AI Risk Management Framework: Initial Draft

Why Values Must Shape AI Design

Neo4j Inc Response

By Kara Doriani O'Shee, Neo4j

Introduction

The National Institute of Standards and Technology (NIST) has requested comments on its first draft of the Artificial Intelligence Risk Management Framework (AI RMF), intended for voluntary use in the design, deployment, and evaluation of AI systems.

Artificial intelligence (AI) is poised to advance quality of life across the globe, bringing new benefits to people, organizations, and society. At Neo4j, we see the enormous potential of AI to enhance human life when used responsibly.

In the sections that follow, we argue that ethics ought to form the foundation of AI risk management. Also known as "ethics by design," this approach makes ethics part of the process of developing AI applications, rather than an afterthought.[1]

Since AI risk comes from its use in decisions that affect human lives, risk management must align with human values based on ethical principles. Businesses should identify values appropriate to the AI at the design stage.

Incorporating values into design is superior to post-hoc intervention, after AI techniques have already caused harm. Examples of AI gone wrong[2] show the dangers of training models on data without guardrails. Because data can reflect real-world bias, AI must be anchored by ethical values to limit automated discrimination and other harms.

NIST has identified fairness, accountability, and transparency as the guiding principles of its risk management framework. We suggest using these values as the basis of AI risk management, just as ethical concepts like privacy and fairness are foundational to US law.

In recent years, ethical codes for AI have taken a similar approach, such as those from the EU High-Level Expert Group and the INEE Global Initiative on Ethics of Autonomous and Intelligent Systems.[3]

What Is Risk?

Managing risk for AI should mean protecting against human harm, on the individual and the aggregate levels.

---

[1] https://www.nature.com/articles/s42256-020-0195-0

[2] An example is Microsoft's Twitter bot, Tay, which learned from conversations with users. Within 24 hours, Tay had begun tweeting racist and sexist insults. Could this outcome have been avoided with a values approach?

[3] https://link.springer.com/article/10.1007/s11023-020-09537-4

Currently, the framework puts forward a three-class taxonomy:

- Technical characteristics
- Sociotechnical characteristics
- Guiding principles

These areas are presented as co-equal, without an explanation of the relationship between them. Businesses that use the framework may have questions about how they interface and what takes precedence. How do concepts like accuracy and reliability (*technical*) interact with the need for safety and interpretability (*sociotechnical*)? And how do the guiding principles relate to these technical and sociotechnical characteristics?

We propose centering the concept of risk on the human person. As ethics is the domain concerned with human harm, it would act as an axis to orient the three-class taxonomy. Through the guiding principles, ethics can play this role in AI risk management.

The Guiding Principles as a Foundation

The guiding principles should act as a foundation for AI risk management by influencing the ethics of technical and sociotechnical design.[4]

Since each AI application has a unique purpose, these principles will show up differently in each use case. For example, fairness can be defined procedurally (treating every person the same) or representationally (ensuring parity of protected groups). Designers should select and define a set of values based on the goal of the AI and assumptions that can be made about the input data. The tradeoffs implied by different values must be well considered and documented so that organizations have the ability to enforce them.

This flexible approach allows for a plurality of values relevant to a use case, as well as their varying dimensions and definitions. It also promotes transparency about what values mean for different use cases.

How Values Inform Technical Decisions

Since the technical realm is less obviously informed by values than the sociotechnical (*explainability, interpretability, privacy, safety, managing bias*), we focus here on how values bear on technical decisions.

Absent an awareness of how values affect technical choices, AI can have harmful consequences. Technical characteristics that appear value-neutral, like model accuracy, often involve value judgments.

Accuracy refers to how well a model captures a pattern that exists in the data. Models extrapolate from statistical patterns in the data, so it's common for them to find predictive

[4] Positioning the guiding principles as the basis of the framework aligns with its goal to frame risk using "characteristics that are aligned with trustworthy AI systems, in conjunction with contextual norms and values" (page 8). Therefore, the RMF would also achieve deeper internal consistency with this approach.

features like race or sex, which are protected under U.S. law. Accuracy metrics should be guided by values to prevent harm from models that have learned to discriminate against protected groups.

Values and Job Matching

For example, in recent years, LinkedIn realized that its job-matching algorithm favored male candidates.[5] The algorithm ranked candidates partly based on how likely they were to apply for a position. Since men tended to be more active on the platform, the system referred more men than women for jobs. The behavioral difference caused an algorithmic bias where being a woman weighed against a candidate.

Thinking through what fairness means is crucial when selecting a suitable evaluation metric for AI. In this situation, group fairness (equal representation of men and women) makes more sense than procedural fairness (treating every person the same way), which would cause a systematic exclusion of women.

With group fairness, a designer might allow for false positives over false negatives. A false positive means offering an opportunity to a woman who was not qualified; a false negative means rejecting a qualified woman candidate. While each involves a tradeoff, the decision must be made because no model is perfectly accurate. In this example, rejecting a qualified woman is more costly since base rates in the data make it less likely that a woman would be selected in the first place.

Values and the COVID-19 Pandemic

The pandemic accelerated AI adoption across healthcare, industry, and government, creating new ethical considerations. AI has helped to detect and prevent disease spread, automate diagnosis, and allocate healthcare resources, all of which require value judgments at different stages.

For example, risk modeling has been used in resource allocation. But the way we conduct risk modeling depends on how we define fairness. Is the goal to maximize the total number of people who benefit (prioritize younger, healthier populations) or to minimize loss of life (prioritize high-risk populations)? Studies warn that AI meant to benefit all patients can worsen racial and economic disparities in healthcare.[6] Defining fairness is not simply a theoretical exercise; it has much higher stakes for vulnerable groups. To focus resources on these groups, we might prioritize sensitivity over specificity, or base decisions on the upper bounds of a confidence limit rather than the median.

Another example is using AI capabilities to foster public trust during a pandemic. Transparency could be built into AI with a feature that generates reports for public accountability. Creating this feature is a technical process with sociotechnical implications, such as what data is provided to the public (*privacy*), how the algorithm uses that data for decisioning (*explainability*), and how it is made understandable (*interpretability*).

---

[5] As reported by MIT Technology Review, https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/

[6] https://academic.oup.com/jamia/article/28/1/190/5893483?login=true

Conclusion

We have sought to shed light on the often-invisible role of values in shaping AI design. At this critical juncture, we urge NIST to use the guiding principles as the basis of all AI risk management activities. In addition to fairness, accountability, and transparency, NIST might also consider principles mentioned in other parts of the RMF, such as human autonomy and dignity, and/or those identified in previous international frameworks. NIST should also advocate for model and AI monitoring, to ensure that AI efforts are continuously reviewed and improved to align with values.

A human-centric approach to AI requires understanding that ethical decision-making is not just another form of technical problem-solving. Ethics is a kind of meta layer for AI development that should influence its technical and sociotechnical aspects. By recognizing the guiding principles as its ethical frame of reference, NIST can foster transparency around the value-driven nature of design decisions and safeguard against avoidable harms.

About Neo4j

Neo4j helps people make sense of data with graph technology by revealing the connections between people, objects, systems, and other entities. Data connections add vital context for ML/AI, improving predictive accuracy and reducing dependence on signals from demographic data.

Further, by providing deep transparency across layers of data, graphs play a key role in the responsible development of AI.

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