Safety. Science. Transformation.™

February 2, 2024

Subject: Request for information related to NIST's assignments under sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

Submitted electronically via regulations.gov

UL Solutions is a leading, global safety science company, continuing our 130-year-old mission of working for a safer world. Our testing, inspection, certification, and software services help address safety and security challenges from across a wide breadth of industries, including energy systems, automation, building products, consumer technology, and automotive. We appreciate this opportunity to provide information to support NIST's assignments under President Biden's October 30, 2023 Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. UL Solutions is committed to leveraging our deep expertise in safety, security, and sustainability to the task of furthering safety science through artificial intelligence (AI). Below, we have provided answers to select topics from the Request for Information.

Risk and harms of generative AI, including challenges in mapping, measuring, and managing trustworthiness

As NIST works to develop a companion resource to the AI Risk Management Framework (NIST AI 100-1), we note the following fundamental challenges to building trustworthy AI systems, as well as challenges that may frustrate risk management and evaluation efforts to assess the trustworthiness of those systems.

Data Challenges

- Data selection for sampling or training may be limited to data available or data generated based on known patterns. In addition, the datasets may contain the inherent and/or inadvertent biases of the author. This manifests as conditions, demographics, statistics, and outcomes that marginalize or eliminate key representative data or key outlier conditions from the hypothesis. As safety specialists, we note that random faults tend to reside in the outlier conditions. Awareness of quantifiable unknowns and known unknowns is critical.
- Lack of ethics and accountability, or lack of inclusiveness, can impact the reliability and safety of outcomes from model execution.
- Data can be misrepresented either through improper conditioning or transformation. Syntactic or semantic abuse impact safety and reliability of results.
- Lack of data governance may result in intellectual property rights violations, retention oversight and confidentiality breach or data loss. For instance,
    - If data used by the AI model is received from third parties, intellectual property rights or contractual restrictions could be violated if data is used in a manner that exceeds permissible use.
    - If inputs or outputs to the AI model include personal information (i.e., data relating to an identified or identifiable natural person), the processing by the generative AI model might violate applicable data protection laws and regulations ("Data Privacy Laws")

Algorithm Challenges

- The algorithms powering generative AI models may be of such complexity that explainability is challenged. Actions can be taken without truly understanding decision criteria or processing logic.
- Algorithms and heuristics may not be accompanied by guidelines regarding usage (e.g., recommended restrictions on use), leading to unintended risk, adverse impact, errors and omissions, limitations, and misinformation.
- Lack of transparency in sources used to sample and train may result in unexplainable results.
- Lack of sampling guidelines, documented methodology for solvers and algorithms and lack of appropriate model refinement may result in a model falling out of date.
- For third party heuristics and machine learning (ML) algorithms, lack of transparency on data and sources and inappropriate use of an organization's data beyond approved policies or contract terms may result in intellectual property infringement, violation of contract terms, or violation of Data Privacy Laws.

Output and Policy Challenges

- Results, output or work product generated by AI/ML models may not be protectible under applicable intellectual property frameworks.
- NIST AI Risk Management Framework (RMF) is focussed on setting up an extensive, voluntary risk management framework. It mentions the different risk levels for AI systems, but it does not define them (unlike the EU AI Act). It also does not set risk management requirements (again unlike the EU AI Act). A data scientist who applies NIST AI RMF to their development cycle will have to use their own judgment as to what level of risk management is sufficient.
- Per CapAI and NIST recommendations, the Internal Review Protocol can be based on ethical standards. One of the ethical values that should be explicitly stated in an organization's charter is that they should "promote transparency". Many AI applications make use of large language models (LLMs). LLMs are notoriously opaque. That begs the question what data scientists should do when their ethical standards clash with AI methodologies. How can they make the right trade-off between ethics and providing the best AI application? Neither NIST, nor the EU AI Act nor ISO/IEC 42001, has guidelines to weigh the pros and cons of AI solutions to make an ethical decision.

The types of professions, skills, and disciplinary expertise organizations need to effectively govern generative AI, and what roles individuals bringing such knowledge could serve

Effectively building and governing generative AI requires highly skilled data scientists, data engineers, model engineers, and an oversight committee (OC) drawn from professional of varied backgrounds. The high-level workflow listed below describes the steps that ideally should be involved in data preparation, preprocessing (transformation), and model development.

