Comments on the Initial Draft of the NIST AI Risk Management Framework

Jonas Schuett
Research Fellow
Centre for the Governance of AI
jonas.schuett@governance.ai

Markus Anderljung
Head of Policy
Centre for the Governance of AI
markus.anderljung@governance.ai

29 April 2022

About the Centre for the Governance of AI (GovAI)

The Centre for the Governance of AI (GovAI) is a nonprofit based in Oxford. It was founded in 2018, initially as part of the Future of Humanity Institute at the University of Oxford, before becoming an independent research organization in 2021. GovAI's mission is to build a global research community, dedicated to helping humanity navigate the transition to a world with advanced AI.

Our comments

We welcome the opportunity to submit comments on the Initial Draft of the NIST AI Risk Management Framework (AI RMF) and look forward to future opportunities to input on the framework. We offer the following submission for your consideration.

Key recommendations:

- Put more emphasis on low-probability, high-impact risks, especially catastrophic risks to society.
- Create a new Socio-Technical Characteristic on "Misuse/Abuse".
- Create a new Guiding Principle on "Alignment with Human Values and Intentions".
- Recommend organizations set up an internal audit function to continually assess whether their AI RMF implementation has improved their ability to manage AI risks.
- Review and update the AI RMF frequently.

1. Whether the AI RMF appropriately covers and addresses AI risks, including with the right level of specificity for various use cases.

We believe that the AI RMF does not appropriately cover and address the following AI risks:

- **Catastrophic risks.** The AI RMF does not put enough emphasis on low-probability, high-impact risks from AI, especially catastrophic risks to society. There is growing concern among AI researchers,[1] public figures,[2] and policymakers[3] that AI systems might pose catastrophic risks to society. We think that, even if the likelihood of catastrophic events is low, their potential impact warrants significant attention. Other NIST publications already take into account catastrophic risks. For example, "catastrophic adverse effects" are part of the impact assessment scale in NIST SP 800-30.[4] The AI RMF Concept Paper also contained a passage on catastrophic risks,[5] but it is not included in the Initial Draft. We suggest adding the original passage from the Concept Paper under Section 4.2 (Challenges for AI Risk Management). We also suggest that NIST creates or commissions further guidance for the assessment of catastrophic risks from AI.

- **Misuse/abuse risks.** The AI RMF risk taxonomy does not put enough emphasis on harms caused by a misuse or abuse[6] of AI systems.[7] For example, language models could be used to create fake news, or systems designed for drug discovery could be used to identify new toxins.[8] We think organizations should try to anticipate malicious third-party actions, and if possible, take measures to prevent them. The AI RMF mentions "unexpected or adversarial use of the model or data" in Section 5.1.4 (Resilience or ML Security). However, we think there is value in addressing a wider set of misuse/abuse risks and making this concern more explicit to encourage a deeper

[1] E.g. Russell (2019), Ord (2020), and Bostrom (2014).
[2] E.g. Stephen Hawking (Cellan-Jones, 2014), Elon Musk (Gibbs, 2014), and Bill Gates (Rawlinson, 2015).
[3] E.g. UK Government (2021, p. 60): "The government takes the long term risk of non-aligned Artificial General Intelligence, and the unforeseeable changes that it would mean for the UK and the world, seriously."
[4] NIST (2012, Table H-3): "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. 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."
[5] NIST (2021, p. 2): "Managing AI risks presents unique challenges. An example is the evaluation of effects from AI systems that are characterized as being long-term, low probability, systemic, and high impact. 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."
[6] The main differentiator between a misuse or abuse of an AI system and an accident is the user's intention (intentional vs. unintentional), while the main differentiator between a misuse and abuse is the user's authorization (authorized vs. unauthorized).
[7] For more information on the misuse and abuse of AI systems, see Brundage et al. (2018) and Weidinger et al. (2021, pp. 25–28).
[8] Urbina et al. (2022).

engagement with misuse/abuse risks, building on emerging best practices.[9] We therefore suggest creating a new Socio-Technical Characteristic on "Misuse/Abuse" in the risk taxonomy, and updating the references in the Functions accordingly.

- **Systemic risks.** The AI RMF does not put enough emphasis on systemic risks caused or exacerbated by AI. For example, AI applications are increasingly fine-tuned versions of so-called "foundation models".[10] If this trend continues, a new type of systemic risk might emerge: some harmful properties of foundation models, such as poor cybersecurity, could propagate through a large number of downstream applications, potentially affecting a large number of people.[11] We welcome that the risk taxonomy mentions "harm to a system",[12] but we are worried that some organizations will interpret this too narrowly. They might focus on harms to an AI system itself, not diffuse harms caused or exacerbated by many AI systems (although "large scale harms to the financial system or global supply chain" are mentioned as examples[13]). Having said that, we are uncertain about the magnitude of existing systemic risks from AI, and what role risk management should play to address them. We suggest adding "systemic risks from foundation models" as a third example in Figure 2. More generally, we recommend to closely monitor the space and potentially include more specific measures to reduce systemic risk in upcoming iterations of the AI RMF.

