Response to NIST AI RMF Concept Paper

25 January 2022

Elham Tabassi, Chief of Staff, Information Technology Laboratory
National Institute of Standards and Technology (NIST)
MS 20899, 100 Bureau Drive, Gaithersburg, MD 20899

Subject: NIST AI Risk Management Framework Concept Paper

Via email to AIframework@nist.gov

To Ms. Tabassi, and the entire NIST team developing the AI Risk Management Framework,

Thank you for the invitation to submit comments in response to the Concept Paper on the NIST AI Risk Management Framework (AI RMF or Framework). We offer the following submission for your consideration.

We are researchers affiliated with UC Berkeley, with expertise on AI research and development, safety, security and ethics. We previously submitted responses to NIST in September 2021 in response to the NIST AI RMF Request For Information.

We understand that the Concept Paper only lays out a high-level structure for the AI RMF, and we appreciate that NIST is working to provide additional detail. In our following comments on the Concept Paper, we focus on issues at the high level that the Concept Paper addresses; we generally do not consider lack of more-detailed guidance to be evidence that something is "missing" per se in the Concept Paper. Many of the recommendations we provided to NIST in September 2021 in response to the AI RMF RFI are at a level of detail beyond what the Concept Paper considers. We look forward to reading the more-detailed first draft of the AI RMF when NIST makes it available, and we will keep our more detailed previous recommendations in mind when preparing comments on the first draft of the AI RMF.

We strongly agree with NIST's statement in Section 1 of the Concept Paper that "Tackling scenarios that can represent costly outcomes or catastrophic risks to society should consider: an emphasis on managing the aggregate risks from low probability, high consequence effects of AI systems, and the need to ensure the alignment of ever more powerful advanced AI systems." We recommend that NIST retain and build upon these points as it formulates the first draft AI RMF. We believe it will be valuable to society, and in the interests of AI developers, for the AI RMF to face these issues and enable constructive management of these risks, starting with the first version of the AI RMF.

We also recommend that NIST plan to create at least one AI RMF Profile specifically oriented towards increasingly multi-purpose AI (including "foundation models"), such as BERT, CLIP, and GPT-3, which can serve as multi-purpose AI platforms underpinning many end-use applications. These increasingly powerful, increasingly multi-purpose advanced AI models have several qualitatively distinct properties compared to the more common, narrower machine learning models, such as potential to be applied to many sectors at once, and emergent properties that can provide unexpected capabilities but also unexpected risks of adverse events. These models could present corresponding catastrophic risks to society, e.g. of correlated robustness failures across multiple high-stakes application domains such as critical infrastructure, which could be constructively addressed by an AI RMF Profile focused on increasingly multi-purpose AI. In addition, a Profile could facilitate the management of important underlying risks of increasingly multi-purpose AI systems, in a way that does not rely on great certainty about each specific end-use application of the technology. The Concept Paper suggests that NIST also may create AI RMF Profiles for specific end-use applications of AI. Presumably, those could include end-use applications in critical infrastructure sectors, and/or other use-case categories that the EU AI Act designates as “high risk”. We believe that would be valuable for supporting risk management in those end-use applications, and could help the AI RMF achieve interoperability with other regulatory regimes such as the EU AI Act. However, the creation of an AI RMF Profile specifically for increasingly multi-purpose AI could provide industry with valuable risk-management best practices, addressing important risk-management issues of increasingly multi-purpose AI that extend beyond particular end-use applications. For example, a profile for increasingly multi-purpose AI could provide guidance on sharing of AI RMF responsibilities between researchers and vendors creating cutting-edge increasingly multi-purpose AI models and offering AI platforms/APIs based on those models in a manner that allows many different end uses, and developers building upon the AI platforms for specific end-use applications using vendor-provided information that may not be customized for their own application area.

