Response to NIST AI RMF Full Draft Playbook, Roadmap and Crosswalks

24 February 2023

Elham Tabassi, Chief of Staff, Information Technology Laboratory
National Institute of Standards and Technology (NIST)
100 Bureau Drive, Gaithersburg, MD 20899

Subject: NIST AI Risk Management Framework January 2023 Full Draft Playbook, Roadmap and Crosswalks

Via email to AIframework@nist.gov

To Ms. Tabassi, and the entire NIST team developing AI Risk Management Framework resources,

Thank you for the invitation to submit comments in response to the NIST AI Risk Management Framework (AI RMF) Full Draft Playbook, Roadmap, and Crosswalks released January 2023. We offer the following submission for your consideration.

We are researchers affiliated with UC Berkeley, with expertise on AI research and development, safety, security, policy, and ethics. We previously submitted responses to NIST in September 2021 on the NIST AI RMF Request For Information (RFI), in January 2022 on the AI RMF Concept Paper, in April 2022 on the AI RMF Initial Draft, and in September 2022 on the AI RMF 2nd Draft and Initial Draft Playbook.

Here is a high level summary of some of our key comments and recommendations on the January 2023 Full Draft Playbook and Roadmap:

- Ensure consistency in the evaluation of both the likelihood and magnitude of identified impacts throughout the mapping function.
- Provide examples of potentially unacceptable risks from the main AI RMF 1.0 guidance document in the Playbook.
- Encourage consideration of the potential for unintended consequences from failures of system objectives specification.
- Enhance the utility of the Playbook by adding publicly available tools and resources to each subcategory.
- Provide examples to help organizations consider potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet.

- Encourage organizations to establish policies and practices to inform users (and allow them to opt out) if they are interacting with an AI system or if a decision that impacts them was made by an AI system.
- Encourage organizations to establish policies and practices to provide recourse or redress to people who experience negative impacts related to the use of an AI system.

In the following sections, we provide detail and additional comments on the NIST AI RMF Full Draft Playbook, Roadmap, and Crosswalks.

Thank you again for the opportunity to comment on the AI RMF Full Draft Playbook, Roadmap and Crosswalks. If you need additional information or would like to discuss further, please contact Anthony Barrett at anthony.barrett@berkeley.edu. In any case, we look forward to further engagement with NIST as you proceed on the AI RMF resource development process.

Our best,

Anthony Barrett, Ph.D., PMP
Visiting Scholar
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Dan Hendrycks, Ph.D.
Berkeley AI Research Lab, UC Berkeley

Jessica Newman
Director
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley
Co-Director
AI Policy Hub, UC Berkeley

Brandie Nonnecke, Ph.D.
Director
CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley
Co-Director
AI Policy Hub, UC Berkeley

Our overarching comments on the NIST AI RMF Full Draft Playbook

Connect the Playbook to available tools

Response/Comment:
Currently, Measure 2.7 includes helpful software resources in the References section, which can help users of the Playbook access state of the art tools to help ensure the security and

resilience of their AI systems. This is incredibly helpful to include, but it begs the question of why other subcategories do not also have helpful tools and resources listed. We think that many if not all of the subcategories would benefit from the addition of a new Tools and Resources section.

Suggested Change:
We recommend adding a new section to all subcategories that lists publicly available tools and resources that can help to operationalize the suggested actions. We understand that not all resources will be equally reliable or fitting across all situations, but they still provide a critical starting point and will make it significantly easier for organizations to make practical changes, especially in the short term. NIST does not need to develop this list from scratch, but can rather lean on the OECD Tools for Trustworthy AI (https://oecd.ai/en/catalogue/tools), and perhaps even explicitly link to the OECD database when that becomes publicly available.

Our comments on specific passages in the NIST AI RMF Full Draft Playbook

Govern 1.1

Response/Comment:
The legal implications of an AI system do not only arise within the deployment context, but throughout the AI lifecycle. For example, decisions about how to train an AI system and which training data to use may have legal implications as evidenced by current lawsuits about illegal data scraping to train image generation models.

