Response to NIST AI RMF 2nd Draft and Initial Playbook

28 September 2022

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

Subject: NIST AI Risk Management Framework 2nd Draft and Initial Draft Playbook

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 2nd Draft of the NIST AI Risk Management Framework (AI RMF or Framework) and accompanying Initial Draft Playbook. We offer the following submission for your consideration.

We are researchers affiliated with UC Berkeley, with expertise on AI research and development, safety, security, policy, and ethics. We previously submitted responses to NIST in September 2021 on the NIST AI RMF Request For Information (RFI), in January 2022 on the AI RMF Concept Paper, and in April 2022 on the AI RMF Initial Draft.

Here are our key high-level comments and recommendations on the AI RMF 2nd Draft:

- We agree with NIST's statements in Section 1.1 and Appendix B of the AI RMF 2nd Draft that **AI innovations have great potential for benefits across society, but that AI systems also can present risks requiring particular approaches and considerations, such as to address emergent properties of AI systems and potential for unintended consequences at both an individual and societal scale.** We also agree with NIST's statements in Section 1.2 that the AI RMF should help organizations to manage those risks by applying the AI RMF, especially from the beginning of an AI system's lifecycle, with aims of reducing the likelihood and magnitude of negative impacts (and increasing the benefits) to individuals, groups, communities, organizations, and society.
    - **We recommend NIST keep these statements in the AI RMF.** These passages highlight distinctive opportunities and risks for AI, and ways in which the AI RMF can help organizations address those risks effectively.
- We agree with NIST's statements in Section 3.2.2 of the AI RMF 2nd Draft that **although the AI RMF "does not prescribe risk tolerance", the AI RMF can be used to prioritize risks** and determine which risks "call for the most urgent prioritization and

most thorough risk management process." **We also agree that "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. Conversely, the lowest-risk AI systems and contexts suggest lower prioritization."**
    - **We recommend NIST keep these statements in the AI RMF.** We understand that specifics of risk tolerance will depend on particular contexts including regulatory considerations. We also believe there is broad agreement on the importance of prioritizing the highest risks to individuals, groups, communities, organizations, and society, and that these include cases "where negative impacts are imminent, severe harms are actually occurring, or catastrophic risks are present". Moreover, there is precedent for NIST framework guidance prompting identification of risks with potentially catastrophic impacts: the NIST Cybersecurity Framework guidance on risk assessment points to NIST SP 800-53 RA-3, which in turn references NIST SP 800-30; the impact assessment scale in Table H-3 of SP 800-30 includes criteria for rating an expected impact as a "catastrophic adverse effect" to individuals, organizations, or a society.
- We recommend refinement of several aspects of the Map, Measure and Manage functions.
    - For the Map function, **we recommend NIST clarify that in addition to describing intended beneficial "use cases" for an AI system as part of Map activities, it is valuable for Map activities to include identification of other potentially beneficial uses of an AI system, as well as negative "misuse/abuse cases".** This would better address both positive and adverse risks of reasonably foreseeable "off label" uses, beyond an AI developer's or deployer's originally intended uses of an AI system. Identification of other potentially beneficial uses should be a clearer part of Map 1.1 on system-use understanding and documentation, and possibly also Map 5.1 on impact identification. Misuse/abuse case identification should be a clearer part of Map 5.1 on impact identification, and possibly also Map 1.1 on system-use understanding and documentation and Measure 2.7 on AI system resilience and security evaluation. We believe it is generally worthwhile to identify reasonably foreseeable uses and misuses of AI systems as part of risk management.
    - For Measure 1.1 on measurement of risks and Manage 1.3 on responses to risks, **we recommend NIST revise their definitions from addressing the "most significant risks" to a broader set of "identified risks".** The equivalent to Manage 1.3 in the AI RMF Initial Draft stated, "Responses to enumerated risks are identified and planned. Responses can include mitigating, transferring or sharing, avoiding, or accepting AI risks." However, Manage 1.3 in the AI RMF 2nd Draft stated, "Responses to the most significant risks, identified by the Map function, are developed, planned, and documented. Risk response options can include mitigating, transferring, sharing, avoiding, or accepting." Measure 1.1 was similarly changed from recommending measurement of enumerated risks to only

recommending measurement of the most significant risks. The change from "enumerated risks" to "the most significant risks" could be counterproductive. It may be that organizations decide to monitor and track, or simply accept, many identified risks that are not the most significant risks. However, in some cases it may be cost-effective and worthwhile to mitigate multiple risks that are not the most significant risks. It seems prudent for organizations to choose how to address all identified risks, rather than to simply ignore identified risks that are not deemed the most significant risks.