We wholeheartedly agree with the statement on p. 3 of the Concept Paper that “if handled appropriately, AI technologies hold great potential to uplift and empower people and to lead to new services, support, and efficiencies for people and society.” We believe that AI offers enormous potential to benefit everyone. We believe this positive potential can be realized with appropriate risk management, and that the NIST AI RMF can play a vital role in helping AI decision makers to do that. We aim to helpfully contribute to NIST’s AI RMF development efforts with our comments.

In the following sections, we provide more in-depth comments, first regarding the questions posed by NIST in the AI RMF Concept Paper, and then on specific passages in the NIST AI RMF Concept Paper.

Thank you again for the opportunity to comment on the AI RMF Concept Paper. If you need additional information or would like to discuss further, please contact Anthony Barrett at anthony.barrett@berkeley.edu. In any case, we look forward to further engagement with NIST as you proceed on the AI RMF development process.

Our best,

Anthony M. Barrett, Ph.D., PMP
Visiting Scholar
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Thomas Krendl Gilbert, Ph.D.
Research Affiliate
Center for Human-Compatible AI, UC Berkeley

Dan Hendrycks
Ph.D. Candidate
Berkeley AI Research Lab, UC Berkeley

Jessica Newman
Director
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Brandie Nonnecke, Ph.D.
Director
CITRIS Policy Lab, CITRIS and the Banatao Institute, UC Berkeley

Richmond Y. Wong, Ph.D.
Postdoctoral Scholar
Center for Long-Term Cybersecurity, UC Berkeley

Our comments on questions posed by NIST in the AI RMF Concept Paper

– Is the approach described in this concept paper generally on the right track for the eventual AI RMF?

Response/Comment:
Yes, broadly speaking, we believe that the proposed approach has valuable potential for enabling the eventual AI RMF to constructively address a number of critical issues, in ways that we expect will meet the expectations for the AI RMF mentioned in Section 4 of the Concept Paper. For example, we agree with the statement in Section 1 of the Concept Paper that “Tackling scenarios that can represent costly outcomes or catastrophic risks to society should consider: an emphasis on managing the aggregate risks from low probability, high consequence effects of AI systems, and the need to ensure the alignment of ever more powerful advanced AI systems.” We believe it would be appropriate for the AI RMF to face these issues and enable constructive management of these risks, starting with the first version of the AI RMF.

Suggested change:
We recommend that NIST continue to include these issues, and provide additional related guidance as appropriate, in the draft AI RMF.

– Are the scope and audience (users) of the AI RMF described appropriately?

Response/Comment:
Yes, broadly speaking, we believe that the scope and audience of the AI RMF are described appropriately. We believe it is valuable that Section 2 of the Concept Paper mentions “people who experience potential harm or

inequities affected by areas of risk that are newly introduced or amplified by AI systems” as an important audience for AI RMF guidance, and “affected communities” as a stakeholder group.

Suggested change:
We recommend that NIST continue to include these groups within AI RMF audiences and stakeholders in the draft AI RMF.

– Are AI risks framed appropriately?

Response/Comment:
Yes, we generally agree with the framing of AI risks, e.g., in the Concept Paper (p. 3, lines 9-10) as “the composite measure of an event’s probability of occurring and the consequences of the corresponding events.”

We also understand that as mentioned on p. 3, lines 10-15 of the Concept Paper, NIST intends to use a definition of “risk” that includes both potentially positive and negative consequences of events, instead of just potentially negative impacts as we typically use "risk" in safety risk management or in many NIST information/cyber security guidance documents. Our sense is that with this broader usage, NIST is aiming for consistency with ISO risk management standards (e.g. ISO 31000), which define risk to include both potentially positive and negative outcomes. We understand that may be valuable for helping to ensure interoperability of the NIST AI RMF with other standards such as ISO 31000. We also believe it should be workable, especially if NIST adds sufficient clarifying terms such as “adverse” to the word “risk” when focusing on risks with negative consequences.

