ESR
Science for Communities

Response to: NIST's AI Risk Management Framework-- Initial Draft

Dr Deepak Karunakaran, Mr Dion J Sheppard, and Dr John S Buckleton.

29 April 2022

Thank you for the opportunity to review and provide feedback on the NIST AI Risk Management Framework – Initial Draft. The authors of this response are part of the Institute of Environmental Science & Research (ESR), Forensic Science group in New Zealand. ESR is a Crown Research Institute of the New Zealand Government and the provider of forensic science services to the New Zealand Police and New Zealand justice system.

ESR is uniquely placed as both a service provider of forensic expertise and a research organisation, which provides both the operational and future development perspectives within forensic science. Our research expertise includes data scientists and statisticians who focus on the development, validation and implementation of novel solutions utilising AI, with a particular focus on law enforcement, justice and forensic science applications.

The NIST AI Risk Management Framework – Initial Draft provides a valuable contribution to support the reliability of AI systems and their implementation. We do see three areas where additional aspects would improve the intended outcome. These are outlined below, along with our recommendations:

1. The *Overview* referred to a yet to be developed Part 3 where guidance would be provided covering the "before, during and after" phases relating to the development and user of AI systems. This is commendable and will support developers, validators, and users in their application of the intended components of this framework.

   ESR has developed such a document which includes a guideline on ethical aspects for AI development and a companion questionnaire guiding researchers through the process. The structure of our approach addresses the design, development, piloting, implementation, and in-use review of AI systems. These five components are structured in a way to support reliability and trust in AI systems while balancing the different levels of oversight required throughout the lifecycle of an AI project. Our document appears to be directly relevant to the yet to be developed "Part 3".

   Recommendation: The authors would be happy to share the ESR AI Ethics Guideline and Questionnaire document with NIST to support the development of Part 3 of the Framework.

2. The *Socio-Technical Characteristics* referred to in Section 5.2 mentions five topics namely explainability, interpretability, privacy, safety and managing bias. The privacy aspect in Section 5.2.3 discusses about the "individual's control of facets of their identities", including data.

However, it does not discuss the right of the individual to control the use of their data to develop AI systems which they might have originally consented to.

Data is one of the most important requirements for development of an AI system. The data is transformed into a machine learning (ML) model using an algorithm. The model is then used as a part of the AI system towards obtaining the desired outcome. It is of vital importance to consider the ethics of using the data obtained from multiple sources, particularly if it comes from human subjects. During the development of the AI system, the objectives of the final product generally evolve over time. Therefore, human subjects should provide informed consent for the purpose for which their data is being used and this approval should be recorded throughout the lifecycle of development of the AI system.

Furthermore, many modern AI systems use an approach which relies on the use of pre-trained models, which are the ML models already developed using massive amount of data and by exploiting high compute power. The pre-trained models are then tuned using new data towards the development of a new model which generally performs better on the standard ML metrics. Most often, the organisations developing the AI system get ethical approval to use the new data while ignoring the ethical considerations associated with data which was used to develop the pre-trained model.

Therefore, it is of high importance to track the limitations and constraints on the use of data, even after it has been transformed into an ML model and is used in the subsequent AI development process.

Recommendation: It is recommended that Section 5.2 is expanded to include a new section titled *5.2.6 Informed Consent* stating that the data acquired from human subjects must be used with responsibility and a blanket consent must not be assumed. It should be the responsibility of the organisation which develops the AI system to ensure that the data used for development of machine learning models towards solving a specific problem or achieving an objective is obtained with full consent.

The consent for the purpose stated in the original contract or form should be strictly matching with the objective of the AI system. Furthermore, if the AI system developed is to be used as a subsystem for another solution, then the consent must be tracked for all the components of the final solution.

3. The *Guiding Principles* in the Section 5.3 discuss the significance of Fairness, Accountability and Transparency. In particular, the section 5.3.2 on Accountability says, “Individual human operators and their organizations should be answerable and held accountable for the outcomes of AI systems, particularly adverse impacts stemming from risks.” The meaning of *operators and evaluators* is defined in Section 3. It includes “academic, public, and private sector researchers; professional evaluators and auditors; system operators; and expert end users.”

