NATIONAL CENTER ON SEXUAL EXPLOITATION

ATTN: AI E.O. RFI Comments,
National Institute of Standards and Technology

FROM: Marcel van der Watt, Ph.D. & Victoria Rousay, M.A.
National Center on Sexual Exploitation

February 2, 2024

Submission:
Request for Information by the National Institute of Standards and Technology (NIST): Assignments under the Executive Order Concerning Artificial Intelligence

The National Center on Sexual Exploitation (NCOSE) welcomes the National Institute of Standards and Technology's (NIST) request for information to assist in carrying out several of its responsibilities under the Executive Order on ***Safe***, ***Secure***, and ***Trustworthy Development*** and ***Use of Artificial Intelligence*** issued on October 30, 2023.

NCOSE also appreciates the Executive Order's instructions to NIST to undertake an initiative to evaluate and audit capabilities relating to Artificial Intelligence (AI) technologies and to ***develop*** various ***guidelines***, including conducting ***AI red-teaming tests*** to ***enable deployment of safe, secure, and trustworthy systems***.

The National Center on Sexual Exploitation (NCOSE)

NCOSE's mission is to expose the interconnections between all forms of sexual abuse and exploitation and to dismantle the systems that perpetuate them. We believe every human being deserves the opportunity to live life to its fullest potential; to pursue dreams and ambitions; express creativity and hone talents; seek beauty, truth, and faith; experience hope, joy, and love with family and friends—to thrive. Such a vision requires not only individuals and institutions that work towards its realization but also a culture that embraces its responsibility to preserve and protect human flourishing. We aspire to create that culture and rely on more than six decades of subject-matter expertise. We do this through our law center, corporate advocacy, public policy advancement, and research. NCOSE spearheads numerous campaigns aimed at eradicating sex trafficking, child exploitation, and the multifaceted ways in which pornography undermines the inherent dignity of all individuals.

The intersection between NCOSE's dedicated mission to eradicate sexual abuse and exploitation and the need for robust AI content moderation systems intuitively lends itself to collaborative efforts for ***Safe***, ***Secure***, and ***Trustworthy Development*** and ***Use of Artificial Intelligence***. It also embodies a commitment to social responsibility and ethical stewardship in developing and applying AI technologies. Ensuring that strategies are grounded in a deep understanding of the dynamics of sexual abuse content requires the influence of those at the forefront of this battle.

Recommendations:

- In all actions geared towards Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, we call for adherence to the five ethical principles[1] identified from the global corpus of principles and guidelines on ethical AI: transparency, justice and fairness, non-maleficence, responsibility, and privacy. We advocate for adopting AI ethics principles that foster AI's responsible development and utilization. This will ensure that generative AI systems function in accordance with accepted ethical standards.

- Advocate for specific federal legislation on deepfakes that removes the burden on victims to prove lack of consent or intent of harm in cases involving deepfakes and AI-generated pornography. Due to the frequent anonymity of perpetrators, it is unjust to place this burden on victim-survivors.

- Mandatory Partnerships for GAN and GenAI Systems. We call for GAN and GenAI systems to be required to partner in efforts like Adobe's Content Authenticity Initiative or similar efforts promoting provenance and transparency.

- Implementation of a Traceability Metadata System. We recommend developing a comprehensive metadata system to ensure responsible use and enhance the accountability of AI-generated content, including images and deepfakes. This system would involve attaching a unique identifier, such as a user ID linked to a verified email address, to every piece of content created by AI technologies. This identifier not only facilitates the traceability of the content back to its creator but also integrates essential details about the creation process itself. This approach fosters a culture of accountability among users and assists platforms in maintaining ethical standards by ensuring that all content can be traced to its source. Implementing such a metadata system is a crucial step towards responsible AI usage, enabling the traceability of content while supporting the creative and constructive potential of AI technologies.

  Key Components of the Metadata System:

  - Unique User ID: Assign a unique identifier to each user, tied to their verified email address, ensuring that all content generated can be traced back to an individual creator.
  - AI System Information: Record the specific AI model or system (LLM, Gen-AI, GAN) used for generating the content, providing insights into the technology behind the creation.
  - Prompts Utilized: Document the exact prompts or instructions provided by the user for creating the content, including details that specify whether the intent was to generate positive (e.g., educational, artistic) or negative (e.g., misleading, unethical) outcomes.
  - Content Classification: Classify the nature of the generated content (positive or negative) based on its intended use and the context of its creation.

[1] Jobin, Anna, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines." *Nature machine intelligence* 1, no. 9 (2019): 389-399.

Implementation Example:

  - Consider a digital platform enabling users to generate various forms of AI-driven content (LLM, Gen-AI, GAN). When a user creates an item, the platform automatically appends a detailed metadata record to it. This record includes the user's unique ID, the AI model used, the provided prompts, and a classification of the content's nature. Such a system promotes transparency and ethical use and aids in content management and copyright considerations, providing a clear lineage of creation that can be invaluable for both creators and consumers in navigating the complexities of AI-generated materials, especially in the consideration of non-consensual and exploitative content.

