Arthur Harvey
CEO & Founder
H&A Luxury Group LTD

Comments on AI Risk Management Framework: Initial Draft

1. Whether the AI RMF appropriately covers and addresses AI risks, including with the right level of specificity for various use cases. Yes, 100%.

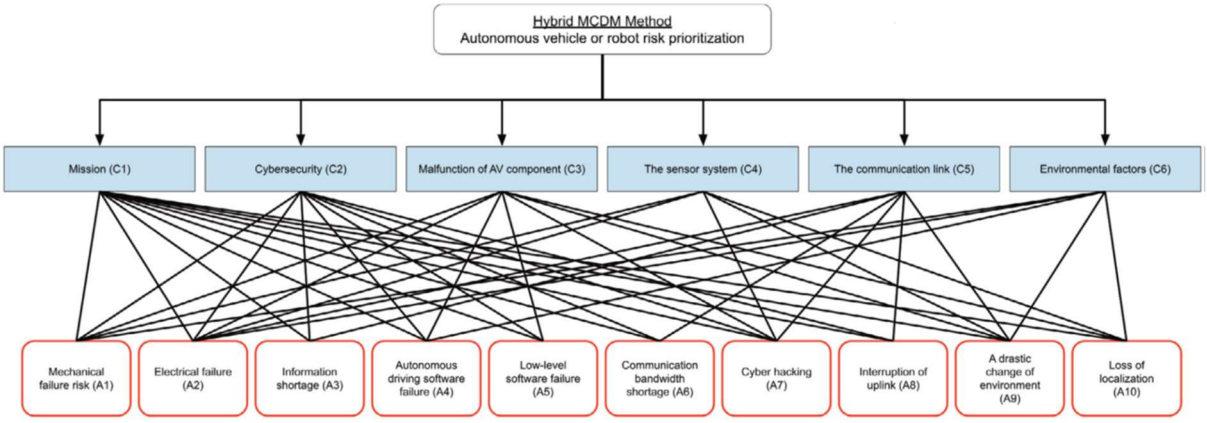
2. Whether the AI RMF is flexible enough to serve as a continuing resource considering evolving technology and standards landscape:

1. "Potential business and societal (positive or adverse) impacts of technical and socio-technical characteristics for potential users, the organizations, or society as a whole are understood.
2. Potential harms of the AI system are elucidated along technical and socio-technical characteristics and aligned with guiding principles.
3. Likelihood of each harm is understood based on expected use, past uses of AI systems in similar contexts, public incident reports or other data.
4. Benefits of the AI system outweigh the risks, and risks can be assessed and managed."

The above requirements might place unreasonable burden on businesses. The AI RMF design should support dynamic/agile risk assessment and mitigation.

This is because the procedure of proving that an AI system is safe to all disturbances/risks (formal verification) is impossible in real life examples. For example, as incidents/risks for an online business or autonomous vehicle are almost never well-defined. A well-defined incident is one in which all (or at least most) of the potential causes of anomalies/risks can be enumerated.

As any gaps identified and exploited by hackers at this point can lead to high-risk security breaches (i.e., outages, failures, nefarious abuse, etc.) as shown in Figures 1.



In addition, it may be that over time, a business can define some incidents or risks—leading to semi-supervised learning techniques.

This typically applies to a closed system with a very limited number of metrics. It might apply, for example, to a relatively simple machine where the product designers or engineers have written documentation of what could go wrong. That list could be used to learn how to detect anomalies and predict similar future risks.

However, for an e-commerce website or IoT/AI- enabled products, it would be a Sisyphean task to try to list what could go wrong and break those things down to tangible incidents where mathematical models could be applied.

If an AI system has well-defined incidents, it is possible to apply supervised learning techniques. There is a well-defined set of incidents; now the system simply must classify the data into these sets. It requires labeled examples of anomalies, but it will be limited to the types of things the company is trying to use its system to find.

If there is a list of a hundred different things that could go wrong, the system is trained about all of them, and then tomorrow the 101st thing happens that was not previously considered, the system will not be able to find it because the system is not trained to do so.

In line with The NIST Framework for Cyber–Physical Systems, we propose the use of simulation and real time anomaly detection tool to find failures of a system caused by disturbances in the environment in real time, and a model of the disturbances can then be used to determine the probability of failure.

AI system might be deemed safe/secure/resilient if two or more conditions are satisfied:

a. No failure has been found after adequate exploration of the space of possible disturbances, and
b. The probability of failure is found to be below an acceptable threshold

This in turn will help build trust and prove compliance. The acceptable threshold that determines the resilience/robustness of the CPS/AI system and its components must be agreed on by the key stakeholders including the regulator in line with the NIST AI Risk Management Framework.

The acceptable threshold may represent the “reasonably likelihood “of system failure. For example, the system is resilient/safe if the probability of failure is found to be below a certain number.

In general, the safety critical software community seem to accept that a failure rate of about 10-3 to 10-4 per hour can be demonstrated prior to release to service, via statistical testing.

Similar best practices can be implemented to prove compliance and develop resilient by design AI systems.