- **Specific risks of certain model types.** The AI RMF is not sensitive enough to the specific risks of certain model types. For example, language models[14] might output toxic language,[15] while image generation models[16] might produce explicit content.[17] Not all of these risks are sufficiently covered and addressed by the risk taxonomy in Section 5; they would benefit from more specific taxonomies.[18] We recommend that such taxonomies be included in the Practice Guide or in Profiles, not the AI RMF itself.

Besides that, we believe that the risk taxonomy has the following weaknesses:

- **The risk taxonomy is not comprehensive.** It does not appropriately cover and address catastrophic risks, misuse/abuse risks, systemic risks, and specific risks of certain model types (see above). We also don't understand why it only includes three Guiding Principles. The selection seems arbitrary and, for example, lacks consideration of aligning AI systems with human values and intentions (see below). It is beyond the scope

[9] See e.g. Microsoft (2021) and OWASP (2021).
[10] Foundation models are large pre-trained models that can serve as the "foundation" for a wide array of down-stream applications (Bommasani et al., 2021).
[11] Bommasani et al. (2021, p. 134): "The one-to-many nature of foundation models, i.e., the same few foundation models being used across many applications, means the intrinsic properties of foundation models pervade to many downstream applications."
[12] NIST (2022, Figure 2): "Harm to an organized assembly of interconnected and interdependent elements and resources, for example, large scale harms to the financial system or global supply chain, which are not sufficiently resilient to adverse AI impacts."
[13] Ibid.
[14] E.g. BERT (Devlin, et al., 2018), GPT-3 (Brown et al., 2020), Gopher (Rae et al., 2021), Chinchilla (Hoffmann et al., 2022), or PaLM (Chowdhery et al., 2022).
[15] Weidinger et al. (2021, pp. 15–16).
[16] E.g. DALL·E 2 (Ramesh et al., 2022).
[17] Mishkin et al. (2022).
[18] See Weidinger et al. (2021) for a taxonomy of risks from language models.

of this submission to suggest a more comprehensive taxonomy, but we wish to highlight an alternative taxonomy which has been proposed by DeepMind[19] and has gained some popularity.[20] At the highest level, it distinguishes between specification, robustness, and assurance.

- **The risk taxonomy does not have the right level of specificity.** We think that the risk taxonomy lacks important details and nuance. As mentioned above, it is not sensitive enough to the specific risks of certain model types. Besides that, the discussion of risk thresholds in Section 4.2.2 is too vague; it would benefit from a few examples. However, we expect the Practice Guide, which has not yet been published, to increase the level of specificity.

2. Whether the AI RMF is flexible enough to serve as a continuing resource considering evolving technology and standards landscape.

We believe that the AI RMF can be flexible enough to serve as a continuing resource. But this will largely depend on the frequency and the extent to which the AI RMF and accompanying documents like the Practice Guide and Profiles are reviewed and updated, as the risk landscape evolves rapidly. We believe that this is one of the most important factors that will determine the framework's success. We therefore suggest including a statement on the frequency of the updates (e.g. "every two years and as appropriate").

3. Whether the AI RMF enables decisions about how an organization can increase understanding of, communication about, and efforts to manage AI risks.

We believe that the AI RMF can enable decisions about how an organization can increase understanding of, communication about, and efforts to manage AI risks. But again, this will largely depend on the Practice Guide and Profiles, which have not yet been published.

4. Whether the functions, categories, and subcategories are complete, appropriate, and clearly stated.

In general, we believe that the functions, categories, and subcategories are complete, appropriate, and clearly stated. We suggest the following minor changes:

- **Incident databases.** We welcome that the AI RMF mentions "public incident reports"[21] as a potential basis for assessing the likelihood of certain harms. However, this formulation might suggest that organizations should focus on individual incident reports (e.g. incidents reported in the media or in research papers). We think organizations can get a much more accurate picture of the actual likelihood of certain harms by looking at

[19] Ortega & Maini (2018).
[20] See e.g. Rudner & Toner (2021a).
[21] NIST (2022, Table 1, ID 4, Subcategory 3): "Likelihood of each harm is understood based on expected use, past uses of AI systems in similar contexts, public incident reports or other data."

well-curated incident databases.[22] This could include the AI Incident Database created by the Partnership on AI,[23] but we expect other actors to create similar databases in the future. We suggest adding "and incident databases" to Table 1, ID 4, Subcategory 3. Although incident databases are already covered by the current formulation ("public incident reports or other data"), we think there is value in mentioning them explicitly.