Data Preparation

- Data scientists assess the information needed for solving a given problem, identifying raw data sources for leading indicators outlier conditions.
- Scientists perform quality checks on raw data for integrity, accuracy, validity, consistency, completeness, and uniqueness. They will also perform checks to confirm that the use of the raw dataset complies with applicable contractual, regulatory, and related requirements. Data scientists must check with the data owner, if any, as well as data privacy and legal functions, to address any questions regarding data use.
- Data engineers prepare representative data (for preprocessing/transformation) by formatting shaping and smoothing the data set.
- Data scientists perform feature and label extraction.

The steps above describe an iterative workflow that should be performed until the data curation satisfies the need for representative data for solving the given problem.

As the process continues to model training, it is critical that training data, instance data, and validation datasets are separate. For small datasets, an often-used ratio is 70-20-10 for training, validation, and tests.

Model Development

Model development is an iterative process that must include validation steps and the definition of metrics to measure the performance of the model. Throughout validation and deployment, model engineers should monitor model performance for concept and data drift, ensuring performance stays within the defined acceptance threshold.

Continuous training should be built into the model lifecycle with a versioning and product management approach for model governance.

Finally, in any organization developing ML/AI systems, or deploying them, there should be a multidisciplinary oversight committee. The OC should be responsible for the policies for covering adoption, leverage, guardrails, quality, and efficiency, as well as the ongoing scrutiny of the model and model performance. The OC should be comprised of professionals with backgrounds in law, privacy, compliance, ethics, cybersecurity, quality, regulatory affairs, policy, data science, and software engineering.

Current techniques and implementations, including their feasibility, validity, fitness for purpose, and scalability, for: model validation and verification

One approach to conformity assessment NIST should review is the CapAI model from the University of Oxford. CapAI enables key conformity assessment needs, like independence, comparability, and quantifiability. It also requires lifecycle coverage, including best practices across each stage from AI ideation to use (i.e., design, development, assessment, operation, and end of life).

The CapAI recommend process includes the following checks and artifacts as outputs from the assessment:

- Internal Review Protocol (IRP)—a tool for quality assurance, risk definition, and risk management
- Summary Datasheet (SDS)—a datasheet that can be submitted at the appropriate time to regulators or other parties to satisfy risk declaration requirements.
- External Scorecard (ESC)—An artifact to provide to external stakeholders to further the organizations' transparency goals or obligations.

This Prevent-Respond-Rectify model provides guidelines for defining errors or adverse events with evidence of substantiation, tracking, and mitigation of the adverse events.

A Responsible AI (RAI) Framework should define risk in three dimensions: data, model, and audience. Just as with the CapAI model, each dimension could get a score between 1 and 5. This would allow a much more fine-grained categorisation than, for example, the one from EU Artificial Intelligence, which only has high, low, and no risk categories. This augmentation to the RAI Framework would allow an organization to understand the particular tests that are needed and at what point they are needed in the development cycle.

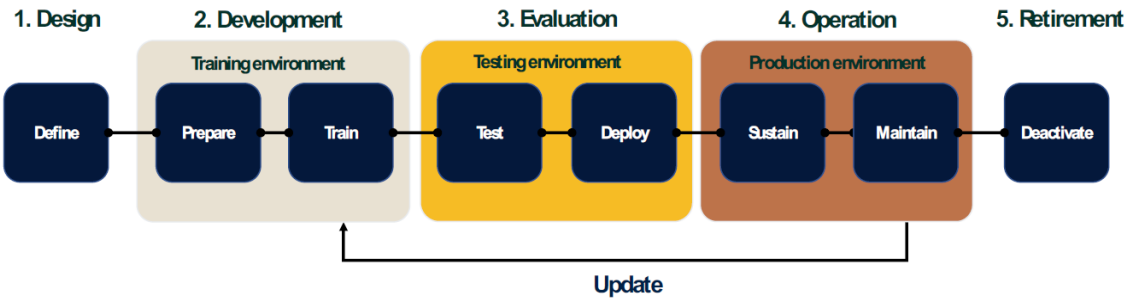

Figure 1. Workflow for Internal Review Protocol (IRP).[1]

As depicted in the figure 1, responsible AI needs to be assured across the development stages. As an example, during Development there could be a need to verify that a data set is fair by ensuring it contains equal representation between genders. During Evaluation, fairness should be re-tested by running Equal Opportunity tests. Different tenets will get different levels of attention in each of the development cycles.

