RFI for NIST AI Executive: ref Order-88 FR 88368

Jointly submitted by

1) Ken Huang
   CEO
   DistributedApps.ai
   [redacted]

2) Mehdi Bousaidi
   Co-Founder and COO
   HORUS Technology Solutions, Inc.
   [redacted]

Submitted via Email to
ai-inquiries@nist.gov

1: Introduction

We are pleased to provide comments to the Department of Commerce's National Institute of Standards and Technology (NIST) in response to its Request for

Information (RFI) on Artificial Intelligence (AI) with document reference Order-88 FR 88368.

This response represents a collaboration between DistributedApps.ai, an AI safety consulting and training company with extensive prior collaborative efforts on generative AI security with OWASP, the Cloud Security Alliance, and the NIST Generative AI Public Working Group, and HORUS Technology Solutions, an IT solutions provider specializing in integrating AI-enabled technologies and cloud solutions for mission critical systems. It contains insights and recommendations drawn from our combined experience consulting, deploying, and auditing generative AI systems across multiple industries.

Our response aims to contribute constructive perspectives on several crucial areas highlighted in NIST's RFI, including defining roles and responsibilities for trustworthy AI development, establishing benchmarks and processes for LLM applications evaluation, suggesting best practices for AI red teaming to uncover vulnerabilities, addressing the emerging threats of synthetic data, and providing guidance on global AI standards.

Across these topics, we emphasize cross-functional collaboration, rigorous testing and auditing, concrete oversight mechanisms, and adaptable policy frameworks as essential ingredients for responsible advancement of AI on local and global scales. Our suggestions balance pragmatic steps with ambitious vision, acknowledging the rapidly evolving landscape of AI progress and associated risks.

We believe this RFI represents a vital first step in consolidating diverse viewpoints from AI practitioners towards crystallizing practical wisdom. The insights can seed ongoing dialogue and eventual coordination between public and private sectors to nurture AI that enhances our society technically, ethically and socially. We are excited to contribute suggestions on this shared endeavor for trustworthy AI.

Our response is based on our experience and we used AI technology to correct grammar errors and polish the contents when needed.

2: Recommendation for AI Roles and Responsibilities

Key recommended roles needed to ensure responsible and ethical AI development and deployment include ethicists to guide principles, policy experts to craft governance frameworks, social scientists to analyze societal impacts, developers to implement safety checks and oversight, deployers to assess risks and determine oversight needs, and users to engage responsibly. Additionally, emerging roles like Chief AI Officers to oversee strategy, AI Security Experts to develop safeguards, Red Teaming Professionals to identify vulnerabilities, AI Engineers to adhere to development lifecycles, and MLOps Professionals to implement DevSecOps in ML development and deployment; are critical across industries utilizing AI. A cross-functional collaboration drawing on this diverse expertise, centered on AI Safety and ethical governance, is vital to address the unique risks introduced by generative AI models and strengthen overall AI governance.

3: Evaluating RAG Based LLM Applications

Evaluating Retrieval-Augmented Generation (RAG) based Large Language Models (LLMs) requires assessing their Accuracy, Completeness, lack of Toxicity, and Reduced Hallucination (ACT), along with the Relevance, Equality, and Legality (REL) of their responses. Accuracy evaluates factual correctness. Completeness measures how comprehensive the responses are. Toxicity Removal checks that responses are not harmful or offensive. Relevance assesses pertinence to the queries. Equality checks for bias across groups. Legality ensures compliance with regulations. Employing this "ACT REL" framework allows for systematic and ethical evaluation of LLM-powered applications across metrics like quality, reliability, and trust. The following are the steps:

3.1. Foundation Model Selection:

Choosing an appropriate foundation model is essential because the base model profoundly impacts the behavior of a derivative system across crucial dimensions like safety, fairness and controllability. Analyzing foundation model candidates against the full ACT REL criteria via red teaming and adjacent techniques will deeply inform selection or rejection for a given use case. The highest potential for beneficial downstream performance arises from foundation models proven safe, legal and equitable from their genesis.

3.2. RAG Pipeline Evaluation:

- In assessing applications utilizing the Retrieve Augmented Generation (RAG) pipeline, it's critical to evaluate the Retriever and Generator components both individually and collectively. This holistic evaluation helps identify potential areas for improvement within the RAG pipeline. See Author Ken Huang's blog at https://cloudsecurityalliance.org/blog/2023/11/22/mitigating-security-risks-in-retrieval-augmented-generation-rag-llm-applications/

3.3. RAG Evaluation Metrics:

The field of RAG evaluation is rapidly evolving, with various approaches and frameworks emerging, such as the
RAG Triad (https://www.trulens.org/trulens_eval/core_concepts_rag_triad/), ROUGE (https://aclanthology.org/W04-1013/), ARES (https://arxiv.org/abs/2311.09476), BLEU (https://dl.acm.org/doi/10.3115/1073083.1073135?ref=blog.langchain.dev), and RAGAs (https://arxiv.org/pdf/2309.15217v1.pdf) metrics. Our framework integrates and extends these frameworks with our "ACT REL" framework.