   Though accountability is a very important guiding principle and key to risk management, it must be carefully considered for AI systems. AI systems are different to conventional automation systems. Consider an example of a complex control system used in flight control of an aircraft. Even though the system is complex, the state variables of the systems are well defined. This supports the development of simulators which can support a thorough testing of the system over a large number of scenarios and determine a reliability measure with high confidence on the performance of control systems.

However, AI systems are associated with higher uncertainty. Typically, the ML models used to develop the AI system are created using data which corresponds to a large number of known scenarios for which the AI system is expected to provide the correct outcome. The data is generally noisy and varies in quality. Once the AI system is developed, it is expected to work on unseen scenarios and present correct outcomes. However, if the unseen scenarios vary from the ones on which the AI system was originally trained, errors are expected, and thus the performance of the system goes down. Due to this uncertainty in the performance of AI system and the fact that determining all the scenarios beforehand is intractable for most applications, measuring the reliability of the system with high confidence is very difficult.

Secondly, it is common for AI systems to be presented with validation documentation and performance metrics indicating the alignment of test scenarios with expected or 'true' outcomes. By its very nature this process indicates that an AI system will not be (and perhaps never be) 100% accurate. Knowledge of the performance and reliability of an AI system supports its intended use and goes to the weight or confidence that an end user has in the output. An incorrect result from an AI system does therefore not necessarily represent a 'failure' of an algorithm. The key aspect that provides confidence that an AI system can be used with confidence is the concordance between the test data set and the population it is designed to be used with.

Thirdly, many AI systems are developed to support a decision-making step undertaken by humans and do not replace the human completely. Results generated by operators of an AI system may also be presented to other decision makers who subsequently incorporate the AI output, along with other information in making their decision. An example of this may be the presentation of an AI generated result in a court trial where the decision making and responsibility of the outcome sits with either a judge or jury, who operate separately from the activities of the AI researchers, validators, and operator. In these situations, the outcome from the decision making (for example a guilty or not guilty verdict) may draw on aspects of an AI output but the ultimate decision and therefore the outcome sits separate to the AI system.

As a more elaborate example, consider the AI system which is used by the on-field police officers to detect illegal drugs. At ESR, we have developed a real-time drug scanning system (Lumi^{TM}) which uses a drug classification machine learning model with a portable spectrometer to scan drugs. It is known that the AI system, though very efficient compared to other expensive laboratory-based drug identification methods has limitations in its accuracy. Therefore, the AI system is only used for initial screening and any cases which are taken up for court trials, a more accurate and reliable laboratory-based testing system is used. Thus, AI system works in conjunction to human decision-making with the potential that with more data and better technology it would continue to become more reliable. This is an example of how risk is managed while the new technology is employed for a critical use case. In this case, both the developers and users of the AI system share the responsibility of risk management.

Therefore, it would be inappropriate to hold operators, evaluators, and their organisations responsible for an *outcome* resulting from an AI system. The current wording within section

5.3.2 creates a burden of accountability that would unreasonably slowdown the adoption of new technology to solve many critical problems by not acknowledging the separation between a system and an outcome.

Recommendation: It is recommended that section *5.3.2 Accountability* is reworded. A balance can be achieved between ensuring researchers, validators and operators are accountable for the development, validation and use of an AI system without extending their responsibilities to cover downstream aspects where third parties make decisions that ultimately result in outcomes.

This can be achieved by making transparent the methods and procedures used by the operators and the evaluators to validate the AI system, encouraging that the data and algorithms be made open source so that the risks and safety considerations are available for everyone to investigate. This is especially important for systems which are used by critical applications in health and law enforcement.

It should be made mandatory for the organisation to make the users aware of the limitations of the AI system including the reliability measures of the system. This would ensure a shared responsibility of risk assessment and management between the organisation which develops and the AI system and the end user.

The operators and evaluators should be continuously engaged with the evaluation of the system even after it has been put into production. This would ensure that unseen but potentially risky scenarios are tracked properly and rectified appropriately.