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.
contact@babl.ai

To: **National Institute of Standards and Technology (NIST)**
100 Bureau Drive
Mail Stop 8900
Gaithersburg, MD 20899–8900

Re: **AI E.O. RFI Comments**

February 2nd, 2024

To Whom It May Concern:

On behalf of the team at BABL AI, I welcome the opportunity to provide public comments in response to the request for information (RFI) related to NIST's assignments under sections 4.1, 4.5, and 11 of the Executive Order Concerning Artificial Intelligence (E.O. 14110). As a firm that audits algorithms for ethical risk, bias, and effective governance, we believe that NIST's activities under the EO are critical for minimizing the risks that these systems can pose, and align with our mission to promote and protect human flourishing in the age of AI.

The team at BABL AI has conducted algorithm audits, bias testing, risk and impact assessments, and is committed to working at the forefront of research[1], policy[2], and education[3] in the fields of AI audit, assurance, and governance. Below we offer our thoughts on how NIST's activities under the EO can further effective risk management within industry and the federal government. We know NIST will receive a wide range of valuable insights from the community, so we will limit our comments to a few key areas where we think our experience auditing the responsible AI practices of companies has given us unique insights.

From the RFI: —----------
"The types of professions, skills, and disciplinary expertise organizations need to effectively govern generative AI, and what roles individuals bringing such knowledge could serve"

[1] https://journals.sagepub.com/doi/full/10.1177/2053951720983865 & https://link.springer.com/article/10.1007/s44206-022-00017-z & https://arxiv.org/abs/2401.14908
[2] https://github.com/algorithmicbiaslab/public-resources/tree/main/policy
[3] https://courses.babl.ai/p/ai-and-algorithm-auditor-certification

"Roles that can or should be played by different AI actors for managing risks and harms of generative AI… "

"Current techniques and implementations… Human rights impact assessments, ethical assessments, and other tools for identifying impacts of generative AI systems and mitigations for negative impacts;"

"Measurable and repeatable mechanisms to assess or verify the effectiveness of such techniques and implementations."
—----------

From our experience, the most important and difficult component of risk management for AI systems (including generative AI) is the **risk assessment process**. In evaluating risk assessments conducted by multidisciplinary teams that include data scientists, lawyers, risk professionals, and subject matter experts, we've noticed common challenges that these teams face, including:

1. An overfocus on risks to internal stakeholders (like the company, employees, or clients as opposed to more vulnerable external stakeholders), likely due to a reflexive adherence to enterprise risk management principles.
2. General difficulty in connecting risks to the technical (or sociotechnical) sources of those risks, and vice versa.
3. A tendency to over-rely on past or known examples of harm as opposed to unencountered yet foreseeable risks.
4. An overfocus on compliance vs. actual risk detection and mitigation.
5. Not connecting the chosen mitigation measures to the actual causes of risk (i.e., not connecting how a control might limit the likelihood, scale, or severity of a risk)
6. Siloing of expert input, e.g., where technical team members weigh in on technical metrics, legal on compliance issues, etc., without true cross-disciplinary reflection.

These issues are complicated when dealing with generative AI, especially 2, 3, and 5, as the complexity of the outputs leads to a kind of "risk assessment paralysis" when it's unclear how to connect the behavior of the system to potential risk. This leads to a defaulting to "known" controls that are not necessarily connected to any real risk of harm or value destruction, like generic benchmarking, ineffective "human-in-the-loop" fixes, or un-considered checklists.

We see several ways in which NIST can contribute to addressing some of these challenges:

A. Promote the development and growth of **algorithmic (or AI) risk assessment as a practice**, which utilizes unique techniques, skillsets, and perspectives that need to be researched, developed, tested, and mastered by practitioners.
B. Encourage the close **integration of algorithmic (or AI) risk assessment** as a practice into the proposed **"measurement science" of AI safety**, especially for generative AI. The siloing of these two emerging practices at this early stage of AI safety research

could potentially exacerbate some of the challenges outlined above and devolve into techno-solutionism and unreflective benchmarking.

C. Emphasize in any supporting companions to the AI RMF the **importance of specific training** and specialization for **practitioners in both AI risk assessment** and **technical testing** (including **red team** members) of AI systems. As we observed above, even multidisciplinary teams of capable practitioners will not necessarily converge on proper risk management without some level of guidance, experience, or training.
D. To further points A-C, we encourage NIST to **start these research and standardization efforts in narrow domain-specific use cases** where the connection between risks and technical metrics can be more closely articulated, then strive to test and generalize the *methodology* as opposed to specific benchmarks. These efforts could engage AI Safety Institute Coalition members in specific domains to maximize their contributions and encourage the co-development of risk assessment and measurement practices as described in B.

I would like to thank NIST for providing us the opportunity to comment on their strategy for fulfilling the obligation of E.O. 14110, and we would be happy to provide further clarification on any of the above comments.

Contact

Shea Brown, Ph.D.
CEO & Founder
BABL AI Inc.
contact@babl.ai