Recommendations on Generative AI risk management and AI evaluation

Author: Gautam Jain
linkedin.com/in/gautamjain
Created on: Feb 1, 2024

No part of the recommendations in this document is created by AI or taken from any existing material written or spoken. The thoughts expressed are original views of the author. Submission is in personal capacity and does not represent the views of the author's employer

Preface

These recommendations are created in response to National Institute of Standards and Technology's (NIST) RFI (Request For Information) as per executive order 14110 Sections 4.1(a)(i)(A) and (C) which directs NIST to establish guidelines and best practices in order to promote consensus industry standards in the development and deployment of safe, secure, and trustworthy AI systems.

Recommendations

Specific and actionable best practices that should be part of AI RMF 2.0 and were missing in AI RMF 1.0 are listed below. To promote a safe, secure and trustworthy AI system in the future, there are 6 unique recommendations namely Disclaimer, Source and Transparency, Human-led testing, Pre-launch readiness, Post-launch feedback and Independent 3rd Party auditor ecosystem. The following section describes each recommendation further:

1. Disclaimer

Every output from a generative AI based product or feature must have a disclaimer which is specific to the intent of the query or the subject within reasonable human understanding. For e.g. if a human requests for financial investment advice to an AI

model, it's output must contain a standard alert message or disclaimer related to the risk associated and that the output is not a legal financial advice and the end user is requested to exercise caution and avail professional financial advice from a registered, professional advisor, tax consultant or accountant - as the case may be. AI developers, deployers and auditors must all adhere to the provision (to be added to AI RMF 2.0) that each response has an associated disclaimer which is relevant within reasonable understanding of the intent of the query.

2. Source & Traceability

In the AI RMF 1.0, it is mentioned that Accountability and Transparency will lead to Trustworthiness. However, there is no mention of Source and Traceability of the generated content or output. If auditors were to pass an AI model for a public launch and wider-availability beyond safe-environment testing, the AI models must be able to either probabilistically or deterministically share the source(s) in creating the output i.e articles read, authors quoted, web pages crawled, systems used, websites referred, books studies, data analyzed, countries associated. In case multiple generic data sets are used to generate the output, the AI model must be able to trace the top 'n' contributing data sets which had the highest percentage role to play in algorithmic learning in the process of generating output.

3. Human-led Testing

Create a community of Trusted Testers.

An AI model, to be publicly available in the hands of humans, or affecting their day-to-day lives, must pass rigorous testing standards

AI models or algorithms, train on data sets and have the computational ability to learn and bypass a test, thus NIST or Secretary of Commerce must actively invest in creating a thriving community of human testers representing different races, regions, religions,

ages, genders, IQs, EQs etc. respecting aspects of diversity, equity, inclusion and belonging.

Create a repository of basic, intermediate and advanced test cases which can be used by the human testers to get started with an AI-model audit, but do maintain > 51% of the tests to be impromptu, human-created, unguided cases

4. Pre-launch readiness

Checklist requirements for a public launch of an AI model should include a Safety score. Based on the human-led testing, each AI model must be given a safety score on a scale of 0-100, which should help the end users of the system to understand directionally how safe it might be to use a particular generative AI based solution

5. Post-launch feedback

NIST must direct AI developers and deployers to continually accept, process, action and inform on each and every piece of feedback received from human interactions. This will help in improving the models based on human expectations of a particular result and maintain a sustainable way to keep AI models in check

The end user must be entitled to receive an update or resolution within reasonable time (SLAs) as is possible for the creators and maintenance personnel of the AI solution

6. Independent 3P audits

The AI RMF 2.0 should have a certification of the highest standard, issued by NIST or other leading regulators to create a 3rd party auditors ecosystem. The auditors may charge a fee for each evaluation based on the AI RMF 2.0 and render an algorithm as safe or unsafe (pass or fail) to be launched

These auditors should be independent entities, not having any vested interests in the companies they are auditing for a given framework

Public launch of AI solutions which do not have a valid certification must be illegal under cybersecurity laws.