

BigBear.ai Holdings, Inc.
6811 Benjamin Franklin Drive
Suite 200
Columbia, Maryland 21046

February 2, 2024

Filed Electronically

James St. Pierre
Information Technology Laboratory
National Institute of Standards and Technology
U.S. Department of Commerce
100 Bureau Drive
Mail Stop 8900
Gaithersburg, Maryland 20899

Dear Director St. Pierre:

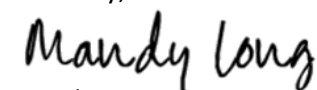
BigBear.ai Holdings, Inc. (BigBear.ai) appreciates the opportunity to submit the following comments to the National Institute of Standards and Technology (NIST) concerning guidelines for auditing artificial intelligence (AI) systems and models, synthetic content labeling, and global technical standards development. Specifically, this letter seeks to provide recommendations and industry perspective in response to the topics put forth in the request for information issued by NIST on December 21, 2023, docket number NIST-2023-0009.

BigBear.ai recognizes and commends NIST's ongoing work to carry out the several responsibilities assigned to the agency under the current Administration's executive order on AI released in October 2023. As one of the few companies supporting the federal government in pioneering AI orchestration, BigBear.ai shares the Administration's goal of ensuring continued U.S. leadership in the development and deployment of trustworthy and secure AI technologies in the public and private sectors.

As NIST continues its work to identify best practices, tools, and methods for evaluating and managing AI capabilities and establish a global engagement plan for the development of AI technical standards, BigBear.ai strongly encourages the development of guidelines, policies, and recommendations that encourage responsible innovation and the advancement of AI software and tools for the benefit of U.S. citizens and businesses.

Given BigBear.ai's two decades of decision intelligence experience with a proven track record of execution, its cutting-edge defense and intelligence business, and a growing commercial footprint, we respectfully submit the following comment letter in response to NIST's RFI relating to the AI executive order. We appreciate your consideration of the following comments and are available to answer any questions you may have.

Sincerely,

A handwritten signature in black ink that reads "Mandy Long". The signature is written in a cursive, flowing style.

Mandy Long
CEO
BigBear.ai

About BigBear.ai

Headquartered in Columbia, Maryland, BigBear.ai is a global, public AI software company and leader in the use of AI and Machine Learning (ML) to support operational decision-making in real-world environments. BigBear.ai delivers AI-powered decision intelligence solutions to customers in three core markets: global supply chains and logistics, autonomous systems, and cyber. BigBear.ai's customers include some of the world's most intricate enterprises, from the U.S. Department of Defense, U.S. intelligence community, and hospitals, to Fortune 500 manufacturing and retail companies. Customers rely on BigBear.ai's AI-powered solutions to address gaps in complex data sets, supply chains, and security networks in an accessible and scalable way, empowering leaders to predict outcomes and make better decisions, faster. With 20 years of experience partnering with the public and private sectors, BigBear.ai is one of the few AI software companies whose products and solutions possess high levels of technological readiness, offering immediate value for customers across complex defense, cyber, and commercial landscapes. BigBear.ai's mission is guiding customers to realize their best possible future by delivering transformative technologies and expert, actionable advice, and the company's leaders and employees are unified in a shared commitment to harness the potential of AI and lead the industry in solving some of society's most complex problems and creating a better world through better choices.

Answers to Specific Topics Posed by NIST in the AI Executive Order RFI

Developing Guidelines, Standards, and Best Practices for AI Safety and Security

1. **Developing a companion resource to the AI Risk Management Framework (AI RMF).**

NIST should consider the following potential risks and harms of generative AI related to trustworthiness, repression, and human rights when developing a companion to the AI RMF:

Misuse of Generated Content. Generative AI can be used to create realistic looking, but fabricated content such as images, videos, and text, which can be exploited for malicious purposes such as spreading misinformation, creating fake news, or impersonating individuals.

Privacy Concerns. Generative AI models have the potential to infringe upon individuals' privacy by generating synthetic data that closely resembles real data. This raises concerns about the potential for unauthorized access to personal information and the erosion of privacy rights.

Ethical Implications. The use of generative AI raises ethical questions regarding the ownership and authenticity of generated content. It challenges traditional notions of authorship and intellectual property rights, leading to debates about who should be held responsible for the content generated by AI systems.

Manipulation and Bias. Generative AI models can perpetuate and amplify biases present in the training data, leading to the generation of biased content. This can exacerbate existing social inequalities and reinforce harmful stereotypes, potentially leading to discrimination and social unrest.