Suggested Change:
Clarify that the legal environment requires attention throughout the AI lifecycle, including during development, and not only at the point of deployment. Clarify that the legal environment should be revisited as ongoing monitoring of the system takes place in case the system ends up being used in a new environment or domain that has new legal implications. Clarify also that part of managing the legal and regulatory requirements includes ensuring that due process and due protection, including whistleblower protection, are provided.

Govern 1.2

Response/Comment:
As currently written, Govern 1.2 does not clearly integrate specific suggested actions with the characteristics of trustworthy AI as defined in the AI RMF Core.

Suggested Change:
In the About section, reiterate what the NIST “characteristics of trustworthiness” are.

In the Transparency and Documentation section add, “To what extent do these policies support the implementation of the characteristics of trustworthiness?”

In the Suggested Actions section add: “Carry out a data privacy or protection impact assessment;” “Incorporate trustworthy characteristics into AI procurement standards;” and add “including vulnerability disclosure” to “Establish the frequency of and detail for monitoring, auditing and review processes.”

Consider adding as a reference, Jessica Newman (2023) “A Taxonomy of Trustworthiness for Artificial Intelligence: Connecting Properties of Trustworthiness with Risk Management and the AI Lifecycle,” UC Berkeley Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf.

Govern 1.3

Response/Comment:
The current activities listed in the Suggested Actions section relate to assessing an AI system’s impacts, but do not relate specifically to determining risk tolerances.

Suggested Change:
Add a Suggested Action that states, “Determine which risks are considered acceptable, which risks require mitigation strategies, and which risks are unacceptable.” (See also our recommendation under Map 1.5 for providing examples in the Playbook of what could be considered unacceptable risks – possibly a similar addition should be made in the Playbook under Govern 1.3.)

In the Transparency and Documentation section add, “How has the organization determined its risk tolerance and how does this inform risk management activities.”

Govern 1.4

Response/Comment:
The list of information to be included in documentation policies should include reference to the risks and impacts of the AI system.

Suggested Change:
Add another sub-bullet that says, “Expected and potential risks and impacts” to the first bullet point in the “Suggested Actions” section, which says, “Establish and regularly review documentation policies that, among others, address information related to:”

Govern 1.5

Response/Comment:
As currently written, the suggested actions listed for Govern 1.5 relate primarily to the assessment of the AI system, but not directly to the review of the risk management process and its outcomes as named in the heading of Govern 1.5.

Suggested Change:
Add a Suggested Action that states “Establish policies to assess and review the risk management process and its outcomes on a regular and ongoing basis.”

Add a Suggested Action that states “Establish policies to define organizational roles and responsibilities to support actions described in Govern 2.1.”

Govern 4.1

Response/Comment:
There are a number of additional actions that may strengthen Govern 4.1 and further help organizations to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize negative impacts.

Suggested Change:
Add a Suggested Action that states “Establish foresight and/or scenario planning exercises to help prepare for uncertainties and an evolving risk landscape.”

Add a Suggested Action that states “Establish policies that encourage consideration of possible uses and misuses of the AI system beyond its expected use.”

Add a Suggested Action that states “Establish policies that incentivize internal reporting of potential challenges or concerns, for example by developing a dedicated phone number and email address.”

Add a Suggested Action that states “Establish policies that encourage continuous monitoring and awareness of contextual changes or shifts in the AI system’s functionality or capabilities over time.”

Govern 4.2

Response/Comment:
The communication of risks and impacts could be expanded upon in the Playbook material on Govern 4.2. Sections 3.3 and 3.4 of our actionable-guidance paper (Barrett et al. 2022) include material on communicating various potential types of impacts (including human rights impacts) as appropriate in context as part of communicating AI system limitations and risks to stakeholders.