We also commend NIST on the many improvements between the AI RMF Initial Draft and the 2nd Draft, including the following:

- Removing the previous, confusing split between "technical characteristics", "socio-technical characteristics", and "principles" of trustworthy AI, and instead confirming that all of them require and involve human judgment, and providing guidance for addressing them
- Clarification of various AI system stakeholders
- Discussion of how AI risks differ from traditional software risks, including the higher degree of difficulty in predicting failure modes for emergent properties of large-scale pre-trained models
- Adding environmental and ecosystem harms to the list of examples of potential harms under "harm to a system"
- Addition of documentation items throughout
- Addressing third-party and supply chain risks as part of each function
- Addition of test, evaluation, verification, and validation (TEVV) activities throughout, including to monitor and assess risks of emergent properties of AI systems

In the following sections, we provide additional detailed comments, first regarding the questions posed by NIST in the AI RMF Initial Draft Playbook, and then on specific passages in the NIST AI RMF 2nd Draft and Initial Draft Playbook.

Thank you again for the opportunity to comment on the AI RMF 2nd Draft and accompanying Initial Draft Playbook. 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 Barrett, Ph.D., PMP
Visiting Scholar
AI Security Initiative, Center for Long-Term Cybersecurity, 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
Co-Director
AI Policy Hub, UC Berkeley

Mark Nitzberg, Ph.D.
Executive Director
Center for Human-Compatible AI, UC Berkeley

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

Our comments on questions posed by NIST in the Initial Draft Playbook

1. Its relative usefulness as a complementary resource to the AI RMF.

Response/Comment:
Broadly speaking, the Playbook clearly provides useful information that complements the main AI RMF document. It provides material about each subcategory that helps readers understand associated activities and outcomes, lists of specific actions to take, and considerations for transparency and documentation, as well as informative references. Many aspects of the Playbook could be refined through use, and we support NIST's efforts to do so.

We particularly commend the inclusion of the transparency and documentation sections, and encourage further refinement and expansion of these sections. We also recommend considering including available tools and toolkits within the resource sections of the playbook, for example drawing upon the soon to be available database of tools for trustworthy AI from OECD.AI.

2. Whether the guidance is actionable, especially as related to organization size.

Response/Comment:
Generally, we expect that the Playbook guidance will be actionable for organizations across a range of sizes.

3. Suggested presentation alternatives for the forthcoming first version of the Playbook, to ensure effectiveness and ease of use.

Response/Comment:
The Playbook's content is currently broken into many web pages, which makes it difficult to download and use or review as a single unified document.

Suggested Change:
Provide an option in the Playbook user interface to download a single pdf or functional equivalent (e.g. provide a way to have all of the Playbook's content on a single web page that can be printed to pdf).

Our comments on specific passages in the NIST AI RMF 2nd Draft

Pages ii, iii

Response/Comment:
We applaud NIST for its approach to the Playbook and Resource Center, which includes inviting stakeholder contributions, updates, and feedback, and planning on a contribution adjudication process using criteria stated here. This would efficiently serve the interests of all stakeholders in continually keeping AI RMF resources up to date, to incorporate new resources addressing new issues and new techniques.

Suggested Change:
It could improve outcomes of the adjudication process for NIST to plan on providing feedback to Playbook and Resource Center material submitters, and to allow re-submission of materials to NIST after revisions that substantively respond to NIST feedback.

Page 1, Section 1.1 (“Trustworthy and Responsible AI”) second paragraph

Response/Comment:
The sentence ”Risks to any software or information-based system apply to AI, including concerns related to cybersecurity, privacy, safety, and infrastructure” is true. However, the AI RMF is not just about risks to an AI system, it is also importantly about risks to individuals, society etc.

Suggested Change:
Modify "Risks to" to "Risks applicable to" and add “also”, so that the sentence in question reads as follows: “”Risks applicable to any software or information-based system also apply to AI, including concerns related to cybersecurity, privacy, safety, and infrastructure.”

Page 3, Section 1.3 and elsewhere throughout the document

Response/Comment:
The phrase ”trustworthy characteristics” appears several times in the document, but it could more clearly refer to characteristics of trustworthy AI, which seems like the concept that NIST intended to refer to.

Suggested Change:
Change "trustworthy characteristics" to either "trustworthiness characteristics" or "trustworthy AI characteristics" throughout the document.

Page 6, Figure 2

Response/Comment:
The column heading “Lifecycle” implies entries will list different AI lifecycle models (e.g. waterfall; iterative) but the entries in the column are a list of phases or stages within a single AI lifecycle.

