Response to NIST on Managing Misuse Risk for Dual-Use Foundation Models

September 9th, 2024

Subject: NIST AI 800-1, Managing Misuse Risk for Dual-Use Foundation Models

Via email to NISTAI800-1@nist.gov

To the NIST team developing guidance for Managing Misuse Risk for Dual-Use Foundation Models,

Thank you for the invitation to submit comments in response to the July 2024 release of the initial public draft (ipd) of the Managing Misuse Risk for Dual-Use Foundation Models guidance (NIST AI 800-1 ipd). We were happy to support the NIST Generative AI Public Working Group, and commend NIST on the creation of this guidance. We offer the following submission for your consideration.

We are researchers affiliated with UC Berkeley, with expertise in AI research and development, safety, security, policy, and ethics. We previously submitted responses to NIST in May 2024 on the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, and at multiple points in 2021-2023 at various stages of NIST development of AI RMF guidance.

Following our recommendations to NIST in 2022 to go beyond the broadly applicable guidance of the AI RMF and to provide an AI RMF profile with guidance specifically for developers and evaluators of foundation models, we undertook our own yearlong effort to create an AI RMF-compatible profile for foundation models: the “AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models” (Barrett, Newman et al. 2023a, 2023b). We sometimes refer to that as the “Berkeley profile” in the following. We have aimed for our Berkeley profile effort to complement and inform the work by NIST and others. Some of our recommendations in the following are based in part on the approach and guidance in the Berkeley profile.

Here is a high-level summary of our key comments and recommendations on NIST AI 800-1:

- NIST AI 800-1 ipd provides a **broadly sensible framework** that parallels a number of important risk management ideas in the NIST AI RMF. In most cases, **we recommend retaining your current draft guidance**, or expanding upon it.
- **For Practice 2.1, we recommend adding a point to address the concept that some risks (e.g., risks of catastrophic and irreversible harms) can be unacceptable regardless of potential benefits**, and acknowledging that while benefits often accrue

primarily to some stakeholders (e.g., company shareholders), risks are borne primarily by others (e.g., members of the public that could be harmed by misuse).

- We recommend **clarifying that the "proxy model" in Practice 4.1 should be treated as "a well understood base case"** (e.g., a foundation model that has been extensively evaluated and released), and is not necessarily "a base case for comparison when assessing marginal risks of release of a new foundation model".
- We recommend strengthening the documentation examples across many of the practices to better support communication of practically useful information while adequately addressing confidentiality concerns. (See Question 4 in the Questions Posed in the Federal Register Request for Comments.)

In the following sections, we provide detail and additional comments.

Thank you again for the opportunity to comment on NIST AI 800-1. If you need additional information or would like to discuss further, please contact Anthony Barrett at anthony.barrett@berkeley.edu or Jessica Newman at jessica.newman@berkeley.edu. In any case, we look forward to further engagement with NIST as you proceed on the Managing Misuse Risk for Dual-Use Foundation Models guidance development process.

Our best,

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

Nada Madkour, Ph.D.
Non-Resident Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

Evan R. Murphy
Non-Resident Fellow
AI Security Initiative, Center for Long-Term Cybersecurity, UC Berkeley

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

Our Overarching Comments

In this section, we provide a number of comments related to cross-cutting topics in the Managing Misuse Risk for Dual-Use Foundation Models guidance.