3.4. Golden Dataset Construction:

A vital component of evaluation is the Golden Dataset, which should include 'ground truth' labels, often derived from human feedback, to gauge the LLM's performance accurately. Constructing such a dataset is an intensive process.

Here are some key considerations for constructing a golden dataset to evaluate large language models:

- Domain Coverage: The dataset should cover the key domains and tasks you want to evaluate performance on.

- Data Collection: Human annotators need to label data and provide ground truth responses.
- Quality Control: As human labeling introduces errors, having mechanisms to filter low quality data is important e.g. using test questions to screen annotators, having multiple annotators per item and using a consensus response.
- Diversity: The data should have diversity in length, complexity, genre etc. to properly evaluate model robustness across different input types.
- Balancing: The proportion of different input types and label categories should match intended real-world use cases to help understand performance in production environments.
- Updating: As models evolve, new datasets may be needed to avoid overfitting. Having processes to continually collect human annotated evaluations on new domains is ideal.
- Analysis Set: A section of the data should be held-out from model training to serve as an unbiased model analysis set.

3.5 Sample Tools for LLM Application Evaluation:

We have used the following tools in the past and would like to recommend them in this RFI.

3.5.1. MLflow 2.9.1

MLflow 2.9.1 introduces tools for evaluating the efficiency of retrievers. This includes the ability to assess embedding models, threshold choices, and chunking strategies using the MLflow evaluate API. See: https://mlflow.org/docs/latest/llms/llm-evaluate/notebooks/question-answering-evaluation.html

3.5.2. DeepEval Framework:

DeepEval is an open-source framework for evaluating language model applications. It provides default metrics for assessing various aspects such as hallucination and relevancy. The framework is designed to evaluate performance based on metrics such as hallucination, answer relevancy, and other aspects using language model models and various other NLP models locally on your machine. It offers simple functions to unit test language model applications and provides support for evaluating existing language model applications built with other frameworks. The default metrics offered by DeepEval are classic metrics, meaning they do not use language model models for evaluation. The framework is designed to make it easy to build and iterate on language model applications, allowing for the customization of evaluation datasets and metrics. DeepEval presents an opinionated framework for the types of tests that are being run, breaking down language model outputs into categories such as answer relevancy, factual consistency, conceptual similarity, bias, and toxicity. The framework also provides support for creating custom metrics to evaluate language model outputs.

https://docs.confident-ai.com

https://github.com/confident-ai/deepeval

3.5.3 Arize: LLM Evaluation

The Arize Phoenix LLM Evals library is an open-source library designed for simple, fast, and accurate evaluations of Large Language Models (LLMs). It is optimized to run evaluations quickly and supports various evaluation approaches depending on available data and use cases. The library provides pretested evaluation templates and convenience functions for a set of common evaluation tasks, such as toxicity evaluation, summarization, and classification. It also offers support for one-click explanations, fast batch processing, and custom dataset creation. The Phoenix LLM Evals library is integrated with LangChain and LlamaIndex, and it is designed to be used in Python pipelines, notebooks, and app frameworks. The library aims to provide deeper capabilities around LLM observability, allowing AI engineers and data scientists to evaluate, troubleshoot, and fine-tune LLM models effectively
https://arize.com/blog-course/llm-evaluation-the-definitive-guide/

3.5.4: OpenAI Eval

The OpenAI Evals framework provides a standardized set of evaluation metrics and tasks to compare different models' performance. An "eval" is a task used to measure the quality of output of an LLM or LLM system by generating an output from an input prompt and evaluating it with a set of ideal answers to find the accuracy. The framework includes a registry of challenging evals and provides a CLI to convert samples to a JSONL file and register the eval. The OpenAI Evals framework is used to evaluate various tasks, such as abstractive summarization, using traditional evaluation methods like ROUGE and BERTScore, as well as novel approaches leveraging LLMs as evaluators. The framework is actively maintained, and there are plans for its future development.

See reference below
https://github.com/openai/evals

Https://towardsdatascience.com/how-to-best-leverage-openais-evals-framework-c38bcef0ec47?gi=9e684448f3c6

3.5.5: UpTrain

The UpTrain project is an open-source LLM (Large Language Model) evaluation toolkit. It provides a framework for evaluating LLMs and LLM systems, offering pre-built metrics such as response relevance, context quality, factual accuracy, and language quality. The toolkit is built with customization at its core, allowing users to configure custom grading prompts and operators with Python functions. It also offers features like seamless logging and evaluation, deeper insights through "evaluate_experiments," and prebuilt evaluations for quick assessment. The project is actively developed, and users can contribute to the repository or create issues for feature requests and integrations.

https://github.com/uptrain-ai/uptrain

4: Comments on Red-Teaming

Red-teaming in generative AI is a critical process, involving structured attempts to challenge, stress-test, and expose potential vulnerabilities in AI systems. In this section, we provide comments on various aspects of AI red-teaming.

