babl

BABL AI Inc.
The Algorithmic Bias Lab
630 Fairchild Street
Iowa City, Iowa 52245
https://babl.ai

From: **Shea Brown, Ph.D.**
Chief Executive Officer
BABL AI Inc.

To: **National Institute of Standards and Technology**
U.S. Department of Commerce
AIframework@nist.gov

Re: **Public Comments on Second Draft of AI Risk Management Framework**

Sept. 29, 2022

To Whom It May Concern:

On behalf of the team at BABL AI, I welcome the opportunity to provide public comments on the second draft of NIST AI Risk Management Framework (RMF) and its companion Playbook. NIST RMF and the playbook is a welcome attempt to establish best practices for organizations to mitigate potential harms of their AI systems, while also setting a standard that external parties such as auditors, regulatory authorities, and civil society can reasonably expect from the organizations in regard to risk management. As a company that audits and assesses algorithms for ethical risks, effective governance, and bias, BABL AI strongly believes that the spirit of this framework furthers our mission to promote and protect human flourishing in the age of AI.

With NIST seeking public comments on the second draft of the RMF and its companion playbook for further development, we recommend the following:

Clarifications – We encourage the authors to clarify the ambiguities regarding:

1. **Status quo**

The framework recommends comparison to the "status quo" to determine whether AI use is appropriate [*sec. 6.2 – Map*]. We recommend NIST to make clear the language on what status quo refers to. Moreover, we urge the authors to emphasize and expand more on this notion of the status quo and why it is important to adopt such a comparative approach.

While we understand that what status quo looks like varies between AI applications and among contexts, it makes sense for NIST to lay out conditions to justify what counts as status quo for purposes of impact and risk assessments. For example, status quo might simply be a previous instantiation of the system which was not AI-enabled. In our own work, we have found that assessing the risks of an AI-free alternative as the status quo helps to reliably establish a useful and robust operational baseline to gauge the extent of risk introduced by the AI.[1] However, in many cases there exists no obvious equivalent. Giving clear guidance on conditions that "status quo" needs to meet would be helpful – i.e., how would an organization justify using time X or time Y as status quo.

2. TEVV

TEVV is cited throughout the framework in the form of tasks that "assess the system relative to technical, societal, legal, and ethical standards or norms, as well as monitor and assess risks of emergent properties." However, given that TEVV has historically been used for products whose characteristics are relatively fixed when they are provided for deployment, it would be important to suggest how the process can be conducted throughout the life cycle of AI-enabled systems.

We recommend giving more details and, if possible, concrete examples, to illustrate what TEVV might entail for AI systems which are dynamic and constantly changing. One way to showcase this specificity is in *sec. 4 – AI Risks and Trustworthiness*, where the authors can show how TEVV contributes to, for example, "Valid and Reliable" as a characteristic of trustworthy AI. In addition, the companion Playbook should provide further details with actionable resources for TEVV tasks for all stages of the AI life cycle.

3. Explainability vs. interpretability

While we recognize the importance in distinguishing the two terms [*sec. 4.6 – Explainable and Interpretable*], we wonder if the usage of these terms might be confusing to users who are not familiar with the debates within the academic communities. We recommend one of these two approaches to clarify the language regarding terminology use:

- **Approach 1:** Using only one term, but clarifying how it is applied to different contexts

We recommend using "explainability" as an umbrella term given its prominence in industry, and providing more discussion and emphasis on the usability of this term as it is applied to different AI actors – e.g., what should be explained to designers, AI operators, or users – as well as contexts – e.g., what should be explained at the development phase vs. the deployment phase.

- **Approach 2:** Using terms in the framework different from those in the playbook

Alternatively, we recommend using scholarly terms in the framework while using plain and simple terms in the playbook. This approach allows the playbook to be more accessible to non-technical users such as compliance officers in SMEs, while retaining the academic tone for the framework document.

[1] Ali Hasan et al., "Algorithmic Bias and Risk Assessments: Lessons from Practice," *Digital Society* 1, no. 1 (August 19, 2022), https://doi.org/10.1007/s44206-022-00017-z.

4. **Design, development, deployment, use, and evaluation**

We recommend reinforcing the consistent use of these phases throughout the document. Furthermore, procurement is an important phase for many organizations, and should therefore be also emphasized in the document.

New considerations – In addition to the clarifications above, we encourage the authors to consider including the following:

1. **Use cases and actionable resources in the playbook**

Overall, we believe that the current version of the playbook is not sufficiently distinct from the framework. We expected the playbook to provide more actionable resources such as questionnaires, illustrative case studies, an example risk profile, or guidance on building a risk profile. All of these resources would be readily leveraged by people at various levels of an organization (e.g., compliance officers, AI ethics/responsible AI leads, product/technical managers) to start conducting risk assessment for their AI systems.

2. **Discussion on trade-offs**

We urge the authors to include some discussion on the trade-offs, for example, between documenting issues in terms of risks vs. impacts [*sec. 5 – Effectiveness of the AI RMF*, *sec. 6.1 – Govern*], or between trustworthy AI characteristics (e.g., explainability vs. accuracy) [*sec. 4 – AI Risks and Trustworthiness*]. It would also be worthwhile to show appreciation for the trade-offs in some of these cases.

3. **More discussion on privacy and global perspectives**

We believe the discussion on privacy and global perspectives is too broad and general, and would benefit from some more in-depth discussion. For privacy, some discussion on the contextual nature of privacy[2] would be appreciated. For global considerations, giving some examples where “perceptions of fairness differ among cultures” [*sec. 4.3 – Fair, and Bias Is Managed*] would be useful for organizations who deploy AI systems in non-U.S. countries.

I would like to thank NIST for providing us the opportunity to comment on the second draft of the AI RMF, and we would be happy to provide further clarification on any of the above questions.

Contact

Shea Brown, Ph.D., CEO & Founder

[2] Helen Fay Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (2009; repr., Stanford, CA: Stanford University Press, 2010).