CIfAI

Cognitive Insights for Artificial Intelligence

Comments on Initial Draft AI Risk Management Framework, National Institute of Standards and Technology.
Submitted by Monica Lopez, PhD, Chief Executive Officer
Organization, Cognitive Insights for Artificial Intelligence (CIfAI)
April 25, 2022

On behalf of Cognitive Insights for Artificial Intelligence (CIfAI), we write in response to the call for comments on the *Initial Draft AI Risk Management Framework* (AI RMF) by the National Institute of Standards and Technology (NIST). We support NIST's efforts to promote the development and use of trustworthy artificial intelligence (AI) technologies and systems, and particularly to address the existing initial draft released March 17, 2022 on the risks in the design, development, use, and evaluation of AI systems.

We at CIfAI provide strategic research-based solutions from a human-centered perspective to ensure the safe and ethical design, development, deployment, and management of AI-enabled autonomous systems across various industries. Our values-based approach is founded on accuracy, consistency, and context-dependency, and supports trusted data across every phase of the AI lifecycle to achieve confident and fair decision making.

We appreciate the opportunity to provide information on the following:
Whether the AI RMF appropriately covers and addresses AI risks, including with the right level of specificity for various use cases.

We believe the AI RMF is incomplete in appropriately covering and addressing AI risks. The AI RMF accepts risk as an inevitable fact and thus presents a framework to manage that inevitable risk. However, we believe this framework insufficient because we advocate the need to make every attempt to *prevent* risk from the onset. While the draft mentions "*thresholds and values can also determine where AI systems present unacceptable risks to certain organizations, systems, social domains, or demographics*"[1] and as such suggests the inclusion of risk thresholds

[1] AI Risk Management Framework: Initial Draft. NIST, p.7, lines 3-4. March 17, 2022. https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf

to prevent the designing, development, and deployment of an AI system capable of "*unacceptable risks*"[2] in the first place, the draft delegates voluntary responsibility to "*AI system owners, organizations, industries, communities, or regulators*"[3] to establish policies and norms regarding those risk thresholds. We argue this approach to be insufficient, and instead highlight the initial steps of the AI lifecycle —the assumptions behind data and algorithms— as paramount to preventing risk.

Motivation for Recommendations

AI systems have the capability to not only improve people's lives, but help manage some of the world's hardest problems. To be able to fulfill this potential, however, we need to first address the major challenge associated with the core of AI technology: data and algorithms. Without data and algorithms there is no AI. For an AI system to function, whether for simple or complex use, an algorithm provides the instructions to perform a task. Algorithms are created from mathematical functions and assumptions, and they are written in different programing languages. These steps provide the instructions the algorithmic model will follow. To run an algorithmic model the data, structured and/or unstructured, must be curated. Algorithms are then trained using the curated data sets via supervised or unsupervised learning, and the desired accuracy of the algorithm's outcome is determined by the training of the algorithm.

While these steps seem straightforward, they are theoretically and technically profoundly complex and ethically worrisome. The starting point begins with human judgment. As a result, risk management must begin from the onset of assumptions made about the data to be used and the algorithms to be designed. Otherwise, we have a domino effect whereby bias and discrimination, surveillance, privacy, and other ethical concerns snowball into evermore larger social implications as AI-enabled systems exponentially grow in complexity and in use cases.

The most controversial areas of AI systems are the algorithms used (e.g. classification, regression, clustering) and their applications (e.g. employment, healthcare, law enforcement, wearable tech). Controversies arise from a lack of algorithmic transparency. The following questions highlight the issues at stake: Who identifies the problem to be solved with AI? Who decides how to solve the problem? Who determines the type of data to use? Who collects the data? Who curates and prepares a data set? Who chooses the programing language to write the algorithm's code? Who decides which mathematical assumptions and functions should be used to build the algorithm? What type of learning should be used to train the algorithm? To underscore the plethora of choices to be made just to create and train an algorithm, there are at least 81 mathematical approaches and functions used in machine learning (ML).[4] On top of this mathematical/statistical complexity before us, mathematicians face an unsolvable problem

---

[2] Ibid.

