Comments on the 2nd draft of the NIST Artificial Intelligence Risk Management Framework (AI RMF)

Author

Adelin Travers
MSc. in Computer Science (Oxon), *Ingénieur diplômé* (Télécom Paris), GPEN, CISM.
For more information visit https://alkaet.github.io/

Table of contents

General comments

The motivation section (Part 1) does need the most rework. This is important since the motivation section explains some key limitations of other standard risk management frameworks like the NIST CSF in the context of AI. The explanation of the limitations of those frameworks in the context of AI systems should be refined and made more explicit.

The core framework (part 2) and the complementary playbook are excellent starting points for a first iteration of the AI RMF and the suggestions below are mostly line edits.

The system level approach is very welcome and is in line with industrial implementation needs and a relatively new line of thought in Adversarial machine learning research (see for instance *ISO/IEC 19792:2009* and *ISO/IEC 24745:2011* on biometrics, *On the Exploitability of Audio Machine Learning Pipelines to Surreptitious Adversarial Examples*, *Rearchitecting Classification Frameworks For Increased Robustness*, *Interpretability in Safety-Critical Financial Trading Systems*).

The revised wording around the definition of risk is welcome since it aligns better with the meaning of the term “risk” in non-AI risk management and avoids potential misunderstandings.

The framework is somewhat putting too much emphasis in certain directions at the expense of others. Bias management is rightfully a clear and essential component of the framework but it is somewhat overly represented in the current draft in proportion to other requirements of trustworthy AI – perhaps due to the concurrent NIST AI bias management efforts. On the contrary, security is not as present as it should be in the current draft. This is unfortunate since the controlled adversarial perspectives proposed in security can be used for interpretability through a controlled exploration of an AI system’s parameters as in *Interpretability in Safety-Critical Financial Trading Systems*.

Lastly, the Govern function and the framework as a whole could benefit from a maturity model *à la* CMMI to guide the improvement process of organizations willing to implement the framework.

Detailed comments and suggested line edits

Suggested additions are marked in red “example addition”, deletions are marked by strikeout ”~~example deletion~~”.

Part 1: Motivation

1. Overview

1.1 Trustworthy and responsible AI

"A useful mathematical representation of the data interactions that drive the AI system’s behavior is not fully known nor is there consensus with regards to its existence or lack thereof.”

“[...] which makes current methods for measuring risks and navigating the risk-benefits tradeoff ~~inadequate~~ difficult to apply as is.“

“AI system, the AI system itself, the use of the AI system, or interaction of other systems’ components or people with the AI system.”

“Trustworthy AI is valid and reliable, safe, fair and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced.” While all of these properties are desirable, all properties are not always required for a system to achieve its objective while limiting the risks related to this objective. Moreover, satisfying multiple of the stated properties is necessarily leading to tradeoffs. Lastly, there could be others not yet stated desirable properties as this is a growing research field with numerous open problems.
It would be good to modify the wording to state that “depending on the target application, the required characteristics among these – and potentially other trustworthiness requirements defined on a need basis – should be clearly identified.”

1.2. Purpose of the AI RMF

“[...] preserve civil liberties and rights, and enhance safety and security”

“The AI RMF is not a checklist and it is not intended to be used in isolation. Organizations may find it valuable to incorporate the AI RMF into broader considerations of enterprise risk management.” It would be valuable to add pointers to these risk management resources here or mention that these references are available in the playbook because some AI RMF users may not be risk managers by training.

“The AI RMF is not a compliance mechanism.” Avoiding the compliance trap is an important element of any risk management framework. However, this sentence could be better worded to show the intent of avoiding compliance as a goal. The framework could argue that “compliance is second to adequate risk management”. This distinction is important since the rest of the paragraph seems to imply that the AI RMF would transform into a compliance framework once laws and regulation have settled.