Suggested Change:
First, in the Playbook section on Govern 4.2 under Suggested Actions, add a bullet that reads as follows: “Report risk factors identified in AI system risk assessment, including on potential types of impacts or harms outside the organization, by time of deployment or at earlier lifecycle stages, as appropriate in context as part of communicating AI system limitations and risks to stakeholders. Incorporate outputs from Map 1.1 and associated impact-assessment and risk-assessment activities, as appropriate. ”

Second, in the Playbook section on Govern 4.2 under “Organizations can document the following”, add “and communicated” to the fourth bullet, so that it reads as follows: “To what extent has the entity documented and communicated the AI system’s development, testing methodology, metrics, and performance outcomes?”

Third, in the AI RMF Playbook, list our actionable-guidance paper (Barrett et al. 2022) as an informative reference for Govern 4.2 for communicating various potential types of impacts.

References:
Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Govern 5.2

Response/Comment:
As currently written, none of the suggested actions for Govern 5.2 directly relate to ensuring that AI actors have meaningful opportunities to provide feedback on system design and implementation, as described in the heading of Govern 5.2.

Suggested Change:
Add a Suggested Action that states “Establish policies that ensure all relevant AI actors are provided with meaningful opportunities to provide feedback on system design and implementation.”

Govern 6.1

Response/Comment:
Govern 6.1 does not currently explicitly mention that the risks associated with third-party entities are relevant to organizational decisions about procuring AI systems. It may be helpful to mention this explicitly. Additionally, third-party data is mentioned, but it may be helpful to clarify that this would include any external data used to train an AI model. Lastly, it may be helpful to recognize supply chain risks of third-party hardware procured and integrated into an organization’s AI system.

Suggested Change:
Add a Suggested Action that states “If relevant, establish policies to support responsible procurement from third-parties.”

Add a Suggested Action that states “If relevant, establish policies to evaluate all third-party training data sources.”

Under Suggested Actions, change “Collaboratively establish policies that address third-party AI systems and data” to “Collaboratively establish policies that address third-party AI systems, including constituent components, such as data, hardware, and software that may be integrated into an organization's own AI system."

Map 1.1

Response/Comment:
First, in the Suggested Actions section of Map 1.1 in the AI RMF Playbook, the third bullet helpfully prompts consideration of "intended AI system design tasks along with unanticipated purposes." As we note in our actionable-guidance paper (Barrett et al. 2022), many AI researchers regard system objectives specification (or alignment of system behavior with designer goals) as an aspect of AI trustworthiness that is already important for AI systems and whose importance will only increase as AI systems grow in scale and capabilities. Specification of an AI system's goals or objectives aims to align the system's behavior with the designer's intentions. Objectives mis-specification risks can include cases where a system meets its literal goals but has unanticipated or unintended behaviors that cause harm. Rudner and Toner (2021) provide brief examples, such as social-media content recommendation machine-learning algorithms that learn to optimize user-engagement metrics by serving users with extremist content or disinformation. Rudner and Toner (2021, p. 10) also suggest accounting for worst-case scenarios, and considering the following questions for an AI system, as part of identifying mis-specification risks: "What objective has been specified for the system, and what kinds of perverse behavior could be incentivized by optimizing for that objective?"

Second, the Map 1.1 Suggested Actions section includes a bullet point for readers to "Consider intended AI system design tasks along with unanticipated purposes in collaboration with human factors and socio-technical domain experts." However, we believe it would be valuable for Map guidance in the Playbook to provide more on identification of other potentially beneficial uses of an AI system, as well as identification of negative "misuse/abuse cases", beyond an AI developer's or deployer's originally intended uses of an AI system. This would better address both positive and adverse risks of reasonably foreseeable "off label" uses. Section 3.1 of our paper Barrett et al. (2022) provides guidance for identifying other potentially beneficial uses of an AI system as well as negative "misuse/abuse cases", and would be useful as an informative reference for Map 1.1 under "Identification of harms" as well as under "Context mapping".

Third, Map 1.1 has now become one of the important places within the AI RMF where users are prompted to consider the potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet. However, in the Suggested Actions section, no further detail is provided about this. We recommend adding, at a minimum, examples of some of the types of issues and dimensions of potential impacts that should be considered at this stage.