- **Optionality.** The Manage function lists four criteria that organizations should use to prioritize identified risks: "impact, likelihood, resources required to address them, and available methods to address them".[24] We think that optionality, the extent to which an individual can decide not to be subject to the effects of an AI system,[25] is another important criterion. We suggest adding the term "optionality" as a fifth criterion to the list.

5. Whether the AI RMF is in alignment with or leverages other frameworks and standards such as those developed or being developed by IEEE or ISO/IEC SC42.

We are uncertain whether the AI RMF is aligned with or leverages other frameworks and standards. Since the AI RMF and other frameworks and standards are still under development, it is too early for a final evaluation. Having said that, we have no reason to believe that they are not aligned and expect that the AI RMF will benefit from continued efforts to increase coherence with related standards and frameworks. We think it will be particularly valuable to align the AI RMF with the OECD AI Framework.[26]

6. Whether the AI RMF is in alignment with existing practices, and broader risk management practices.

We believe that the AI RMF is in alignment with existing practices, and broader risk management practices. In particular:

- **Enterprise Risk Management (ERM).** The AI RMF seems to be aligned with popular ERM frameworks, such as ISO 31000[27] and COSO ERM 2017.[28] Although they occasionally use different terms, the core risk management process looks very similar across the three frameworks.

- **Three Lines of Defense (3LoD).** The AI RMF also seems to be aligned with the 3LoD model.[29] In particular, the Governance function mentions the need to coordinate roles and responsibilities,[30] which is exactly what the 3LoD model is for. Due to the model's

[22] See McGregor (2020).
[23] AI Incident Database.
[24] NIST (2022, Table 3, ID 1, Subcategory 2).
[25] OECD (2022, p. 67).
[26] Ibid.
[27] ISO 31000:2018 Risk management — Guidelines.
[28] COSO Enterprise risk management — Integrating with strategy and performance (2017).
[29] For more information, see IIA (2013, 2020), Davies & Zhivitskaya (2019), Bantleon et al. (2021).
[30] NIST (2022, Table 4, ID 2, First Subcategory): "Roles and responsibilities and lines of communication related to identifying and addressing AI risks are clear to individuals and teams throughout the organization."

popularity and widespread use in practice, we recommend adding an explicit reference to the model in Table 4, ID 2, first subcategory.

7. What might be missing from the AI RMF.

We believe that the most important thing that is missing from the AI RMF is discussion of catastrophic risks to society (see above). In addition to that, we think that the following aspects are also missing:

- **Alignment with human values and intentions.** The AI RMF does not address the need to align AI systems to human values and intentions. There is a substantial body of literature on the so-called "alignment problem", the problem of ensuring that an AI system reliably acts in accordance with human values and intentions.[31] Solving the alignment problem is highly complex and involves a number of subproblems.[32] It is also a significant part of the unique risk profile of AI systems, as they frequently find ways to maximize their reward in ways that human designers did not intend.[33] Although there has been some recent progress in aligning AI systems to human values and intentions,[34] more work is required. By acknowledging the problem, the AI RMF could incentivize more of such work. We therefore suggest adding "Alignment with Human Values and Intentions" as another Guiding Principle. (Alternatively, the AI RMF could call the principle "Specification", which is a closely related framing of the problem.[35]) We believe that value and intent alignment should be a fundamental principle for the development and deployment of AI systems, especially for more capable and general systems that might get developed in the future. In practice, we think that asking organizations to align systems to the intentions of their users or operators will be most actionable, whereas aligning systems to human values would be more difficult to operationalize.

- **Internal audit.** The AI RMF does not specify a mechanism for assessing whether the framework has been implemented effectively. We think this is problematic, because some organizations may only superficially engage with certain risks, especially low-probability, high-impact risks, and could implement the framework in a box-ticking fashion. We welcome that the AI RMF recommends that organizations should "periodically evaluate whether the AI RMF has improved their ability to manage AI risks".[36] However, we think that the framework should specify a mechanism for how to do this. Internal audit[37] seems to be the best solution, because many organizations already