“The research community may find the AI RMF to be useful in evaluating various aspects of trustworthy and responsible AI and related impacts.” This sentence needs to be reviewed to specify the types of uses (quantitative or qualitative only etc). Note that this sentence currently contradicts the opening statement of 3.2.1 which states that “AI risks and impacts

that are not well-defined or adequately understood are difficult to measure quantitatively or qualitatively".
In light of this, the statement could be rewritten as "The research community may find the AI RMF to be useful in identifying new and remaining challenges in evaluating various aspects of trustworthy and responsible AI and related impacts."

"Using the AI RMF may reduce the likelihood and degree of negative impacts and increase the benefits to individuals, groups, communities, organizations, and society." It should also be noted that this is sound risk management which has been surprisingly absent from many AI projects to this point in time. Not implementing the AI RMF or similar AI governance systems would thus amount to an organization not performing due diligence.

"Applying the Framework at the beginning of an AI system's lifecycle [...]" As above, it should also be noted that it is sound risk management to include risk controls from the very infancy of a project whether AI related or not.

2. Audience

"The broad audience of the AI RMF is shown in Figure 1." The legend of Figure 1 states that it represents the lifecycle and there is therefore a risk of reader confusion. This sentence needs to be revised to make the link with the figure clearer.

3. Framing Risk

"AI risk management is about offering a path to minimize potential negative impacts of AI systems, such as threats to civil liberties and rights, harm that comes from the inadequation between the purported capabilities of an AI and its real capabilities[...]" This addition is aimed at addressing the common disconnect between an AI system output and users' interpretation of this output, for instance when a correlation link is reported as a causation link.

"While some AI risks and benefits are well-known, it can be challenging to assess the degree to which a negative impact is related to actual harms." This is a really important remark but it is missing some context for all audiences to understand. This is particularly the case since components interaction within the AI system may accentuate or reduce potential impact of AI risks (see for instance *ISO/IEC 19792:2009* and *ISO/IEC 24745:2011* on biometrics, *On the Exploitability of Audio Machine Learning Pipelines to Surreptitious Adversarial Examples*, *Rearchitecting Classification Frameworks For Increased Robustness*, *Interpretability in Safety-Critical Financial Trading Systems*).

3.2. Challenges for AI Risk Management

3.2.1. Risk Measurement

"AI risks measured in a laboratory or a controlled environment may differ from risks that emerge in operational setting or the real world. This is partially due to interactions within the AI system and with its environment whose complexity would not be tractable in a controlled laboratory environment."

3.2.2. Risk Tolerance

“To the extent that challenges for specifying risk tolerances remain unresolved, there may be contexts where a risk management framework is not yet readily applicable for mitigating AI risks. In the absence of risk tolerances prescribed by existing law, regulation, or norms, the AI RMF equips organizations to define reasonable risk tolerance, manage those risks, and document their risk management process.” The wording of those two sentences introduces an apparent contradiction where there is none. This is mostly because the first sentence mentions risk management frameworks and the AI RMF is a risk management framework. “Compared to a non AI-centric risk management framework, the AI RMF provides a more appropriate methodology to apprehend [...]” is a preferable wording which limits the apparent contradiction.

“In some cases where an AI system presents the highest risk – where negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently mitigated.” There term ***risk avoidance*** should be mentioned as it is the one used in risk management to describe a no-go decision due to risks outweighing benefits.

3.2.3. Risk Perspectives

This section should mention that the aim is to bring the ***residual risk*** below an acceptable risk threshold for negative impacts. Mentioning these two risk management terms (risk avoidance & residual risk) explicitly will help ground the AI RMF and ensure a common vocabulary across communities.

3.2.4. Organizational Integration of Risk

“[...] confidentiality, integrity and availability of training and output data and general security of the underlying technical stack (both software and hardware) for AI systems.” For references in how specific optimizations made to enhance AI system speed introduce a new attack surface that can be exploited by adversaries, see *Sponge Examples: Energy-Latency Attacks on Neural Networks* or *https://github.com/alkaet/LobotoMl*.

4. AI Risks and Trustworthiness

“Trustworthy AI is: valid and reliable, safe, fair and bias is managed,
secure and resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced to a level sufficient to maintain the risks faced by the given use of the AI system below acceptable thresholds.” It should be insisted that there is no *one-size-fits-all* in risk management. Intended use and similar circumstances identified in the Map function drive the appropriate level of all of the aforementioned characteristics.

