Dear NIST,

Thank you for your valuable work producing the AI Risk Management Framework, as well as for this opportunity to provide feedback. As AI has much potential to affect society, we believe it is important that its associated risks are proactively managed, and we are glad NIST is putting forth effort to do so.

We would like to highlight a few aspects of the framework that we particularly commend. First, we are pleased to see that the framework is taking a flexible and adaptable approach. As AI is a fast-changing field, we believe proper management of risks will necessarily involve adapting as the landscape changes, and we are pleased to see NIST shares this view.

Second, we are happy to see the framework acknowledge the beneficial potential of AI. We believe that beneficial uses of AI should not be unnecessarily hindered while appropriate guardrails are being determined.

Third, we are pleased to see the framework recognize potential shortcomings and tradeoffs involved in various risk-management desiderata. For example, in the section on explainability, the framework mentions risks that may arise from explainability methods themselves, such as due to a lack of fidelity in explanation methodology. We agree that risks here may be introduced insofar as users assume an AI explanation necessarily corresponds to the internal workings of the AI model on an algorithmic level, as current explainability methods do not typically give such guarantees.

If you would be interested, we would also like to offer a few recommendations for how the framework could be further improved:

- In the Overview section, the framework mentions that AI systems sometimes operate in unintended ways due to such systems “making inferences from patterns observed in data rather than a true understanding of what causes those patterns.” While this phenomenon is certainly a concern that should be emphasized, there is a separate concern that we believe should additionally be emphasized: that of AI systems correctly understanding the patterns in data, but solving problems in ways their programmers didn’t intend, or solving subtly different problems from what their programmers intended (effectively doing what their programmers “say” instead of what they “mean”).
- In the section on safety, the framework highlights AI systems that interact directly with humans, such as in factories and on roads. Such direct interactions provide clear safety concerns, but we believe other systems may also raise indirect safety concerns which the framework may additionally want to highlight. For example, AI systems integrated into crucial infrastructure such as the electrical grid may present risks to safety if they have certain failure modes. Additionally, systems such as large language models may interface with other AI systems or broader society in ways their producers don’t initially imagine, and it would be beneficial for producers and operators of such systems to consider what safety risks they may introduce.

- In Table 1, under ID 4 (“Risks and harms…”), the framework states “Likelihood of each harm is understood based on expected use, past uses of AI systems in similar contexts, public incident reports or other data.” While focusing on the rate of past incidences is likely appropriate for most use cases, we believe that such an approach may underestimate the potential harms from failure modes for emerging technologies with limited track records, especially if such technologies introduce failure modes that are both particularly unlikely and particularly severe. We believe this section would be strengthened by adding an additional sentence saying something along the lines of, “In contexts where severe tail risks may occur, past uses will underestimate expected harms, and other methods – such as analyses to lower bound the likelihoods of such tail risks – may be necessary to estimate the expected harms from the AI system.”

Again, we would like to thank you for the opportunity to provide input on this framework.

Sincerely,
Daniel Eth