- **NIST AI 800-1 ipd provides a broadly sensible framework** that parallels a number of important risk management ideas in the NIST AI RMF. Many of the recommended practices in NIST AI 800-1 also align with the recommendations in our Berkeley profile (Barrett, Newman et al. 2023a, 2023b).
    - **In most cases, we recommend retaining your current draft guidance,** or expanding upon it. We highlight a few such cases in the following (e.g. in our comment on Practice 5.1), but not all such cases.
- **We recommend adding greater clarity and more examples in Section 2 about the risks that are within scope of this guidance, including the amplification of violence, radicalization, and extremism online**.
    - This could be achieved by adding bullet points or a short paragraph with more examples, or at a minimum by expanding the final sentence of the first paragraph such that it includes more examples and reads something like, "This document addresses both emerging misuse risks, such as a foundation model facilitating the development of a novel biological weapon, as well as current harms from misuse, such as a foundation model generating CSAM or NCII, *generating disinformation, automating phishing attacks, or amplifying the spread of violence, radicalization, and extremism in online discourse*. (Italics indicate new proposed language.)
- NIST AI 800-1 ipd often discusses assessment and management of misuse “risk”, which seems broadly appropriate. However, in some parts of the document, it could be helpful to add more specific and actionable language, e.g., about assessing and managing risks in terms of model “capabilities of concern” as well as in broader “risk” terms.
    - One key example is the guidance on defining risk thresholds in Practice 2.1, e.g. to “Define a risk threshold for each identified threat profile”. **We recommend adding language for Practice 2.1, clarifying that it can be appropriate to define risk thresholds in terms of model capabilities of concern**.
        - Our sense is that the uncertainties in estimated probabilities of model-misuse events are greater than uncertainties in evaluated capabilities of models that could be misused. Thus, defining risk thresholds in terms of a combination of event probabilities and model capabilities, instead of focusing on just model capabilities, seems likely to substantially compound uncertainties for risk management. Moreover, many companies’ risk management approaches currently include defining model capability thresholds that would trigger risk-management actions, as noted by the recent GovAI report on risk thresholds (Koessler et al. 2024).
    - Clarifying that risk thresholds can be defined in terms of model capabilities of concern also could improve consistency with the guidance on measuring model capabilities of concern in Practice 4.1.
- One key issue is **risk baselines**, which is most clearly related to Objective 4 and Practice 4.1 but can affect assumptions across many other objectives and practices including creation of threat profiles under Objective 1 and Practice 1.1.

    - The current wording of material under Practice 4.1 does not clearly indicate whether the misuse risks from a new model should be compared to a) information from other sources, e.g. internet, or b) other foundation models.
    - If assessing marginal risks of releasing a new foundation model, it could be valuable to compare risks of a threat actor's misuse of that new foundation model to risks of that threat actor's misuse of already-available foundation models. However, using existing models as a baseline could easily lead to an exponential growth in risk from AI models overall, year over year. That would be a slippery slope, and bad risk management public policy. We therefore recommend that more time-tested baseline comparisons also be used where feasible, such as information available from the Web, while recognizing that this is also a moving target.
        - **This may be most important for models planned for open-weights release.** Models with closed-weights releases can be rolled back, they can be made unavailable very quickly via the provider's control of an API, but open-weights models cannot effectively be made unavailable after release of their weights. Model developers should not be provided with an easy path to irreversible release of ever more powerful frontier models via an incrementally-rising marginal-risk comparison to their last release.
    - We recommend **clarifying that the "proxy model" in Practice 4.1 should be treated as "a well understood base case"** (e.g., a foundation model that has been extensively evaluated and released), and is not necessarily "a base case for comparison when assessing marginal risks of release of a new foundation model" (e.g., information available on the internet).

Our Comments on Questions Posed in the Federal Register Request for Comments

In this section, we provide answers to the specific questions posed in the Federal Register RFC (89 FR 64878) on NIST AI 800-1.

1. What practical challenges exist to meeting the objectives outlined in the guidance?

Below are our comments on practical challenges for some of the outlined objectives.

Anticipate Misuse Risk:

Some key challenges of anticipating misuse risk include anticipating the misuse risk of novel or unknown misuse scenarios. This is particularly important for models deployed with open weights, for which effectively-irreversible release decisions must be made despite uncertainty about currently-unknown model risks.

Establish Plans for Managing Misuse RIsk:

Lack of transparency can be a barrier to establishing effective plans for managing misuse risk. Most plans are not made available by the developer/deployer or are made available in a very limited capacity. Overall, general concepts and best practices are available, but context-specific and explicit guidance has room for improvement.

2. How can the guidance better address the ways in which misuse risks differ based on deployment (e.g., how a foundation model is released) and modality (text, image, audio, multimodal, and others)?

Comment:
One of the listed key challenges (point 6) addresses the dependence on experts for the evaluation of some risks, yet it is included as a consideration in practice 1.1.
The guidance effectively covers high-priority practices for managing misuse risk, but does not distinguish between open-weight models and models deployed with structured access.