Security Risks. Generative AI models are vulnerable to attacks and adversarial manipulation, which can compromise their integrity and reliability. Adversarial actors may exploit vulnerabilities in AI systems to generate malicious content or manipulate the output of AI models for nefarious purposes.

Regulatory Challenges. The rapid advancement of generative AI technology poses challenges for regulatory frameworks designed to safeguard against potential harms. Regulators face the task of developing effective policies and guidelines to mitigate risks associated with the proliferation of generative AI while fostering innovation and economic growth.

Trustworthiness and Transparency. Ensuring the trustworthiness and transparency of generative AI systems is crucial for building user confidence and facilitating widespread adoption. However, achieving transparency in AI systems can be challenging due to the complexity of underlying algorithms and the opacity of decision-making processes.

The companion to the AI RMF should address best practices for AI actors. NIST should consider the following examples of best practices when developing recommendations:

Ethical Guidelines and Standards. AI actors, including researchers, developers, and industry stakeholders, should establish or adopt ethical guidelines and standards governing the development, deployment, and use of generative AI systems. This includes principles such as fairness, transparency, accountability, and privacy by design.

Risk Assessment and Impact Analysis. Both AI developers and end users should conduct comprehensive risk assessments and impact analyses to identify potential risks and harms associated with generative AI systems throughout their lifecycle. This involves evaluating the social, ethical, and legal implications of AI technologies and implementing appropriate mitigation measures.

Transparency and Explainability. AI actors should prioritize to the greatest extent possible or practicable transparency and explainability in the design and implementation of generative AI systems. This includes providing clear documentation of AI algorithms, data sources, and decision-making processes to facilitate understanding and scrutiny by external stakeholders.

Data Governance and Privacy Protection. Robust data governance frameworks and privacy protection mechanisms are crucial for ensuring the responsible collection, storage, and use of data used to train and operate generative AI systems. This involves implementing data anonymization, encryption, access controls, and other security measures to safeguard sensitive information.

Human Rights and Social Impact Assessments. Conducting human rights and social impact assessments can be key tools for evaluating and understanding the potential implications of generative AI technologies on individuals, communities, and society as a whole. This includes assessing the risks of discrimination, bias, exclusion, and other adverse effects on vulnerable populations and marginalized groups.

Multi-stakeholder Engagement and Collaboration. AI actors should engage in multi-stakeholder dialogue and collaboration with government agencies, civil society organizations, academia, and other relevant stakeholders to address the complex governance challenges associated with generative AI. This involves fostering open and inclusive discussions, sharing best practices, and co-developing policy solutions.

Regulatory Compliance and Accountability. AI actors should comply with applicable laws, regulations, and industry standards governing the responsible use of generative AI technologies.

This includes establishing mechanisms for accountability, oversight, and redress in cases of misconduct, negligence, or harm caused by AI systems.

The necessary professions, skills, and disciplinary expertise that organizations require to govern generative AI effectively is a complex ethical issue. The specific aspects of effective AI governance vary based on the type of generative AI in question, and organizations require a diverse range of talent to execute their missions. However, the following professions and expertise are key to effectively governing generative AI:

AI Ethics and Policy Experts. Professionals with expertise in AI ethics, policy analysis, and regulatory compliance are essential for developing and implementing ethical guidelines, standards, and regulatory frameworks governing the responsible use of generative AI technologies.

Legal and Regulatory Specialists. Lawyers, legal scholars, and regulatory experts play a critical role in interpreting and applying existing laws and regulations relevant to generative AI, as well as advocating for new legislation to address emerging challenges and risks.

Data Scientists and Machine Learning Engineers. Data scientists and machine learning engineers possess the technical expertise required to develop, train, and evaluate generative AI models, as well as identify and mitigate risks associated with data quality, bias, and algorithmic transparency.

Cybersecurity and Privacy Professionals. Cybersecurity experts and privacy specialists are responsible for assessing and mitigating security threats, vulnerabilities, and privacy risks associated with generative AI systems, including data breaches, unauthorized access, and information leakage.

Human Rights and Social Impact Researchers. Professionals with backgrounds in human rights, social sciences, and impact assessment are instrumental in evaluating the potential societal, cultural, and ethical implications of generative AI technologies, as well as advocating for the protection of individual rights and collective well-being.

