To the Information Technology Laboratory,
National Institute of Standards and Technology,
100 Bureau Drive, Mail Stop 8900,
Gaithersburg, MD 20899–8900.

Subject: NIST AI Executive Order

Dear NIST Review Committee,

I am writing on behalf of AI2030, an initiative launched by FinTech4Good, with a commitment to fostering responsible Artificial Intelligence (AI) development for a better world. Our approach is rooted in the AI 2030 Responsible AI Framework, which is built upon six foundational pillars— fairness, transparency, accountability, privacy, sustainability, and safety & security— designed to steer the holistic design, development, utilization, and deployment of AI. We engage with a broad range of experts, policymakers, and stakeholders in our efforts, as outlined on our website.

In response to the Request for Information (RFI) issued by the National Institute of Standards and Technology (NIST) pertaining to the Executive Order on Safe, Secure, and Trustworthy Development and Use of AI, AI2030 aims to offer insights based on principles such as Shared Responsibility, Differentiated Responsibility, Enhanced Responsibility for AI infrastructure providers, SME(Small and Medium Enterprises) Empowerment in AI Integration, Balanced Innovation and User Protection in AI, Flexible Mechanism and Sandbox Approach in AI Regulation, and the Collaboration and Ecosystem Principle. Our response touches upon the following key areas in alignment with AI 2030 mission. To avoid redundancy, the principles highlighted are implied within each section of our response.

- **AI Model and System Evaluation (MSE)**
    - *Establish Operational AI MSE Guidelines for Developers:*
        - Develop a regulatory framework that is adaptable to the specific contexts and risks of different applications, such as healthcare, human resources, finance etc. This framework should outline MSE exercises for AI at the application level, emphasizing unique and diverse human interactions within each sector.
        - Include requirements for application-specific testing against a broad spectrum of potential harms. These guidelines must be broad yet adaptable, designed to address societal impacts, ethical concerns, and unintended use cases across different fields.
        - Foster collaboration among academic and non-academic institutions to create open-source tools, streamlining compliance with AI MSE guidelines by reducing technical complexities and associated costs

Suggested Implementation Mechanisms:

- Establish a regulatory committee composed of AI ethicists, legal experts, and technologists, alongside practitioners experienced in developing and deploying MSE methodologies across the AI lifecycle.
    - This committee will define MSE scope and methods, tailoring them to the distinct needs of those creating new base models versus those developing applications on top of these models[1].
- Facilitate sector-specific workshops to gather insights on diverse human interactions and potential AI harms, encouraging the exploration of multiple approaches beyond existing MSEs to enhance AI safety.

    - *Mandate Continuous MSE Throughout AI Development:*
        - Implement policies requiring AI developers to conduct MSE at multiple stages of AI system development, not just post-deployment.
        - Require developers to integrate feedback from MSE exercises into ongoing AI system refinement and updates.

Suggested Implementation Mechanisms:
- Introduce legislation requiring periodic MSE across the AI development lifecycle.
- Set up compliance monitoring bodies to ensure integration of MSE feedback into AI updates.

    - *Transparency and Reporting Requirements:*
        - Enforce disclosure norms for AI developers to publicly report the methodology, findings, and actions taken as a result of MSE exercises.
        - Create a public repository for these reports to enhance transparency and facilitate cross-industry learning.

Suggested Implementation Mechanisms:
- Implement regulations mandating the disclosure of MSE processes and findings.
- Establish a central public repository for MSE reports to facilitate learning and transparency.

    - *Support Research and Methodological Improvements:*
        - Allocate funding for research into advanced methodologies for AI MSE, focusing on addressing broader societal risks and ethical challenges.
        - Foster partnerships[2] between academia, industry, and government agencies to develop best practices and innovative approaches to AI MSE.

---

[1] For example, OpenXAI is a versatile, lightweight library aimed at evaluating the reliability of post hoc explanation methods with state-of-the-art implementations and user-friendly APIs, supporting new datasets, methods, and metrics, though it has not yet reached widespread adoption; proposals for enhancing its scale include open-source models, responsible databases, and improved labeling mechanisms.(https://open-xai.github.io/).