Recommendations:

- We recommended removing “consider” from point 3 of practice 1.1, to emphasize the importance of consulting external experts.
- It may be beneficial to explicitly list practices that apply to open-weight models where appropriate in the guidance.

3. How can the guidance better reflect the important role for real-world monitoring in making risk assessments?

Comment:
Gradual or staged releases of models can have many benefits including: identifying risks early in a controlled environment, limiting the impacts of previously unidentified risks, and gathering feedback on specific use cases. Additionally, it is important to utilize periodic evaluations of applied safeguards post-deployment to verify their effectiveness in protecting against existing and emerging risks. For third-party entities with access to the model, required periodic performance reports to the model developer may support more accurate risk assessments of deployed models. It is also beneficial to provide internal and external stakeholders the ability to report incidents of model misuse after model deployment.

Recommendations:

- Under practice 5.1, we recommend adding a point “Consider deployment with gradual, phased releases, and/or structured access through an API or other mechanisms, with efforts to detect and respond to misuse or problematic anomalies.”

- Under practice 6.1, we recommend adding a point “Provide third-party distribution channels with standardized reporting procedures that include reporting templates, and predetermined reporting frequencies. Consider running periodic audits on third-party distribution channel reports to evaluate validity.”
- We recommend retaining or expanding on the draft language for Practice 6.3. As one expansion, we recommend adding a point “Establish a process for anonymous misuse reporting available to internal and external stakeholders” to emphasize the importance of user and community feedback.

4. How can the guidance's examples of documentation better support communication of practically useful information while adequately addressing confidentiality concerns, such as protecting proprietary information?

Comment:
Overall, the guidance covers the baseline for high-priority documentation and communication for managing misuse risk. We have made recommendations regarding documentation under several practices (1.2, 1.3, 2.1, 3.3, 4.1, 6.2, 7.3) that may support the communication of practically useful information.

Recommendations:

- Under practice 1.2 documentation, we recommend adding “Documentation on the comparison between past assessment predictions and scores and actual model impact.” to emphasize the importance of comparing actual outcomes to predicted outcomes to evaluate the accuracy of the assessment methods.
- Under practice 1.3, We recommend adding a point that addresses the importance of making evaluation results and documentation related to capabilities of concern available to the public, or at least other model providers. This recommendation is particularly meaningful considering the recommended practice in point number 1, which states “Identify similar models for which capabilities of concern have already been measured”.
- Under practice 2.1, we recommend changing point “a” under documentation to “Reasoning and process for determining each threshold level in the context of model misuse risk, along with whether or not each threshold level triggers a pause or halt of model development, training, or deployment.” and adding point “b” “The documentation protocols for each level of identified misuse risk.”
- Under practice 3.3, We recommend adding point “b” under documentation to address documenting rejected or failed security measures. It may be valuable to document a summary of the security measures that were considered, used, or tested, but ultimately determined to be ineffective for reducing the risk of model theft to help the model developer, or other model developers, avoid ineffective security controls.
- Under practice 4.1, Under documentation, we recommend adding a point “Correlation analysis results between evaluation method results that include the resources required

to implement each method". This documentation can help provide transparency, and efficiency, for the practice outlined in point number 2 (Barrett et al. 2024).

- Under practice 6.2 documentation, we recommend changing point "a" to "A summary of the incident response process and the organizational roles and responsibilities in the process for each category or misuse and/or level of risk" to emphasize the importance of having different procedures for different types of risks as well as different levels of risk (e.g. high impact vs low impact risks).
- Under practice 7.3, we recommend changing point 3 under documentation to "Share verified reports of misuse with relevant third parties, such as AI incident databases and risk repositories (Slattery et al. 2024)" to emphasize the importance of documenting risks, not only incidents.

5. How can the guidance better enable collaboration among actors across the AI supply chain, such as addressing the role of both developers and their third-party partners in managing misuse risk?

Comment:
Practices that may better enable collaboration among actors across the supply chain include emphasizing the importance of sharing model evaluation results, providing clear documentation instructions for third-party distribution channels, establishing a process for anonymous misuse reporting, and sharing verified reports with relevant third parties. We have provided recommendations under several practices (1.3, 6.1, 6.3, 7.3) that address each of these points.

