From: Mica Endsley
Sent: Thursday, April 28, 2022 4:38 PM
To: aiframework
Subject: Comments on AI Framework

Thank you for the opportunity to review the NIST AI Framework. As a researcher in the field of human-AI interaction field for over 35 years, former Chief Scientist of the Air Force, and recent Chair of the National Academy of Sciences Panel on Human-AI Teaming, I appreciate the degree to which the framework address issues associated with how the AI can affect human performance and decision making (e.g. the socio-cultural aspects) as well as the reliability and robustness of the AI itself.

Overall, it covers many relevant areas and is tackling an area of significant national importance. A few comments are provided below, specifically with respect to the importance of developing AI systems that will effectively in conjunction with human decision makers charged with overseeing the AI, who will have ultimate responsibility and accountability for decisions and actions associated with the AI. I hope you are able to take these issues into consideration as you further develop your framework to address this important topic.

sincerely,

Mica R. Endsley, PhD
President, SA Technologies

(1) Include considerations provided the recent in National Academies of Sciences report on Human-AI Teaming:

https://nap.nationalacademies.org/catalog/26355/human-ai-teaming-state-of-the-art-and-research-needs

This document provides a detailed discussion of the technical issues associated with AI trust, bias, situation awareness of AI systems, AI transparency and explainability, AI coordination/communication, and models and metrics for human-AI integration.

(2) Include consideration of Human Readiness Levels (HRL) – which quantifies risks associated with integration of human with the system, across development life-cycle, including a consideration of human-system prototyping and testing with relevant populations. See ANSI/HFES Standard 400-2021. This will significantly help with risk mitigation in terms of the socio-technical aspects of the AI.

https://www.hfes.org/Publications/Technical-Standards#ANSI

(3) Other specific comments.

Figure 2 – additional harms: Effects on skills of labor force (e.g. degradation of manual skills).

Figure 4 – Socio-technical aspects of human-AI interaction go far beyond "explainability" and bias. (See NAS report).

1. AI, like other forms of automation can negatively affect human situation awareness and ability to oversee/monitor the AI (e.g. low vigilance and lack of active engagement puts people out-of-loop reducing their ability to understand what the system in doing). (Endsley, 2017).
2. There is a significant issue associated with display transparency for presenting SA information required to oversee and interact with the AI in real-time. See (NAS, 2021)
3. In addition, there needs to be a consideration of meaningful control (ref), which in addition to this need for display transparency and situation awareness of AI actions, also includes mechanisms and training that would allow people to be able to be effective at actually intervening and correcting for problems. (Boardman and Butcher, 2019).

Page 10. Robustness needs to include the ability of the AI model to handle a wide variety of situations and cases that it can be expected to encounter in a domain.

Page 10. Resilience also needs to be considered from the standpoint of the AI-human interaction. It is a joint effect.

Page 11 – Interpretability: (This is generally covered in the literature as display transparency. Some key features that are critical for human understanding of AI systems include: (1) what the AI thinks is happening – its perceptions of relevant world information, interpretation of data, projections, and level of confidence in that information, and (2) information on the AI itself – its state or mode, ability to perform tasks (capabilities/limitations) in the present context, its task status and performance, its projected actions and its impact on others, and the sufficiency of its situational representation. (NAS, 2021)

Page 12 – Bias – while the biases listed are very relevant, a broader definition of bias is that the system can generate recommendations or decisions (based on its rules or training sets) that are not generalizable to a broader class of situations it may be applied to. Some examples of this have been racial or gender bias, but it also can be biased in many different (and difficult to detect) ways. (NAS, 2021).

References:

ANSI/HFES. (2021). Standard 400 - Human Readiness Level Scale in the System Development Process.

Boardman, M., & Butcher, F. (2019). An Exploration of Maintaining Human Control in Ai Enabled Systems and the Challenges of Achieving it: NATO: Technical report

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. Human Factors, 59(1), 5-27.

National Academies of Sciences Engineering and Medicine. (2021). Human-AI teaming: State-of-the-art and re needs. Washington, DC: The National Academies Presss.