AI Risk Management Framework
U.S. National Institute of Standards and Technology (NIST)
100 Bureau Drive
Gaithersburg, MD 20899
AIframework@nist.gov

September 28, 2022

Dear NIST Staff,

I am an independent researcher with an interest in artificial intelligence. I am excited by the work NIST is doing with the AI Risk Management Framework (RMF or Framework) and Playbook, as well as NIST's call for comments on the second draft of the Framework. I recommend that NIST consider revising the Framework to include two best practices: (1) use compute as a measure of AI riskiness and (2) use data cards.

Use Compute as a Measure of Riskiness

On page 20, the RMF mentions that some applications of AI may be "deemed to be high-risk" but the Framework does not define "high-risk." This is unsurprising, as defining "high-risk" is extremely difficult. But one cause of AI riskiness is their increased capabilities. For example the increased scale of AI models increases the believability of content they generate, which will allow them to be used in new applications that empower disinformation actors or harass individuals, in a manner not possible before.[1] [2] [3]

And while "riskiness" is subjective, capabilities are more quantifiable. As noted by OpenAI:[4]

[1] https://arxiv.org/abs/2108.07258: "Advances in the scale (§4.2: training), multimodality (§4.1: modeling), and adaptivity (§4.3: adaptation) of generative foundation models will allow them to be misused to generate high-quality, cheap, and personalized content for harmful purposes."

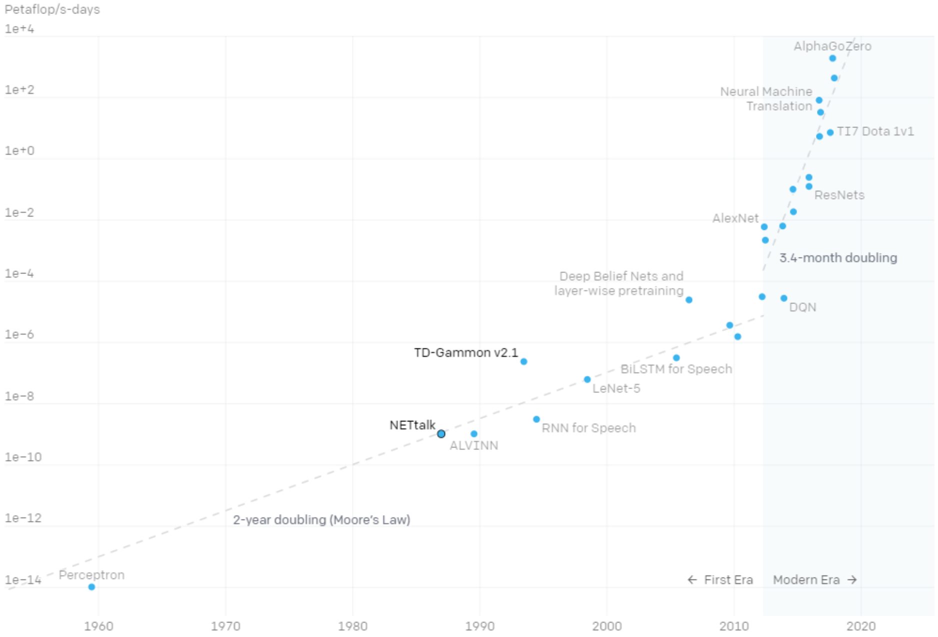
[2] https://arxiv.org/abs/2108.07258: "Foundation models are capable of automatically generating much higher-quality, human-looking content than prior AI methods. They may empower disinformation actors, where states, for example, create content to deceive foreign populations without being transparent that the content is linked to a state."

[3] https://arxiv.org/abs/2108.07258: "In addition to deceiving foreign populations, foundation models' ability to generate high quality synthetic images (deepfakes) or text may be abused to harass individuals. Deepfakes have already been used for the purpose of harassment. For example, Rana Ayyub, an Indian investigative journalist, was targeted by a high-quality deepfake that superimposed her face onto a pornographic video, leading her to leave public life for months. Because foundation models are often multimodal (§4.1: modeling), they could similarly impersonate speech, motions, or writing, and potentially be misused to embarrass, intimidate, and extort victims."

[4] https://openai.com/blog/ai-and-compute/

> “Algorithmic innovation and data are difficult to track, but compute is unusually quantifiable”.
>
> “Improvements in compute have been a key component of AI progress, so as long as this trend continues, it’s worth preparing for the implications of systems far outside today’s capabilities.” (see Figure 1).

Figure 1: The Increasing Compute Usage of More Capable AI Systems



Source: https://openai.com/blog/ai-and-compute/

As a result, the Framework should recommend using the compute used to train the model, as one measure of an AI model’s risk of misuse.

Use Data Cards

Data cards (also known as model cards or system cards) are a standardized system for providing essential facts about AI datasets that are needed for responsible AI development.[5] According to the US military’s Joint Artificial Intelligence Center (JAIC, now renamed to CDAO)[6]:

> “The JAIC recently adopted the industry-proven use of data cards and has begun developing the process workflow and framework that will support AI projects and requirements across the Department”

[5] https://arxiv.org/abs/2204.01075: “Data Cards are structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a dataset's lifecycle for responsible AI development.”

[6] https://www.ai.mil/blog_09_03_21_ai_enabling_ai_with_data_cards.html

> “The use of data cards is key to accelerating the adoption of safe and ethical AI into DoD operations.”

While the Playbook does include two indirect references to data cards, I suggest that their use be directly advocated and in the Framework.

Conclusion

Thank you for this opportunity to contribute to the Framework. I am happy to provide additional details or information at the future workshop or other activities that NIST is using to develop the Framework.

Best Regards,

Andre Barbe, PhD