployed AI systems and related terminologies and concepts from disciplines outside of AI practice such as law, sociology, psychology, anthropology, public policy, systems design, **national security,** and engineering.

Transparency and Documentation

Organizations can document the following:

- To what extent has the entity addressed stakeholder perspectives on the potential negative impacts of the AI system on end users, **national security,** and impacted populations?
- **What stakeholder outreach has been established for feedback on emerging national security risks?**
- **If the system being developed is providing AI-as-a-service, or an available component for other developers, to what extent has the entity identified and documented clients or users of their responsibilities and provided them necessary resources to develop technology using the service in a manner that safeguards national security?**

References (new)

- Greg Allen & Taniel Chan, Artificial Intelligence and National Security, Belfer Center for Science and International Affairs (2017).
- Miles Brundage, et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, Future of Humanity Institute, et al. (2018).

Map 1.5

Organizational risk tolerances are determined and documented.

Suggested Actions

- Establish risk criteria in consideration of different sources of risk, (e.g., financial, operational, safety and wellbeing, **national security,** business, reputational, and model risks) and different levels of risk (e.g., from negligible to critical).
- Review uses of AI systems for "off-label" purposes, especially in settings that organizations have deemed as high-risk. **Include in this review scenarios of unintended consequences for harm to national security from off-label use by anticipated users and by adversaries**. Document decisions, risk-related trade-offs, and system limitations.

Transparency and Documentation

Organizations can document the following:

- **What potential business, financial, and reputational risks are introduced by potentially creating or neglecting to consider national security risks in the system development and deployment?**

Map 1.6

System requirements (e.g., "the system shall respect the privacy of its users") are elicited from and understood by relevant AI actors. Design decisions take socio-technical implications into account to address AI risks.

About

Eliciting system requirements, designing for end users, and considering societal **(including national security)** impacts early in the design phase is a priority that can enhance AI systems' trustworthiness.

Suggested Actions

- Include potentially impacted groups, communities and external entities (e.g. civil society organizations, national security stakeholders, research institutes, local community groups, and trade associations) in the formulation of priorities, definitions and outcomes during impact assessment activities.
- **Analyze potential of the system being developed to reveal identity, location, features, entity characteristics, steal data, change behavior of a system, or influence sentiment.**
- **Consult with others developing and deploying similar technologies in order to maintain situational awareness of relevant and emerging national security risks and challenges.**

Transparency and Documentation

- How will the relevant AI actor(s) address changes in accuracy and precision due to either an adversary's attempts to disrupt the AI system or unrelated changes in the operational/business environment, which may impact the accuracy of the AI system?
- **What are specific ways an adversary may attempt to exploit or disrupt the system?**

References

- Sarah Kreps, Democratizing Harm: Artificial Intelligence in the Hands of Nonstate Actors, Brookings (2021).
- Miles Brundage, et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, Future of Humanity Institute, et al. (2018).
- Greg Allen & Taniel Chan, Artificial Intelligence and National Security, Belfer Center for Science and International Affairs (2017).

- John Villasenor, Artificial Intelligence, Geopolitics, and Information Integrity, Brookings at 131-142 (2019) (paper within The Global Race for Technological Superiority: Discover the Security Implications).
- Rebecca Klar, AI 'Wild West' Raises National Security Concerns, The Hill (2023).
- "The internet of things (IoT), cars, phones, homes, and social media platforms collect streams of data, which can then be fed into AI systems that can identify, target, and manipulate or coerce our citizens." Final Report, National Security Commission on Artificial Intelligence at 45 (2021).

Map 2.2

Information about the AI system's knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI actors when making informed decisions and taking subsequent actions.

Suggested Actions

- Plan and test human-AI configurations under close to real-world conditions, **including examples of "off-label" use, "on-label" use with unintended consequence, and potential "misuse cases" by bad actors**, and document results.
- Document connections the AI system or data will have to external networks (including the internet), **for example**, financial markets, **social media, news outlets, academic databases, health care institutions, logistics,** and critical infrastructure that have potential for negative externalities.
- **Given the AI system's outputs, how does the end user know what to be confident in (i.e., treat as true) so as to not over-trust or under-trust the system and, for example, promote resilience to disinformation and influence operations?**

Transparency and Documentation

- **Does the AI system provide sufficient information to the end user so that they know what to be confident in (i.e., treat as true) so as to not over-trust or under-trust the system?**

References (new)

- Miles Brundage, et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, Future of Humanity Institute, et al. (2018).

