To: aiframework
Subject: NIOSH Comments on NIST AI Risk Management Framework
Date: Friday, April 29, 2022 2:53:52 PM

Good afternoon!

Thank you for the opportunity to review and provide comments for your draft AI Risk Management Framework. I am glad to discuss these comments or to assist in the editing process so they are adequately incorporated into the final product.

Comments on "Question 1. Whether the AI RMF appropriately covers and addresses AI risks, including with the right level of specificity for various use cases", on "4. Whether the functions, categories, and subcategories are complete, appropriate, and clearly stated" and on "7. What might be missing from the AI RMF".

General response:
This document would benefit from recognizing explicitly unique risks that AI implementation in the workplace poses to workers. These individuals may be exposed or have to respond to the outputs of the AI system for extended periods of time and as a condition of employment (without the means to avoid or remove themselves from the negative impacts of the system).

Specific responses:
Page 4, Figure 1, Where do workers fit into this model? I recommend Operators and Evaluators, but there may need to be an adjustment in definition to this pool of individuals in the text below.
Page 5, line 21, recommend adding "or series of events".
page 6, Figure 2, left hand box "Harm to people":

-- Workers represent a unique category of people which can be negatively impacted by the deployment of AI systems. The harm to workers could differ from the harm to general population. For example, worker health and safety could be negatively impacted by AI systems (e.g. use of AI systems to make employee termination decisions). Therefore, I would suggest creating a separate bullet to recognize "workers" as a separate category of people uniquely affected by AI systems. For example, such a bullet could read: "Worker harm: Workers are negatively affected (e.g. psychological harm) by an AI system deployed in the workplace".

-- Additionally, harm could include health (think impacts of fatigue or inability to use the restroom) and should include psychological health. Therefore, recommend replacing "physical safety" with "health, safety, and well-being"

page 6, Figure 2, left hand box "Harm to people", bullet "Group/Community harm": Foreseeable harm of AI to groups and communities is not limited to discrimination. I suggest broadening this box and rephrase as follows "A class or group of people is negatively affected (e.g. discriminated against) as a result of an AI system".
Page 12, 5.2.4, Encourage a broader definition of "safety", inclusive of potential impact on health (fatigue, alertness) and well-being (mental health). Practical approaches should include designing the system or tool with the human in mind.
page 12, 5.3 Guiding Principles:

-- Suggest adding to 5.3.2, Accountability: There should be a connection between accountability and responsibility. Human control of these systems should be a grounding principle and those who are accountable, should also be responsible for maintaining control of these systems.

--Suggest adding a new subsection 5.3.4 "Effectiveness" as follows: "Technology should be used to improve productivity or working conditions and should not be used haphazardly. While the improper use of a particular AI system may not directly cause harm, it may ultimately impact trust in other AI-based systems. Therefore, AI should be implemented only if it is the right tool to address the problem/concern." This text is based on a NIOSH blog on AI (https://blogs.cdc.gov/niosh-science-blog/2021/05/24/ai-future-of-work/). Line 35 on page 12, Figure 3 on page 8 and Figure 4 on page 9 would need to be adjusted accordingly as well.

page 16, Table 1, row 4, right column:

-- Suggest adding "workers" as a very distinct category of affected people in two cells as follows "Potential business and societal (positive or adverse) impacts of technical and socio-technical characteristics for potential users, workers, the organizations, or society as a whole are understood" and "Benefits of the AI system outweigh the risks, and risks can be assessed and managed. Ideally, this evaluation should be conducted by an independent third party or by experts who did not serve as front-line developers for the system, and who consults experts, workers, stakeholders, and impacted communities."

-- Encourage in this section (and others as appropriate) that investigations and public reporting of adverse outcomes use a systems approach (much like NTSB and aircraft accidents).

Comment on "6. Whether the AI RMF is in alignment with existing practices, and broader risk management practices".

Specific response:

page 5, lines 24-25: Sentence "According to the International Organization for Standardization (Guide 73:2009; IEC/ISO 31010), certain risks can be positive" is confusing as it mixes two ISO definitions of risk: "effect of uncertainty" and "combination of the probability of occurrence of harm and the severity of that harm". While the first "risk" can be both positive and negative, the second "risk" can be only negative. In 2021 ISO recognized the substantial consequences of the confusion resulting from the existence of the two definitions of "risk" and their inconsistent use and created the ISO/IEC Joint Task Force on the Concept of Risk and Associated Terms via TMB Resolution 81/2021.

This NIST document defines and uses (with the exception of the quoted above sentence) "risk" along the lines of the second ISO definition, while the quoted sentence uses the first definition. I suggest removing this sentence as it does not add value, but rather creates confusion. Alternatively, please elaborate further on the existence of two definitions and clearly state which definition of "risk" is used in which occurrence in the document.

v/r,

Jay Vietas

Jay A. Vietas, PhD, CIH, CSP

Chief, Emerging Technologies Branch
National Institute for Occupational Safety & Health