Response to: AI Risk Management Framework: Second Draft

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

29 September 2022

Thank you for the opportunity to review and provide feedback on the NIST AI Risk Management Framework – Second 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 – Second Draft is a noticeable improvement on the first draft and appropriately responds to suggestions and recommendations provided by the authors in our response to NIST's earlier draft document.

From our review of this Second Draft, we have identified three topics where we seek clarification for our own purpose, and where we believe the document could be improved by providing that clarity for other organisations. These three aspects are outlined below, along with our recommendations.

1. Geographical Considerations

The section labelled *Govern 1.1* describes requirements relating to "legal and regulatory" expectations. We agree that an organisation developing or implementing AI solutions should be aware of the legal framework in which the solution is being implemented. However, the NIST AI RMF document would benefit from additional information to help clarify where responsibility for addressing legal and regulatory alignment might best sit for projects or application that span different jurisdictions. This multi-jurisdictional issue may arise in situations where the submission of data and the analysis occur in different locations (a common occurrence with cloud computing) or where an AI solution is implemented across a multiple jurisdictions.

> ***Recommendation 1:*** The authors suggest that section *Govern 1.1*, and the earlier related commentary within the document address the potential for legal and regulatory differences between different jurisdictions and the potential that this creates for differences in response.

Additionally, legislative expectations relating to AI continue to evolve and it would be beneficial to include commentary on the implications of future expectations that may be applied to systems deployed today. This is especially relevant where the outcomes of AI informed decisions are assessed at a future date, for instance in law enforcement and judicial applications.

2. Third Party Risks

A number of sections of the document describes expectations relating to the awareness and monitoring of, and responses to, third party risks. In particular sections *Govern 6.1* and 6.2, *Map 4.1* and *4.2*, and *Manage 3.1*. An understanding of the external components (software, libraries, data) that are incorporated into an AI system is an important consideration.

These sections would benefit from additional information that describes the difference between managing risk and anticipating risk. It is unlikely that an AI solution developer could anticipate and therefore manage all risks, but they should have processes that support responses to a previously unknown risk when it is identified. An example of this difference between anticipating and responding when it comes to risk is the Log4J issue. This library had been extensively incorporated into software solutions by many different developers. However, no one anticipated the risk and therefore was able to manage it until it had been identified. Once identified prompt actions are warranted.

One of the frequently used methodologies to develop AI based applications is to leverage an existing neural network model which has been trained using a large dataset and industry grade computational infrastructure (transfer learning). The AI developer uses their custom dataset to further train the model focused on their application. Considering the complex architectures of the neural networks, and the large number of parameters, it is well known that they are difficult to explain. In such scenarios, where third party pre-trained models are used as a basis for further development, it is very difficult to determine the root cause of potential undesirable behaviour of the resulting AI system. This is quite different to the risk associated with using third-party components in traditional software development.

Recommendation 2: We recommend firming up what is meant by third party software (we have assumed that things like Apache libraries are meant). We recommend a justified emphasis on procedures for remediation but do not really see how we could identify these risks in advance.

We also recommend that pre-trained models which are used for development are separately monitored as a part of continuous development and maintenance of AI systems. It is of vital importance that the behaviour of both the AI system and its components are tracked independently when in production.

3. Automation vs Human Oversight

The document describes the importance of human oversight throughout the AI process which we strongly support. One section, however, appears to imply an additional level of human intervention which may not have been intended by the document. Section *Map 2.2* describes an expectation that AI systems are overseen by humans. We seek clarity on where this oversight should reasonably be expected to be applied. In our experience, AI systems can provide considerable benefit when they

automate and standardise the analysis of a data set that was previously reviewed by a human, or where they provide an automated result where previously no data insights existed. An example of the former is machine analysis of large data sets that would not be possible for a human to undertake. An example of the later is the automated analysis of field generated data, providing a result to supplement what was previously a decision-making process that relied only on human experience and circumstances.

In both of these examples the AI system would be operating autonomously, generating a result without human oversight while performing analysis of the data and computing inferences. Human oversight is likely to have been included in the design, development, validation, and implementation of the AI system. And the AI generated output may contribute to and inform human decision making. But in either case, there would be no human oversight of the data analysis performed by the AI system.

This should not be seen as a limitation of an AI system, as it goes to the heart of why an AI system is likely to have been designed, developed, and implemented. Specifically, to 'remove' the human from the burden of reviewing data, or to enable analysis of data that was not accessible for human review.

Self-driving cars are an example to illustrate this point. Though the current self-driving cars, during trial, strictly require human oversight, it is unlikely that they could be put into production unless the human involvement is completely avoidable. A few unfortunate fatalities associated with self-driving cars have been reported to be due to the failure of continuous human oversight. Therefore, mandating an oversight of AI system, and expecting the human to reliably do it could be counterproductive.

Furthermore, an undesirable outcome from AI system might be considered riskier than a similar outcome from a human being. In other words, even if AI systems perform better than human beings in a real-time environment they cannot be employed confidently unless their performance supersedes those of human beings considerably. In such cases, the requirement of human oversight could be unethically made necessary by the AI developers to hide the fact that they are yet to prove the reliability of their system. AI-based face recognition or other automated biometric matching systems used in forensic context are such examples.

> ***Recommendation 3:*** It is recommended that reference to human oversight describes the difference between human-in-the-loop (HITL) workflows and workflows where the human contribution happens prior to deployment, or as a result of the outcome from the AI but not necessarily within the workings of the AI system as it undertakes its operational function.
> When AI systems with a requirement of human oversight are deployed it must be advised that additional risks could arise. In critical systems, it is of vital importance to evaluate these risks before AI system is productionised. Even though the effectiveness of HITL AI systems might be higher, their behaviour in the real-world environment might lead to issues and undesirable outcomes.