Suggested Change:
Change the column heading “Lifecycle” to “Phase” or “Lifecycle Phase”.

Page 8, Section 3.2.1 “Risk Measurement”, third paragraph

Response/Comment:
The current sentence “Organizations will want to identify and track emergent risks and consider techniques for measuring them” may not be the most accurate.

Suggested Change:
Change the phrase "Organizations will want to identify and track emergent risks...." to "Organizations' risk management efforts will be enhanced by identifying and tracking emergent risks...."

Page 15, Section 4.4 (“Secure and Resilient”) last paragraph

Response/Comment:
The phrase here, “unexpected or adversarial use of the model or data” seems to point to consideration of potential abuse and/or misuse cases. Mentioning those terms here could helpfully prompt consideration of abuse and/or misuse cases. That could also build on best practices for consideration of adversarial misuse potential of an AI system such as in Microsoft (2021) and related software development guidance such as OWASP (2021), as well as in Section 3.1 of our paper Barrett et al. (2022).

Suggested Change:

Change "use" in this passage to "use (or abuse/misuse)".

References:
Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Microsoft (2021) Foundations of assessing harm. Microsoft, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

OWASP (2021) Abuse Case Cheat Sheet. OWASP, https://cheatsheetseries.owasp.org/cheatsheets/Abuse_Case_Cheat_Sheet.html

Page 15, Section 4.5; Page 21, Table 3; Page 25, Table 5 and elsewhere throughout the document

Response/Comment:
Comment: These passages include one or more of the following phrases when discussing an AI system: "intended use case", “intended purpose”, “use”, or “task”. We believe there can be drawbacks in employing singular terms such as "use", "use case", “purpose”, “task”, etc. in ways that could suggest only considering a single intended use of an AI system. AI systems can have multiple uses and it is worth identifying these as part of the Map and Manage functions, e.g. to enable policies disallowing specific uses that would present unacceptable risks. The drawbacks of assuming a single intended use would be especially important for increasingly general-purpose AI that can be employed in many end-use applications. For any AI system, and especially for increasingly multi-purpose or general-purpose AI systems such as BERT, CLIP, and GPT-3, focusing on a single intended use could overlook many important beneficial opportunities as well as risks of adverse events.

Suggested Change: We recommend that in these passages and throughout AI RMF documents, NIST generally either employ terminology such as AI system “uses” or “use cases” instead of “use” (and “purposes” instead of “purpose”, “tasks” instead of “task”, etc.) or include related notes, to avoid implying that all AI systems would have a single intended use.

Page 22 and elsewhere

Response/Comment: In several places in the AI RMF 2nd Draft (e.g. the definition of Map 5.3 on p. 22), NIST seems to use “impact” as interchangeable with “risk” or “potential impact”. This can be confusing to readers. A risk of an impact (i.e. a “risk” or “potential impact”) does not necessarily lead to an actually-observed impact.

Suggested Change: For each instance of “impact”, consider changing to “risk” or “potential impact” if those are really what NIST meant.

Page 24, Measure 2.5

Response/Comment: Two parts of the definition of Measure 2.5 could result in mistaken impressions or confusion. First, the phrase “Deployed product is demonstrated to be safe….” could imply that safety testing does not need to be conducted before deployment. It seems much more prudent to conduct safety testing before deployment. Second, the term “implicate” in the phrase “Safety metrics implicate system reliability and robustness….”, seems unclear and confusing.

Suggested Change:
First, instead of “Deployed product is demonstrated to be safe….”, use a phrase such as “The product to be deployed is demonstrated to be safe….” so it is clear that safety testing should be conducted before deployment. Second, use clearer wording than “implicate”, such as “indicate sufficient performance for”.

Our comments on specific passages in the Initial Draft Playbook

Map 1.1, regarding objectives mis-specification risks

Response/Comment:
In the “Actions” section of Map 1.1 in the AI RMF Playbook, the fourth bullet helpfully prompts consideration of “intended AI system design tasks along with unanticipated purposes.” As we note in our actionable-guidance paper (Barrett et al. 2022), many AI researchers regard system objectives specification (or alignment of system behavior with designer goals) as an aspect of AI trustworthiness that is already important for AI systems and whose importance will only increase as AI systems grow in scale and capabilities. Specification of an AI system's goals or objectives aims to align the system’s behavior with the designer’s intentions. Mis-specification risks can include cases where a system meets its literal goals but has unanticipated or unintended behaviors that cause harm. Rudner and Toner (2021) provide brief examples, such as social-media content recommendation machine-learning algorithms that learn to optimize user-engagement metrics by serving users with extremist content or disinformation. Rudner and Toner (2021, p. 10) also suggest accounting for worst-case scenarios, and considering the following questions for an AI system, as part of identifying mis-specification risks: “What objective has been specified for the system, and what kinds of perverse behavior could be incentivized by optimizing for that objective?”