Lastly, one of the important elements of understanding the deployment context as called for in Map 1.1 is to consider whether the AI system may reasonably be expected to be used in a high-stakes setting such as government, education, health, or law enforcement, or if the AI system may be used in critical infrastructure or safety critical systems. Additional actions are likely to be warranted in such cases.

Suggested Changes:
First, in the Suggested Actions section of Map 1.1 in the AI RMF Playbook, under the third bullet that prompts consideration of “intended AI system design tasks along with unanticipated purposes”, add or adapt the following as a sub-bullet:

- “For the intended AI system tasks or objectives, what unintended perverse or adverse behaviors could be incentivized by over-optimizing for those objectives? Incorporate any new impacts or risks identified into other impact assessment or risk assessment steps such as in Map 5.1.”

Second, we recommend listing our actionable-guidance paper (Barrett et al. 2022) as an informative reference for Map 1.1 under “Identification of harms”, and under “Context mapping”.

Third, in the Suggested Actions section, we recommend adding to the sub-bullet “potential positive and negative impacts to individuals, groups, communities, organizations, and society,” so that it says, “potential positive and negative impacts to individuals, groups, communities, organizations, and society. For example, consider positive and negative impacts related to issues or dimensions such as harassment, stereotyping, addiction, manipulation, equity, discrimination, accessibility, physical and psychological safety, security, privacy, labor rights, civil rights, human rights, democratic values and processes, human autonomy and freedom, human dignity, wellbeing, environmental impacts, potential for harms from mis-specified goals or implementation in “off-label” uses, or other systemic impacts such as to critical infrastructure or essential services.”

Lastly, we recommend adding a suggested action that states “If the AI system may be used in a high-stakes setting (such as government, education, health, or law enforcement) or if the AI system may be used in critical infrastructure or safety critical systems, determine what additional context and risks are at stake.”

References:

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Tim G.J. Rudner and Helen Toner (2021) Key Concepts in AI Safety: Specification in Machine Learning. CSET, https://cset.georgetown.edu/wp-content/uploads/Key-Concepts-in-AI-Safety-Specification-in-Machine-Learning.pdf

Jessica Newman (2023) “A Taxonomy of Trustworthiness for Artificial Intelligence: Connecting Properties of Trustworthiness with Risk Management and the AI Lifecycle,” UC Berkeley Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf

Map 1.5

Response/Comment:
The Map 1.5 guidance in the AI RMF Playbook suggests that organizations set risk tolerance considering regulations, sector requirements, and other factors. Many organizations would find it helpful to see some examples of what could be considered unacceptable risks. In discussion of risk prioritization and risk tolerance, the main AI RMF 1.0 guidance document (NIST 2023, p. 8) states that “In cases where an AI system presents unacceptable negative risk levels – such as where significant 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 managed.” This passage could be adapted into Map 1.5 Playbook examples of unacceptable risks. (It also could be adapted into Govern 1.3 Playbook guidance on risk tolerance.)

Suggested Change:
In the AI RMF Playbook Map 1.5 Suggested Actions section under the fourth bullet “Identify maximum allowable risk tolerance above which the system will not be deployed, or will need to be prematurely decommissioned, within the contextual or application setting” add another sentence or a sub-bullet as follows: “Examples of cases where an AI system presents unacceptable negative risk levels can include: where significant negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present.” (Also consider making a similar addition in Govern 1.3 Playbook guidance on risk tolerance.)

References:
NIST (2023) AI Risk Management Framework (AI RMF 1.0). AI 100-1. National Institute of Standards and Technology, https://doi.org/10.6028/NIST.AI.100-1.

Map 5.1

Response/Comment:
We appreciate that NIST has listed our actionable-guidance *arXiv* paper (Barrett et al. 2022) as an informative reference for Map 5.1 in the AI RMF Playbook, which includes evaluating magnitude of identified impacts. Section 3.2 of our paper provides an impact magnitude rating scale that includes consideration of societal-scale impact factors, which would usefully inform prioritization and go/no-go decisions as part of Map activities.

