10 September 2021

Response to Request for Information on NIST-2021-0004: Artificial Intelligence Risk Management Framework

Submitted By: Mary Kathryn (Kathy) Rondon (https://www.linkedin.com/in/kathy-rondon-7a057b21/)

Introduction

The National Institute of Standards and Technology (NIST) aims to create a Risk Management Framework that provides voluntary standards for the design, development and use of AI that minimizes risk to individuals, organizations, and society. It is likely that many respondents to this public Request for Information (RFI) will focus on the development and use of the technology itself, and this is—of course—a critical area for the Framework to address. However, as a data governance professional and thought leader with a strong interest in data ethics, my comments and suggestions focus on standards that apply to the underlying data upon which AI technology is trained.

The NIST RFI specifically notes that "While there is no objective standard for ethical values, as they are grounded in the norms and legal expectations of specific societies or cultures, it is widely agreed that AI must be designed, developed, used, and evaluated in a trustworthy and responsible manner to foster public confidence and trust. Trust is established by ensuring that AI systems are cognizant of and are built to align with core values in society, and in ways which minimize harms to individuals, groups, communities, and societies at large." While these statements are accurate, despite the lack of an "objective" standard for ethical values, this does not mean that there are not ethical standards and frameworks that can usefully be applied to the NIST AI Risk Management Framework. The Data Ethics Framework developed as part of the U.S. Federal Data Strategy Action Plan was drafted by a panel of Federal Government Data Executives, draws from other NIST frameworks, as well as U.S. law and regulation and specifically addresses the ethical use of AI. In particular, it advocates for the following ethical standards:

- Uphold Applicable Statutes
- Respect the Public, Individuals, and Communities
- Act With Honesty, Integrity, and Humility
- Hold Oneself and Others Accountable
- Promote Transparency
- Stay Informed of Developments in Data Management and Data Science

(https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf)

Conceptual Foundation

There is a statement bandying about the internet that now borders on becoming a truism: that artificial intelligence is neither artificial nor intelligent. Digging deeper into this idea, however, yields the underlying conceptual approach to the response to this RFI. That is, that AI—regardless of the topic or the sophistication of the algorithm, or the expertise of the designer, or the process by which it is developed—is based upon the use vast amounts of data (or at least statistically significant samples of vast amounts of data). And that data is what drives the "decision making" ability of AI. When insufficient attention is paid to the collection of data, the post hoc analysis of data, the documentation of contextual metadata about datasets, the quality standards and thresholds against which the data is evaluated, then trust in the results of AI application can be illusory. Therefore, in an effort to help define the "Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk," the comments and suggestions below align to the overall conceptual approach that: ***trust in the design, development, and use, of AI requires that the data upon which AI is built and trained be subject to a rigorous data governance program and that the key tenets and procedures of that data governance program be transparent to both developers of AI and their end users***.

Engendering Trust Through Data Governance

Any AI Ethical Framework or Risk Management Framework that focuses solely or primarily on the development of the AI coding and the process by which the technology is evaluated and tested is incomplete. In a very real sense, AI is only an insanely effective pattern recognition application, so AI—regardless of the sophistication of the algorithm itself--is no better than the data upon which it is predicated. Therefore, the primary focus of the AI Risk Management Framework should be on standards associated with the collection, sharing, evaluation, and use of the data upon which the AI application is based.

Recognizing that data is key to transparency, trust, and ultimate effectiveness of artificial intelligence, the Advance Technology Academic Research Center's (ATARC) AI Working group has developed a working draft of a model that assesses the transparency of an AI project, largely from a data perspective. (https://atarc.org/project/information-technology-artificial-intelligence-machine-learning-ml-model-transparency/) While an interesting conceptual model, the ATARC effort only seeks to assess; it provides no guidance on HOW to affect the change from lower to greater transparency. This is where a rigorous Data Governance approach can be leveraged. Below are some proposed standards for Data Governance as it applies to AI development projects, that—if implemented—would go a long way toward creating the trusted environment that the NIST Framework seeks to enable.

