August 19, 2021

National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

Re: NIST Artificial Intelligence Risk Management Framework (NIST-2021-0004)

On behalf of Buoy Health, Inc. ("Buoy"), we are pleased to submit a comment in response to the National Institute of Standards and Technology ("NIST") RFI for *Artificial Intelligence Risk Management Framework (NIST-2021-0004)*.

Enclosed is the following:

- Comment re: Artificial Intelligence Risk Management Framework

The contact for this comment is Cory Lamz, Esq., Vice President, Legal & Data Privacy Officer, legal@buoyhealth.com.

Buoy appreciates the opportunity to submit this comment.

Warmly,

Buoy Health, Inc.

Darin Baumgartel, PhD
Andrew Dumit
Cory Lamz, Esq., CIPP/US/E
Kun-Hua Tu, PhD

Introduction

Buoy Health, Inc. ("Buoy," "we," "our") leverages artificial intelligence throughout our business. Given the accelerated pace of the advancement of AI solutions for healthcare and the potential value for leveraging data for public health, we readily acknowledge the importance of establishing a risk management framework related to artificial intelligence.

Buoy's AI Health Assistant service asks end users questions related to their symptoms during a short, conversational exchange. Based on these answers, our service provides relevant medical information that empowers end users to self-diagnose and take further action, when necessary. The engine that determines what medical information is provided to end users based on their answers, as well as what next steps are relevant, is powered by AI.

Our comment addresses topics 1-4, 6, and 9, as numbered in the NIST RFI. On the following page, the topics are identified in blue text, followed by our commentary. Our comments are rooted in our team's practical and academic expertise with respect to Buoy's AI, as well as AI more broadly. Our hope is that, by contributing to this important conversation from a place of expertise and operational experience with respect to Buoy's own existing risk management program, we can assist NIST in its aim to develop a standardized risk management framework that improves the management of artificial intelligence – for AI system owners, users, data subjects, and beyond – not only in the healthcare space, but more broadly, too.

Thank you for the opportunity to contribute to this important conversation.

Buoy’s Comment

1. The greatest challenges in improving how AI actors manage AI-related risks – where “manage” means identify, assess, prioritize, respond to, or communicate those risks;

- AI has less rigid behavior compared to traditional software, and, because of AI's extrapolative nature, the risk management of AI should be treated differently. Where software’s risks can be addressed (mitigated, remediated, etc.) at all stages of the software development life cycle, risks associated with AI should be identified and itemized up front, before data processing generation and feature engineering, since these steps can also provide sources of risk, such as risk due to bias in the underlying training data. AI risk identification can also be more difficult than in traditional software engineering, because software is often built to accommodate a known set of user actions, whereas AI systems will, with greater certainty, have to encounter inputs that deviate from the content of the training set. Additionally, since AI often requires more time and iteration than discrete software engineering, risk identification may also need to be an iterative process. Further, any AI outputs may be distributed far more disparately than software – once the AI output is out of the box, it may be difficult to put back in, so to speak.

- Risk assessment (and therefore management) of AI should be reviewed on a continual basis. Such assessment must be completed routinely over time as the AI, and the purpose of the AI, evolves over time.

- Different audiences may have different levels of risk tolerance. Risk management must be customarily prioritized based on a specific audience’s interests. The same must be true for the risk management framework: such a framework must be tailored to the level of risk tolerance held by the audience(s) associated with the AI. Audiences may include, among others, the user of the AI, the organization sponsoring the AI, and any data subjects of the AI.

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

- In addition to the characteristics identified by NIST, the following additional important characteristics, and related questions, should be considered: audience and purpose.

    - **Audience**: in managing risk, we must assess which groups are involved with the AI, i.e., the audience. Groups may include, among others, the user of the AI, the system owner sponsoring the AI, and any data subjects of the AI. To whom are risks communicated? For each of these parties, what risk do they bear? How do you properly manage their risk? For example, the system owner bears the risk of the AI producing outputs inconsistent with the established **purpose**, so the system owner should take steps to manage that risk and communicate the steps to manage the risk to the appropriate audiences(s) (**explainability, transparency**). In another example, the system owner also bears the risk of the AI not being sufficiently **transparent**, so the system owner should take steps to manage that risk and communicate the steps to manage the risk to the appropriate audience(s) (which goes to **explainability**, **transparency**). In managing such risk, we also must assess whether the training data, evaluation criteria and/or change history of the AI are made transparent to audiences in ways that are appropriately tailored to the audience so as to facilitate their understanding (**explainability**, **transparency**).