Interdisciplinary Teams and Collaborative Networks. Effective governance of generative AI requires interdisciplinary collaboration and knowledge exchange among diverse stakeholders, including researchers, policymakers, industry leaders, civil society organizations, and community representatives. Interdisciplinary teams and collaborative networks enable holistic problem-solving and informed decision-making in complex and rapidly evolving regulatory environments. This also includes pure philosophy specialist as well in order to understand the core behaviors.

Ethical Designers and User Experience Specialists. Designers and user experience specialists focus on designing intuitive, inclusive, and ethical user interfaces and interactions for generative AI systems, as well as promoting user empowerment, autonomy, and informed consent in decision-making processes.

Public Engagement and Stakeholder Relations Professionals. Public engagement specialists and stakeholder relations professionals facilitate transparent communication, community engagement, and stakeholder participation in the governance of generative AI, fostering trust, accountability, and public legitimacy in AI policymaking and implementation processes.

2. **Creating guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities and limitations through which AI could be used to cause harm.**

Creating guidance and standards for evaluating and auditing AI capabilities is a critical aspect of responsible AI development and deployment, as it helps ensure the reliability, fairness, and safety of AI systems across various domains and applications.

NIST should consider the following definitions, types, and designs of test environments, scenarios, and tools when developing guidance and benchmarks to evaluate AI:

Test Environments. It is important to establish diverse and representative test environments that simulate real-world conditions and scenarios relevant to the intended use of AI systems. This includes considering factors such as data diversity, environmental variability, and task complexity to accurately evaluate AI capabilities and performance.

Scenarios. Designing comprehensive test scenarios involves defining clear objectives, criteria, and metrics for assessing AI capabilities, limitations, and safety across different use cases and deployment contexts. This may include testing for robustness, generalization, adversarial robustness, safety-critical functionality, and ethical compliance.

Tools for Evaluation. Developing standardized tools and methodologies for evaluating AI capabilities enables consistent and systematic assessment of AI systems' performance, reliability, and safety. This may involve leveraging techniques such as benchmarking, validation, verification, and simulation to measure and compare the effectiveness and efficiency of AI algorithms and models.

Benchmarks for evaluating bias in data, models, and AI lifecycle practices should focus on the following elements:

Data Biases. Addressing biases in data requires careful examination of data collection methods, sampling techniques, and data preprocessing pipelines to identify and mitigate sources of bias that may lead to unfair or discriminatory outcomes. This involves promoting diversity, inclusivity, and representativeness in training datasets and adopting bias detection and mitigation strategies during the data preprocessing stage.

Model Biases. Evaluating biases in AI models involves analyzing model performance across different demographic groups, socioeconomic contexts, and cultural backgrounds to identify disparities and inequities in prediction accuracy and decision-making outcomes. This requires adopting fairness-aware evaluation metrics, fairness-enhancing algorithms, and model explainability techniques to mitigate biases and enhance transparency.

AI Lifecycle Practices. Implementing structured mechanisms for bias detection and mitigation throughout the AI lifecycle involves integrating fairness, accountability, and transparency (FAT) principles into AI development, deployment, and monitoring processes. This includes establishing governance frameworks, audit trails, and quality assurance mechanisms to ensure responsible AI practices and mitigate potential risks of bias and discrimination.

Guidance and standards for evaluating AI capabilities should also address structured mechanisms for gathering human feedback, including:

Human-Subject Trials. Conducting human-subject trials and user studies enables researchers and developers to gather qualitative and quantitative feedback on AI systems' usability, effectiveness, and user satisfaction. This involves engaging diverse user populations in interactive sessions, surveys, interviews, and usability tests to elicit insights, preferences, and concerns about AI-driven applications and interfaces.

AI Red-Teaming. Implementing AI red-teaming exercises involves simulating adversarial attacks, edge cases, and failure modes to stress-test AI systems' robustness, resilience, and security. This includes deploying skilled red teams or ethical hackers to identify vulnerabilities, exploit weaknesses, and uncover potential risks and vulnerabilities in AI algorithms, models, and infrastructure.