“These characteristics are inextricably tied to human social and organizational behavior, the datasets used by AI systems, the choice of AI models/algorithms, AI technology stack and the decisions made by those who build them, and the interactions with the humans who provide insight from and oversight of such systems.”

“Human judgment must be employed when deciding on the specific metrics irrespective of whether the metrics are qualitative or quantitative [...]”

“Addressing AI trustworthy characteristics individually will not assure AI system trustworthiness, and tradeoffs are always involved. Trustworthiness is greater than the sum of its parts.” The wording should be revised to include cases for which there is a clear priority ordering of the aforementioned characteristics. In such cases, the tradeoffs should be informed by the characteristics’ priority ordering as is done in standard cybersecurity and risk management. For instance, priority can be given to a single component of the Confidentiality, Integrity and Availability triad if Availability requirements are deemed to trump confidentiality requirements for the considered system use case (See for instance Bruce Schneier’s blog for such an example in the Internet of Things:*Integrity and Availability Threats*). The follow up sentence starting at “Ultimately, it is a social concept, [...]” is not sufficiently explicit to address this point in its current wording. See also *Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations* for an example of existing tradeoffs within these characteristics.

“Increasing the breadth and diversity of stakeholder input throughout the AI lifecycle [...] “ While introducing additional stakeholders is often a net benefit, it does come with some challenges. The AI RMF should not shy away from outlining the tradeoffs coming with a larger pool of stakeholders such as dilution of opinion representation and added management difficulties like increased communication overheads. The AI RMF should additionally pose itself as a guide in ensuring a common language between this very diverse pool of stakeholders.

“In many settings such experts provide their insights about particular domain knowledge, and are not necessarily able to perform intended oversight or governance functions for AI systems they played no role in developing.” This should be complemented by a sentence outlining that “Systemic decisions in how the oversight is designed should be carefully considered to avoid a box ticking approach by the expert. For instance, the oversight will be inadequate if the domain knowledge of the “experts” chosen by the AI system designer does not match the domain knowledge needed to review the task. Similarly the oversight will be inadequate if insufficient time and/or reward are allocated to the review process.”

4.1. Valid and Reliable

“Accuracy measurements should always be paired with clearly defined test sets and details about test methodology; both should be included in associated documentation. The methodology should specify how the test set is expected to relate/represent real use case examples – to avoid a test set that is constructed to mirror the training set distribution but not the real world input distribution. The methodology should not hide away behind fabricated complexity, be easily understood by and avoid misleading other stakeholders in particular end users. ”

“[...] a goal for overall correctness of ~~model~~ AI system operation” This is prefered as it goes beyond the model and includes the whole deployment stack.

“[...] but also that it should perform in ways that minimize potential harms to people if it is operating in an unexpected environment.” Mention that this is both a **fail-safe and fail-secure** mechanism rather than an all encompassing performance guarantee. Generalization/robustness in all possible circumstances is an open research problem.

“Validity and reliability for deployed AI systems is often assessed by ongoing ~~audits~~ testing or monitoring that confirm a system is performing as intended.” “Audits” seems to be a term stronger than current industry practices for AI system management since it implies some formalism and systematic methodology. Such a formalized and systematic methodology is **not yet commonplace practice** for AI risk management in organizations but will become the norm as the AI RMF progresses. To reflect current practices, “testing” would be a better term.

“Measurement of accuracy, reliability, and robustness contribute to trustworthiness and ~~should consider that certain types of failures can cause greater harm – and risks should be managed to minimize the negative impact of those failures~~ should prioritize the minimization of negative impacts based on the potential impact with higher harm treated first.“

4.2. Safe

“Safe operation of AI systems requires responsible design and development practices, clear information to deployers on how to use a system appropriately, and responsible decision-making by deployers and end-users – in parallel to a clear explanation of the risks involved with mishandling the system. This explanation of the risks entailed is crucial to responsibilize the deployers and end-users.”