Suggested Change:

In the “Actions” section of Map 1.1 in the AI RMF Playbook, under the fourth bullet that prompts consideration of “intended AI system design tasks along with unanticipated purposes”, add or adapt the following as a sub-bullet:

- “For the intended AI system tasks or objectives, what unintended perverse or adverse behaviors could be incentivized by using those objectives? Incorporate any new risks identified into other risk management steps such as in Map 5.1.”

References:

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Tim GJ Rudner and Helen Toner (2021) Key Concepts in AI Safety: Specification in Machine Learning. CSET, https://cset.georgetown.edu/wp-content/uploads/Key-Concepts-in-AI-Safety-Specification-in-Machine-Learning.pdf

Map 1.1 and Map 5.1, regarding identification of both positive and adverse risks of reasonably foreseeable “off label” uses

Response/Comment:

The draft Playbook guidance for the Map 1.1 Action section includes a bullet point for readers to “Collaboratively consider intended AI system design tasks along with unanticipated purposes.” However, we believe it would be valuable for Map guidance in the Playbook to more clearly prompt identification of other potentially beneficial uses of an AI system, as well as identification of negative “misuse/abuse cases”, beyond an AI developer’s or deployer’s originally intended uses of an AI system. This would better address both positive and adverse risks of reasonably foreseeable “off label” uses. Our recent paper (Barrett et al. 2022) could be listed as an informative reference for Map 1.1 and Map 5.1 for identification of other potentially beneficial uses of an AI system as well as negative “misuse/abuse cases”. The Playbook material for Map 1.1 and Map 5.1 also could adapt material in Section 3.1 of Barrett et al. (2022) to provide more direct guidance for identifying other potentially beneficial uses of an AI system as well as negative “misuse/abuse cases”.

In addition, we recommend adding high level consideration of the AI system's security vulnerabilities into Map 1.1, for example as part of the "limitations" of an AI system. NIST likely will touch on this in forthcoming Playbook guidance for Measure 2.7; however, it also informs the impacts and limitations of an AI system's purpose and may be useful to consider in brief at this stage.

Suggested Change:

First, as a small step for the AI RMF Playbook, we recommend listing our actionable-guidance paper (Barrett et al. 2022) as an informative reference for Map 1.1 and Map 5.1 for identification

of other potentially beneficial uses of an AI system and identification of negative “misuse/abuse cases”.

Second, as a more extensive step for the AI RMF Playbook, consider adding or adapting the following into the AI RMF Playbook sections for Map 1.1 or Map 5.1; the following is directly copied or adapted from Section 3.1 of our actionable-guidance paper (Barrett et al. 2022):

In addition to identifying an intended use case of an AI system, also:

- **Identify other potentially beneficial use cases or applications of an AI system,** if your organization has not already done that (e.g., as part of assessing business opportunities) in a reasonably systematic manner.
    - Use appropriate methods in order to:
        - Identify other potentially beneficial uses or applications that your organization might want to intentionally pursue
        - Identify other potentially beneficial uses or applications that your intended users might attempt, which your organization might not intentionally pursue
    - For methods to identify other potentially beneficial uses or use cases, consider methods such as:
        - Brainstorming
        - Reviewing publications that discuss current and potential uses of other AI systems. This can include industry publications and references such as ISO/IEC TR 24030.
        - Reviewing available information on competitors’ uses of AI systems with similar characteristics
- **Identify misuse/abuse cases and types of adversary attack of an AI system**, using threat modeling, red team methods or other procedures as appropriate.
    - Identify applicable misuse/abuse case types as listed in the following resources or in other relevant resources as appropriate:
        - MITRE ATLAS (MITRE 2021a) for AI-related adversary tactics and techniques
        - Microsoft (2021a) for machine learning-related failure modes and threat taxonomy, and Microsoft (2021b, 2022a) for additional considerations in AI system threat identification and threat modeling
        - MITRE ATT&CK® (MITRE 2021b) for general cybersecurity adversary tactics and techniques
        - NIST SP 800-30 Appendix E for general information security threat events for both adversarial tactics, techniques and procedures as well as non-adversarial (i.e. accidental) threat events
        - For illustrative examples of AI misuse cases, see, e.g., the “Policy Implications” section of the OpenAI (2019a) announcement of GPT-2, the “Case Studies” section of the MITRE ATLAS (MITRE 2021a) home page, or Brundage et al. (2018).
        - Provide mechanisms such as email, web forms, or other hotlines for internal and external stakeholders to report concerns about potential

types of AI misuse/abuse, or to report incidents of misuse/abuse, vulnerabilities discovered, etc. along with appropriate protections for stakeholders making reports. For example, see the vulnerability reporting procedures and safe-harbor policy of OpenAI (2020).