Red teaming of high-risk generative AI models is one of the most important requirements of the executive order. The term “red teaming” is loosely defined as “a structured testing effort to find flaws and vulnerabilities in an AI system.” AI red teaming plays a crucial role in enabling the deployment of safe, secure, and trustworthy AI systems by identifying and mitigating risks, vulnerabilities, and failure modes. By addressing use cases, best practices, capabilities, limitations, risks, and harms associated with AI red-teaming, organizations can enhance their ability to assess and manage AI-related risks effectively while promoting the responsible and ethical development and deployment of AI technologies. For an AI system, however, red teaming might not involve actual hacking at all. For example, one way to attack an LLM is to prompt it in a way that bypasses any restrictions or guardrails that its developers may have placed on it. Most LLM chatbots are purposefully designed not to output harmful or toxic content such as hate speech. However, many users have discovered various prompt hacks, or “jailbreaks,” that subvert these controls. These prompt hacks take the form of natural-language instructions that are given to the AI model, usually once it has been fully trained and deployed for use in a software application like a chatbot. For users without access to the inner workings of the models themselves, experimenting with prompts is an accessible, attention-grabbing way to explore the systems’ capabilities and limitations.

NIST should consider the following use cases, best practices, capabilities, limitations, supplemental practices, risks, and harms when developing guidance for red teaming:

Use Cases. AI red-teaming can be particularly valuable in use cases where the deployment of AI systems introduces significant risks to safety, security, and trustworthiness. Examples may include autonomous vehicles, medical diagnostics, financial fraud detection, and cybersecurity applications where the consequences of AI failures can be severe.

Best Practices for AI Safety. Red-teaming exercises should involve the identification and exploration of realistic threat models that encompass a wide range of potential risks and vulnerabilities associated with AI systems. This involves adopting a proactive and adversarial mindset to anticipate and mitigate potential attack vectors, exploitation techniques, and failure modes.

Capabilities. AI red-teaming can help identify weaknesses, vulnerabilities, and failure modes in AI systems that may not be apparent through conventional testing methods. This includes exploring edge cases, adversarial inputs, and unforeseen interactions that could compromise system safety, security, and reliability.

Limitations. Red teaming exercises have certain limitations, including inherent uncertainties, assumptions, and simplifications inherent in modeling complex AI systems and adversarial behaviors, as well as resource constraints, time limitations, and the inability to fully replicate real-world conditions requiring careful interpretation and validation of results. Red teaming may also fail to detect unknown unknowns or novel attack vectors that have not been previously considered or anticipated.

Supplemental Practices. To address the limitations of red teaming, organizations may employ supplemental practices such as penetration testing, vulnerability assessments, adversarial training, and continuous monitoring to enhance the effectiveness and comprehensiveness of AI risk assessment and management efforts. These practices help identify, prioritize, and address vulnerabilities and threats throughout the AI lifecycle.

Risks and Harms. AI red teaming may inadvertently introduce risks and harms, such as disrupting production systems, causing unintended consequences, or compromising user privacy and security. It is essential to carefully manage and mitigate these risks by conducting red teaming exercises in controlled environments, following ethical guidelines, and implementing safeguards to protect against potential harm.

Reducing the Risk of Synthetic Content

NIST should consider the following tools, technology, and techniques for reducing the risk of synthetic content:

|| Content Authentication

Verification Mechanisms. Developing robust methods for authenticating content is crucial for distinguishing between genuine and synthetic media. This may involve the use of digital signatures, watermarking techniques, and cryptographic protocols to verify the integrity and provenance of digital assets.

Blockchain Technology. Leveraging blockchain technology can provide decentralized and tamper-proof ledgers for tracking the origin and history of digital content, enhancing transparency and accountability in content authentication processes.

|| Detecting and Labeling Synthetic Content

Machine Learning Algorithms. Implementing machine learning algorithms for detecting and labeling synthetic content can help identify patterns, anomalies, and artifacts indicative of content manipulation or generation by AI systems. This involves training models on diverse datasets containing both authentic and synthetic media to improve detection accuracy and generalization.

Multimodal Analysis. Incorporating multimodal analysis techniques, including image, video, audio, and text analysis, enables comprehensive examination of content features and characteristics across different modalities, enhancing the effectiveness of synthetic content detection and labeling approaches.

Adversarial Detection Techniques. Developing adversarial detection techniques enables proactive identification of ways in which malicious actors may attempt to circumvent content

manipulation protections. This involves analyzing adversarial attacks, evasion strategies, and exploitation techniques to anticipate and mitigate potential threats to content integrity and authenticity.

Robustness Testing. Conducting robustness testing and adversarial validation helps evaluate the resilience of content authentication and detection systems against various attack scenarios and adversarial inputs, enhancing their effectiveness and reliability in real-world environments.

|| Auditing and Maintaining Tools for Analyzing Content

Compliance Audits. Performing compliance audits and quality assurance checks ensures that content labeling and authentication tools adhere to established standards, guidelines, and regulatory requirements. This involves evaluating the accuracy, reliability, and consistency of labeling algorithms and authentication mechanisms through systematic testing and validation processes.

