Dear NIST,

Thank you for continuing your valuable work crafting the AI Risk Management Framework, as well as for another opportunity to provide feedback on the framework. Overall, we believe this second draft is quite good, and notably, we think it is improved compared to the first draft.

There are a several aspects of the current framework that we are particularly pleased with, which we would like to highlight:

- The framework takes a nuanced view of AI, acknowledging both risks and benefits. As both the risks and benefits of AI are likely to be substantial, we agree with NIST's decision to consider both instead of strictly focusing on one side of the equation.

- The framework appropriately discusses the important-yet-thorny topic of prioritization between different risks, both by defining "risk" as "the composite measure of an event's probability of occurring and the magnitude (or degree) of the consequences of the corresponding events" and later calling for the highest risks to have "the most urgent prioritization and most thorough risk management process." We would in particular like to highlight the following line as valuable, "In some cases where an AI system presents the highest risk – where negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present – development and deployment should cease in a safe manner until risks can be sufficiently mitigated." We agree with the decision to emphasize that development and deployment should be avoided when risks are sufficiently high, and we further support the inclusion of potential catastrophic risks among the category of risks that warrant such caution.

- In several places, the framework mentions emergent properties of AI systems as warranting attention. Concerns about emergent properties are relatively new (heightened especially with recent advances in large language models), yet associated risks may be particularly severe due to lower likelihood of being anticipated.

- The section "Appendix B: How AI Risks Differ from Traditional Software Risks" is a valuable addition. Many users of the framework may be unaware of some of these differences, and we believe this section will be helpful to them. In particular, we could imagine that many users may be unfamiliar with concerns about emergent properties of large-scale pre-trained models, as well as opacity of many current AI systems, as these concerns are very different from concerns typical of traditional software.

There are also a couple areas where we believe the framework could be strengthened, and we therefore offer the following advice:

- Increase emphasis on how risk measurement can be qualitative when quantitative measurements are impossible or impractical. The framework mentions that measurement can be either quantitative or qualitative, yet this point is not emphasized

particularly strongly, and we believe many readers may default to thinking of measurement as an inherently quantitative endeavor and thus neglect any measurement of risks that are not easily quantifiable.

- In Appendix B, include a bullet point on how strong optimization pressure during training can lead to worse behavior (i.e., “Goodhart's law”), such as due to overfitting or specification gaming. As traditional software doesn’t typically involve optimizing behavior the way many types of AI systems do, users who are familiar with traditional software but not with AI might overlook associated failure modes.

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

Sincerely,
Daniel Eth