- Immediate Removal of Harmful Imagery: Establish policies and measures of enforcement for the immediate removal of deepfake and AI-generated pornography – without question, delay, and irrespective of its commercial status or user uploading content.

- Regulations and Reporting Mechanisms on Development Platforms: Implement regulations on platforms (I.e., GitHub, HuggingFace, GitLab) to facilitate the reporting of problematic and exploitative AI models, ensuring a safer environment for content creation and sharing.

The following policy enforcement approaches are recommended to ensure the responsible use and integration of AI systems (API access), with a focus on safeguarding against the generation and dissemination of harmful or illicit content:

1. Coreference Resolution Enforcement:

   → Implementation Requirement: Mandate the deployment and ongoing improvement of coreference resolution algorithms within content moderation systems. This effort aims to boost the precision in identifying sexually explicit or illicit content through regular updates that accommodate evolving language use and content nuances.

   → Justification: Enhancing coreference resolution is crucial for maintaining responsible usage of AI systems and secure access to APIs. It not only aids in preventing the production of explicit or illicit materials but also ensures the correct differentiation between instances of sexual violence, other sensitive content, or innocuous content, enabling appropriate management of such cases.

   → NCOSE's contribution: NCOSE can contribute its expertise in understanding the nuances of sexual exploitation and abuse, sexually explicit and illicit content, predatory behavior, and requirements for child online safety.

2. Red Teaming and Adversarial Testing:

   → Establishment of Continuous Testing: Initiate a consistent and systematic red teaming protocol that tests the ability of policies to withstand adversarial efforts aimed at bypassing sexually explicit content restrictions for user-system interactions and API access. This includes creating benchmark assessments, routine evaluation of the model's reaction to high-risk prompts, and suspension of deployment pending the assurance of predominantly safe responses.

   → Justification: Implementing red teaming after coreference resolution enables the simulation of real-world adversarial attacks, reflecting a proactive stance on safeguarding user safety and model integrity. This approach anticipates and counters emerging threats and solidifies the system's defenses, ensuring ongoing secure user-system interaction and API integration.

   → NCOSE's contribution: NCOSE can contribute its expertise in identifying potential vulnerabilities and risks associated with generating and disseminating harmful or illicit content. NCOSE can assist in designing rigorous testing scenarios and evaluating the model's response to high-risk prompts, helping to ensure the policies' effectiveness and robustness.

3. User-Reported Feedback Mechanism:

   → Feature Integration: Introduce a mechanism for user-reported feedback within the AI system and API access framework, allowing users to highlight any content they deem questionable or inappropriate. This facilitates the capture of real-time insights into the model's interaction with the public, uncovering nuanced or unforeseen risks that may pose challenges to the existing sexually explicit content policy.

   → Justification: Strengthening safety mechanisms and fostering a comprehensive understanding of AI interactions necessitate active user involvement. This strategy empowers the model to evolve and adjust based on direct, human-centric feedback, promoting a robust and effective approach to enforcing content standards and ensuring the safety of user-system interactions and API integrations.

   → NCOSE's contribution: NCOSE can help guide the types of content that users should flag and report. NCOSE can assist in defining clear guidelines and criteria for identifying questionable or inappropriate content, ensuring that the feedback mechanism effectively captures relevant insights and supports the ongoing improvement of the AI system and API access framework.

These suggestions are aimed at bolstering the effectiveness of content moderation systems, enhancing the safety and reliability of AI technologies, and aligning with NIST's mission to foster innovation and industrial competitiveness in AI through advancing measurement science, standards, and technology in ways that enhance economic security and improve quality of life.

NCOSE eagerly anticipates any opportunity to support NIST's endeavors to promote Artificial Intelligence's safe, secure, and trustworthy development and utilization. This may include:

1. *Active Participation in Policy Formation and Implementation*: Engage with NIST policy experts and provide informed perspectives that shape effective and enforceable standards. This assures that AI system guidelines resonate with the deep-rooted commitment to eradicating sexual exploitation. NCOSE can provide assistance in ethical decision-making within the policy-making process, ensuring that recommendations align with unwavering dedication to social responsibility and the protection of human dignity.

2. *Advocacy and Awareness*: Helping with targeted campaigns emphasizing the importance of integrating ethical and responsible considerations into AI to combat sexual exploitation and bridging the gap between technological advancement and social welfare. Awareness Initiatives may include educational programs that shed light on the complex challenges and solutions in content moderation related to sexual exploitation while rallying broader support for responsible AI deployment in this critical area.

In conclusion, the critical issue of sexual abuse and exploitation persists as a stain on our society that demands attention. We implore NIST to consider these recommendations to pave the way for a future untainted by sexual exploitation while simultaneously enhancing economic security and improving quality of life.

Marcel van der Watt, Ph.D.
Director, Research Institute

Victoria Rousay, MA
Corporate Advocacy
Program Manager & Analyst