moz://a

Comment: AI Risk Management Framework

Developing a trustworthy AI ecosystem will require a major shift in the norms that underpin our current computing environment and society. Since 2019, Mozilla Foundation has focused a significant portion of its internet health programs on AI. Building on existing work, Mozilla published a white paper, Creating Trustworthy AI, that analyzes the current AI landscape and offers potential solutions.

We're pleased to offer up the following comment on the AI Risk Management Framework:

---

2. How organizations currently define and manage characteristics of AI trustworthiness and whether there are important characteristics which should be considered in the Framework besides: Accuracy, explainability and interpretability, reliability, privacy, robustness, safety, security (resilience), and mitigation of harmful bias, or harmful outcomes from misuse of the AI;

3. How organizations currently define and manage principles of AI trustworthiness and whether there are important principles which should be considered in the Framework besides: Transparency, fairness, and accountability;

With respect to items 2 and 3:

1. **Require that AI systems be designed such that tools for public oversight and auditing are key to a transparency strategy.**

AI systems must be developed in a way that enables third party validation and audit. In some contexts, companies or platforms may be compelled to develop data archives or

public APIs that researchers, journalists, and other watchdogs can use to study patterns of discrimination or harm.

Several social media platforms have developed open political ad libraries that provide detailed information about the advertisements appearing on its platform, a first step towards empowering third parties to audit the platforms. However, when Mozilla assessed Facebook's Ad API ahead of the 2019 EU elections, many researchers told us that the API did not allow them to download machine-readable data in bulk, nor was the data comprehensive and up-to-date.

Clear, accurate, and meaningful information about the AI system must be provided, which may include: detailed documentation about the model, information about the source code and training data, normative descriptions of decisions made about the system, and the release of public transparency tools.

2. Definitions of terms like "bias" and "fairness" must be developed in consultation with those communities that are most impacted by AI systems.

Even when steps have been taken to reduce bias in a model, that system can still make decisions that have a discriminatory effect. For instance, when Facebook changed its ad platform to prevent advertisers from targeting attributes like "ethnic affinity" for categories like housing or jobs, it was determined that the platform still enabled discrimination by allowing advertisers to target users through proxy attributes.[1]

When it comes to racial bias and inequity in our technologies, the scholar Ruha Benjamin observes that an intention to "do good" can also "coexist with forms of malice and neglect."[2] Good intentions do not stand up when weighed against documented harms.

As such, any efforts to address bias and discrimination in AI must work with those communities most impacted by such systems. The Algorithmic Justice League (AJL) has developed new strategies for more effective algorithmic auditing of AI systems in order to assess a system's readiness for deployment, with a goal of pushing for improvements or abolition of the system, prioritizing affected populations.

8. How organizations take into account benefits and issues related to inclusiveness in AI design, development, use and evaluation—and how AI design and development may be carried out in a

[1] Till Speicher et al., "Potential for Discrimination in Online Targeted Advertising," in *FAT 2018 - Conference on Fairness, Accountability, and Transparency*, vol. 81 (NY, US, 2018), 1–15, https://hal.archives-ouvertes.fr/hal-01955343.

[2] Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*, Polity, 2019.

way that reduces or manages the risk of potential negative impact on individuals, groups, and society.

With respect to item 8:

3. Require enhanced processes for inclusive AI design.

More participatory processes will need to be developed for consulting with diverse communities throughout the AI product life cycle. This will require teams to adopt a more open approach to how AI systems and products are developed, using frameworks and tools such as participatory design, co-design, or design justice.[3] Particular features of the AI system should be thoroughly tested with diverse user groups across geographic regions and languages before being deployed.

Mozilla's Common Voice is an example of a participatory voice AI project with distributed governance across geographic regions and languages. Common Voice is an open source, multi-language dataset of voices that anyone can use to train speech-enabled AI, representing 40+ languages worldwide. New approaches to data governance are also being explored in Mozilla's Data Futures Lab, which takes an inclusive design approach to developing reusable data infrastructure.

11. How the Framework could be developed to advance the recruitment, hiring, development, and retention of a knowledgeable and skilled workforce necessary to perform AI-related functions within organizations.

With respect to item 11:

4. Incorporate considerations for computing education, training, and ongoing development of AI workforce.

Engineers, product managers, designers, and other members of the cross-functional teams building AI wield a great degree of decision-making power. Some universities have moved toward making ethical computing courses required for CS/Eng students, and also making these courses more practical. In a recent landscape analysis of 115 university courses in tech ethics, researchers conclude that while CS as a discipline has been slow to adopt ethical principles, it has made a great deal of progress in recent years. They recommend

[3] Sasha Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need*, Information Policy, MIT Press, 2020.

that students hear the message that "code is power" when they first start learning how to code and that this message should be reinforced throughout coursework.[4]

There are many initiatives underway aimed at helping developers think critically about their work, such as Mozilla's Responsible Computer Science Challenge, a project which brings an ethics lens into university STEM coursework. One output of the project, Mozilla's Teaching Responsible Computing Playbook, serves as a resource and a blueprint for STEM educators seeking to transform CS/Eng curricula.

Developing a more effective AI risk assessment framework requires multistakeholder engagement across diverse sectors. To improve the trustworthy AI landscape, rules and standards must reflect a clear, socially and technically grounded vision.

Mozilla Foundation appreciates the opportunity to comment on the standards efforts that are underway, and we look forward to future opportunities to weigh in on the framework.

[4] Casey Fiesler, Natalie Garrett, and Nathan Beard, "What Do We Teach When We Teach Tech Ethics?: A Syllabi Analysis," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (SIGCSE '20: The 51st ACM Technical Symposium on Computer Science Education, Portland OR USA: ACM, 2020), 289–95, https://doi.org/10.1145/3328778.3366825.