Suggested change:
We recommend that NIST try to use terms such as "adverse consequences" or “adverse risks” instead of just "consequences" or “risks”, whenever it would not seem sufficiently clear that NIST AI RMF terms such as “catastrophic risks to society” focus on negative consequences of risks.

– Will the structure – consisting of Core (with functions, categories, and subcategories), Profiles, and Tiers – enable users to appropriately manage AI risks?

Response/Comment:
Yes, we expect that the structure as laid out in Section 5 of the Concept Paper could enable users to appropriately manage AI risks. For example, creating Profiles could enable NIST to target in-depth guidance in the form of Profiles for particular types of AI such as multi-purpose AI, or high-stakes end-use applications such as in critical infrastructure, while minimizing AI RMF usage costs for a larger number of AI systems that do not present the same types of issues. Please also see our other comments below regarding passages on Page 7 of the Concept Paper and categories or subcategories that NIST could add.

– Will the proposed functions enable users to appropriately manage AI risks?

Response/Comment:
Yes, broadly speaking, we expect that the functions as laid out in Section 5 of the Concept Paper (in combination with appropriate guidance in other parts of the AI RMF) could help users to appropriately manage AI risks.

One caveat is that for some audiences, the name of the function “Measure” may seem to imply requirements for levels of quantification and/or certainty that would be unrealistic to expect for novel AI systems, and that would be unnecessary for some reasonable risk management steps.

Suggested change:
We recommend adding text clarifying that the function “Measure” can include types of risk analysis, risk rating, etc. that do not require levels of quantification and/or certainty that would be unrealistic to expect for novel AI systems of activities, and that would be unnecessary for some reasonable risk management steps.

– What, if anything, is missing?

Response/Comment:
At a high level, we have not identified major items as missing per se from the AI RMF as outlined in the Concept Paper. Please see our other comments below for more specific suggestions on how to expand on the Concept Paper for the draft AI RMF, e.g. regarding passages on Page 7 of the Concept Paper and categories or subcategories that NIST could add.

Our comments on specific passages in the NIST AI RMF Concept Paper

Page 2, Line 12
Response/Comment:
The sentence “AI risk management follows similar processes as other disciplines” seems a bit unclear.

Suggested change:
Change “as other disciplines” to "...as risk management for other disciplines" or a similar wording to clarify your intent.

Page 2, Lines 14-17
Response/Comment:
We strongly agree with NIST’s statement in Section 1 of the Concept Paper that “Tackling scenarios that can represent costly outcomes or catastrophic risks to society should consider: an emphasis on managing the aggregate risks from low probability, high consequence effects of AI systems, and the need to ensure the alignment of ever more powerful advanced AI systems.” (These statements are in line with key recommendations from our submissions in response to the NIST AI RMF RFI.) As many leading AI researchers have argued, in some foreseeable situations AI systems could pose catastrophic risks to society,

and potential for these risks will likely grow with increasing power of advanced AI (see, e.g., Russell 2019 and Bommasani et al. 2021). We believe it will be valuable to society, and in the interests of AI developers, for NIST to include these issues in the scope of the AI RMF, and to enable constructive management of these risks, starting with the first version of the AI RMF.

Suggested change:
We recommend that NIST retain and build upon these points as NIST formulates the first draft AI RMF, so that the scope addressed by the AI RMF clearly includes both 1) catastrophic risks to society, and 2) the need to ensure the alignment of ever more powerful advanced AI systems.

References:

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Kohd PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K, and Liang P (2021), On the Opportunities and Risks of Foundation Models. *arXiv*, https://arxiv.org/abs/2108.07258

Russell S (2019) Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Page 3, Lines 4-6
Response/Comment:
We believe it is valuable that NIST states in Section 2 of the Concept Paper that “All stakeholders should be involved in the risk management process”, and that stakeholders may include “affected communities.” This is consistent with best practices for effective risk identification; see e.g., guidance on including stakeholders during project risk identification (PMI 2017, section 11.2), as well as guidance on the ranges of types of stakeholders to include when identifying potential types of AI harm (Microsoft 2020). The diversity of perspectives from such approaches can help identify a greater breadth and depth of risks and risk-mitigation measures that otherwise could be missed by a team without the same perspectives.