Recommendations:

- Under practice 1.3, we recommend adding a point that addresses the importance of making evaluation results and documentation related to capabilities of concern available to the public, or at least other model providers. This recommendation is particularly meaningful considering the recommended practice in point number 1, which states "Identify similar models for which capabilities of concern have already been measured".
- Under practice 6.1, we recommend adding a point "Provide third-party distribution channels with standardized reporting procedures that include reporting templates, and predetermined reporting frequencies. Consider running periodic audits on third-party distribution channel reports to evaluate validity."
- Under practice 6.3, we recommend adding a point "Establish a process for anonymous misuse reporting available to internal and external stakeholders" to emphasize the importance of user and community feedback.
- Under practice 7.3, We recommend changing point 3 to "Share verified reports of misuse with relevant third parties, such as AI incident databases and risk repositories (Slattery et al. 2024)" to emphasize the importance of documenting risks, not only incidents.

Our Comments on Specific Sections or Passages

Objective 1

Practice 1.1

One of the listed key challenges (point 6) addresses the dependence on experts for the evaluation of some risks. **We recommended removing “consider” from point 3 to emphasize the importance of consulting external experts.**

Practice 1.2

We recommend including assessments of misuse in the context of impact on socioeconomic systems and ecosystems (e.g. the environment, job displacement, labor market disruption) in addition to impact on public safety. Expanding on the point to describe public safety at large to include elements such as sociotechnical systems and ecosystems would also be sufficient.
We also recommend adding a point to address the importance of assessment method validation.

In the documentation sections, **we recommend adding “Documentation on the comparison between past assessment predictions and scores and actual model impact.”** to emphasize the importance of comparing actual outcomes to predicted outcomes to evaluate the accuracy of the assessment methods.

Practice 1.3

We recommend adding a point that addresses the importance of making evaluation results and documentation related to capabilities of concern available to the public, or at least other model providers. This recommendation is particularly meaningful considering the recommended practice in point number 1, which states “Identify similar models for which capabilities of concern have already been measured”.

Objective 2

Practice 2.1

Point 1 recommends defining risk thresholds for each identified threat profile. While this is incredibly valuable, it may also be limited due to numerous factors, including potential disincentives to consider a wide variety of threat profiles. **We recommend also defining overarching risk thresholds that an organization considers unacceptable that either 1) are not tied to specific threat profiles, or 2) are applicable in a broader cross-cutting way to multiple threat profiles.** We additionally recommend adding that these risk thresholds should be determined in consultation with external communities, including relevant experts and communities likely to experience the greatest impact and harm.

The practice in point 3 recommends weighing potential benefits against misuse risks, but some risks can be unacceptable regardless of potential benefits. It may not be appropriate to consider potential benefits when determining thresholds of certain catastrophic or unacceptable risks. Moreover, benefits often accrue primarily to some stakeholders (e.g., company shareholders) while risks are borne primarily by others (e.g., members of the public that could be harmed by misuse). Such issues are often considered by regulators when setting risk management rules for industry. **We recommend adding a point to address the concept that some risks (e.g., risks of catastrophic and irreversible harms) can be unacceptable regardless of potential benefits.** It could be worth adapting language from p. 8 of the NIST Generative AI Profile, AI 600-1 (Autio et al. 2024), that suggests that unacceptable risks include cases "where significant negative impacts are imminent, severe harms are actually occurring, or large-scale risks could occur"; similar language is also on p. 8 of the NIST AI Risk Management Framework, AI 100-1 (NIST 2023): "In cases where an AI system presents unacceptable negative risk levels – such as where significant 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 managed." We also note that any process of weighing benefits and risks should hold accounts of both to the same standard; if *marginal risks* are considered, then they should be weighed against *marginal benefits*, e.g. only the novel benefits above and beyond existing sources or tools should be taken into account.

We also recommend changing point "a" under documentation to "Reasoning for each threshold level in the context of model misuse risk, along with whether or not each threshold level triggers a pause or halt of model development, training, or deployment." and adding point "b" "The documentation protocols for each level of identified misuse risk."

In addition, we recommend, as mentioned in our "Overarching Comments" section above, **clarifying that it can be appropriate to define risk thresholds in terms of model capabilities of concern**.