[31] E.g. Christian (2020), Gabriel (2020), Kenton et al. (2021), Hendrycks et al. (2021).
[32] For example, we can distinguish between the normative challenge, what values or principles, if any, we ought to encode in artificial agents, and the technical challenge, how to formally encode values or principles in artificial agents so that they reliably do what they ought to do (Gabriel, 2020, pp. 412–413). For an overview of other decompositions of the alignment problem, see Kenton et al. (2021, pp. 2–5).
[33] Krakovna et al. (2020) contains an explanation and a list of examples.
[34] E.g. Ouyang et al. (2022), Bai et al. (2022), Leike et al. (2018), Christiano et al. (2018).
[35] Ortega & Maini (2018): "Specification ensures that an AI system's behavior aligns with the operator's true intentions." See also Rudner & Toner (2021b).
[36] NIST (2022, p. 20).
[37] IIA (2022): "Internal auditing is an independent, objective assurance and consulting activity designed to add value and improve an organization's operations. It helps an organization accomplish

have an internal audit function and there are established best practices.[38] We suggest recommending the implementation of an AI-specific internal audit function in Section 8. The AI RMF should also specify how this function differs from other risk management functions, especially the second line of defense.

- **Vulnerability assessment.** The AI RMF does not mention the need to assess an organization's vulnerabilities. However, some industries like aviation and nuclear have moved away from a focus on likelihood, and instead focus more on vulnerabilities. This approach to risk assessment makes it easier to identify corresponding mitigations. We recommend that NIST investigates how vulnerability assessments could play a role in the AI RMF.

8. Whether the soon to be published draft companion document citing AI risk management practices is useful as a complementary resource and what practices or standards should be added.

The document has not yet been published.

9. Others?

We wish to highlight a few parts of the AI RMF that we find particularly valuable. We encourage NIST not to change the following parts:

- **Risks to individuals, groups and society.** We welcome that the AI RMF does not just focus on risks to organizations ("business risks"), but also includes risks to individuals, groups and society.[39] We think this is one of the greatest strengths of the framework.

- **Unacceptable risks.** We welcome that the AI RMF encourages organizations to consider "whether an AI system [that presents unacceptable risks] should be designed, developed, or deployed at all."[40] We are worried that organizations will deploy unsafe systems, because they have already spent significant resources on their development. We therefore think there is value in highlighting the non-deployment option.

- **Socio-technical perspective.** We welcome that the AI RMF takes a socio-technical perspective on risk.[41] We think that risks can only be managed effectively if systems are viewed in their socio-technical context.

- **Participatory approaches.** We strongly agree with the recommendation to get "input from a broad and diverse set of stakeholders […] throughout the AI lifecycle"[42] and we welcome that the AI RMF mentions participatory methods in subcategories of the

its objectives by bringing a systematic, disciplined approach to evaluate and improve the effectiveness of risk management, control, and governance processes."

[38] See e.g. the certification program Certified Internal Auditor (CIA) by the Institute of Internal Auditors.

[39] NIST (2022, Figure 2.

[40] Ibid., p. 6.

[41] Ibid., Section 5.2.

[42] Ibid., p. 10.

Measure[43] and Manage function.[44] We think it is particularly important for organizations to engage with the groups most affected by the risk, especially marginalized groups.[45]

[43] Ibid., Table 2, ID 3, Second Subcategory.
[44] Ibid., Table 3, ID 3, Second Subcategory.
[45] DeepMind (2021, pp. 2–3) makes a similar recommendation. See also Mohamed et al. (2020).

References

Bai, Y., Jones, A., Ndousse, K., et al. (2022). *Training a helpful and harmless assistant with reinforcement learning from human feedback*. arXiv. https://arxiv.org/abs/2204.05862

Bantleon, U., d'Arcy, A., Eulerich, M., et al. (2021). Coordination challenges in implementing the three lines of defense model. *International Journal of Auditing, 25*(1), 59–74. https://doi.org/10.1111/ijau.12201

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). *On the opportunities and risks of foundation models*. arXiv. https://arxiv.org/abs/2108.07258

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Brown, T. B., Mann, B., Ryder, N., et al. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165

Brundage, M., Avin, S., Clark, J., et al. (2018). *Malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv. https://arxiv.org/abs/1802.07228

Cellan-Jones, R. (2014). Stephen Hawking warns artificial intelligence could end mankind. *BBC*. https://perma.cc/GA8A-BT5M

Chowdhery, A., Narang, S., Devlin, J., et al. (2022). *PaLM: Scaling language modeling with Pathways*. arXiv. https://arxiv.org/abs/2204.02311

Christian, B. (2020). *The alignment problem: Machine learning and human values*. W. W. Norton & Company.