- **For each potential misuse case, consider whether the misuse case could be a way for an adversary to attack your AI system, or a way for an adversary to use your AI system to attack something/someone else.**
    - Compile a list of all misuse cases you identify, along with their key characteristics, such as whether the misuse case could be a way for an adversary to attack your AI system or a way for an adversary to use your AI system to attack something/someone else.
    - For more detailed example procedures on how to identify and analyze abuse cases, see OWASP guidance on identifying and prioritizing abuse cases for web-application software development (OWASP 2021). However, keep in mind that some adaptations of OWASP guidance will likely be appropriate for analyzing AI applications. For example, instead of basing abuse cases on the OWASP Top 10 web application security risk types, consider basing abuse cases on AI-relevant security risk types such as listed in MITRE ATLAS (MITRE 2021a).

When to identify use cases and misuse/abuse cases of an AI system:

- Identify potential use cases during early stages of your AI system lifecycle, such as plan and design, at minimum.
- Identify misuse cases during all major stages of your AI system lifecycle (or approximate equivalents in Agile/iterative development sprints), such as: plan, data collection, design, train/build/buy, test and evaluation, deploy, operate and monitor, and decommission.
- Plan to revisit use and misuse case identification at key intended milestones, or at periodic intervals (e.g., at least annually), whichever comes first.

For staffing to identify potential use cases and misuse cases of an AI system:

- Include members of each of the following functional teams (or equivalents) as appropriate:
    - For identifying ***both* potentially beneficial use cases *and* for identifying misuse cases** to consider in risk assessment or other assessments:
        - Product development, operations, human-computer interaction, user experience, policy, and ethics professionals
    - In addition, **for identifying misuse cases** to consider in risk assessment or other assessments:
        - Security
    - In addition, **for identifying other potentially beneficial use cases** to consider in risk assessment or other assessments:
        - Marketing and sales
- Consider including members of other teams as appropriate, such as:
    - Research and development (for additional technically-informed perspective on AI system capabilities and limitations)

    - External-facing teams and/or external stakeholders (for additional early identification of potential stakeholder concerns and other stakeholder perspectives)

After identifying use cases and misuse cases of an AI system:

- **For each identified potentially beneficial use case of an AI system:**
    - **If it seems plausible that your organization or your intended users might pursue/attempt those uses, consider carrying out impact assessments, risk assessments, or other assessments of those use cases, or incorporating these new uses into your overall risk assessment processes, as appropriate.** For example, if there is significant potential for using an AI system in a more sensitive application than originally intended, that could modify the overall assessment of risk of the AI system in question, even if its originally intended use case seems low risk.
        - E.g., consider characterizing each use case according to the OECD framework for the classification of AI systems (OECD 2022)
- **For each identified misuse/abuse case of an AI system:**
    - **Incorporate identified misuse/abuse cases and their key characteristics into your organization's harms modeling, security threat modeling, risk assessments, risk register, or related risk management processes as appropriate** (e.g., as threat events in context of risk assessments per NIST SP 800-30).
        - As part of AI system threat modeling and risk assessment, consider resources such as Microsoft (2022a) and Microsoft (2021b).

References:

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. ArXiv abs/1802.07228 (2018) https://arxiv.org/abs/1802.07228

Microsoft (2021a) Failure Modes in Machine Learning. Microsoft, https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning

Microsoft (2021b) AI/ML Pivots to the Security Development Lifecycle Bug Bar. Microsoft, https://docs.microsoft.com/en-us/security/engineering/bug-bar-aiml

Microsoft (2022a) Threat Modeling AI/ML Systems and Dependencies. Microsoft, https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml

MITRE (2021a) ATLAS, Adversarial Threat Landscape for Artificial-Intelligence Systems. https://atlas.mitre.org/

MITRE (2021b) ATT&CK. https://attack.mitre.org/

OECD (2022) OECD Framework for the Classification of AI Systems. OECD Digital Economy Papers, No. 323. Organisation for Economic Co-operation and Development, https://doi.org/10.1787/cb6d9eca-en