Suggested change:
We recommend that NIST retain and build upon these points as NIST formulates the first draft AI RMF.

References:

Microsoft (2020) Foundations of assessing harm, Microsoft,
https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

PMI (2017) Guide to the Project Management Body of Knowledge, Sixth Edition, Project Management Institute, Newtown Square, PA

Page 3, Lines 17-18

Response/Comment:

We agree with NIST that AI risk management should include both minimizing negative impacts, and identifying opportunities to maximize positive impacts. However, this framing may imply assumption of the development and deployment of an AI system. It may be worth noting here that sometimes the outcome of a risk management process may also lead to an “off ramp” that precludes the development or deployment of a particular AI system or in a particular context. (We revisit this idea in our comment below on Page 4, Lines 31-34.)

Suggested change:

We recommend that NIST retain or expand on the language on Page 3, Lines 17-18 of the Concept Paper, to encourage organizations to consider entirely avoiding AI systems that pose unacceptable risks to rights, values, or safety.

Page 3, Lines 21-23

Response/Comment:

NIST has often stated the importance of measurability of aspects of AI system trustworthiness for the AI RMF. Here in Section 3 (Page 3, Lines 21-23), the Concept Paper seems to provide a rough definition of measurability, i.e. something that can "indicate AI system trustworthiness in meaningful, actionable, and testable ways." Presumably that means something that can be checked or verified, e.g. by a programmer's automated code test or by an auditor reviewing relevant information, but not necessarily quantitative.

Suggested change:

We have two main recommendations related to measurability. First, we recommend that NIST expand on the implicit meaning of measurability provided in Section 3 of the Concept Paper, to provide a clearer definition and/or examples of NIST’s intended meaning of measurability in the AI RMF.

Second, we recommend that NIST also provide guidance in the AI RMF on reasonable risk management steps for cases where risk may not be measurable yet, or at least not with precise quantitative methods. Some other stakeholders made similar recommendations in submissions to NIST, responding to the NIST AI RMF RFI. For example, the U.S. Chamber of Commerce's Technology Engagement Center ("C_TEC") stated that "The RMF should specifically address situations where risk cannot be measured and offer guidance on reasonable steps for mitigating that risk without limiting innovation and investments in new and potentially beneficial AI technologies." (C_TEC 2021) In addition, the Future of Life Institute stated that "As a matter of risk management, it would be especially negligent to ignore uncertain risks if the possible harms could be irreversible and catastrophic." (FLI 2021)

References:

C_TEC (2021) Re: Artificial Intelligence Risk Management Framework Request for Information. U.S. Chamber of Commerce's Technology Engagement Center. https://www.nist.gov/document/ai-rmf-rfi-comments-us-chamber-commerces-technology-engagement-center-ctec

FLI (2021) Response to the request for Information: Artificial Intelligence Risk Management Framework. Future of Life Institute. https://www.nist.gov/system/files/documents/2021/09/17/ai-rmf-rfi-0106.pdf

Page 3, Lines 32, 39

Response/Comment:
Among other attributes, we agree that the AI RMF should "be clear" and "be easily usable." Enabling ease of use of AI RMF materials for diverse stakeholders will help stakeholders to realize the great potential benefits of AI while constructively managing adverse risks and minimizing risk-management costs. For example, NIST should consider including implementation guides, and ensuring that AI RMF documentation adequately explains all terms such as "Implementation Tiers" for audiences not familiar with the Cybersecurity Framework.

Suggested change:
We recommend that NIST retain and build upon these points, working to maximize usability and user-friendliness of guidance materials, as NIST formulates the first draft AI RMF. For example, check that the AI RMF documentation adequately explains all terms such as "Implementation Tiers" for audiences not familiar with the Cybersecurity Framework.