“Employing safety considerations during the whole lifecycle and starting as early as possible with planning and design can prevent failures or conditions that can render a system dangerous. “

“AI safety measures should take cues from measures of safety used in other fields, such as transportation, ~~and~~ healthcare and finance (for damages to property).”

4.3. Fair – and Bias Is Managed

No comments as of this version.

4.4. Secure and Resilient

“[...] adversarial attacks [...]” **An attack is by nature adversarial.** This is therefore a tautology in the general case. The prefered term should be one of: (a) attacks, (b) adversarial events or (c) adversarial examples if referring only to that particular subfield of Adversarial machine learning. In the case of (c), the wording should make it clearer that possible attacks on AI systems are much wider than adversarial examples, e.g. sponge examples, membership inference or poisoning. The same vocabulary issue (“adversarial attacks”) is present on page iii under the The NIST Trustworthy and Responsible AI Resource Center description paragraph.

“[...] to maintain their functions and structure in the face of internal and external change, and to degrade gracefully when this is necessary (Adapted from: ISO/IEC TS 5723:2022) may be said to be resilient.” The difference with the robustness or generalization property definition in 4.1 is not clear. It can be further argued that, even after the line

“Resilience has some relationship to robustness except that it goes beyond the provenance of the data to encompass unexpected or adversarial use of the model or data.“ in the next paragraph, the difference is not explicit since an attack is simply **a special case** of adverse event or threat.

“AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be secure.” This statement should be reviewed as all CIA components are not always required for a system to be secure. For instance, priority can be given to a single component of the Confidentiality, Integrity and Availability triad if Availability requirements are deemed to trump confidentiality requirements for the considered system use case. (See for instance Bruce Schneier’s blog for such an example in the Internet of Things:*Integrity and Availability Threats*).

“[...] protocols to avoid or protect against, to respond to in-progress and recover from attacks.” This addition helps address all five components of the NIST CSF (Detect, Protect, Identify, Respond and Recover).

“Other common security concerns relate to data poisoning and the exfiltration of models, training data, or other intellectual property through AI system endpoints.” Despite referring to ***AI systems***, the framework and this sentence in particular focus very strongly on the data and model. The framework should nonetheless take a larger stance rather than limit itself to well known data and model centric aspects of adversarial machine learning. To this end the NIST AI RMF should be proactive in identifying areas of concerns in the entire AI supporting stack like availability attacks (see for instance *Sponge Examples: Energy-Latency Attacks on Neural Networks* or *https://github.com/alkaet/LobotoMl*).

4.5. Transparent and Accountable

The presentation of the second paragraph of this section is somewhat unclear with regard to the specificity of the AI ecosystem and how this impacts accountability. More specifically, model reuse and fine tuning are widespread practices. It is debatable whether a public model provider should be responsible for misuse of the provided models. This accountability attribution is rendered even more complex since AI systems’ models can be copied for a fraction of their design cost by model extraction attacks. Note however that fine tuning is described in Appendix b and this concern could thus be addressed by a forward pointer.

“[...] supporting attribution of decisions of the AI system to subsets of training data can assist with both transparency and accountability.” An AI system can be used for tasks other than decision making or classification which would not be covered here. For instance, image generation or so-called Deep Fakes would not be covered by the current wording of the AI RMF despite being arguably an important source of risk for AI systems. This sentence in particular should be extended to include methods such as entangled watermarking (see *Entangled Watermarks as a Defense against Model Extraction*) that enable attribution of a model despite extraction and copy attempts. Moreover, this sentence is tying the entirety of transparency and accountability to training data despite the accountability problem spanning both data and model, if not the entire AI system stack (see for instance the discussion of ML logging in *SoK: Machine Learning Governance*).

4.6. Explainable and Interpretable

“Risk from lack of explainability may be managed by descriptions of how ~~models~~ AI systems work tailored to individual differences such as the user’s knowledge and skill level.” Prefer the broader “AI system” wording.