Proposed Standard: A Data Governance Plan should be part of the design phase of all AI projects.

Laser focus on data at the design phase of an AI project would enable better transparency throughout the lifecycle of AI development and use. This focus could be articulated in a Data Governance Plan that would become a standard artifact of AI projects that adhered to the NIST standard. Such a plan should address the following issues:

- A detailed description of the data to be used for the AI training and development and why such datasets were chosen;
- If certain datasets were explicitly disregarded, an explanation of why those datasets were not used;
- Methods employed in the collection of training datasets;
- Policies and methods for protection of personally identifiable information (if applicable);
- Methods for documentation of data provenance and lineage;
- Data sharing and retention policies;
- Data handling policies and procedures;
- Accountable parties for implementation of the Data Governance Plan.

Proposed Standard: When the development of AI is based upon the post hoc analysis of data collected for other purposes, specific documentation of the reasons why the data were chosen for this purpose, any issues of bias that were considered, and how bias risk was mitigated in the design of the AI project.

Almost by definition AI training datasets are post hoc. That is, the data in training sets have already been seen, reviewed, or even evaluated before the statistical analysis takes place. There is nothing inherently wrong with post hoc analysis, as long as the associated risks are explicitly noted and mitigated. Academic standards recognize that, in post hoc analysis, there is a greater risk for "cherry picking" data or data dredging to produce desired results. This risk can be multiplied exponentially if coded into an AI algorithm for which transparency into the training data and methods used for selecting it is limited or non-existent. Therefore, a detailed explanation of this process and any biases considered and mitigated should be part of the design phase, either in the project Data Governance Plan or as a stand-alone artifact. This artifact should be made available to researchers and AI end users.

Proposed Standard: Source data for the development of AI should be transparent to researchers and end users; if the data itself is protected for security or privacy reasons, then a robust metadata record describing the data should be transparent to researchers and end users.

Injecting the discipline of digital data curation into the development of AI would also increase transparency, trust, and reusability of data. Data curation is discipline that enables data discovery and retrieval, maintaining its quality, adding value, and providing for reuse over time (University of Illinois School of Library and Information Sciences). Essentially it is the documentation of contextual metadata about data assets according to standard operating procedures and consistent terms of reference. Such context can provide extensive information about the data used in training datasets, without providing unfettered access to the data itself. This discovery process with access to the detailed metadata record would be a first step in providing transparency to researchers and end users, and—at times—could be sufficient in and of itself to increase the trust in the AI product.

The curation process, qualifications of data curators, and data cataloguing tools used (if any), and the manner that such catalogue records could be publicly accessed could be included in the Data Governance Plan or drafted as a separate artifact.

Proposed Standard: Required data quality dimensions and thresholds should be articulated in the design phase of the AI project, and any departures from the quality standards noted and justified.

It is now conventional wisdom that quality data is required for maximizing the development and use of AI. But what does quality data mean? In many cases, there is no rigorous consideration about what "quality" means in the specific context of the question or topic the AI seeks to address. Data profiling and cleansing tools almost always default to consistency and completeness of datasets, but these are fairly rudimentary definitions of "quality" and are not always the overriding concerns for a given AI project. A more robust discussion of what is required of data quality standards for the particular project is needed in the design phase.

MIT Professor Richard Wang and Worcester Polytechnic Institute Professor Diane Strong published *"Beyond Accuracy: What Data Quality Means to Data Consumers,"* in 1996 (Journal of Management of Information Systems, Spring 1996), laying out 20 data quality dimensions that include other dimensions such as precision, objectivity, relevancy, timeliness, and traceability. AI projects that default to completeness and consistency without specific considerations of how different data quality dimensions apply to the particular question at hand risk using data that appears "high quality," but is of questionable quality for the specific use case. The design phase of an AI project should include specification of the quality standards considered, why those quality standards are applicable to the AI project, and the thresholds required for the data to be considered of sufficient quality to be employed. This can be part of the Data Governance Plan or drafted as a separate artifact.