Objective 3

Practice 3.2

We recommend changing point 1 to "Prior to developing a model, assess overall misuse risk from model theft by combining the estimated capabilities of concern with an estimate of the probability of model theft and an estimate of the misuse risk impact." to emphasize the importance of including the impact of misuse risk in the risk assessment.

Practice 3.3

We recommend adding point “b” under documentation to address documenting rejected or failed security measures. It may be valuable to document a summary of the security measures that were considered, used, or tested, but ultimately determined to be ineffective for reducing the risk of model theft to help the model developer, or other model developers, avoid ineffective security controls.

Objective 4

Practice 4.1

Under documentation**, we recommend adding a point “Correlation analysis results between evaluation method results that include the resources required to implement each method”.** This documentation can help provide transparency, and efficiency, for the practice outlined in point number 2 (Barrett et al. 2024).

We also recommend, as we mention in our “Overarching Comments” section above, **clarifying that the “proxy model” in Practice 4.1 should be treated as “a well understood base case”** (e.g., a foundation model that has been extensively evaluated and released), **and is not necessarily “a base case for comparison when assessing marginal risks of release of a new foundation model”** (e.g., information available on the internet).

Practice 4.2

We recommend retaining language around incentives in point 2 (“Clearly specify what goal the red team is trying to achieve in advance, provide incentives and accountability for achieving those goals, and select red teams based on their ability to achieve those goals”) and point 3, (“Rely on red teams comprised of external experts that are meaningfully independent from the model developer and who do not have incentives that conflict with their red-teaming goal”). All else equal, we expect that red-team findings would be most reliable if the compensation structures for red teamers provide incentives (e.g. cash bonuses) for red teamers to find meaningful problems in a foundation model, rather than to give the foundation model a pass as quickly as possible and avoid upsetting their model developer client by finding problems that could delay a model’s release.

Objective 5

Practice 5.1

We recommend retaining the two current points under Practice 5.1, on considering levels of access to a model that a malicious actor would gain via planned deployment approaches. For example, we agree that “sharing the model’s weights can significantly limit options to monitor for misuse (Practice 6.1) and respond to instances of misuse (Practice 6.2).” We have similar statements in our Berkeley profile (Barrett, Newman et al. 2023a) under Manage 2.4.

We also recommend adding a point “Consider deployment with gradual, phased releases, and/or structured access through an API or other mechanisms, with efforts to detect and respond to misuse or problematic anomalies.”

Practice 5.2

We recommend adding a point to address the importance of a process for periodic evaluation of safeguards. This point can be added as a recommended practice or as a recommended documentation procedure.

Objective 6

Practice 6.1

In support of point 4, **we recommend adding a point “Provide third-party distribution channels with standardized reporting procedures that include reporting templates, and predetermined reporting frequencies. Consider running periodic audits on third-party distribution channel reports to evaluate validity.”**

Practice 6.2

We recommend retaining or expanding on the draft language for Practice 6.2. As one expansion, under documentation, **we recommend changing point “a” to “A summary of the incident response process and the organizational roles and responsibilities in the process for each category or misuse and/or level of risk”** to emphasize the importance of having different procedures for different types of risks as well as different levels of risk (e.g. high impact vs low impact risks).

Practice 6.3

We recommend retaining or expanding on the draft language for Practice 6.3. As one expansion, **we recommend adding a point “Establish a process for anonymous misuse reporting available to internal and external stakeholders”** to emphasize the importance of user and community feedback.

Objective 7

Practice 7.1

We recommend maintaining the guidance to regularly publish transparency reports including key details regarding misuse risks and how those are managed. We recommend adding discussion of realized instances of misuse to these transparency reports as such cases are discovered.

Practice 7.3

We recommend changing point 3 to "Share verified reports of misuse with relevant third parties, such as AI incident databases and risk repositories (Slattery et al. 2024)" to emphasize the importance of documenting risks, not only incidents.

Example Safeguards

Below are our recommendations for recommended safeguards.