In the About section of Map 5.1 in the AI RMF Playbook, it appears that NIST intended for each instance of "likelihood" to be "likelihood and magnitude" in this passage. Making that correction to add "and magnitude" would make the passage consistent with the description of the Map 5.1 subcategory: "Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented." That also would make the description of Map 5.1 consistent with widely accepted best practices for risk assessment and prioritization, which include considering magnitude of potential impacts as well as likelihood of impacts. Prioritizing potential impacts in a way that considers only their likelihood and ignores their magnitude could result in overlooking risks of events that may not occur on a daily basis but can have severe and irreversible impacts for individuals, organizations and society when they occur.

Suggested Change:
Add "and magnitude" to the About section of Map 5.1 in the AI RMF Playbook, so that the "About" section of Map 5.1 reads as follows: "AI actors can evaluate, document and triage the likelihood and magnitude of AI system impacts identified in Map 5.1. Likelihood and magnitude estimates may then be assessed and judged for go/no-go decisions about deploying an AI system. If an organization decides to proceed with deploying the system, the likelihood and magnitude estimates can be used to assign TEVV resources appropriate for the risk level."

References:
Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. "Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks". ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Map 5.2

Response/Comment:
First, our actionable-guidance paper Barrett et al. (2022) could be listed as an informative reference for Map 5.2 for identification of various types of potential impacts to individuals, groups, organizations and society. Material in Section 3.2 of Barrett et al. (2022) prompts consideration of various factors that could lead to high consequences at a societal scale, and Section 3.3 prompts consideration of impacts to human rights.

Second, we note that in the Playbook section Map 5.2 under Suggested Actions, the fifth bullet suggests assessing impact likelihood but not also magnitude. As we discussed in our previous point, widely accepted best practices for risk assessment include assessing magnitude of potential impacts as well as likelihood of impacts. Prioritizing potential impacts in a way that considers only their likelihood and ignores their magnitude could result in overlooking risks of events that may not occur on a daily basis but can have severe and irreversible impacts for individuals, organizations and society when they occur.

Suggested Change:
First, we recommend listing our actionable-guidance paper (Barrett et al. 2022) as an informative reference for Map 5.2 for identification of various types of impacts to individuals, communities, organizations and society.

Second, in the Playbook section Map 5.2 under Suggested Actions, in the fifth bullet, change "likelihood" to "likelihood and magnitude" so that the fifth bullet reads as follows: "Identify a team (internal or external) that is independent of AI design and development functions to assess AI system benefits, positive and negative impacts, and their likelihood and magnitude."

References:

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Measure 2.8

Response/Comment:
Measure 2.8 is the subcategory in which the risks associated with transparency and accountability are examined and documented. One critical component of transparency and accountability is ensuring that people are made aware if they are interacting with an AI system or if critical decisions that impact their lives or livelihoods were made by an AI system. However, this is not currently explicitly mentioned in Measure 2.8.

Suggested Change:
Add a suggested action that states “Develop policies and practices to inform users if they are interacting with an AI system or if a decision that impacts them was made by an AI system.”

Manage 1.3

Response/Comment:
In Manage 1.3, organizational response options for identified risks can include avoiding those risks. Following on our recommended Map 1.1 guidance on identifying potential uses and misuses beyond an AI developer’s or deployer’s originally intended uses of an AI system, we believe it would be helpful for Manage 1.3 guidance to better address both positive and adverse risks of reasonably foreseeable “off label” uses. Section 3.1 of our paper Barrett et al. (2022) provides guidance for defining and communicating to key stakeholders whether any potential use cases (or categories of use cases) would be unacceptable, disallowed, or another category for which an organization would provide specific risk management guidance.

Suggested Change:
Add a suggested action that states “Consider defining and communicating to key stakeholders whether any potential use cases (or categories of use cases) would be unacceptable, or would be treated as “high risk” or another category for which your organization would provide specific risk management guidance or other risk mitigation measures.”