[3] Ibid. p. 7, line 10.

[4] Deisenroth, M. P., Faisal, A. A., & Ong, C. S. 2020. Mathematics for Machine Learning. Cambridge University Press, https://mml-book.github.io/book/mml-book.pdf

known as the mathematical logic paradox[5] which points out the learnability problem in machine learning of generalizing knowledge from limited data.

As a result, it is paramount to assess and devise risk management strategies that start at the beginning of the AI lifecycle before development. This is of particular urgency for systems classified as high-risk.[6] High-risk systems include AI technology used in critical infrastructures; educational or vocational training; safety components of products; employment, management of workers, and access to self-employment; essential private and public services; law enforcement that may interfere with people's fundamental rights; migration, asylum, and border control management; administration of justice and democratic processes;[7] and emotion recognition systems.[8] Many of these high-risk technologies are already in use. Take, for example, the use of AI to threaten democracy through the creation of deep fakes, propagation of disinformation, dissemination of propaganda, and manipulation of public discourse.[9] The correction of biases, discriminatory profiling, and privacy violations, among other threats, is not a switch to be turned off and on at will. Guardrails must be built from the onset to prevent such threats to human rights.[10] These high-risk AI systems must be created under strict supervision, inspection, and regulation and be subjected to proper risk management procedures across the design process. While ethical principles are being proposed and adopted across the AI industry, an approach based on human rights is a far more robust framework for the ethical and lawful development and use of AI systems.[11] Given that data sets and the generation of algorithms are foundational to creating AI systems, it is without doubt that responsibility falls squarely on the scientists (i.e. mathematicians, programmers, engineers) building these systems.

The *Initial Draft AI RMF* does not mention the initial building blocks —data and algorithms— of the AI lifecycle. Moreover, three critical components remain unaddressed: i) determination of the type of AI system to create, ii) appropriate selection of data sets, and iii) development and

---

[5] Castelvecchi, D. 2019. Machine learning leads mathematicians to unsolvable problem. Nature *565*(7737), 277-278.

[6] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. European Commission. April 21, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[7] Regulatory Framework Proposal on Artificial Intelligence. European Commission. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[8] Malgieri, G., & Ienca, M. 2021. The EU regulates AI but forgets to protect our minds. European Law Blog. https://europeanlawblog.eu/2021/07/07/the-eu-regulates-ai-but-forgets-to-protect-our-mind/

[9] Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report. Stanford University. https://ai100.stanford.edu/sites/g/files/sbiybj18871/files/media/file/AI100Report_MT_10.pdf

[10] Urgent Action Needed Over Artificial Intelligence risks to human rights. United Nations. September 15, 2021. https://news.un.org/en/story/2021/09/1099972

[11] Berthet, A. 2019. Why do emerging AI guidelines emphasize 'ethics' over human rights? Open Global Rights. https://www.openglobalrights.org/why-do-emerging-ai-guidelines-emphasize-ethics-over-human-rights/

training of algorithms. If these three components are not controlled and regulated, and their respective risks assessed and mitigated, AI's impact will become irreversibly chaotic. Regulation and oversight is paramount at this level of development and is imperative to ensure that everyone is developing AI safely. Otherwise, we have after-the-fact risk mitigation in light of catastrophic outcomes.

Risks generated by AI systems, their negative impacts, and mitigation strategies have been proposed.[12,13] Risks can be created if:

- the reason to build the AI system is wrongly established,
- transparency is lacking on the sourcing, collecting, cleaning, and selecting of data,
- data are insufficient and unrepresentative,
- built algorithmic models drift due to inappropriate training,
- the algorithm training is carried out with the least explainable deep neural networks,
- the team building the AI system lacks diversity,
- there is no oversight over the team building the AI system,
- monitoring throughout the AI lifecycle is lacking, and
- the overall system violates fundamental human rights.

These points underscore how organizational strategy, technical methods, people involved, processes followed, transparency and explainability, and regulatory frameworks and compliance thereof are together the various parts of the AI lifecycle that necessitate control to prevent the building of major risk.

We suggest the following recommendations to prevent the above. These recommendations attempt to address the initial stages of an AI system's development with the goal of achieving their transparency, oversight, and compliance.