4.7. Privacy-Enhanced

“Evaluations of AI RMF effectiveness – including ways to measure bottom-line improvements in the trustworthiness of AI systems – will be part of future NIST activities, in conjunction with stakeholders.” This sentence is extremely important and should be further emphasized to show how researchers should use the NIST AI RMFT to highlight areas that cannot currently be performed adequately and require further academic research. For instance, an important unexplored area in the academic research realm are methods for estimating the work space and ultimately scoping a technical AI system audit engagement.

Part 2: Core and Profiles

6. AI RMF Core

6.1. Govern

The Govern function and the framework as a whole could benefit from a maturity model *à la* CMMI to guide the improvement process of organizations willing to implement the framework.
See also the detailed comments in the playbook section.

6.2. Maps

See detailed comments in the playbook section.

6.3. Measure

The Measure Function is a very important framework function and is adequate in its general current form. Nonetheless, the effectiveness of a number of metrics – that the Measure function should rely upon – for AI systems risk evaluation are still debated in the community. Moreover, as pointed out above, assessment of AI systems is currently limited due to the need to research and develop new tools to perform functions such as scoping.
Lastly, the Measure function should be the function that specifically addresses the potential issue of box ticking in AI risk management. To this end, it should provide insight into the limitations of the different currently available metrics and evaluations methods.

“Independent review by sufficiently qualified personnel both technically and from a governance standpoint[...]”

“Where tradeoffs among the trustworthy characteristics arise, measurement provides a traceable basis to inform management decisions.” The current wording does not reflect that this is also and primarily subordinate to the context provided by the Map function. This context is crucial in that it informs the prioritization of individual trustworthiness properties.

“Options may include recalibration, impact mitigation, removal of the system from production as well as a range of compensating, detective, deterrent, directive and recovery controls.”

6.4. Manage

No comments as of this version.

7. AI RMF Profiles

No comments as of this version.

Appendix A

The AI Design, AI development and AI deployment tasks currently do not explicitly include technical security experts. These tasks should include technical security experts in quality of security architects, DevSecOps and overall risk anticipation personnel. Risk management is a peculiar thought process. While this thought process can be acquired, it is at the present time often not yet part of the training curriculum of the listed following technical actors: “machine learning experts, data scientists, developers, domain experts, socio-cultural analysts, data engineers, data providers”.
TEVV should operate in parallel and as a complement to security and trustworthiness architects to challenge and validate the design choices.

“Tasks can be incorporated into a phase as early as design, where tests are planned in accordance with the design requirement.” The wording of this sentence is somewhat misleading. As stated in the following bullet points, the tests should **challenge** the design requirements as needed rather than accommodate them. This is crucial since TEVV is an oversight function and thus needs sufficient independence.

“Third-party entities are responsible for AI design and development tasks, in whole or in part. Note however that ~~B~~by definition, they are external to the design, development, or deployment team of the organization that acquires its technologies or services.” Connector inserted to make the sentence less prone to misreading.

Appendix B

This section is missing the risks linked to the underlying stack both for hardware and software. (see for instance *Sponge Examples: Energy-Latency Attacks on Neural Networks* or *https://github.com/alkaet/LobotoMl*)

“» comprehensively address security concerns related to evasion, model extraction, membership inference, availability or other machine learning attacks;”

Playbook comments

Suggested additions are marked in red “example addition”, deletions are marked by strikeout ”~~example deletion~~”.

General comments

In a significant number of the Playbook subsections, the transparency subsubsection actually contains elements or questions that would benefit from being integrated to the action subsubsection. This is because these elements or questions go beyond the transparency function – as they go beyond the realm of communication with internal and external parties. In fact, they are crucial to ensure **sufficient depth** for the framework to be more than a box ticking tool. Considering the items listed in those questions as actions should thus be necessary for the framework‘s due implementation. We have explicitly marked these elements or questions in the detailed comments below.
Rather than rewriting each of these questions as actions, a similar effect could alternatively **more easily but less impactfully** be obtained by changing the wording of the transparency section to “**Organizations should ~~can~~ document the following:”**

Detailed playbook comments

Suggested additions are marked in red “example addition”, deletions are marked by strikeout ”~~example deletion~~”.