Map 3.2

Potential costs, including non-monetary costs, which result from expected or realized AI errors or system functionality and trustworthiness - as connected to organizational risk tolerance - are examined and documented.

About

Anticipating negative impacts of AI systems is a difficult task. Negative impacts can be due to many factors, such as system non-functionality or use outside of its operational limits, and may range from minor annoyance to serious injury, financial losses, **threats to national security**, or regulatory enforcement actions. AI actors can work with a broad set of stakeholders to improve their capacity for understanding systems' potential impacts – and subsequently – systems' risks.

Transparency and Documentation

Organizations can document the following:

- **How can the intentional misuse of the system by adversaries be detected and mitigated?**

References (new)

- Miles Brundage, et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, Future of Humanity Institute, et al. (2018).

Map 3.3

Targeted application scope is specified and documented based on the system's capability, established context, and AI system categorization.

Suggested Actions

- Consider narrowing contexts for system deployment, including factors related to:
    - How outcomes may directly or indirectly affect users, groups, communities, **national security**, and the environment.
    - Geographical regions in which the system operates **and potential national security implications of use in different geopolitical environments**. **This includes the ways the system could be used by an autocratic government to surveil or oppress their own people or to export these capabilities.**
    - How AI system features and capabilities can be utilized within other applications, **and the national security risks from examples of system composition**.
- Engage AI actors from legal and procurement functions **and national security expertise** when specifying target application scope.

References (new)

- Dahlia Peterson & Samantha Hoffman, Geopolitical Implications of AI and Digital Surveillance Adoption, Foreign Policy at Brookings (2022).
- Miles Brundage, et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, Future of Humanity Institute, et al. (2018).

Map 4.1

Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third-party's intellectual property or other rights.

Suggested Actions

- Inventory third-party material (hardware, open-source software, foundation models, open source data, proprietary software, proprietary data, etc.) required for system implementation and maintenance. **Determine national security risks from critical third-party resources that are impacted by foreign entities as a supply chain vulnerability or dependence (e.g., banning a foreign online platform).**
- Review redundancies related to third-party technology and personnel to assess potential risks due to lack of adequate support **or deliberate suspension of availability of data or services**.

Map 4.2

Internal risk controls for components of the AI system including third-party AI technologies are identified and documented.

About

In the course of their work, AI actors often utilize open-source, or otherwise freely available, third-party technologies – some of which may have privacy, bias, and security risks **introduced either accidentally or intentionally.**

Suggested Actions

- **Identify any third-party technology, components, or data that is produced by or controlled by entities or nations of concern that may introduce a national security threat. The Consolidated Screening List (CSL) is a list of parties for which the United States Government maintains restrictions on certain exports, reexports, or transfers of items, and can be checked for issues regarding a third party provider.**

References (new)

- Consolidated Screening List, U.S. International Trade Administration (last accessed 2023).
- Tianyu Gu, et al., BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, IEEE (2019).
- Menghan Xiao, Popular Machine Learning Framework PyTorch Compromised with Malicious Dependency, SC Media (2023).

Map 5.2

Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.

About

AI systems are socio-technical in nature and can have positive, neutral, or negative implications that extend beyond their stated purpose. Negative impacts can be wide-ranging and affect individuals, groups, communities, organizations, and society, as well as the environment and national security.

Organizations can create a baseline for system monitoring to increase opportunities for detecting emergent risks. After an AI system is deployed, engaging different stakeholder groups – who may be aware of, or experience, benefits or negative impacts that are unknown to AI actors involved in the design, development and deployment activities – allows organizations to understand and monitor system benefits and potential negative impacts more readily.

Suggested Actions

- Establish and document stakeholder engagement processes at the earliest stages of system formulation to identify potential impacts from the AI system on individuals, groups, communities, organizations, **national security,** and society.
- **Consult with others developing and deploying similar technologies in order to maintain situational awareness of relevant and emerging national security risks and challenges.**

Transparency and Documentation

Organizations can document the following:

- **What stakeholder outreach has been established for feedback on emerging national security risks during deployment?**

References (new)

- Miles Brundage, et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, Future of Humanity Institute, et al. (2018).