We also recommend listing our actionable-guidance paper (Barrett et al. 2022) as an informative reference for Manage 1.3 documentation and communication of whether potential uses would be unacceptable.

References:
Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Manage 4.1

Response/Comment:
One component of Manage 4.1 is capturing and evaluating input from users and enabling appeals. However, providing people with the ability to opt out of the use of the AI system when possible is not currently mentioned. This is central to one of the principles included in the White House AI Bill of Rights and is an expectation that many users will have.

Suggested Change:
Add a suggested action that states “Establish mechanisms for people to have specific and clear opportunities to opt out of the use of the AI system whenever possible.”

Manage 4.3

Response/Comment:
Manage 4.3 importantly includes guidance on how to share information about errors, incidents, and negative impacts with users and impacted parties, but it does not currently mention providing recourse or redress for those harms. This should be part of the consideration organizations make when they expect that their AI system could cause harm to people.

Suggested Change:
Add a Suggested Action that states “Establish mechanisms to provide recourse or redress to people who experience negative impacts related to the use of the AI system.”

Our comments on the NIST AI RMF Roadmap

Response/Comment:
We broadly agree with the current AI RMF Roadmap, including on expanded TEVV efforts, profiles, and guidance related to explainability and interpretability.

Regarding AI RMF profiles, we are currently leading an effort to create an AI RMF profile for increasingly multi-purpose or general-purpose AI, such as cutting-edge large language models. We aim to publish Version 1.0 by the end of 2023, preceded by draft versions for feedback. We have a brief project overview and call for stakeholders to provide input and feedback at: https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/.

For the Roadmap, we recommend adding the following: Characterization and measurement of AI system objectives mis-specification risk. DeepMind Safety Research (DSR 2018) includes “specification” as one of three key types of safety characteristics; the other two are robustness and assurance, which have relatively more associated guidance in the AI RMF 1.0. Specification represents the process of specifying AI system goals, objectives, or proxy metrics so that the system’s behavior aligns with the designer’s or deployer’s intentions. Specification problems occur when a system meets its literal goals, but also causes harms or has other behaviors that the designer or deployer did not anticipate or intend. Rudner and Toner (2021) provide brief examples, such as social-media content recommendation machine-learning algorithms that learn to optimize user-engagement metrics by serving users with extremist content or disinformation. An active area of AI safety research aims to develop methods for aligning AI systems during model training, and for validation and verification of AI system objectives alignment. These methods will be increasingly important as AI systems grow in capability.

Suggested Change:
Add the following to the AI RMF Roadmap: Characterization and measurement of AI system objectives mis-specification risk.

References:
DSR (2018) Building safe artificial intelligence: specification, robustness, and assurance. DeepMind Safety Research, https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1

Tim G.J. Rudner and Helen Toner (2021) Key Concepts in AI Safety: Specification in Machine Learning. CSET, https://cset.georgetown.edu/wp-content/uploads/Key-Concepts-in-AI-Safety-Specification-in-Machine-Learning.pdf

Our comments on the NIST AI RMF Crosswalks

Response/Comment:
The currently available crosswalks to the AI RMF, such as to ISO/IEC FDIS 23894, seem quite useful. However, their current presentation design does not necessarily include instances where another standard or framework has topics that the AI RMF does not. This can give the impression that the AI RMF has no gaps in comparison to other standards and frameworks.

Suggested Change:
Modify the design of the AI RMF crosswalks to make it clearer whether and how another standard or framework has topics that the AI RMF does not. For example, one of the Crosswalks is, “An illustration of how NIST AI RMF trustworthiness characteristics relate to the OECD Recommendation on AI, Proposed EU AI Act, Executive Order 13960, and Blueprint for an AI Bill of Rights.” It would be helpful to see that one of the principles from the Blueprint for an AI Bill of Rights (“Human Alternatives, Consideration, and Fallback,” which states, “You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter”) does not neatly map onto any of the NIST AI RMF trustworthiness characteristics.