[2] For instance, Microsoft, Anthropic, Google, and OpenAI have launched the Frontier Model Forum, a coalition dedicated to promoting the safe and responsible development of advanced AI models through safety research, establishing best practices, facilitating knowledge sharing, and leveraging AI to tackle

Suggested Implementation Mechanisms:

- Allocate government funds to support research into AI MSE guidelines, methodologies, and frameworks, while fostering collaboration among academia, industry, and government entities to establish and refine MSE best practices.
    - Establish a Partnership Framework: Create a formal structure for collaboration between academic and non-academic institutions, defining roles, responsibilities, and mutual benefits to encourage the development of open-source tools.
    - Consider employing a research and funding approach inspired by the Defense Advanced Research Projects Agency (DARPA) to foster innovation and successful practices in AI development and deployment, drawing on prior examples of successful collaborations for guidance.
    - Create a Shared Resource Repository: Fund initiatives where institutions can share, access, and contribute to a repository of open-source tools, resources, and best practices for AI MSE compliance.
    -

- **Generative AI Risk Management**

    - *Develop Ethical Guidelines for Generative AI:*
        - Formulate a comprehensive set of ethical guidelines and technical standards, incorporating insights from the NIST AI Risk Management Framework[3]. Address data privacy, bias prevention, content integrity, algorithmic transparency, and user consent, enhancing data sourcing standards with data cards and data nutritional labels[4], and detailing dataset diversity and annotation contributors.
        - Differentiate between ethical guidelines and actionable standards, using model cards[5] for algorithmic transparency and interpretability, and include product design principles for privacy by design, user permissions for ML training, and transparent AI disclosures in line with regulatory expectations.
        - Extend guidelines to cover risk management strategies, advocating for partnerships with social media and other platforms to address generative AI risks, and propose stricter access regulations to prevent malicious use, ensuring a balanced approach between ethical considerations and technical challenges.

---

societal challenges, while encouraging participation from other organizations in the field (https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/).

[3] https://www.nist.gov/itl/ai-risk-management-framework

[4] For instance, consider collaborative AI community platforms like Hugging Face (https://huggingface.co/docs/datasets/dataset_card) to develop data cards and dat nutritional labels.

[5] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T., 2019, January. Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229), https://arxiv.org/abs/1810.03993.

        - Craft specialized guidance for Gen AI infrastructure providers, clearly distinguishing their unique responsibilities from those of other stakeholders throughout the GenAI development lifecycle.

Suggested Implementation Mechanisms:

- Convene a diverse working group including ethicists, data scientists, legal professionals, and stakeholders from sectors relevant to generative AI to refine ethical guidelines and standards, referencing the NIST AI Risk Management Framework and incorporating Model Cards and Data Nutritional Labels principles.
- Host sector-specific and inclusive public consultations, complemented by workshops, to integrate a wide range of perspectives, ensuring the guidelines are nuanced, actionable, and aligned with privacy by design principles, algorithmic transparency, and regulatory expectations for AI use disclosure.

- *Implement Mandatory Labeling Systems:*
    - Mandate clear labeling of all generative AI-produced content, enabling consumers to easily identify AI-generated outputs.
    - Develop user-friendly and informative labeling guidelines, incorporating a concise description of the AI's functions and limitations, and integrate these requirements with existing technology standards, akin to the ISO/IEC 27000 series for information security[6], to ensure consistency and reliability in disclosures.
    - Allocate funding to support the development of innovative solutions that ease the integration of labeling systems within SME operations.

Suggested Implementation Mechanisms:

- Collaborate with technology companies and industry bodies to define standard labeling formats and protocols.
- Integrate labeling requirements into existing technology standards and enforce them through regulatory bodies or industry self-regulation.

- *Certification Process for Generative AI Systems:*
    - Create a certification process that evaluates generative AI systems for compliance with ethical standards, transparency, and data privacy norms.
    - The certification could include a comprehensive review of the AI's development process, data handling practices, and output analysis.
    - Tailor the certification process distinctly for infrastructure providers, large enterprises, and SMEs, ensuring it meets the specific needs and capabilities of each group.

[6] https://www.iso.org/news/ref2266.html

Suggested Implementation Mechanisms:

- Establish a regulatory body or an independent agency tasked with AI system certification, focusing on audits covering AI development, data management, algorithmic decisions, and post-launch performance. To mitigate innovation bottlenecks and cost concerns that could advantage only large companies, propose limiting mandatory certification to entities with a significant user base, excluding academic research.
- Direct funds to empower existing certification bodies to undertake AI system certification responsibilities effectively.
- Introduce a tiered certification process, with the rigor of requirements scaling according to company size. This strategy aims to balance thorough evaluation for larger companies against a more manageable approach for smaller firms, ensuring a fair and adaptable framework as the AI sector matures.

- **Reducing Risk of Synthetic Content**

    - *Develop and Implement Detection Mechanisms:*
        - Develop state-of-the-art algorithms to identify synthetic content, with a specific focus on content generated by Generative Adversarial Networks (GANs) and their variants, continuously improving as AI technologies evolve[7].
            - Additionally, prioritize the development of algorithms and technologies that can not only identify synthetic content but also detect instances where synthetic content is being used in a coordinated fashion or by malicious actors.
        - Encourage ongoing research to adapt these algorithms to the rapidly changing landscape of AI-generated content, including deepfake images and videos produced by GANs.
        - Explore application-specific mechanisms for synthetic content prevention, addressing issues like fake social media content, automatic translations, fake job postings, and fake news, potentially through audit mechanisms like the NYC anti-bias HR law[8].