Page 3, Line 41

Response/Comment:
We agree with the statement in Section 4 of the Concept paper that the AI RMF should be "appropriate for both technology agnostic (horizontal) as well as context-specific (vertical) use cases." However, it seems unclear whether NIST intends "technology agnostic" in the sense that a government procurement official might mean (where they specify requirements and leave it up to vendors to propose technologies to satisfy requirements), or whether NIST intends this to at least partly address increasingly multi-purpose AI that can pose important risks extending beyond a specific use case. In any case, we believe it would be valuable for the AI RMF to support management of risks from a perspective that is relatively agnostic regarding end-use applications of an AI system, but may be oriented toward classes of technologies such as multi-purpose AI. For example, some types of increasingly powerful, multi-purpose and advanced AI pose risk-management issues that extend beyond particular end-use applications, and we believe it could be appropriate for NIST to create AI RMF Profiles specifically oriented towards managing important underlying risks of increasingly multi-purpose AI or other technologies in a way that does not rely on great certainty about each specific end-use application of the technology.

Suggested change:
We recommend that NIST rephrase the Concept Paper passage in question, to include "end use application agnostic" (either in addition to, or instead of, "technology agnostic").

Page 4, Lines 31- 34
Response/Comment:
We believe it is valuable that Section 5.1.1 of the Concept Paper suggests that model management decision-making include “decisions that an AI solution is unwarranted or inappropriate versus the status quo, per a qualitative or more formal quantitative analysis of benefits, costs, and risks, and to stop development or to refrain from deployment.” We recommend that the Framework encourage organizations to consider entirely avoiding AI systems that pose unacceptable risks to rights, values, or safety.

Suggested change:
We recommend that NIST retain or expand on the language in Section 5.1.1 of the Concept Paper when drafting the AI RMF, e.g., to encourage organizations to consider entirely avoiding AI systems that pose unacceptable risks to rights, values, or safety.

Page 4, Line 35
Response/Comment:
Note 1 of the Concept Paper states that “Context refers to the domain and intended use, as well as scope of the system….” We believe there can be drawbacks in employing singular "use" or "use case" language in ways that imply that consideration of a single intended use of an AI system, especially when assessing increasingly multi-purpose AI that can be employed in many end-use applications. For such AI, focusing on a single intended “use” could overlook many important beneficial opportunities as well as risks of adverse events.

Suggested change:
We recommend that NIST either employ terminology such as AI system “uses”, “use cases” instead of “use”, or make other edits or add notes, to avoid implying that all AI systems would have a single intended use.

Page 4, Lines 36-37
Response/Comment:
Note 1 of the Concept Paper also states that AI system context could be “associated with a timeframe, a geographical area, social environment, and cultural norms within which the expected benefits or harms exist, specific sets of users along with expectation of users….” However, a status quo “context” could treat social context in an overly static way. If the status quo is interpreted as current day cultural norms and ethical standards, and AI systems are built to reflect those standards, it could be difficult to change those systems years down the road if the ethical standards shift.

Suggested change:
We recommend that material on Mapping Context include recognition that social contexts can change over time, which might mean that the mapping of context is an activity that needs to occur on a repeated or iterative basis, and that risk analysis activities such as scenario analysis should consider how contexts might change over time.

Page 5, Lines 4-6
Response/Comment:
We agree with Note 3 in Section 5.1.1 of the Concept Paper, which states that the AI RMF “Map” function (to establish context and enumerate risks related to the context) “should be performed by a team who is sufficiently diverse and multidisciplinary, representing multiple departments of the organization, and ideally includes a sufficiently diverse set of stakeholders from outside the organization.” The diversity of perspectives from such approaches can help identify a greater breadth and depth of risks that otherwise could be missed by a team without the same perspectives.