AI red-teaming

As a standard approach, AI red-teaming is constrained by scale to the end of the development cycle. This "structured testing effort to find flaws and vulnerabilities in an AI system" is specific to each use case, model, training dataset and deployment platform. Approaches include creating threat vectors that are injected into code, prompts, augmented datasets, RAG processes, etc. However, rather than limiting red-teaming to post-build, there should be a lifecycle approach to the various threat vectors.

A recommended approach is to create a lifecycle process beginning with ideation, intake process, hypothesis, model assumptions, data sampling, through to build and assessment (validation and verification), and to inject unintended consequences into each phase of this lifecycle. This approach also leverages an OC to validate outlier conditions and introduce bias for detecting flaws in model behavior. A typical OC consists of Legal, Privacy, Data, Infrastructure, Ethics, Compliance representatives in addition to the Data Science organization. Each use case should be validated for risk and impact before this exercise.

Once done, the process should be documented with results, findings, and learnings to aid in transparency. This process is best done using internal resources, given the deep knowledge of edge conditions, and intended use.

In the science-based safety and security domain, adverse conditions are those that impact quality and functional safety of products or processes in the manufacturing spectrum. However, without clear definition of risk and liability topology, red-teaming is not comprehensive and residual risk remains (i.e., unknown unknowns).

Reducing the risk of synthetic content

- While the RFI is seeking input on approaches to reducing the risk of synthetic content, as listed below, synthetic data has many useful applications.

[1] Floridi, Luciano, *et al*. capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act (March 23, 2022), p17. (https://ssrn.com/abstract=4064091 or http://dx.doi.org/10.2139/ssrn.4064091)

- Synthetic data enables model development process without relying on data copy from one environment to another.
- Data sharing issues can be mitigated with synthetic data, particularly access controlled data such as sensitive personally identifiable information, including medical or health information.
- Synthetic data is superior to deidentified data as the correlation of variables is maintained.
- Performance testing requires large volumes of data which can be synthetically generated.
- The leverage of pattern matched correlations in synthetic data generation enables compliance to data privacy stipulations.
- Synthetic data can mitigate data gaps when used within a semantic context.
- Randomly generated synthetic data may not have the same statistical properties and could skew results, but fidelity to patterns in real life data make it possible to use synthetic data for inference and content generation.

Advance Responsible Global Technical Standards for AI Development/Guidelines and standards for trustworthiness, verification, and assurance of AI systems

Below is a list of standards for AI in the certification and accreditation domain. While these standards touch on many applications, there are gaps when it comes to addressing safety, security, and sustainability.

- ISO/IEC TR 5469:—2), Information Technology — Artificial intelligence — Functional safety and AI systems.
- ISO/IEC/TR 24028:2020, Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
- ISO/IEC/TR 29119-11:2020, Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems
- ISO/IEC/TR 24029-1, Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
- ISO/IEC 24029-2, Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods
- ISO/IEC 25059:**2023-06**, Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems
- ISO/IEC 5259 (series), Information technology — Artificial intelligence — Data quality for analytics and machine learning (ML)
- ISO/IEC/IEEE 15939, Systems and software engineering — Measurement process
- ISO/IEC/TS 4213, Information technology — Artificial intelligence — Assessment of machine learning classification performance
- ISO/IEC TR 24368, Information Technology — Artificial intelligence — Overview of ethical and societal concerns
- ISO/IEC/TR 24027, Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making
- ISO/IEC 23894, Information technology — Artificial intelligence — Guidance on risk management
- ISO/IEC 42001 – AI Management System

We see the potential of leveraging AI to advance science-based safety, security, and sustainability. Our 130-year history of testing, inspection, and certification experience, as well as our successes in Outlines of Investigation and Safety Standards development, give us a unique perspective in the manufacturing and componentry space. There are several existing standards that help define intended use and governance of AI systems; we feel there is an opportunity to augment the current list with respect to functional safety, security and sustainability needs.

We appreciate the opportunity to provide input to NIST's ongoing efforts to further trustworthy AI through guidelines, standards, and other resources. Please direct any questions to Derek Greenauer (derek.greenauer@ul.com; 202-296-8092).

Sincerely,

/s

Sreelatha Surendranathan
Chief Digital Officer
UL Solutions