We recommend adding a 6th category to the recommended safeguards "Test for misuse capability" to address the measurement and monitoring of model capabilities. The safeguard category may include the following implementation methods:

- Evaluate the model's capability to upskill threat actors by using benchmarks and red teaming (Barrett et al. 2024).
- Utilize internal and external evaluation teams that include experts in the field, particularly in fields that require scarce domain expertise.

Under "improve model training" we recommend the following:

- Change the first method to "Filter training data to exclude examples that could result in capabilities that increase the likelihood of misuse, such as biological sequence data, cyber weaponization information, persuasion tactics, or known CSAM/NCII images."
- Add an implementation method to address the importance of evaluating data origins and data gathering practices and evaluating the risks of synthetic training data. Using synthetic training data can reinforce biases, and increase the likelihood of model error (Hao et al., 2024; McDuff et al., 2023; Whitney & Norman, 2024), if not utilized in compliance with recommended responsible practices (De Wilde et al., 2024).

Under "Stop development if a model displays significant misuse risk" we recommend the following:

- Adding an implementation method to address establishing different responses to different levels of risk. Certain levels of risk may trigger the need to pause or halt model development or training, while other levels of risk may trigger the need to stop development or de-deploy the model (see OpenAI 2023).

Endnotes Section

It appears that Endnote 49 citation in Appendix B (Table 1 on p. 19) of NIST AI 800-1 ipd seems intended to refer to endnote 48. There is no Endnote 48 citation in the Appendix. We recommend fixing the Endnote 49 citation in the Appendix, and changing it to become an Endnote 48 citation.

References

Chloe Autio, Reva Schwartz, Jesse Dunietz, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall, Kamie Roberts (2024) NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. National Institute of Standards and Technology, https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, Krystal Jackson (2023a) AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models, Version 1.0. UC Berkeley Center for Long-Term Cybersecurity, https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

Anthony M. Barrett, Jessica Newman, Brandie Nonnecke, Dan Hendrycks, Evan R. Murphy, Krystal Jackson (2023b) AI Risk Management Standards Guidance for General Purpose AI and Foundation Models. In AAAI 2023 Fall Symposium, *Assured and Trustworthy Human-Centered AI (ATHAI)*, Arlington, VA, Oct 26, 2023. https://drive.google.com/file/d/1pdSUYGs7dEjvwrbJKOodMoQ7VMfi2_Td/view

Anthony M. Barrett, Krystal Jackson, Evan R. Murphy, Nada Madkour, Jessica Newman (2024) Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models. *arXiv*, https://arxiv.org/abs/2405.10986

Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, & Zhangjun Zhou (2024) Synthetic Data in AI: Challenges, Applications, and Ethical Implications. *arXiv*, https://arxiv.org/abs/2401.01629

Leonie Koessler, Jonas Schuett, Markus Anderljung (2024) Risk thresholds for frontier AI. *arXiv*, https://arxiv.org/pdf/2406.14713

Daniel McDuff, Theadore Curran, & Achuta Kadambi (2023) Synthetic Data in Healthcare. *arXiv*, https://arxiv.org/abs/2304.03243

NIST (2023) AI Risk Management Framework (AI RMF 1.0). AI 100-1. National Institute of Standards and Technology, https://doi.org/10.6028/NIST.AI.100-1

OpenAI (2023) Preparedness Framework (Beta). OpenAI, https://cdn.openai.com/openai-preparedness-framework-beta.pdf

Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson (2024) The AI Risk

Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. Massachusetts Institute of Technology, https://cdn.prod.website-files.com/669550d38372f33552d2516e/66bc918b580467717e194940_The%20AI%20Risk%20Repository_13_8_2024.pdf

Cedric Deslandes Whitney, Justin Norman (2024) Real Risks of Fake Data: Synthetic Data. Diversity-Washing and Consent Circumvention. *arXiv*, https://arxiv.org/abs/2405.01820

Philipe De Wilde, Payal Arora, Fernando Buarque, Yik Chan Chin, Mamello Thinyane, Serge Stinckwich, Eleonore Fornier-Tombs, & Tshilidzi Marwala (2024) Policy Guideline: Recommendations on the Use of Synthetic Data to Train AI Models. Tokyo: United Nations University, https://collections.unu.edu/eserv/UNU:9480/Use-of-Synthetic-Data-to-Train-AI-Models.pdf