24 January 2022

Elham Tabassi, Chief of Staff, Information Technology Laboratory
National Institute of Standards and Technology (NIST)
MS 20899, 100 Bureau Drive, Gaithersburg, MD 20899

Subject: NIST AI Risk Management Framework Concept Paper

Via email to AIframework@nist.gov

Dear Ms. Tabassi, and the NIST team developing the AI Risk Management Framework,

We write to submit comments on the Concept Paper on the NIST AI Risk Management Framework (AI RMF or Framework). We are a group of researchers at the University of Cambridge, and at the Leverhulme Centre for the Future of Intelligence – a leading international centre in AI ethics. We have published dozens of technical papers on machine learning and artificial intelligence, along with reports, white papers and peer-reviewed articles on the ethics and governance of artificial intelligence.

We welcome the NIST AI Risk Management Framework Concept Paper, and commend NIST on your ongoing work to address risks in the design, development, use, and evaluation of AI products, services, and systems. We emphasize that the approach set out in the Concept Paper is the right one. With this framework, NIST can demonstrate global leadership and shape global standards. It will reduce harm, encourage adoption, and provide business certainty. **Our key piece of feedback is to keep this proposed framework and not water it down**.

Below we offer input on the questions in the ‘Note to Reviewers’.

Yours sincerely,

Haydn Belfield
Academic Project Manager - Centre for the Study of Existential Risk, University of Cambridge
Associate Fellow - Leverhulme Centre for the Future of Intelligence

Matthijs Maas
Research Associate - Centre for the Study of Existential Risk, University of Cambridge
Associate Fellow - Leverhulme Centre for the Future of Intelligence

Shahar Avin
Senior Research Associate - Centre for the Study of Existential Risk, University of Cambridge

Seán Ó hÉigeartaigh
Executive Director - Centre for the Study of Existential Risk, University of Cambridge
Programme Director - Leverhulme Centre for the Future of Intelligence

− Is the approach described in this concept paper generally on the right track for the eventual AI RMF?

Yes.

1.
We believe the paper strikes the appropriate balance between acknowledging the extensive benefits of AI technologies (Page 3), as well as the potential risks to be guarded against. Appropriate risk management is a necessary and normal step in unlocking that potential for people, businesses and society.

2.
We also appreciate the commitment to collaborative working: that the RMF should "be consistent or aligned with other approaches" - and "be law- and regulation-agnostic to support organizations' abilities to operate under applicable domestic and international legal or regulatory regimes" (Page 4, Lines 4-8). When the EU AI Act passes, the focus will shift to developing 'common specifications and harmonised standards'. With these two parallel processes happening on both sides of the Atlantic, we believe there is a real opportunity here to work in a coordinated way to set global standards that reflect our shared values. Importantly, this can also ensure adequate cohesion and harmonization of standards and policies, and avoid the costs (both operational and normative) of policy fragmentation amongst different bodies of AI regulation.

− Are the scope and audience (users) of the AI RMF described appropriately?

Yes.

1.
We appreciate the statement (Page 3, Lines 4-6) that "All stakeholders should be involved in the risk management process", and the detail that such stakeholders should include "affected communities". It is vitally important that those who "experience potential harm or inequities" (and their representatives) are included in the conversation. Standard setting cannot just be industry setting standards for itself.

2.
One particular audience that might be worth highlighting for attention is "people who are responsible for procuring or acquiring AI systems and services". This is a category of people in companies and government who face particular challenges around working out whether the systems they are procuring actually operate as advertised, whether they are safe and secure enough, etc. Procurement is often where the rubber hits the road when it comes to risk management. We talk about some of the specific challenges of procurement around AI systems in defence contexts here.

Our understanding is that NIST is already thinking about the particular needs of government procurement officials. See e.g. the mention of "requirements representative" (Page 3 Lines 5-6). Such representatives might be considered as falling under audience (3) ‘evaluating or governing’. But it may be worth considering whether to break this out as a fifth audience.

− Are AI risks framed appropriately?

Yes.

1.
We strongly agree with NIST’s statement that “An example is the evaluation of effects from AI systems that are characterized as being long-term, low probability, systemic, and high impact. Tackling scenarios that can represent costly outcomes or catastrophic risks to society should consider: an emphasis on managing the aggregate risks from low probability, high consequence effects of AI systems, and the need to ensure the alignment of ever more powerful advanced AI systems” (Page 2, Lines 13-17).