Beneficial Use Cases for AI Red-Teaming in Risk Assessment and Management:

AI red-teaming is particularly beneficial in scenarios where AI systems have significant impacts on decision-making, privacy, or safety. For instance, in healthcare, red-teaming can test AI diagnostics for biases or errors. In finance, AI systems used for credit scoring can be red-teamed to identify unfair biases. Red-teaming is also crucial in autonomous vehicle development, where the stakes of AI failure are high. The followings are additional beneficial use cases:

- Algorithmic Recruiting Systems: As AI tools are increasingly used to screen resumes, conduct interviews, and evaluate candidates, red teaming can help surface issues like biases, gaming vulnerabilities, or violations of equal opportunity hiring practices before deployment.
- Predictive Policing: With data-driven risk assessment tools informing police resource allocation, red teams can audit for "feedback loops”, where over-policing of minority areas reinforces biased data collection and unfair arrests predictions.
- Government Services: Evaluating AI eligibility and approval systems for welfare, food stamps, housing assistance, etc. via red teaming helps safeguard ethical access to vital public services.
- Child Safety: AI techniques for detecting child exploitation imagery or grooming patterns online should undergo rigorous red team review to protect privacy and avoid false accusations.
- Infrastructure Monitoring: Before relying on AI to autonomously evaluate risks or anomalies in power grids, dams, water systems etc., red teaming helps validate safe failure modes.

Limitations of AI Red-Teaming and Mitigation:

Based on our experience, we identities the following limitations of AI Red-teaming

- Resource intensive. Conducting rigorous red teaming of complex AI systems requires substantial technical expertise, computing resources, and time. This limits the number of AI systems that can be thoroughly evaluated.
- Difficulty replicating real-world conditions. While red teams try to simulate real-world conditions, it is impossible to perfectly replicate the diversity, complexity, and evolving nature of how AI systems are ultimately deployed. This can lead to over or under estimation of risk.
- Arms race stalemate. Sophisticated AI developers continuously adapt systems and detection approaches in response to known red team attack tactics. However, the rapid evolution of AI capabilities also leads to new potential vulnerabilities and attack surfaces faster than red teams can adequately simulate and test. Red teams may lack the resources, visibility, and time to effectively probe every new advancement. This can create a volatile threat landscape that outpaces the capacity for comprehensive security testing, allowing unknown risks to emerge. Proactive collaboration between red teams, developers and regulators can help address the gaps, but some degree of uncertainty is inevitable.

- Ethical considerations around permission and deception. Certain more aggressive red team techniques could violate terms of use or require deception, putting testers in an ethical quandary. Standards around permission and transparency requirements are still evolving in this relatively new domain.

As mitigation strategy, ongoing monitoring of user feedback and field reports as one supplementary practice. Others that can help plug red teaming blind spots include:

- Algorithmic audits to assess model fairness, explainability and accountability
- Moral machine experiments modeling human values preferences
- Participatory design incorporating perspectives from impacted communities
- Ethics boards and external watchdog oversight beyond technical testing

By recognizing red teaming's constraints, and bridging gaps with complementary evaluation approaches, organizations can get greater assurance of reliable and ethical AI outcomes. The union of techniques provides defense-in-depth.

Best Practices in AI Red-Teaming for Safety:

While AI Red-Teaming is a very complex and evolving domain, we have followed the following best practices in the past:

- Define realistic threat models based on intended use cases: Prior to testing, outline relevant adversarial threats like data poisoning, model stealing, prompt injection, etc. For example, an AI medical diagnosis system will prioritize robustness against data manipulation, while a chatbot may focus on guarding against extracting its model parameters.
- Employ a wide range of attack approaches: Testing strategies should cover both white-box attacks that have internal system access, and black-box attacks that interact externally. Strategies could include corrupted training data, perturbed inputs, model inversion, and more.
- Test with appropriate safeguards: Build testing protocols to prevent actual harm, like data destruction or service disruption, while allowing insightful attacks. For example, cancer prediction models were red teamed by research groups by attempting adversarial medical image contamination, but with safeguards against impacting real patients.
- Complement with other evaluation paradigms: Red teaming provides valuable but narrow insights that focus on intentionally breaking systems. Holistic assessments should also analyze beneficial use cases, human-AI

collaboration, explainability, accountability, ethics and related factors to enable trustworthy AI via defense-in-depth.