Suggested change:
We recommend that NIST retain and expand upon the Concept Paper Note 3 when writing guidance for the draft AI RMF. NIST may find it useful to adapt guidance on including stakeholders during project risk identification (e.g., PMI 2017, section 11.2), as well as guidance on the ranges of types of stakeholders to include when identifying potential types of AI harm (e.g., Microsoft 2020).

References:

Microsoft (2020) Foundations of assessing harm, Microsoft, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

PMI (2017) Guide to the Project Management Body of Knowledge, Sixth Edition, Project Management Institute, Newtown Square, PA

Page 5, Lines 24-34
Response/Comment:
We agree with the Concept Paper’s inclusion of governance in the AI RMF, as well as with the framework approach where governance is defined as a function of its own and is also integrated into the other functions.

Suggested change:
We recommend that NIST retain and expand upon this in the draft AI RMF.

Page 5, Lines 32-34
Response/Comment:
We agree with the statement in Note 6, Section 5.1.1 of the Concept Paper, that “Effective risk management cannot occur where governance is robust only in the early stages of an AI system, and not as the AI system evolves or is updated over time.” Our only concern is that the use of double negatives in this sentence (“cannot…and not”) may be slightly confusing.

Suggested change:
We recommend editing this passage to avoid double negatives, e.g., “For greatest effectiveness of risk management, governance should be robust in the early stages of an AI system, and should continue as the AI system evolves or is updated over time.”

Page 6, Lines 2-5
Response/Comment:
The Concept Paper suggests that NIST may create AI RMF Profiles for specific end-use applications of AI. Presumably, those could include end-use applications in critical infrastructure sectors, and/or other use-case categories, such as those that the EU AI Act designates as "high risk." We believe that would be valuable for supporting risk management in those end-use applications, and could help the AI RMF achieve interoperability with other regulatory regimes such as the EU AI Act.

We also believe it would be valuable for NIST to create at least one AI RMF Profile specifically oriented towards managing the broad context and associated risks of increasingly multi-purpose AI (including "foundation models"). These models, such as BERT, CLIP, and GPT-3, can serve as multi-purpose AI platforms underpinning many end-use applications. These increasingly powerful, increasingly multi-purpose advanced AI models are the focus of cutting-edge research and have several qualitatively distinct properties compared to the more common, narrower machine learning models, such as potential to be applied to many sectors at once, and emergent properties that can provide unexpected beneficial capabilities but also unexpected risks of adverse events. An AI RMF Profile for increasingly multi-purpose AI could address important underlying risks and early-development risks of such technologies in a way that does not rely on great certainty about each specific end-use application of the technology. For example, the Profile could provide Map or Measure function guidance, such as on assessing potential for catastrophic risks to society such as correlated robustness failures across multiple high-stakes application domains such as critical infrastructure. (For more on the capabilities and risks of increasingly multi-purpose and advanced AI, see e.g. Bommasani et al. 2021 and Russell 2019.)

Guidance in the AI RMF Profile for increasingly multi-purpose AI could be based in part on examples of assessments and/or risk management controls already implemented by market leaders such as DeepMind and OpenAI. For example, OpenAI's 2019 announcement of GPT-3 included enumeration of several categories of potential misuse cases (OpenAI 2019), which apparently informed OpenAI's decisions on disallowed/unacceptable use-case categories of applications based on GPT-3 (OpenAI 2020). DeepMind's 2021 announcement of their large language model Gopher also included consideration of safety risks and mitigation (Rae et al. 2021).