Govern

1.1

No comments as of this version.

1.2

“Organizational policies and procedures will vary based on available resources and risk profiles, but ~~can~~ help systematize AI actor roles and responsibilities throughout the AI model lifecycle.”

“[...] Lack of clear information about responsibilities and chains of command ~~will~~ limits the effectiveness of risk management.”

The bullet points listed in the actions section under the organizational policies requirements should be reviewed. The current bullet points tend to merge together standards, procedures, guidelines and policies. This risks yielding monolithic and confusing documents that fail to achieve the desired outcome in less mature organizations.
Make the different categories clearer or modify the wording into “Organizational policies, guidelines, procedures and standards should:”

1.3

No comments as of this version.

1.4

Adding a reference to threat hunting and other ways of proactively looking for potential incidents and compromises would be beneficial in this section.

“Establishing and maintaining incident response plans can reduce the likelihood of additive impacts during an AI incident, by alleviating stress and providing clear responsibilities.”

“Establish policies and procedures for monitoring AI system performance, and to address bias, safety and security problems, across the lifecycle of the system.”

“Establish policies and procedures for AI system incident response, or confirm that existing incident response policies address AI systems.”

“Establish mechanisms to enable the sharing of feedback from impacted individuals or communities and relevant authorities about negative impacts from AI systems.

2.1

“Establish policies that separate management of AI system development functions from AI system testing and auditing functions, to enable independent course-correction of AI systems”

2.2

The playbook should provide general directions for relevant training. Traditional risk management training, technical training on AI systems and potentially specialized courses to be developed should be outlined in this section.
The training should provide adequate technical depth to understand the specificities of AI compared to other traditional software.

“How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

2.3

“Some organizations grant authority and resources (human and budgetary) to a designated officer who ensures adequate performance of the institution’s AI portfolio (e.g. predictive modeling, machine learning). Similarly a designated officer should be responsible of the adequate risk management and trustworthiness of the systems in the AI portfolio”

“Organizational management should ~~can~~:”

“Organizations should ~~can~~ establish board committees for AI risk management and oversight functions and integrate those functions within the organization’s broader enterprise risk management approaches.”

“How does ~~Did~~ your organization’s board and/or senior management sponsor, support and participate in your organization’s AI governance?”

“What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“Do AI solutions provide sufficient information to assist the personnel to make an informed decision and take actions accordingly?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

3.1

“Organizational management should ~~can~~:”

4.1

“Have you documented and explained that machine errors may differ from human errors?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

4.2

No comments as of this version.

4.3

“Establish policies and procedures to facilitate and equip AI system testing and auditing.”

5.1

No comments as of this version.

5.2

“When risks arise, resources are allocated based on the assessed risk of a given AI system.“ This wording tends to suggest a reactive risk approach. A **proactive risk management** approach and corresponding wording like ”ahead of risk occurrence” is preferable.

“Does the AI solution provide sufficient information to assist the personnel to make an informed decision and take actions accordingly?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

6.1

This section would benefit from putting more emphasis on the specificities of AI third party risk management compared to traditional (non-AI) third party risk management. These specificities are informed by the stochastic nature of an important proportion of AI systems and by specific practices in the AI development realm such as model reuse and fine tuning.

To further address this, include a statement in the “Actions” section for organizations which already have strong traditional (non-AI) third party risk management procedures which would not readily port to AI systems third party risk management: “consider and address the potential limitations of the currently existing third party management policies and procedures in light of the differences and specificities of AI with regard to other systems”

“Did you ensure that the AI system can be audited by independent third parties?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“Did you establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

6.2

“To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“Did you establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“If your organization obtained datasets from a third party, how did your organization assess and manage the risks of using such datasets?”

MAP

1.1

“Track, inventory and document existing AI systems held by the organization, and those maintained or supported by third-party entities.”

1.2

“ How did your organization address usability problems and test whether user interfaces served their intended purposes? Consulting the community or end users at the earliest stages of development to ensure there is transparency on the technology used and how it is deployed.”