The goal is tailored, ethical probing that uncovers problems early while preventing real-world damage. Combined with broad functional testing and risk analysis, red teaming helps discover unknown issues to create safer, more robust AI.

Review Across AI Lifecycle for Effective Red-Teaming:

We suggest to have red teaming efforts across the AI lifecycle:

Design Stage:
At the initial design phase, red teams should scrutinize the foundational architectures, data sources or model selections and guardrails intended to ensure safe, beneficial generative AI systems. Testing methodologies at this stage may focus more on theoretical vulnerabilities rather than live experiments.

Development Stage:
As core components get built, more rigorous security probing and adversarial attacks can be conducted in sandboxed environments. The goal is discovering flaws early before real-world deployment.

Pre-Deployment Stage:
Prior to public release, red teams should conduct tests on production-equivalency systems under simulated real-world conditions across a range of scenarios. For large language models like GPT4 and Claude, this may involve techniques like fine-tuning copies on unfiltered internet data to assess unsafe response tendencies.

Post-Deployment Stage:
Once generative AI systems are working in the field, ongoing security reviews are critical to address evolving threats. Real traffic analyses, black box prodding for failure modes and monitoring for evidence of malicious use are all deployment phase review tactics. For example, Google constantly red teams its public APIs and cloud offerings using planned attacks and automated vulnerability scanners.

By weaving red team perspectives throughout the generative AI lifecycle, from architecture to post-deployment, with both internal and third-party testing, organizations can surface the most impactful issues early on. This maximizes safety while minimizing cost.

Sequence of Actions and Documentation in AI Red-Teaming:

A typical sequence involves planning, execution, analysis, and response. Documentation is crucial at each step, from the initial plan detailing objectives and methods to the final report outlining findings and recommendations. For instance, documenting the response of a generative AI system to various test scenarios helps in future risk mitigation strategies.

Planning Stage
- Clearly define objectives and constraints
- Develop hypothetical threat scenarios/adversarial models
- Detail testing methods and procedures
- Establish metrics for success
- Document in formal planning memo

Execution Stage
- Perform authorized attacks and probes
- Collect data on system responses across scenarios
- Track adherence to documented procedures
- Flag any deviations or test expansions
- Record both quantitative metrics and qualitative observations

Analysis Stage
- Organize and interpret test results
- Identify successful attacks and unintended behaviors
- Assess against pre-defined success criteria
- Highlight most severe vulnerabilities or failures
- Document full technical analysis report

Response Stage
- Present findings to system designers and owners
- Provide ranking and recommendations
- Collaborate on mitigations and design improvements
- Update vulnerabilities tracking as issues are addressed
- Issue final red team outcome report

With meticulous documentation throughout, red teaming yields maximum security and safety insights while providing evidence and accountability for the AI development team to implement fixes - leading to more robust systems.

Information Sharing Best Practices:

Sharing information about vulnerabilities and attack vectors must balance transparency with security. One approach is anonymizing data before sharing or using secure channels. For instance, sharing findings with other AI developers can foster industry-wide improvements without compromising proprietary information.

Here are some suggestions for information sharing best practices:

Anonymize Data

When sharing data that contains sensitive or proprietary information, remove any personally identifiable information or details that could compromise confidentiality. Anonymizing data allows information to be shared safely.

Use Secure Channels

When sharing sensitive information, use secure communication channels such as encrypted email, secure file transfer services, or password-protected documents. This helps prevent unauthorized access.

Coordinate Disclosure

If sharing details on a security vulnerability, coordinate disclosure with involved parties by giving them advance notice before publicly releasing details. This allows time to fix issues before exploits occur.

Share Minimum Necessary

Only share the minimum level of detailed information needed to convey the issue or findings. Oversharing risks exposing unnecessary proprietary information.

Establish Guidelines

Have clear guidelines on what type of information can be shared and with whom. Get legal/compliance advice when needed.

Sign NDAs

Consider non-disclosure agreements if sharing unpublished research or sensitive proprietary details with third parties. NDAs legally protect confidentiality.

Build Trust Networks

Cultivate trusted circles for sharing special access programs, sensitive research or confidential initiatives. Sharing with vetted partners can enable more transparency.

Optimal Composition of AI Red Teams:

Red teams should comprise individuals with diverse backgrounds, including technical experts, domain specialists, and ethicists. This diversity ensures a comprehensive assessment. For example, a red team for a medical AI system might include data scientists, medical professionals, and bioethicists.

Here are some suggestions for optimally composing AI red teams:

Include a mix of domain experts and creative thinkers. Domain experts like doctors, lawyers, engineers bring real-world perspectives. Creative thinkers like ethicists, researchers, philosophers approach problems in novel ways.

Cover both technical and ethical expertise. Technical experts evaluate system architecture, code, data quality, and performance metrics. Ethicists evaluate potential harms, bias issues, and policy implications.

