Stuart Millar
AI RMF RFI Response 14/9/21

1. The greatest challenges in improving how AI actors manage AI-related risks—where “manage” means identify, assess, prioritize, respond to, or communicate those risks;

A NIST framework would be ideal to manage (identify, assess, prioritise, respond to and communicate) AI risks. I think broadly those risks relate to bias, ethics and the security of AI against adversaries. This framework could be similar in nature to those previously issued by NIST. Alongside that should sit actual explanations of technical solutions to mitigate those risks. A pre-design stage also needs to consider the dataset itself being used for the use-case. Is it being made from scratch vs. being inherited, is it labelled, if so who labelled it. How accurate are those labels? What does the class balance look like? The lineage of the data should be tracked, consistently captured and have an audit trail. A broader question is for the given dataset in question, figure out how we define, measure and deal with bias in that data.

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

One can also consider any legal requirements already existing to mitigate against bias, depending on the use-case. If there is a legal requirement for the algorithm to be explainable/interpretable (in case of court case triggered by bias), that law is likely there for a reason. If this is the scenario, consider very carefully whether this is the right-use case. Be aware early on of any and all legal requirements across geographical regions that will affect an algorithm.

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: Transparency, fairness, and accountability;

This depends on the use-case, if it has impact on human beings. More benign and basic use-cases may not need that level of scrutiny, it could hinder the adoption and hence the benefits.

4. The extent to which AI risks are incorporated into different organizations' overarching enterprise risk management—including, but not limited to, the management of risks related to cybersecurity, privacy, and safety;

In practice, this is in its infancy, generally.

5. Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

Addressed in other comments.

6. How current regulatory or regulatory reporting requirements (e.g., local, state, national, international) relate to the use of AI standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles;

I would highy recommend reviewing the output from the ETSI industry group for securing AI: https://www.etsi.org/committee/1640-sai?jjj=1631696466212

7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts;

I feel like we need to recognise that algorithms are software. We are building models and in essence writing software. Best-practice for software is that is needs to go through rigorous, documented QA, meaning:

1. The algorithm does what it is meant to.
2. The algorithm learns what it is meant to.
3. The algorithm does not do what it is not meant to do.
4. The algorithm does not learn what it is not meant to learn.

I think point 4 is particularly poignant and we should be cognizant of that.

I can suggest a detailed and documented (in case of audit) discussion with stakeholders on:

What is the main intended objective?
What are any other intended objectives?
What possible learning outcomes do we need to mitigate against?
What is the specification of the algorithm?

Also decide what would a human review entail exactly to check for bias (at any point in the project) if artefacts (such as the data, or a trained model) were available for the closest of inspections.

In the deployment stage, extra tooling will be needed in dashboards etc to measure selected metrics that are identified as being indicators of bias, from class balance in datasets through to the distribution of predicts, data/concept and everything in-between. This needs designed and evaluated before deployment, not afterward.

Two interesting papers that are relevant:

Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction
Christina Wadsworth, Francesca Vera, Chris Piech

Does Object Recognition Work for Everyone?
Terrance DeVries, Ishan Misra, Changhan Wang, Laurens van der Maaten

8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation—and how AI design and development may be carried out in a way that reduces or manages the risk of potential negative impact on individuals, groups, and society.

9. The appropriateness of the attributes NIST has developed for the AI Risk Management Framework. (See above, “AI RMF Development and Attributes”);

I believe the suggested attributes are indeed appropriate.

10. Effective ways to structure the Framework to achieve the desired goals, including, but not limited to, integrating AI risk management processes with organizational processes for developing products and services for better outcomes in terms of trustworthiness and management of AI risks. Respondents are asked to identify any current models which would be effective. These could include—but are not limited to—the NIST Cybersecurity Framework or Privacy Framework, which focus on outcomes, functions, categories and subcategories and also offer options for developing profiles reflecting current and desired approaches as well as tiers to describe degree of framework implementation; and

I know previous NIST frameworks for risk/privacy have sections on prioritising risks, followed by controls. I think the language is important and the term 'control' does feel use for AI security and bias. I expect a similarly structured framework could be effective, with explicit listing of AI risks and then explicit controls for those risks, all documented.

11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.

I suppose interview questions could be designed against the framework to assess a candidate's mindset and awareness of these issues.

12. The extent to which the Framework should include governance issues, including but not limited to make up of design and development teams, monitoring and evaluation, and grievance and redress.

Again this depends on the use-case, but the make-up of development, data-engineering, data labelling teams could have an influence on how much bias, conscious or unconscious, may be encoded into a dataset, and by extension then how much the model itself could be biased.