Christiano, P., Shlegeris, B., & Amodei, D. (2018). *Supervising strong learners by amplifying weak experts*. arXiv. https://arxiv.org/abs/1810.08575

Davies, H., & Zhivitskaya, M. (2018). Three lines of defence: A robust organising framework, or just lines in the sand? *Global Policy, 9*(1), 34–42. https://doi.org/10.1111/1758-5899.12568

DeepMind (2021). Request for information: Artificial intelligence risk management framework. *NIST*. https://perma.cc/N365-6YBV

Devlin, J., Chang, M.-W., Lee, K., et al. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://arxiv.org/abs/1810.04805

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines, 30*, 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gibbs, S. (2014). Elon Musk: artificial intelligence is our biggest existential threat. *The Guardian*. https://perma.cc/8WLM-LN4G

Hendrycks, D., Carlini, N., Schulman, J., et al. (2021). *Unsolved problems in ML safety*. arXiv. https://arxiv.org/abs/2109.13916

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). *Training compute-optimal large language models*. arXiv. https://arxiv.org/abs/2203.15556

IIA (2013). *The three lines of defense in effective risk management and control: IIA position paper*. https://perma.cc/NQM2-DD7V

IIA (2020). *The IIA's three lines model: An update of the three lines of defense*. https://perma.cc/GAB5-DMN3

IIA (2022). *The definition of internal auditing*. https://perma.cc/B34B-WRMC

Kenton, Z., Everitt, T., Weidinger, L., et al. (2021). *Alignment of language agents*. arXiv. https://arxiv.org/abs/2103.14659

Krakovna, V., Uesato, J., Mikulik, V., et al. (2020). Specification gaming: the flip side of AI ingenuity. *DeepMind*. https://perma.cc/QVW5-6RAJ

Leike, J., Krueger, D., Everitt, T., et al. (2018). *Scalable agent alignment via reward modeling: a research direction*. arXiv. https://arxiv.org/abs/1811.07871

McGregor, S. (2020). *Preventing repeated real world AI failures by cataloging incidents: The AI incident database*. arXiv. https://arxiv.org/abs/2011.08512

Microsoft (2021). *Foundations of assessing harm*. https://perma.cc/426W-5SQ7

Mishkin, P., Ahmad, L., Brundage, M., et al. (2022). DALL·E 2 preview: Risks and limitations. *Github*. https://perma.cc/4Z7C-26TY

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*, 659–684. https://doi.org/10.1007/s13347-020-00405-8

NIST (2012). *Guide for conducting risk assessments* (NIST Special Publication 800-30 Revision 1). https://perma.cc/6VKS-9SW2

NIST (2021). *AI risk management framework concept paper*. https://perma.cc/J9NL-8RD6

NIST (2022). *AI risk management framework: Initial draft*. https://perma.cc/FGM8-5TTG

OECD (2022). *Framework for the classification of AI systems*. https://doi.org/10.1787/cb6d9eca-en

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.

Ortega, P. A., & Maini, V. (2018). Building safe artificial intelligence: specification, robustness, and assurance. *Medium*. https://perma.cc/L7PK-LC46

Ouyang, L., Wu, J., Jiang, X., et al. (2022). *Training language models to follow instructions with human feedback*. arXiv. https://arxiv.org/abs/2203.02155

OWASP (2021). *Abuse case cheat sheet*. https://perma.cc/9JXT-KV62

Rae, J. W., Borgeaud, S., Cai, T., et al. (2021). *Scaling language models: Methods, analysis & insights from training Gopher*. arXiv. https://arxiv.org/abs/2112.11446

Ramesh, A., Dhariwal, P., Nichol, A., et al. (2022). *Hierarchical text-conditional image generation with CLIP latents*. arXiv. https://arxiv.org/abs/2204.06125

Rawlinson, K. (2015). Microsoft's Bill Gates insists AI is a threat. *BBC*. https://perma.cc/N6Y8-CG2S

Rudner, T. G. J., & Toner, H. (2021a). Key concepts in AI safety: An overview. *Center for Security and Emerging Technology*. https://doi.org/10.51593/20190040

Rudner, T. G. J., & Toner, H. (2021b). Key concepts in AI safety: Specification in machine learning. *Center for Security and Emerging Technology*. https://doi.org/10.51593/20210031

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin Random House.

UK Government (2021). *National AI strategy*. https://perma.cc/RYN4-EEBR

Urbina, F., Lentzos, F., Invernizzi, C., et al. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence, 4*, 189–191. https://doi.org/10.1038/s42256-022-00465-9

Weidinger, L., Mellor, J., Rauh, M., et al. (2021). *Ethical and social risks of harm from language models*. arXiv. https://arxiv.org/abs/2112.04359