Aim for diversity of thought, background, gender, ethnicity and age. Diverse teams minimize groupthink, leverage unique perspectives, and surface more issues.

Select some red team members through adversarial AI and algorithm auditing competitions. High competition performers demonstrate skill in breaking systems.

Establish a rotating membership with limited term lengths. Rotating members minimizes insider threats and stagnant thinking over time through infusion of fresh perspectives.

Maintain separation between red team testers and developers. Independence avoids conflicts of interest and improves objectivity.

Appoint senior red team leadership with past auditing experience and high status in the organization. Leadership buy-in signals importance and increases transparency.

Overall composition should balance technical capabilities with creative, ethical and critical thinking skills backed by team diversity and independence. This helps ensure comprehensive, unbiased review of AI systems.

Economic Feasibility for Different Organizations:

For large organizations, comprehensive red-teaming is generally feasible and necessary due to the scale of impact. For smaller organizations, cost-effective methods like collaborative red-teaming (sharing resources with other organizations) or focused red-teaming (targeting critical areas) can be adopted.

Large companies (Fortune 500/Big Tech) - Comprehensive red teams are feasible and standard practice given substantial budgets and high risk tolerance. Teams can include 12-50+ full time ethicists, auditors, technologists.

Mid-size companies (over 500 employees) - Smaller red teams (5-10 employees) or outsourced red team audits on major initiatives are economically reasonable. Focus on ethics, legal compliance and technical soundness.

Startups/Small companies (<100 employees) - Budget limitations may preclude internal red teams so external audits could be done on a project basis. Collaborative models like shared red team resources across similar startups can distribute costs.

Academic labs - Pooling university resources across departments for shared red teams brings down costs. Students get real-world experience while labs get increased oversight.

Independent researchers - Leveraging hackerspaces, open source communities, volunteer groups of interdisciplinary experts provides affordable access to external red teaming.

Government agencies - Red teams are mandated for evaluating law enforcement, military and intelligence programs where public trust matters most. Funding is secured given regulatory requirements.

Non-profits - Grants focused specifically on ethics/oversight can pay for external red team audits against criteria like algorithmic fairness, transparency, interpretability etc.

The feasibility analysis balances the downside risks the organization faces against time and funding constraints. Prioritizing red team access for high risk AI systems is key.

Appropriate Unit of Analysis:

The unit of analysis depends on the AI application. For standalone models, the model itself may be the focus. In system-wide deployments, the entire ecosystem, including user interfaces and data pipelines, should be considered. For instance, in a social media recommendation system, analyzing just the algorithm without considering the data inputs and user interface might overlook significant risks.

AI red-teaming is essential for the safe and ethical deployment of generative AI systems. It requires careful planning, diverse teams, and should be complemented with other risk assessment techniques to ensure comprehensive coverage of potential vulnerabilities

The appropriate unit of analysis for AI red teaming depends greatly on the scope and context of the AI system under review. Here are some additional considerations:

For narrow AI models, the model algorithms and training data may warrant the deepest inspection. However, the end-to-end pipeline should be evaluated including internal APIs, data ingestion systems, model integration processes, monitoring tools, and model outputs.

For enterprise-wide AI implementations, red teams need to assess the broader ecosystem including user interfaces, training infrastructure, pipeline orchestration systems, deployment architectures, model management systems, and more.

For consumer AI products, red teaming should simulate real-world user contexts. Testing focus may expand to client software, cloud APIs, mobile apps, websites, devices, manuals, marketing claims as well as algorithmic components.

For AI-as-a service offerings, audits should span developer APIs, SDKs, platforms,, documentation,educational materials in addition to core models.

For AI systems involving robots, autonomous vehicles, IoT devices etc the physical components, sensors, actuators, and manifestation of decisions into the real world should factor into red teaming.

Defining an appropriate scope starts with mapping the end-to-end AI assembly line. This helps identify high risk components and their intersections. Setting analytical boundaries too narrowly or widely can cause blindspots. The context and use cases should guide appropriate scope.

5: Comments on Reducing the Risk of Synthetic Content

In this section, we provide comments on a broad spectrum of concerns and methodologies in reducing the risk of synthetic content based on our previous experiences with our customers.

Authenticating Content and Tracking its Provenance:

Existing methods for content authentication and provenance tracking often rely on digital watermarking and metadata embedding. These techniques can be enhanced with blockchain technology to create immutable records of content creation and distribution. Future tools might include more sophisticated cryptographic methods or AI-driven analysis to verify the origin and authenticity of content. The challenge lies in developing universally accepted standards and ensuring interoperability across different platforms and media types.

Techniques for Labeling Synthetic Content:

Labeling synthetic content, such as deepfakes or AI-generated text, is crucial for transparency. Watermarking is a prevalent method, but future approaches might involve more subtle and robust techniques that are difficult to remove or alter without degrading the content itself. This could include steganographic methods or the integration of AI algorithms that can embed and detect unique content signatures.