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

OpenAI (2020) Coordinated Vulnerability Disclosure Policy. OpenAI, https://openai.com/security/disclosure/

OWASP (2021) Abuse Case Cheat Sheet. OWASP, https://cheatsheetseries.owasp.org/cheatsheets/Abuse_Case_Cheat_Sheet.html

Map 5.1, regarding identification of potential impacts to individuals, groups, organizations and society

Response/Comment:
The Playbook material for Map 5.1 could provide much more direct guidance for identifying various types of potential impacts. Material in Section 3.2 of Barrett et al. (2022) prompts consideration of various factors that could lead to high consequences at a societal scale, and Section 3.3 prompts consideration of impacts to human rights. Barrett et al. (2022) could be listed as an informative reference, and also could be adapted into Playbook guidance for Map 5.1 for identification of various types of potential impacts to individuals, groups, organizations and society.

Suggested Change:
First, as a small step for the AI RMF Playbook, we recommend listing our actionable-guidance paper (Barrett et al. 2022) as an informative reference for Map 5.1 for identification of various types of impacts to individuals, communities, organizations and society.

Second, as a more extensive step for the AI RMF Playbook, consider adding or adapting the following to the AI RMF Playbook sections for identifying various types of impacts as part of Map 5.1; the following is directly copied or adapted from Sections 3.2 and 3.3 of our actionable-guidance paper (Barrett et al. 2022), and includes passages adapted from NIST SP 800-30 and NISTIR 8062 as well as UN (1948) and UN (2011):

- **Identify or assess reasonably foreseeable potential impacts to individuals, groups, organizations and society, as appropriate.**
    - Identify or assess reasonably foreseeable potential adverse impacts or harms of the following types:
        - To organizational operations, including:
            - Missions and functions
            - Image and reputation, including:
                - Loss of trust and reluctance to use the system or service
                - Internal culture costs that impact morale or productivity
        - To organizational assets, including legal compliance costs arising from problems created for individuals
        - To other organizations
        - To individuals, including impacts to health, safety, well-being, or fundamental rights
        - To groups, including populations vulnerable to disproportionate adverse impacts or harms
        - To the Nation or other societal impacts, including:
            - Damage to or incapacitation of a critical infrastructure sector
            - Economic and national security
            - Impacts on democratic institutions and quality of life
            - Environmental impacts
        - If appropriate for your context, also identify or assess more specific types of harms within the above categories (e.g., per NIST SP 800-30 Table H-2) or other types of harms (e.g., as outlined in Microsoft 2021a, 2021b, or in other resources in Appendix III of PAI 2022)
- **Aim to identify scenarios with high consequences for society, or with factors that could lead to high consequences**
    - **"High"** impact category description: The threat event could be expected to have a **severe or catastrophic** adverse effect on organizational operations, organizational assets, individuals, other organizations, or the Nation or society as a whole.
        - A severe or catastrophic adverse effect means that, for example, the threat event might: (i) cause a severe degradation in or loss of mission capability to an extent and duration that the organization is not able to perform one or more of its primary functions; (ii) result in major damage to organizational assets; (iii) result in major financial loss; or (iv) result in severe or catastrophic harm to individuals involving loss of life or serious life-threatening injuries.
        - **Factors that could lead to severe or catastrophic consequences for society include:**

- **Potential for correlated robustness failures or other systemic risks across high-stakes application domains** such as critical infrastructure or essential services[1]
- **Potential for other systemic impacts, which can be accumulated, accrued, correlated or compounded at societal scale**[2], e.g.:
    - **Potential for correlated bias** across large numbers of people or a large fraction of a society's population[3]
    - **Potential for many high-impact uses or misuses beyond an originally intended use case**, e.g., if an AI system is a cutting-edge large-scale language model, "foundation model" or another highly multi-purpose / general-purpose AI system[4], or if it enables recursive improvement of capabilities of cutting-edge AI system algorithms or architecture through code generation, architecture search, etc.[5]
- **Potential for large harms from mis-specified goals** (e.g., using over-simplified or short-term metrics as proxies for desired longer-term outcomes)[6]

---

[1] See, e.g., discussion of correlated failures of "foundation models" spanning multiple critical functions in Section 4.9.3 of Bommasani et al. (2021).

[2] This draws upon the following types of systemic risks to society as described by Mallah (2022) and FLI (2022):

- Accumulated risk: small harms accumulating over time to form a major harm;
- Accrued risk: where events that are low-probability in the short-term, but high-impact, can accrue and build to significant-probability in the medium term;
- Correlation risk: where there are adverse events that are not evident in unit tests or accuracy tests, but can be expected to emerge from correlated decisions or correlated actions with a large number of users, instances, or executions of the system;
- Latent risk: where harms that will not manifest significantly or at all on system training or release may still be expected to appear with distributional shift, new use cases, or qualitative shifts in capabilities arising from quantitative scaling;
- Compounding risk: where harms would be expected to manifest only when either other problems occur or unexpected, but conceivable conditions or interactions manifest.