Continuous Monitoring and Updates. Implementing continuous monitoring and updates for content labeling and authentication tools helps address emerging threats, vulnerabilities, and evolving techniques used by malicious actors. This includes integrating feedback mechanisms, vulnerability assessments, and threat intelligence feeds to enhance the resilience and adaptability of content authentication and detection systems over time.

Advance Responsible Global Technical Standards for AI Development

The initiative outlined by NIST to advance responsible global technical standards for AI development is essential for promoting interoperability, trustworthiness, and responsible innovation in the global AI ecosystem.

NIST should focus on the following elements while developing AI-related consensus standards:

Consensus Building. Developing AI-related consensus standards requires broad-based collaboration and engagement among diverse stakeholders, including industry, academia, government agencies, standards organizations, and civil society groups. This involves facilitating open dialogue, knowledge sharing, and consensus-building processes to address complex technical, ethical, and regulatory challenges associated with AI development and deployment.

International Cooperation and Coordination. Promoting international cooperation and coordination is critical for aligning AI standards, guidelines, and best practices across different regions and jurisdictions. This involves fostering partnerships, information sharing, and capacity building initiatives to harmonize regulatory frameworks, streamline certification processes, and facilitate cross-border collaboration in AI research and innovation.

Design Standards to Align with the AI RMF and National Standards Strategy. Ensuring alignment with the principles set forth in the NIST AI Risk Management Framework and the U.S. Government National Standards Strategy for Critical and Emerging Technology helps promote consistency, coherence, and interoperability in AI standardization efforts. This involves integrating risk management principles, ethical considerations, and regulatory compliance requirements into the design, development, and implementation of AI-related standards.

NIST should work to ensure that AI-related consensus standards address the following topics:

AI Nomenclature and Terminology. Establishing clear and standardized nomenclature and terminology for AI technologies helps promote common understanding, communication, and interoperability among stakeholders. This includes defining key concepts, terms, and terminology related to AI development, deployment, and governance.

Best Practices for Data Capture, Processing, and Protection. Developing best practices for data capture, processing, and protection involves addressing issues related to data quality, privacy, security, and regulatory compliance throughout the AI lifecycle. This includes establishing guidelines for data collection, annotation, anonymization, and consent management to ensure responsible and ethical use of data in AI applications.

Trustworthiness, Verification, and Assurance of AI Systems. Establishing guidelines and standards for assessing the trustworthiness, verification, and assurance of AI systems helps build confidence and reliability in AI-driven technologies. This involves defining criteria, metrics, and evaluation methodologies for assessing the robustness, reliability, fairness, transparency, and accountability of AI algorithms, models, and systems.

Human-Computer Interface Design for AI Systems. Incorporating human-computer interface design principles into AI standards helps enhance usability, accessibility, and user experience in AI-driven applications. This includes designing intuitive, inclusive, and user-centric interfaces that facilitate human-AI interaction, understanding, and decision-making.

Inclusivity of Stakeholder Representation. Ensuring inclusivity of stakeholder representation in standards development processes promotes diversity, equity, and transparency in decision-making. This involves engaging a diverse range of stakeholders, including underrepresented groups, marginalized communities, and non-traditional actors, in standards-setting activities to reflect a broad spectrum of perspectives and interests.

Adoption and Implementation Strategies. Developing strategies for driving adoption and implementation of AI-related international standards involves raising awareness, building capacity, and providing technical assistance to stakeholders. This includes promoting education and training programs, facilitating technology transfer and knowledge exchange initiatives, and incentivizing voluntary compliance with international standards and guidelines.

Implications for Competition and International Trade. Anticipating and addressing the potential implications of standards for competition and international trade is essential for fostering a level playing field and promoting fair competition in the global marketplace. This involves conducting impact assessments, monitoring market dynamics, and fostering collaboration with industry stakeholders and trade partners to mitigate trade barriers, harmonize regulatory requirements, and promote interoperability in AI markets.

In summary, advancing responsible global technical standards for AI development requires concerted efforts to promote collaboration, coordination, and consensus-building among stakeholders worldwide. By addressing key topics such as AI nomenclature, data practices, trustworthiness, inclusivity, adoption strategies, and trade implications, NIST will help facilitate the development of interoperable, trustworthy, and ethically aligned AI technologies that benefit society while addressing emerging challenges and opportunities in the global AI landscape.