Detecting Synthetic Content:

AI models are increasingly being used to detect synthetic content. These models are trained to recognize the subtle inconsistencies and artifacts that differentiate AI-generated content from authentic material. The ongoing challenge is the constant evolution of generative AI, which makes detection a moving target. Enhanced machine learning models, coupled with human expertise, are essential for effective detection.

Resilience of Labeling Techniques to Content Manipulation:

The resilience of labeling techniques is a critical concern. As synthetic content becomes more sophisticated, labeling methods must evolve to withstand manipulation attempts. This requires ongoing research and development, focusing on creating labels that are intrinsically bound to the content and degrade it if tampering is attempted.

Economic Feasibility for Organizations:

The adoption of these techniques varies in economic feasibility for small and large organizations. Larger organizations might have the resources to implement sophisticated systems, but for smaller entities, cost-effective and easy-to-use tools are necessary. Open-source solutions and scalable cloud services could play a significant role in bridging this gap.

Preventing Harmful Synthetic Content:

Preventing the generation of harmful synthetic content, such as child sexual abuse material or non-consensual intimate imagery, is of paramount importance. This involves not only technical solutions, such as filters and monitoring systems, but also legal and ethical frameworks to govern AI content generation. Collaboration with law enforcement and regulatory bodies is essential in this aspect.

Circumvention by Malign Actors:

Malign actors continually seek ways to circumvent detection and labeling techniques. This necessitates a proactive approach, where security measures are continuously updated and improved. It also involves educating users about the risks and signs of synthetic content.

Risk Profiles for Different Model Types:

Open-source models, due to their accessibility, have different risk profiles compared to closed models. With open-source models, there's a higher risk of misuse, necessitating more stringent monitoring and control mechanisms. Risk assessment should be an ongoing process throughout the AI development lifecycle.

Comprehensive AI Life Cycle Approaches:

Approaches that span the entire AI development and deployment lifecycle are crucial. This includes careful curation and filtering of training data, incorporating both automated and human feedback in the training process, and thoughtful consideration during model release. At different levels of the AI system, from the model to the application, security and authenticity measures must be integrated.

Testing Software:

Software used for detecting and analyzing synthetic content needs rigorous testing to ensure reliability and effectiveness. This involves not just technical testing but also ethical and legal considerations to ensure that the software does not infringe on privacy or other rights.

Auditing and Maintenance:

Finally, the tools and methods used for labeling and authenticating synthetic content must be regularly audited and maintained. This ensures they remain effective against new forms of synthetic content and are compliant with evolving legal and ethical standards.

In conclusion, addressing the risks associated with synthetic content in AI requires a holistic approach that encompasses technical, ethical, and legal dimensions. It demands continuous collaboration and innovation across various sectors to stay ahead of the evolving challenges.

6: Comments on Advance Responsible Global Technical Standards for AI Development

Global Ethical Norms and AI

We strongly recommend the development of a globally accepted set of ethical guidelines for AI, building upon initiatives like the European Union's Ethics Guidelines for Trustworthy AI. NIST can work with the UN to build a platform for synthesizing these diverse guidelines into a single, globally applicable framework. Such an endeavor would involve consultations with member states, academia, the private sector, and civil society to ensure a comprehensive and multi stakeholder approach.

Additionally, we suggest the NIST to work with the UN to spearhead an initiative to create standardized technical auditing protocols for AI systems. These protocols would be designed to monitor compliance with the established ethical norms. Given NIST’s reputation in global IT and security standards, and UN's global reach and authority, jointly, NIST and UN could facilitate the sharing of these protocols across member states, thereby enhancing the ethical compliance of AI systems worldwide.

To operationalize these technical auditing measures, NIST and UN could establish an international AI Ethics Auditing Body. This body would be responsible for certifying AI systems based on their adherence to global ethical norms and could offer training and resources for in-house auditing by professionals.

Moreover, we recommend that the NIST and UN actively promote machine learning models specifically trained to detect ethical discrepancies in AI systems.
This multi pronged approach ensures that ethical considerations are not just theoretical aspirations but embedded in the practical governance and auditing of AI systems globally.

Data Privacy in the Age of AI

Existing data protection frameworks, such as the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States, have laid important groundwork in their respective jurisdictions. However, the global nature of AI technologies requires a more harmonized approach to data privacy.

We strongly recommend that NIST and UN take the lead in crafting a Global Data Privacy Framework for AI. This framework should aim to standardize best practices in data anonymization and encryption, making these universally applicable and accepted. Such an endeavor would benefit from multi stakeholder input, including contributions from member states, privacy experts, technology companies, and civil society, to ensure a comprehensive and universally adaptable policy.