Creation of an AI RMF Profile specifically for increasingly multi-purpose AI could provide industry with valuable risk-management best practices addressing their unique issues. For example, the Profile could provide guidance on sharing of AI RMF responsibilities between researchers and vendors creating cutting-edge increasingly multi-purpose AI models and offering AI platforms/APIs based on those models in a manner that allows many different end uses, and developers building upon the AI platforms for specific end-use applications using vendor-provided information that may not be customized for their own application area. For the increasingly multi-purpose AI being developed by market leaders like OpenAI and DeepMind, the structural choices at stake in AI development are just as significant and distinct from those that manifest in particular use cases. For example, the performance of systems of this scale generates normative indeterminacies that imply a need for more stakeholder feedback (Dobbe et al. 2021). We also believe it would be appropriate to carry out more in-depth risk assessment with longer time horizons, at more points in the AI system life cycle, and to implement other more extensive risk-mitigation controls, etc. for increasingly multi-purpose models than for AI with more limited capabilities. For example, it could be valuable for red teams to conduct more extensive interaction with AI systems to identify emergent properties of such systems, which are more likely with large-scale models, though it also may be more difficult to detect emergent hazardous capabilities of increasingly advanced AI (Hendrycks et al. 2021 p. 7).

We believe that most AI systems would be easily identified either as “narrow AI” systems, or as end-use applications that adapt multi-purpose AI models, for which it would be appropriate for the end-use application developer to utilize AI RMF guidance applicable to those specific systems or end-use applications without requiring the end-use application developer to perform assessment of an entire multi-purpose AI model. Creation of an AI RMF Profile for increasingly multi-purpose AI would allow the AI RMF to provide appropriately targeted guidance for a small number of increasingly multi-purpose AI while minimizing costs for a large number of other, more narrow AI systems.

Although some other AI risk frameworks, such as the EU AI Act, seem highly focused on AI end-use application areas, we believe NIST has already created precedents for analogues to NIST AI RMF Profiles aimed at issues that cut across end-use application areas. For the Cybersecurity Framework, NIST (2021) provides several example Profiles for industry sectors (e.g. position, navigation and timing services) that seem analogous to AI end use application categories. However, other NIST Cybersecurity Framework Profiles focus on critical issues (e.g. ransomware risk management) that extend beyond specific software application end-use categories.

Suggested change:
We recommend that NIST plan to create one or more AI RMF Profiles for increasingly multi-purpose AI systems, for managing important underlying risks of such technologies in a way that does not rely on great certainty about each specific end-use application of the technology. For more of our suggestions on how to do that, please see our “Page 6, Lines 2-5 Response/Comment” above.

References:

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Kohd PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K, and Liang P (2021), On the Opportunities and Risks of Foundation Models. *arXiv*, https://arxiv.org/abs/2108.07258

Dobbe R, Gilbert TK, and Mintz Y (2021) Hard choices in artificial intelligence. *Artificial Intelligence* 300:103555, 2021.

Hendrycks D, Carlini N, Schulman J, and Steinhardt J (2021) Unsolved Problems in ML Safety. *arXiv*, https://arxiv.org/abs/2109.13916

NIST (2021) Examples of Framework Profiles. National Institute of Standards and Technology. https://www.nist.gov/cyberframework/examples-framework-profiles

OpenAI (2019) Better Language Models and Their Implications. OpenAI, https://openai.com/blog/better-language-models/

OpenAI (2020) Usage guidelines. OpenAI, https://beta.openai.com/docs/usage-guidelines

Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, Aslanides J, Henderson S, Ring R, Young S, Rutherford E, Hennigan T, Menick J, Cassirer A, Powell R, van den Driessche G, Hendricks LA, Rauh M, Huang P-S, Glaese A, Welbl J, Dathathri S, Huang S, Uesato J, Mellor J, Higgins I, Creswell A, McAleese N, Wu A, Elsen E, Jayakumar S, Buchatskaya E, Budden D, Sutherland E, Simonyan K, Paganini M, Sifre L, Martens L, Li XL, Kuncoro A, Nematzadeh A, Gribovskaya E, Donato D, Lazaridou A, Mensch A, Lespiau J-P, Tsimpoukelli M, Grigorev N, Fritz D, Sottiaux T, Pajarskas M, Pohlen T, Gong Z, Toyama D, de Masson d'Autume C, Li Y, Terzi T, Mikulik V, Babuschkin I, Clark A, de Las Casas D, Guy A, Jones C, Bradbury J, Johnson M, Hechtman B, Weidinger L, Gabriel I, Isaac W, Lockhart E, Osindero S, Rimell L, Dyer C, Vinyals O, Ayoub K, Stanway J, Bennett L, Hassabis D, Kavukcuoglu K and Irving G (2021) Scaling Language Models: Methods, Analysis & Insights from Training Gopher. DeepMind, https://storage.googleapis.com/deepmind-media/research/language-research/Training%20Gopher.pdf