Note that these types of systemic risks are not necessarily exclusive of one another.

[3] E.g., as discussed by Schwartz et al. (2022, p. 32): "The systemic biases embedded in algorithmic models can … be exploited and used as a weapon at scale, causing catastrophic harm."

[4] We believe that most AI systems could be readily identified as being in one of the following categories:

A. One of a few large-scale, cutting-edge, increasingly multi-purpose or general-purpose AI system platforms (including "foundation models"), such as BERT, CLIP, GPT-3, DALL-E 2, and PaLM.
B. A relatively narrow-purpose end-use application that builds on a multi-purpose AI model platform.
C. One of many small-scale and/or stand-alone narrow-purpose AI systems.

Category A presents substantial potential for systemic impacts to society (see, e.g., Bommasani et al. 2021).

[5] As the DeepMind paper on the software code-generation AI system AlphaCode stated, "Longer term, code generation could lead to advanced AI risks. Coding capabilities could lead to systems that can recursively write and improve themselves, rapidly leading to more and more advanced systems." (Li et al. 2022) For discussion of related issues, see, e.g., Russell (2019).

[6] For examples of mis-specified objectives, such as social-media content recommendation machine-learning algorithms that learn to optimize user-engagement metrics by serving users with

- **Identify potential or actual human rights impacts, per the Universal Declaration of Human Rights or UDHR (UN 1948), UN Guiding Principles on Business and Human Rights or UNGP (UN 2011) and related guidance (Nonnecke and Dawson 2021). Potential example questions and UDHR Articles to consider include:**
    - UDHR Article 2, including non-discrimination and equality before the law
        - How could an AI system's bias in data or unfair algorithmic decisions affect rights to equal protection and non-discrimination?
    - UDHR Article 3, including right to life and personal security
        - How could an AI system's algorithmic decisions affect the right to life and personal security?
    - UDHR Article 12, including privacy and protection against unlawful governmental surveillance
        - How could an AI system be used for surveillance, leading to loss of privacy or inadequate protection of personally identifiable information?
    - UDHR Articles 18 and 19, including freedom of thought, conscience and religious belief and practice, and freedom of expression and to hold opinions without interference
        - How could an AI system affect rights to express opinions or practice religion?
    - UDHR Articles 20 and 21, including freedom of association and the right to peaceful assembly
        - How could an AI system affect rights to association, peaceful assembly, and democratic participation in government?
    - UDHR Articles 23 and 25, including rights to decent work and to an adequate standard of living
        - How could an AI system affect rights to decent work, including effects on adequate standard of living via displacement of human workers?

References:

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang,

extremist content or disinformation, see, e.g., Rudner and Toner (2021). Identifying mis-specification risks can also be aided by considering the following questions for an AI system: "What objective has been specified for the system, and what kinds of perverse behavior could be incentivized by optimizing for that objective?" Rudner and Toner (2021, p. 10)

Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang , Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. ArXiv abs/2108.07258 (2021) https://arxiv.org/abs/2108.07258

FLI (2022) Subject: First Draft of the NIST AI Risk Management Framework. Future of Life Institute, https://www.nist.gov/system/files/documents/2022/05/19/Future%20of%20Life%20Institute.pdf

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu and Oriol Vinyals. (2022) Competition-Level Code Generation with AlphaCode. DeepMind, https://storage.googleapis.com/deepmind-media/AlphaCode/competition_level_code_generation_with_alphacode.pdf

Richard Mallah (2022) Remarks in *Panel 1 of Building the NIST AI Risk Management Framework: Workshop #2*, March 23-31, 2022, virtual, https://www.nist.gov/news-events/events/2022/03/building-nist-ai-risk-management-framework-workshop-2

Microsoft (2021a) Foundations of assessing harm. Microsoft, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

Microsoft (2021b) Types of harm. Microsoft, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/type-of-harm

Brandie Nonnecke and Philip Dawson. Human Rights Implications of Algorithmic Impact Assessments: Priority Considerations to Guide Effective Development and Use. (2021) Carr Center Discussion Paper Series

https://carrcenter.hks.harvard.edu/publications/human-rights-implications-algorithmic-impact-assessments-priority-considerations