To ensure effective implementation of this framework, we suggest that NIST work with the UN to establish an international body focused on Data Privacy in AI. This entity would oversee the compliance of AI technologies with the global privacy standards, offering certifications for systems that meet the criteria. Moreover, this body could serve as a repository for best practices, guidelines, and resources related to data privacy in AI, thus serving as a valuable resource for all stakeholders, especially cybersecurity professionals.

Additionally, we propose that NIST and the UN facilitate the creation of a global standard for data sharing agreements. Such a standard could govern the ethical and secure sharing of data across borders, particularly critical for AI systems that rely on large and diverse datasets. This would involve complex legal and technical solutions, such as federated learning techniques that respect data sovereignty while allowing for cross border data utilization.

Through these initiatives, the UN's High Level Advisory Body on Artificial Intelligence can act as a catalyst in forming a more secure and privacy respecting global landscape for AI. By spearheading the development of universal data privacy norms and ensuring their global adoption, the UN can address one of the most pressing challenges in AI governance today.

National Security Concerns and AI

The intersection of Artificial Intelligence (AI) and national security presents a complex set of challenges with global implications. These challenges extend beyond ethical considerations to encompass national security issues, including asymmetrical warfare, espionage, and the integrity of critical infrastructure. To address these challenges effectively, a comprehensive framework is needed for global governance.

Firstly, NIST can recommend the formulation of a global agreement specifically focused on the deployment of AI in national security contexts. This agreement should establish the legal and operational boundaries for the use of AI in military and intelligence operations. It must also offer guidelines for engagement in asymmetrical warfare scenarios involving AI technologies.
To ensure compliance and effective enforcement, NIST and UN could model this governance framework on the lines of the International Atomic Energy Agency (IAEA). Much like the IAEA's role in nuclear non proliferation, an International Oversight Body for AI in National Security could be established under the UN's aegis. Utilizing the mandates under Chapter VII of the United Nations Charter, this body would have the authority to maintain or restore international peace and security. In the case of violations of the multilateral agreement, the Oversight Body could invoke sanctions under Article 41, which could range from trade embargoes on AI related technologies to financial restrictions aimed at curtailing AI research for national security purposes.

Given the sensitive and often classified nature of national security initiatives, a mechanism for secure third party auditing is also imperative. Auditors with the requisite security clearances could be authorized to conduct periodic evaluations of AI systems employed in national security contexts. This would serve as an additional layer of oversight and mitigate risks such as espionage and unauthorized data access, thereby ensuring that AI applications are in compliance with international norms and the multilateral agreement.

Furthermore, the establishment of a repository for best practices, technical standards, and countermeasures concerning AI in national security is recommended. This centralized resource would help member states align their national security strategies involving AI with global guidelines, thereby promoting a more secure and stable international environment.

Through these initiatives, NIST and UN can work together to provide a robust governance structure for the deployment of AI in national security contexts. By modeling this framework on successful precedents like the IAEA and integrating comprehensive multilateral agreements with stringent oversight and auditing mechanisms, we can strive for a future where AI technologies are deployed in a manner that enhances global peace and security, rather than undermining it.

Cross Border Data Flows and Sovereignty

The issue of cross border data flows and data sovereignty is an essential component of global AI governance. AI systems, especially those relying on machine learning algorithms, often require access to vast amounts of data, which frequently cross national borders. While this international flow of data can foster innovation and global collaboration, it also raises complex questions about data sovereignty, jurisdictional rights, and the security of data.

Given these challenges, we strongly recommend NIST and UN spearhead an initiative to establish a universal framework governing cross border data flows in the context of AI. This framework should lay down the legal and ethical guidelines for data transfer across jurisdictions, ensuring that it aligns with international norms and respects the sovereignty of nations. It should address key issues such as data localization requirements, mechanisms for lawful data transfer, and procedures for resolving jurisdictional disputes.

One way to address the technical challenges involved in secure and ethical cross border data flows is through federated learning and differential privacy techniques.

These technologies allow for the utilization of data for AI training and operation without physically transferring data across borders, thereby respecting data sovereignty while still enabling global collaboration. Therefore, the universal framework should include technical standards for federated learning and differential privacy to ensure that cross border data flows are both ethical and secure.

To monitor and enforce this framework, we propose the creation of a UN affiliated body dedicated to overseeing global data flows and data sovereignty in the context of AI. This body could be modeled on existing international governance structures, such as the World Trade Organization's framework for trade, but would focus exclusively on data issues as they relate to AI. This body would have the authority to arbitrate disputes, issue guidelines, and even impose sanctions in accordance with UN mandates if necessary.

AI in Global Critical Infrastructure

The integration of AI into critical infrastructure such as power grids, healthcare systems, and transportation networks is a burgeoning issue that demands stringent governance. While the use of AI in these sectors promises efficiency and optimization, it also introduces novel vulnerabilities that could be exploited, leading to potential disruptions with severe societal and economic ramifications. Given the critical nature of these sectors, any form of cyber attack or malfunction could have catastrophic consequences, not just for individual nations but potentially for global stability.