RECOMMENDATIONS

1. Achieve visibility, transparency, and accountability of mathematical functions and models used to create AI systems.

Create a controlled database for the registration of mathematical functions and algorithmic models, including the organization and regulation of open-source AI libraries.[14] This necessitates a standardization system to record, for example, high-risk AI systems, their intended use case, the techniques and technologies utilized, relevant source code, and regular updates on changes. The assignment of distinct serial numbers, like that required for motor vehicles, appliances, and

---

[12] Cheatham, B., Javamardian, K., & Samandari, H. 2019. Confronting the risks of AI. McKinsey & Company. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence

[13] Baquero, J. A., Burkhardt, R., Govindarajan, A., & Wallace, T. 2020. Derisking AI by design: How to build risk management into AI development. McKinsey & Company. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Derisking%20AI%20by%20design/derisking-ai-by-design-build-risk-management-into-ai-dev.pdf

[14] 7 Best Open Source AI Libraries. Analytics Steps. July 21, 2021. https://www.analyticssteps.com/blogs/7-best-open-source-ai-libraries

other mechanical devices, can be helpful for tracking and cross-checking mathematical functions and algorithmic models used with related and upgraded AI systems. Research teams must comply by registering their system in the database.

2. Enable beneficial AI R&D.
Issue guidance to scientists developing AI systems on the various features related to data and algorithm building that necessitate avoidance to prevent untrustworthy and adhere to human rights-abiding AI. This establishes a benchmark for scientists to actively integrative human rights principles, and indicates that any process or method that does not confirm such principles would be in violation of such thereof.

3. Incentivize cross-disciplinary team collaboration.
Ensure that the guidance issued in Recommendation 2 demands that the team of scientists developing AI systems must work side-by-side with diverse professionals from different ethnicities, genders, professions, and socio-economic levels. This will ensure that the computer science-centric model of AI system building is eliminated, and that the entire AI lifecycle —from inception to market introduction and usage at scale— is truly collaborative and integrates a plethora of skills and perspectives.

4. Require algorithmic transparency.
Ensure that the guidance issued in Recommendations 2 and 3 also requires specification on and justification for the type of AI system to be developed, the kind of data to be collected and curated, the logical mathematical functions to be used for creating the algorithmic model, and the selected training approach for ML.

5. Include interdisciplinary, independent oversight.
Create an interdisciplinary, independent oversight committee and a public body responsible for overseeing and inspecting the AI systems developed before they are deployed. Interdisciplinary means including professionals from different ethnicities, genders, and professions, and disciplines.

6. Strengthen society with human rights advocacy.
Ensure human rights experts are included in the inspection and investigation of AI systems. A diverse population of people from different ethnicities, gender, and professions must be included the discussions and/or debates to ensure AI systems comply with all aspects of human rights and are not a threat to democracy

7. Enforce company transparency.
Impose on companies creating AI systems to disseminate information about their high-risk AI systems. Such information must include the organization’s name, the start and end dates of manufacturing, and the specific use purpose of their high-risk AI system.

8. Consider the concerns of workers.
Require the development of mechanisms for whistleblowers and establishment of safeguards against retaliation. Company workers must be protected when they challenge an AI-driven automated decision and/or when the AI system does not follow compliance of the above recommendations during development.

9. Supply training resources.
Provide resources for appropriate training to support the independent staff operating, overseeing and inspecting AI systems.

10. Enforce compliance.
Impose fines and penalties to companies for non-compliance with human rights' impact assessments and risk mitigation requirements. Fines should depend on the nature and severity of the regulatory non-compliance.

11. Encourage international cooperation.
Ensure AI systems comply with international standards given that AI systems are bound to be for international use. Such standards include the OECD AI principles and recommendations,[15] G20 AI principles,[16] and the proposed European law on artificial intelligence.[17]

[15] Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments. May 21, 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#backgroundInformation

[16] G20 AI Principles. OECD.AI. June 9, 2019. https://oecd.ai/en/wonk/documents/g20-ai-principles

[17] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. European Commission. April 21, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206