Suggested Implementation Mechanisms:

- Collaborate with universities, tech companies, and telecommunications providers to integrate AI-based detection algorithms directly into devices and systems used to consume media, such as internet browsers, mobile phones, and televisions, to

[7] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., & Thies, J. (2019). Face Forensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. arXiv preprint arXiv:1912.06673. https://arxiv.org/abs/1912.06673

[8] https://news.bloomberglaw.com/daily-labor-report/new-york-city-targets-ai-use-in-hiring-anti-bias-law-explained

provide comprehensive protection against unknown synthetic media consumption.
- Ensure that the detection of synthetic content is decoupled from content creators and distributors, establishing a trustworthy and impartial system that prevents potential abuse, whether by distributors or malicious actors.
- Fund and support ongoing research projects focused on the development of advanced detection technologies for GAN-generated synthetic content and explore methods to enhance privacy preservation, including the incorporation of differential privacy mechanisms into the training of GANs[9].

    - *Launch Educational Campaigns:*
        - Initiate comprehensive educational campaigns and resources, both online and through global conferences, to raise awareness about synthetic content.
        - Focus on educating the public about the nature and risks of synthetic content, using diverse mediums for outreach.

Suggested Implementation Mechanisms:
- Collaborate with educational institutions, NGOs, and media organizations to create comprehensive educational materials that explain the technologies behind synthetic content generation, such as GANs, while emphasizing the significance of device-level protection.
- Utilize online platforms, social media, and community outreach programs to disseminate information effectively, ensuring the public understands the nature and risks associated with synthetic content, including GAN-generated content.

- **Advancing Responsible Global Technical Standards**

    - *Establish International AI Standards Body:*
        - Advocate for the creation of an international body dedicated to AI standards, whether within existing international frameworks like the UN or G20 or as an independent entity, ensuring it commands sufficient global influence and participation[10].
        - Focus on including ethical considerations in the standard-setting process, ensuring respect for privacy, non-discrimination, transparency, and accountability.
        - Consider adding provisions for self-certification by AI companies to demonstrate compliance with applicable laws and standards.

---

[9] Frigerio, E., Qiu, Q., & Aste, T. (2020). Differentially private generative adversarial networks for time series, continuous, and discrete OpenAI data. Scientific reports, 10(1), 1-13. https://www.nature.com/articles/s41598-020-64977-1

[10] Consider leveraging and building upon existing AI principles like the G7 Hiroshima AI principles (http://www.g8.utoronto.ca/summit/2023hiroshima/231030-ai-principles.html) to inform and guide the development of global AI standards.

Suggested Implementation Mechanisms:

- Initiate diplomatic engagements and international policy dialogues to form the proposed standards body.
- Develop a framework for membership, governance, and operation of this body, ensuring fair and diverse representation.

    - Promote Ethical Integration in AI Standards:
        - Emphasize the importance of integrating ethical considerations into global AI standards.
        - Suggest a proactive approach to ethical AI development, encompassing a range of societal and cultural contexts.

Suggested Implementation Mechanisms:

- Organize international symposiums and workshops with ethicists, technologists, and policymakers to discuss ethical integration in AI.
- Create an open-access portal for sharing emerging practices and standards in ethical AI, encouraging global participation.
- Direct financial resources towards the creation of innovative researches, solutions and tools aimed at simplifying the adoption process and reducing the costs associated with adoption of these standards.
- Allocate funding and resources to empower training organizations to offer relevant, low-cost training programs for SMEs and initiate campaigns to enhance awareness among SMEs.

We hope these contributions will assist NIST in developing effective guidelines and standards for AI. AI2030 is committed to contributing towards shaping the future of AI development and deployment, balancing innovation benefits with safeguarding against risks.

Thank you for considering our input in this vital initiative.

Sincerely,
The AI2030 Team

Lead contributors:

Daniela Muhaj
Chair, AI2030 Global Fellow Program

Xiaochen Zhang
Executive Director, AI2030
CEO, FinTech4Good

Expert reviewers and contributors:

1. Ryan Gurney
2. Vinicius Moura
3. Carlos Pinto

4. Smita Rajmohan
5. Christopher Smedberg
6. Nedelina Teneva
7. Kyle Walter

Appendix:

1. **About AI 2030**
   The AI 2030 Initiative is a member-based program aiming to create the world's largest community focused on responsible AI. Its goal is to harness AI's power for humanity's benefit while minimizing negative impacts. Presently, the initiative prides itself on a diverse and vibrant network of over 250 experts in responsible AI, hailing from a myriad of industries across more than 30 countries. This collective expertise and global perspective empower AI 2030 to lead, innovate, and collaborate in the responsible development and application of artificial intelligence.

2. **AI 2030 Responsible AI Framework**