PAI (2022) Publication Norms for Responsible AI. Partnership on AI, https://partnershiponai.org/workstream/publication-norms-for-responsible-ai/

Tim GJ Rudner and Helen Toner (2021) Key Concepts in AI Safety: Specification in Machine Learning. CSET, https://cset.georgetown.edu/wp-content/uploads/Key-Concepts-in-AI-Safety-Specification-in-Machine-Learning.pdf

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

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. (2022) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. SP 1270. National Institute of Standards and Technology, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

UN (1948) Universal Declaration of Human Rights (UDHR). United Nations, https://www.un.org/en/about-us/universal-declaration-of-human-rights

UN (2011) UN Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework. United Nations Office of the High Commissioner on Human Rights, https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

Map 5.2, “About” section

Response/Comment:
In the “About” section of Map 5.2 in the AI RMF Playbook, it appears that NIST intended for each instance of "likelihood" to be "likelihood and magnitude" in this passage. Making that correction to add "and magnitude" would make the passage consistent with the description of the Map 5.2 subcategory: "Likelihood and magnitude of each identified impact based on expected use, past uses of AI systems in similar contexts, public incident reports, stakeholder feedback, or other data are identified and documented." That also would make the description of Map 5.2 consistent with widely accepted best practices for risk assessment, which include assessing magnitude of potential impacts as well as likelihood of impacts. Ignoring the magnitude of potential impacts, and considering only likelihood of potential impacts, could result in overlooking risks of events that may not occur on a daily basis but can have high consequences for individuals, organizations and society when they occur.

Suggested Change:

Add "and magnitude" to the “About” section of Map 5.2 in the AI RMF Playbook, so that the “About” section of Map 5.2 reads as follows: “The likelihood and magnitude of AI system impacts identified in Map 5.1 should be evaluated. Potential impacts should be documented and triaged. Likelihood and magnitude estimates may then be assessed and judged for go/no-go decisions about deploying an AI system. If an organization decides to proceed with deploying the system, the likelihood and magnitude estimates can be used to assign oversight resources appropriate for the risk level.”

Map 5.2, References section

Response/Comment:
We appreciate that NIST has listed our actionable-guidance paper (Barrett et al. 2022) as an informative reference for Map 5.2, which includes evaluating magnitude of identified impacts. Section 3.2 of our paper provides an impact magnitude rating scale that includes consideration of societal-scale impact factors, which would usefully inform go/no-go decisions as part of Map activities. (This also should provide benefits of consistency with cybersecurity event impact rating in NIST SP 800-30. The impact magnitude rating categories in Section 3.2 of Barrett et al. 2022 document closely follow the impact magnitude rating Table H-3 of NIST SP 800-30, except that we use “Nation or society as a whole” instead of “Nation”, and we add a subheading with “Factors that could lead to severe or catastrophic consequences for society include” as well as associated material under that sub-heading. Moreover, we have updated our actionable-guidance paper to reflect changes in AI RMF terms from the AI RMF Initial Draft to the 2nd Draft, and we plan to continue updating it, e.g. for AI RMF 1.0 when released in 2023.)

Suggested Change:
No change to the Playbook Map 5.2 listing of our actionable-guidance paper (Barrett et al. 2022) as an informative reference, except perhaps to list our paper specifically for rating magnitude of identified impacts.

References:
Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966

Govern 4.2

Response/Comment:
The communication of risks and impacts could be expanded upon in the Playbook material on Govern 4.2. Sections 3.3 and 3.4 of our actionable-guidance paper (Barrett et al. 2022) include material on communicating various potential types of impacts (including human rights impacts) as appropriate in context as part of communicating AI system limitations and risks to stakeholders.

Suggested Change:
First, in the Playbook section on Govern 4.2 under “Actions”, add a bullet that reads as follows: “Report risk factors identified in AI system risk assessment, including on potential types of impacts or harms outside the organization, by time of deployment or at earlier lifecycle stages, as appropriate in context as part of communicating AI system limitations and risks to stakeholders. Incorporate outputs from Map 5.1 and associated impact-assessment and risk-assessment activities, as appropriate. ”

Second, in the Playbook section on Govern 4.2 under “Organizations can document the following”, add “and communicated” to the fourth bullet, so that it reads as follows: “To what extent has the entity documented and communicated the AI system’s development, testing methodology, metrics, and performance outcomes?”

Third, in the AI RMF Playbook, list our actionable-guidance paper (Barrett et al. 2022) as an informative reference for Govern 4.2 for communicating various potential types of impacts.

References:
Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. ArXiv abs/2206.08966 (2022) https://arxiv.org/abs/2206.08966