2022 04 25 _ Comments on AI Risk Management Framework Initial Draft from Sam Hilton at (CSER)

Introduction

I am Research affiliate at the Centre for the Study of Existential Risk (CSER) at the University of Cambridge and secretariate to the All-Party Parliamentary Group for Future Generation in the UK Parliament. My background is in risk policy having led the team in the UK government responsible for civil nuclear safety and worked on financial stability policy in the UK Treasury.

I was recently recommended to read your AI Risk Management Framework: Initial Draft. I am sending you this email with a few comments in case it is helpful to your work. In case helpful these notes are also in the attached document.

Overall I was hugely impressed with this document. You have acknowledged the need for flexible processes that are outcome driven and can adapt to changes in technology. I found the framing and the explanation of the challenges and the basics of the RMF Core (Govern, Map. Measure, Manage) to be well done and well thought out. I would provide the following constructive comments:

On the AI RMF taxonomy:

It was not clear to me that the categories being used were comprehensively exhaustive. Some things that may have been missing are:

Technical Design characteristics
Missing characteristic: alignment.
Explanation: Some AIs have been known to accidentally be directed towards goals that are not the ones their developers wanted, in particular through specification gaming. "For example, an agent performing a grasping task learned to fool the human evaluator by hovering between the camera and the object." See more on this at: https://deepmindsafetyresearch.medium.com/specification-gaming-the-flip-side-of-ai-ingenuity-c85bdb0deeb4.

Socio-Technical Characteristics
Missing characteristic: hard to misuse.
Explanation: Developers and AI companies should be assessing the risks that their technology can be misused by criminals and others. For more on this see the paper on the topic here: https://www.cser.ac.uk/news/malicious-use-artificial-intelligence/
(Note, that paper is on the malicious misuse of AI. There can also be cases of accidental and non-malicious misuse where an algorithm is used outside of its specification. E.g. using a self driving car AI in a car with a trailer attached without realising some of the algorithms would not work correctly)

These are just two examples. There may well be other missing characteristics too

Solutions

Potentially it would be worth expanding the categories or consulting in more depth with technical experts to try to ensure that the categories are as comprehensively exhaustive (and mutually exclusive) as possible to the various problems of AI.
Alternatively I could see a case for saying that at this stage in the technological development of AI it might be very hard to map out the full range of trustworthy AI characteristics. In this case it might be worth adding "other" / "other alignment issues" categories to each section

On the AI RMF Core:

Map and Measure sections: Vulnerability assessments
These sections were decent. The focus of the Map and Measure sections were on understanding the AI system itself. I expect there is also value for risk managers in understanding the key vulnerabilities of the broader systems into which the AI will be deployed, be it defence or energy or medical diagnostic systems. For example the financial system is vulnerable to systematic errors where many actors are making the same false judgements at the same time. This identification of broader vulnerabilities could lead to an identification of further ways that the AI could be inherently risky, as well as leading to additional solutions to manage those risks. Building on the previous example, a trading algorithm AI that is low risk when it is only used by a small % of the market might be high risk if it is used by a large % of the market, so should perhaps be restricted in its use.

The Manage section
This section was good. Perhaps something could be added on sharing risk data with other organisations. There is decent evidence that information sharing about cyber-vulnerabilities can reduce cybersecurity risks. It seems plausible that information sharing of bias found in data sets or other technical problems could help reduce AI risks across organisations.

The Govern section
This section was a bit lacking in the details of the level accountability mechanisms I would expect to see in an organisation that deals well with risks. I think more could be said in this section, especially on "Accountability structures are in place to ensure that the appropriate teams and individuals are empowered, responsible, and trained for managing the risks of AI systems". This seems crucial to the overall governance process. Things that could be added here include:

- Executive leadership of the organization has sufficient expertise and skills to be able to make decisions on AI system development and deployment
- Risk analysis, risk plans and the risk governance process should be audited by an independent reviewer. Risk functions should follow the three lines of defence model (risk owners, senior management, intendent audit)
- There should be whistleblowing regimes in place to allow staff to speak up if concerned about risks
- Organisations should aim to have risks plans that they can justify to outside actors.