From: Aleksander Madry
Subject: Comments on 1st Draft of the AI RMF

Overall, I like the direction this framework is going in (and can see how much work putting it together required)! In particular, I like the way the whole landscape has been structured and the taxonomy it puts forth (modulo two comments---see below).

Now, my biggest worry is how approachable this document is to someone that is a layperson in this domain. Yes, to me, everything was crystal clear (and I kept nodding a lot as I read). But I wonder if the considerations/definitions this document brings up might be too abstract for someone who does not have much experience in this domain. Having some simple examples in each subsection could help with this.

Also, the next (and maybe the biggest) challenge will be provide some example implementations and developing all the mentioned companion documents. I wonder what the plan to go about it is and where these examples would be gathered from.

Finally, the idea of making this be a living document is great (and necessary given how quickly things evolve)---I wonder how it will be work out logistically.

More specific remarks:

p. 9. I am not sure I agree with your delineation between accuracy and reliability. To me, what reliability discusses corresponds to "just" statistical justification of the measurement of accuracy/standard performance, while reliability corresponds more to what the document calls robustness (i.e., making sure the model's performance gracefully degrades once the deployment conditions start to vary). But, of course, this is a matter of individual opinion---I just prefer to not have more terms than necessary.

p. 10, l. 33. I absolutely agree that it is important to get input from all the stakeholders, but the key challenge here is: how to operationalize that? I think at the very least it is important to acknowledge that this is a difficult process, as it is not just about giving all the stakeholders an opportunity to provide input but also making sure they are informed enough/are in a position to provide an input that is meaningful.

p. 11, I like a lot how the document delineates the (tricky) difference between explainability and interpretability, but this is a place where I thought some example would be especially needed to make these definitions and the difference between them accessible to a non-expert reader.

p. 12., I think it would be again worthwhile to provide some examples of bias in the models and, in particular, make it explicit this is not only about the (very important) bias around protected attributes but also more "benign"---while still damaging---biases (including the "recognizing wolves by the presence of snow" type of biases).

p. 12. l. 22-24. I think it is important to say that it is not always possible to get both high degree of risk control and high level of performance and there might necessarily be a tradeoff there.

p. 13, I think talking solely about fairness (and not also other potential societal impacts of model deployment, like feedback loops that encourage certain behavior, e.g., spread of misinformation) might be a too narrow framing here, especially as many of these broader problems are a big challenge right now.

p. 13, It seems to me that starting with Transparency and only then talking about Fairness and Accountability might be more natural as the other two topics build in a sense on transparency.