1.3

“AI systems should present a business benefit beyond the status quo when considering inherent risks and implicit or explicit costs. If the implicit or explicit risks outweigh the advantages, organizations should feel confident in performing *risk avoidance*, i.e., refusing to implement an AI solution whose risks surpass potential benefits.”

“How do the technical specifications and requirements align with the AI system’s goals and objectives?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“To what extent is the output appropriate for the operational context?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

1.4

“~~Reconsider~~ Approach the design, implementation strategy, or deployment of AI systems by including ~~with~~ potential impacts that do not reflect institutional values.”

1.5

“For systems deemed “higher risk,” such decisions should include approval from relevant technical or risk-focused executives or AI risk steering committee.”

1.6

No comments as of this version.

1.7

No comments as of this version.

2.1

“AI actors should define the technical learning or decision-making task an AI system is designed to accomplish, along with the benefits that the system will provide.“ This sentence does not address AI systems that do not perform decision making such as generative systems (e.g. DeepFake generators). Consider revising the wording of this sentence to address these types of AI system functionalities.

“How do the technical specifications and requirements align with the AI system’s goals and objectives?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

2.2

“Does the AI solution provide sufficient information to assist the personnel to make an informed decision and take actions accordingly?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“To what extent is the output of each component appropriate for the operational context?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“How will the accountable 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 operational/business environment, which may impact the accuracy of the AI system?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

2.3

This section is missing a sentence outlining risks linked to specialized AI hardware which are prone to specific attacks such as availability attacks (see *Sponge Examples: Energy-Latency Attacks on Neural Networks*).
It is also missing a sentence considering new types of attacks based on data reordering that leverage lack of randomness in data pipelining to achieve poisoning (see *Manipulating SGD with Data Ordering Attacks*).

“This may have a disproportionately negative impact on minorities, vulnerable and disadvantaged groups such as black, indigenous, and people of color, women, LGBTQ+ individuals, people with disabilities, or people with limited access to computer network technologies.” **If the framework aims at being used globally, the framework should consider and outline that minorities, vulnerable and disadvantaged groups are not the same in every country or region of the planet**. Local circumstances should be considered and documented by the organizations aiming at implementing the framework.

“How does the entity ensure that the data collected are adequate, relevant, and not excessive in relation to the intended purpose?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

3.1

“Have the appropriate training material and disclaimers about how to adequately use the AI system been provided to users?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

3.2

To follow a proactive rather than reactive AI risk management mapping, suggest using scenario-based risk analysis as is commonly performed in non-AI risk management processes of mature organizations. Threat modeling in security is an example of such a successful process that organizations can draw inspiration from. (See for instance https://www.threatmodelingmanifesto.org/)

Insist that organizations should pay specific attention in their analysis to:

1. compounding risk factors
2. apply a system level perspective to inter and intra system component interactions which may mitigate or exacerbate risks

3.3

“To what extent has the entity clearly defined technical specifications and requirements for the AI system?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“How do the technical specifications and requirements align with the AI system’s goals and objectives?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

4.1

“Did you establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“If your organization obtained datasets from a third party, how did your organization assess and manage the risks of using such datasets?”

“How and by whom will the results be independently verified?”

4.2

“Did you ensure that the AI system can be audited by independent third parties? If so, how will this audit be conducted?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

“Did you establish mechanisms that facilitate the AI system’s auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system’s processes, outcomes, positive and negative impact)? If so, how are these mechanisms implemented?” This is a crucial question for implementation depth and should probably be an actual “Actions” requirement of the framework rather than only a transparency element.

5.1

“The Map function provides an opportunity for organizations to assess potential AI system impacts based on identified risks. This enables organizations to create a baseline for system monitoring and to increase opportunities for detecting emergent risks.” This set of sentences is hard to understand.
A proposed rewording is “Organizations should create a baseline for system monitoring and to increase opportunities for detecting emergent risks based on the AI system impact assessment from the Map function.”

5.2

“ Will or c~~C~~an the AI system be audited by independent third parties?”

5.3

No comments as of this version.