Russell S (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking

Page 6, Lines 12-18

Response/Comment:

We believe it is valuable that Figure 1 in the Concept Paper indicates an AI system life cycle that is cyclical, roughly consistent with iterative/Agile etc. development methodologies. We also believe it is valuable that the caption for Figure 1 in the Concept Paper states that example activities in the Deployment phase include user feedback and override, and post deployment monitoring. However, it is unclear whether the Deployment stage includes independent auditing and user redress. It also could be valuable to encourage deployment with gradual, phased releases and efforts to detect misuse or problematic anomalies. For example, OpenAI has used a staged-release approach to roll-outs of large language models such as GPT-2, partly to minimize risks of misuse (OpenAI 2019).

Suggested change:

We recommend that NIST retain and expand these aspects in the draft AI RMF. For example, we suggest that NIST explicitly mention independent auditing and user redress, as well as gradual, phased releases and detection of misuse or problematic anomalies, as part of the Deployment stage.

References:

OpenAI (2019), GPT-2: 6-Month Follow-Up. https://openai.com/blog/gpt-2-6-month-follow-up/

Page 7, regarding categories under the "Map" and/or "Measure" functions
Response/Comment:
For categories or subcategories under the Map and/or Measure functions, we believe it would be valuable for the AI RMF to assess important aspects that may extend beyond NIST's currently planned characteristics of trustworthy AI, and to prompt participation by end-users, affected communities, and stakeholders in the development process.

Suggested change:
First, we recommend categories or subcategories that more clearly denote assessment of NIST's planned list of characteristics of trustworthy AI. We also recommend considering adding categories or subcategories for other potential characteristics of trustworthy AI, or perhaps for subcategories of the AI capabilities, such as assessment of the extent to which an AI qualifies as multi-purpose or generally-capable AI. In addition, we recommend categories or subcategories that prompt more active forms of participation by members of civil society (e.g., being consulted or having input into aspects of the design). Last, we recommend adding categories about identifying known and expected limitations, what data it was trained on, and procedures used in making tradeoffs between functional and safety properties.

Page 7, regarding categories under the "Govern" function
Response/Comment:
For categories of risk management activities and outcomes related to the Govern function, we believe it would be valuable for the Framework to include a comprehensive set of governance mechanisms to help organizations mitigate identified risks. For example, as part of risk communication, it could be appropriate to report information using one or more of the "model cards", "datasheets" and "reward reports" frameworks (see Gebru et al. 2018, Mitchell et al. 2018, and Gilbert et al. forthcoming). It also would be valuable to provide associated guidance, e.g. on how to have people with appropriate expertise performing governance functions, and how to involve stakeholders appropriately.

Suggested change:
We recommend retaining the categories currently under the Govern function to include determining who should be responsible for implementing the Framework within each organization, and ongoing monitoring and evaluation mechanisms that protect against evolving risks from continually learning AI systems. We also recommend adding categories or subcategories for support for other governance mechanisms, including incident reporting, risk communication, complaint and redress mechanisms, independent auditing, and protection for whistleblowers.

References:

Gebru T, Morgenstern J, Vecchione B, Vaughan J W, Wallach H, Daumé III H, and Crawford K (2018), Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Gilbert T K, Dean S, Lambert N, and Zick T (forthcoming), Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems. UC Berkeley Center for Long Term Cybersecurity.

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji I D, and Gebru T (2019), Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* 2019, pp. 220-229.