  - **Purpose**: in managing risk, we must assess whether the inputs and outputs of the system are intended or not. Any inputs and outputs inconsistent with the purpose, i.e., any unintended or unexpected usage, would bear higher risk. Any unintended usage of an AI system, or unintended results, constitutes a new source of risk which may not have been assessed unless the **purpose** of such usage was deliberately considered in the risk assessment. For example, as we have seen in market, AI used in hiring has had the unintended consequence of introducing or perpetuating certain biases in the hiring process. In another example, AI systems that generate human-like text could generate harmful misinformation, whether intentionally or unintentionally, if the AI system owner did not adhere to the **purpose** of the system, appropriately manage risks, and have appropriate safeguards in place to mitigate that risk. In both cases, balancing the purpose, established prior to processing, against the outputs would enable different **audiences** to manage the risks of the AI (in these cases, the AI's unintended consequences).

- It is worth noting that, although we advocate that explainability is a critical aspect of the risk management framework, we acknowledge that explainability must be kept in check. Explainability should keep in mind audience and content. Additional risk may be introduced in how materials are communicated and to which audience. Further, and per the Wharton risk management framework,[1] too much explainability creates re-creation (i.e., intellectual property) risk.

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: transparency, fairness, and accountability;

- In addition to the principles identified by NIST, **communicability** should be considered. A lacking or incorrect understanding of the AI by audiences would increase the likelihood and/or severity of risk. Further, communicability would serve as the foundation to these other principles of transparency, fairness, and accountability – i.e., effective communication fosters the establishment of each of these principles. This is

[1] "Artificial Intelligence Risk & Governance," Artificial Intelligence/Machine Learning Risk & Security Working Group (AIRS), The Wharton School, University of Pennsylvania, https://ai.wharton.upenn.edu/artificial-intelligence-risk-governance/ (last accessed Aug. 12, 2021).

evident in the example of a bank loan. If an applicant is rejected, rejection should include communication on how the system used to assess the applicant works, what considerations were made, and what the result means. Simply, we would propose that the new standard, communicability, should indicate that the AI can appropriately and sufficiently explain both the system and its results to relevant audiences.

4. The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management – including, but not limited to, the management of risks related to cybersecurity, privacy, and safety;

- There are several straightforward mechanisms to incorporate AI risk management into enterprise risk management. This may include, among other things:
  - **Logging** of inputs and outputs
    - **Inputs**: a digital trail of all inputs that went into a decision, so that outputs can be audited and understood. This is similar to software best practices for logging, but more comprehensive for AI systems as they can be dynamic in nature. Inputs should extend to the training data used to train a model in addition to the inputs to a particular decision from that model.
    - **Outputs**: a digital trail of all outputs and how those outputs have been extrapolated across the enterprise. This supports the importance of the characteristic proposed above, purpose – e.g., if healthcare data is collected and processed for a healthcare-related purpose but then used later for employment decisions, such would be inconsistent with the initial purpose and create significant risk.
  - **Monitoring**, including metrics related to risk such as use engagement and following of recommendations, as well as model performance metrics which may degrade over time, indicating newly evolving risks from inappropriate predictions or results.
  - **Risk registry** – A risk assessment inventory that includes both the risk level (e.g. is this low risk like an algorithm that recommends television shows; medium risk like an algorithm that sorts a social media feed; or high risk like self-driving automobiles?), as well as risk likelihood (e.g. is your risk unlikely like a skilled malicious actor reverse-engineering your model, or highly likely like having one or more customers unhappy with your recommendation system?).
  - **Review of new applications** of a model – this is similar to purpose, as no model should be generically repurposed. Legal/security/risk teams should sign off on new model applications.

6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;

- All standards should be made consistent across jurisdictions to every extent possible. Such consistency would facilitate the support of developing and managing AI at scale. At the very least, inconsistent standards would freeze or weaken innovation. At most, inconsistent standards could lead to harm of a data subject.

9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, "AI RMF Development and Attributes");

- We agree with the appropriateness of the attributes that NIST has developed for the AI Risk Management Framework. With respect to Attribute 8, "Be a living document," we would suggest that NIST better contextualize expectations on what "living" means here. How is it living? Who modifies it? How often, and with what input?

- We also pose the following question to NIST with respect to these attributes: how might NIST **incentivize** organizations to subscribe to the AI Risk Management Framework? As such a framework is not mandated by law, how might NIST create buy-in? One suggestion would be for NIST or a third party to issue a certification, to make it easy to identify those organizations that subscribe to the AI Risk Management Framework. Otherwise, there is little to no incentive, at least in the private market, to bear the operational burden of maintaining a living document year-over-year, much less a program to support the AI Risk Management Framework as a whole.