With this in mind, we strongly recommend that the NIST and the UN take the lead in developing robust cybersecurity protocols specifically tailored for AI implementations in critical infrastructure. These protocols should define the security measures, including but not limited to data encryption, multi factor authentication, and intrusion detection systems, that must be in place when AI technologies are integrated into essential services. Moreover, these protocols should be designed to adapt to the evolving nature of both AI technologies and cybersecurity threats, ensuring that they remain effective in safeguarding critical systems.

Given the international implications of a security breach in critical infrastructure, a multilateral approach is imperative. We propose the establishment of a specialized UN body, perhaps modeled on the lines of the International Atomic Energy Agency (IAEA), which has been effective in monitoring nuclear technology. This new body, which could be termed the International Agency for AI Security in Critical Infrastructure (IAAISCI), would be responsible for the oversight of AI security in critical sectors across member states. It would certify systems that comply with the established cybersecurity protocols and could have the authority to issue warnings or even sanctions under UN mandates for non compliance.

To ensure the efficacy and adaptability of these cybersecurity protocols, periodic auditing is essential. We recommend that third party cybersecurity firms with specialized knowledge in AI be authorized to conduct these audits under the oversight of the IAAISCI. These audits would assess the resilience of AI integrated systems against a range of potential cyber attacks and ensure compliance with international standards.

AI in Global Humanity Safety

The escalating advancement of AI technologies, which are increasingly capable of autonomous decision-making, underscores the vital need for rigorous governance focused on safety and alignment with human values. This critical aspect of AI development is not just about preventing minor errors or inefficiencies; it involves averting potential catastrophic outcomes that could arise from misaligned AI actions. The notion that AI systems, if not properly aligned, might execute actions harmful to human interests, or in worst-case scenarios, pose existential threats, is a profound concern that demands immediate and comprehensive attention.

To address these challenges, a cooperative approach involving major international bodies is essential. NIST, with its expertise in standards and technology, is well-positioned to collaborate with global organizations like the UN. Together, they could develop what might be termed a "Global Alignment Framework." This framework would not only establish universal guidelines for AI safety and value alignment but would also represent a significant step towards harmonizing AI governance across different nations and cultures.

This Global Alignment Framework would serve multiple functions. Primarily, it would define the core principles and standards to ensure that AI systems are developed with inherent safeguards against actions that could be detrimental to human welfare. It would also emphasize the alignment of AI systems with universally accepted human values, a complex yet crucial aspect given the diverse range of cultures and ethical perspectives worldwide.

Enforcement of such guidelines is equally important. The United Nations, through its extensive global influence and reach, could establish an International Safety and Alignment Body. This body's role would be multifaceted: conducting audits of AI technologies to assess their compliance with the Global Alignment Framework, certifying systems that meet the required standards, and possibly overseeing ongoing monitoring to ensure continued adherence. Such a body would not only foster trust in AI technologies but also work as a deterrent against the development and deployment of systems that could pose risks to humanity.

The collaboration between NIST and the UN to establish a Global Alignment Framework and an International Safety and Alignment Body represents a proactive and necessary measure. It acknowledges the immense potential of AI while simultaneously addressing the imperative need to ensure these technologies are developed and utilized in ways that are safe, ethical, and aligned with human values. This approach is not just about safeguarding our present; it's about securing a future where AI acts as a benefactor, not a threat, to human existence.

7: Contracting to the DistributedApps.ai and HORUS Team

We recommend that if NIST is interested in engaging the services of the DistributedApps.ai and HORUS Technology Solutions team, the issuance of a sole-source award would be a streamlined and efficient approach. HORUS Technology Solutions is an 8(a) SBA program-certified small business and as such can receive sole source awards. Based on Section 8(a) of the Small Business Act (15 U.S.C. 637 and FAR 19.8), HORUS may receive sole-source contracts for up to $7 million for acquisitions assigned manufacturing North American Industry Classification System (NAICS) codes and $4.5 million for all other acquisitions. In accordance with FAR 19.804-3(c), NIST may work with the SBA to issue a sole source award to our team, nominating 8(a) participant HORUS Technology Solutions as the intended recipient. Alternatively, NIST could work with GSA 8(a) STARS III to issue a sole source task order to HORUS Technology Solutions. Sole source orders on GSA 8(a) STARS III typically utilize the exception at FAR 16.505(b)(2)(i)(E) citing the Small Business Act (15 U.S.C. 637(a)) as the statutory authority. Whether through collaboration with the SBA or utilizing GSA 8(a) STARS III, our team is ready to help NIST realize the ethical and safe utilization of AI.