This is a key focus for our academic and policy work. Some scenarios do indeed represent catastrophic societal risks, and in this context it is very important to focus on aggregate risks, and on ensuring frameworks are flexible enough to ensure the alignment of increasingly powerful AI systems.

These are concerns that have been highlighted by many:

- Leading AI researchers (such as the 8,000 signatories of this 2015 Open Letter)
- Leading companies (such as OpenAI and Google DeepMind), and
- Allied nations - (e.g. the UK Government's 2021 National AI Strategy states "AI risk, safety, and long-term development - The government takes the long term risk of non-aligned Artificial General Intelligence, and the unforeseeable changes that it would mean for the UK and the world, seriously.").

We recommend that this language be retained in the first draft RMF. These issues should be developed upon, and and additional related guidance should be provided, as appropriate

2.
We agree with the RMF’s approach to framing ‘risk’ through a “broader definition that offers a more comprehensive view” (Page 3, Line 12). We should therefore not restrict our conception of risk to e.g. a "specific physical risk to an individual". Such a narrow framing could risk ignoring or underplaying both:

- significant effects for the rights of an individual or company, damages in the form of significant economic loss or reputational effects, and damages relating to data protection and privacy, non-discrimination, defamation and freedom of expression.
- broader societal risks. Including potential uses (such as recommender or information curation systems) that could effect potentially harmful shifts in our markets, democracies,

and information ecosystems that might be substantial, yet hard to discern in terms of decisive impacts on individual health, safety, or rights.

− Will the structure – consisting of Core (with functions, categories, and subcategories), Profiles, and Tiers – enable users to appropriately manage AI risks?

Foundation models (profiles)
Yes.

1.
The AI RMF Profiles (and Tiers) provide a strong basis for appropriately managing AI risks. We do recommend a number of particular areas in which Profiles would be helpful. In particular, we recommend that Profiles be created for two areas:

- Foundation models. A foundation model is one that is increasingly multi-purpose or general-purpose, and can underly many other models - such as GPT-3 or BERT. These models have specific characteristics that make a specific Profile useful. This includes often being trained on large-scale scraped data sets, and the risk of correlated failures.
- Certain application areas. The EU has done useful work which the NIST RMF can build upon. For instance, the EU AI Act, in its Annex III, identifies eight application areas that it describes as “high-risk”. NIST may not want to use their specific language, but this list clearly identifies areas in which additional guidance - and a Profile - would be useful. Specifically, this includes
  - Biometric identification and categorisation of natural persons
  - Management and operation of critical infrastructure
  - Education and vocational training
  - Employment, workers management and access to self-employment
  - Access to and enjoyment of essential private services and public services and benefits
  - Law enforcement
  - Migration, asylum and border control management
  - Administration of justice and democratic processes

− Will the proposed functions enable users to appropriately manage AI risks?

Yes.

1.
On the ‘Map’ function, the Paper notes that “Context refers to the domain and intended use, as well as scope of the system….” (Page 4, Lines 35-38). However, it is important to note that AI systems (such as foundation models) are becoming increasingly general and do not have a single “intended use”. Increasingly, AI systems can be employed in many end-use applications. Furthermore, they can sometimes adapt and learn online, e.g. through interaction with users. Too much of a focus on singular “intended use” could fail to “find, recognize, and describe risks”.

2.
On the ‘Measure’ function, it is important that risks that are not (yet) measurable, or at least not with precise quantitative methods, are not allowed to fall through the cracks. We note that the U.S. Chamber of Commerce's Technology Engagement Center ("C_TEC") stated that "The RMF should specifically address situations where risk cannot be measured and offer guidance on reasonable steps for mitigating that risk without limiting innovation and investments in new and potentially beneficial AI technologies" (C_TEC 2021).

3.
One thing to consider is whether ‘Map, Measure, Manage, Govern’ might frame risk management as somewhat individualized or atomised. It might suggest that this is done by specific groups on their own - mapping, measuring, managing and governing only their own risks, by themselves.

It might be worth considering expanding this list to include a more collaborative function, for example: "Share: information, best practice, new techniques etc". Alternatively, this collaborative aspect could be highlighted across the four proposed functions.

[end]