AI Risk Management Framework: Second Draft

On the Importance of Simpson's Paradox and Systems Theory in AI

Neo4j Inc Response
By Kara Doriani O'Shee

To our fellow citizens, leaders, and whom it may concern,

The National Institute of Standards and Technology (NIST) has requested comments on its second draft Artificial Intelligence Risk Management Framework (AI RMF). Neo4j welcomes the opportunity to support the agency in these efforts. We feel that this draft version of the AI RMF is on the right track, especially NIST's work in defining the key characteristics of trustworthy AI. To further advance this draft, we offer three recommendations:

- Include guidance on the importance of measuring accuracy for subgroups of data, in addition to aggregate measures
- Use a systems approach to AI safety to account for the risk arising from interactions between AI actors, the user base, and the system
- Adopt terms from program evaluation to enhance the language and usability of the function tables in "Core and Profiles"

The comments that follow expand on these recommendations in the order of their appearance in the draft AI RMF.

Trustworthy and Responsible AI

We applaud NIST's new framing of AI risk management, which defines the key characteristics of trustworthy AI and recognizes that they will play out differently in various contexts. Explaining these key characteristics (***valid and reliable, safe, fair and bias is managed, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced***) helps to create a common taxonomy around the ethical principles to be used in the design of AI systems, which we had suggested in our previous comments. We argued that ethical principles should form the foundation of AI risk management because even highly technical decisions are influenced by values and assumptions.

By defining trustworthy AI in socio-technical terms, NIST provides a human-centered understanding of AI systems and avoids the limitations of a strict technocentric view. We believe that this framing will foster more effective risk management because it considers

AI risk as a function of human values in a way that differs from other types of risk management. In sum, we concur with this statement:

"Trustworthiness is greater than the sum of its parts. Ultimately, it is a social concept, and the characteristics listed in any single guidance document will be more or less important in any given situation to establish trustworthiness" (AI Risk Management Framework: Second Draft, page 11).

Valid and Reliable

The draft RMF does an excellent job of defining the "valid and reliable" characteristics of trustworthy AI, and we commend the emphasis on monitoring reliability throughout the life cycle. Future conditions may not match the conditions of the training data, which underscores the necessity of continuous monitoring.

To better account for risk in this arena, Neo4j recommends adding discussion around the value of analyzing accuracy for data subgroups. **Even when models appear to be performing well on general accuracy metrics, there is no guarantee of equally strong performance for all groups.** Aggregate measures of accuracy can easily hide the poor performance of data subgroups, creating risk for the individuals belonging to these groups.

A statistical phenomenon known as Simpson's Paradox suggests the dangers of relying on aggregate measures alone. The paradox occurs when a trend appears for two separate groups but disappears or reverses when the groups are combined. **Consequently, it is possible to draw opposite conclusions from the same data, depending on how it is segmented**. To address the paradox, we need to understand the context of data, as the factors producing disparate results can only be found outside the dataset. How was the data collected? What confounding factors could cause a trend to appear for one group, but not the entire population? The art of data science is to see beyond the data and apply real-world knowledge to make meaningful interpretations.

At Neo4j, our expertise in graph technologies affords us unique insight into the relevance of context in AI risk management. Because graph databases store data along with the relationships between data, they can provide context to inform design decisions in data collection, cleaning, and ML techniques. This is especially the case when graphs are used to track the provenance of data, as both data and metadata (information about data sourcing) can be stored in a knowledge graph. A knowledge graph becomes a repository not only for the training data, but also information about how that data was collected, changed, and analyzed. **Without a way to check the history of data, it can be impossible to answer questions about the larger context and investigate problems that arise, like Simpson's Paradox**.

To ensure the performance of subgroups in a dataset, we need to be able to segment data into these groups. Within a graph database, a single data point can be given multiple labels, **facilitating data segmentation along multiple demographic and behavioral lines**. Improving accuracy for groups in the data can help organizations build more inclusive AI systems that perform well for diverse populations.

Multiple labeling also permits side-by-side comparison of data from various sources **so that it's possible to detect when different types of data have been merged inappropriately**. For example, one dataset might represent answers to a certain question, while another dataset represents answers to a slightly different question. When two types of data are not equivalent (apples to oranges), they can be removed from the analysis and training data. In situations like the COVID crisis, this capacity can be critical. As reported by The New York Times, regulators had to piece together data from individual hospital systems when Omicron began to emerge. Health officials struggled to make decisions because their data systems consisted of "a big jumble of different studies and different subsets that were stitched together;" in other words, non-equivalence of data.

We hope that these insights are useful in informing the AI RMF on validity and reliability, with our general points as follows: 1) accuracy metrics should encompass analysis of subgroups as well as the entire dataset; 2) identifying the confounding factors that affect accuracy for different groups requires strong data provenance practices; and 3) comparative analysis of data types can aid in accuracy by assuring the quality and equivalence of data.

Safe

We appreciate NIST's attention to safety for stakeholders at every stage of the life cycle, including designers, deployers, and end-users. Given the complex interplay of actors and the AI system itself that affects safety over time, **we advocate a systems theory approach that admits the potential for cumulative harm from the dynamics between the user base and the system.**

Organizations must consider safety not only across the separate spheres of designer, deployer, and user, but also how they interact with one another within the AI environment. An example is AI systems that utilize ratings, which often require additional oversight or design features to guard against discrimination and other harm.

Core and Profiles

NIST has organized AI risk management activities into four major functions: ***govern, map, measure, and manage*** risks. The functions can be performed in any order and at any time in the AI lifecycle to achieve the outcomes. While we understand that NIST will provide more guidance in the AI RMF Playbook, we feel that the usability of the tables could be improved with terms from program evaluation: objectives, expected outcome/result, activities, outputs, etc.

In this taxonomy, “Category” could be “Expected Outcome” or “Result,” while “Subcategory” would become “Activity,” a process objective. Each “Activity” text would require rephrasing and should begin with a verb to convey its purpose as a process objective. For example, Govern 1.2 would be changed to “Integrate the characteristics of trustworthy AI into organizational policies, processes, and procedures.” For some activities, reframing may necessitate further breakdown into discrete units so that each activity represents a single process indicator.

This will have the benefit of clarifying how these categories relate to one another and what they accomplish, making it easier for organizations to consider how they may best be implemented within a governance structure.

Conclusion

Context is key for building AI systems that are inclusive, safe, and situationally accurate because all data problems must be addressed in light of their real-world conditions. Graph technologies offer state-of-the-art methods for seeing data in context, which is essential to designing, deploying, and monitoring AI systems that are trustworthy. Neo4j commends NIST for its work in developing these guidelines and looks forward to engaging in future iterations.

Please do not hesitate to contact us at government.relations@neo4j.com if we can be of further assistance.