UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

Developing Robust Solutions, Policies, and Safeguarding Responses to AI-Generated CSAM

Response to OHCHR Call for Input on Existing and Emerging Sexually Exploitative Practices against Children in the Digital Environment (submitted May 14, 2024)

*The **National Participatory Action Research for Children's Safety: Artificial Intelligence-generated Child Sexual Exploitation and Abuse** project is a new research initiative funded by the Social Sciences and Humanities Research Council of Canada. It is led by Principal Investigator **Dr. Jia Xue**, in collaboration with **Dr. Rhonda McEwen** and **Dr. Sara M. Grimes** at the University of Toronto, in Toronto, Canada. Our research aims to identify technological, regulatory, and social service solutions to the growing problem of AI-generated child sexual abuse materials in Canada and beyond. We bring to this response our globally recognized and cross-disciplinary relevant expertise and our extensive combined experience conducting cross-sector research on children's safety, rights, and wellbeing in the digital environment.*

In April 2023, a Canadian man was convicted for using Artificial Intelligence (AI) to generate child sexual abuse materials (CSAM) for the first time in our country's history. It will surely not be the last. CSAM are images or videos that show a child engaged in or depicted as being engaged in explicit sexual activity.[1] While the term child pornography is still used by the public, scholars and practitioners have shifted to the term CSAM because while pornography depicts adults who have consented to be filmed, children cannot legally consent to sex or consent to have images of their abuse recorded and distributed. An explicit photo or video of a child is evidence they have been a victim of sexual abuse. Studies show that prepubescent children are at the greatest risk of being depicted in CSAM and 84.2% of these videos and images contain severe abuse.[2]

Children shown in CSAM experience negative outcomes and are usually victimized twice: first by the person committing the sexual abuse, and again by those who view it. Victims of CSAM report feelings of shame, and blame, and struggle with low self-esteem that affect many different areas of their lives such as relationships, careers, and overall health, post-traumatic stress disorder, substance and alcohol abuse, obesity, eating disorders, depression, and problematic sexual behaviors.

There is an urgent need to establish consensus on definitions, measures, and approaches to AI-generated CSAM and develop robust safeguarding responses. We applaud the Special Rapporteur's call for input on existing and emerging sexually exploitative practices against children in the digital environment for her forthcoming report to the UN General Assembly. We appreciate the opportunity to share insights from existing and emerging research on the role of AI in facilitating CSAM in our responses to the Special Rapporteur's questions 1, 3,4, 7, and 8.

---

[1] Child Sexual Abuse Materials: The Facts, National Children's Advocacy Center, 2023. Access from https://calio.org/wp-content/uploads/2023/02/Child-Sexual-Abuse-Materials-Fact-Sheet.pdf

[2] ECPAT International (2018). Trends in online child sexual abuse material. Bangkok: ECPAT International.

Question 1: How technologies are used to facilitate the sexual exploitation and abuse of children.

AI development has introduced a number of technologies that facilitate the sexual exploitation and abuse of children. AI tools, such as deepfakes, avatars, immersive sex games, de-aging technologies, and voice cloning amplify and extend existing methods already being used to exploit minors while potentially circumventing existing protections. For example, deepfakes, which are built on Deep Learning AI, can create hyper-realistic visual depictions of a person without their consent or participation.[3] Studies show that between 96% and 98% of deepfakes videos available online feature sexual content.[4] Industry research suggests there has been a tenfold increase in the number of deepfakes detected globally between 2022 and 2023.[5] In 2020, over 100,000 women and underage girls had explicit images generated by a publicly available deepfake bot on the Telegram messaging app.[6] The availability and ease of deepfake technology pose significant risks associated with their growing use as tools for threats and harassment, blackmail, and sexual abuse.[7]

The nature of digital content already allows for easy replication and widespread dissemination, making containment challenging. Furthermore, the anonymizing capabilities of technologies like encryption complicate offender identification. The rapid development of AI capable of autonomously generating non-consensual sexual abuse materials highlights the pressing need to address the escalating issue of AI-generated CSAM, calling for more proactive prevention and control strategies.[8]

Additionally, widespread access to both internet and AI tools expose children to increased risks through the consumption, creation, or sharing of abusive materials, often leading to grooming and sexual exploitation. This escalating challenge has gained attention on the national and international fronts, with countries such as the US, UK, Norway, and the European Union making it a top legislative priority. [9] Understanding the implications of AI in this context is crucial before it becomes a widely used tool for malicious purposes.

---

[3] Brandon, J. (2018). Terrifying high-tech porn: creepy 'deepfake' videos are on the rise. Fox news, 20.

[4] Sommers, M. (2020). Deepfakes, explained. Ideas Made to Matter. MIT Sloan; Hurst, L. (2023). Generative AI fueling spread of deepfake pornography across the internet. EuroNews.com. https://www.euronews.com/next/2023/10/20/generative-ai-fueling-spread-of-deepfake-pornography-across-the-internet

[5] Sumsub (2023, November 28). Sumsub Research: Global Deepfake Incidents Surge Tenfold from 2022 to 2023. PR Newswire. https://www.prnewswire.com/news-releases/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023-301998891.html

[6] Hao, K. (2021). Deepfake porn is ruining women's lives. Now the law may finally ban it. MIT Technology Review. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/

[7] Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. The International Journal of Evidence & Proof, 23(3), 255-262.

[8] DeKeseredy, W. S. (2021). Image-based sexual abuse: Social and legal implications. Current Addiction Reports, 8(2), 330–335.

[9] Dorotic, M., & Johnsen, J. W. (2023). Child Sexual Abuse on the Internet: Report on the analysis of technological factors that affect the creation and sharing of child sexual abuse material on the

Question 3: What are the remaining gaps that limit the effective implementation and application of existing laws, policies and guidelines to prevent, detect, report and protect children from sexual exploitation and sexual abuse online?

The global commitment to combat and protect against various forms of sexual abuse is exemplified in various legislative measures and international agreements, such as the United Nations General Assembly (UNGA) Resolution on the Rights of the Child in the digital environment (adopted by consensus in November 2023), the United Nations Declaration on the Elimination of Violence Against Women, and the Council of Europe Convention on the Protection of Children against Sexual Exploitation and Sexual Abuse. The UK has developed a tailored legal framework, the Revenge Pornography Guideline to combat image-based offenses. These measures are crucial and timely. Canada and other countries that have not yet taken meaningful action on this front, can advance this global movement by introducing new laws and policies and strengthening existing ones aimed at preventing, reporting, and protecting children from CSAM and other forms of sexual exploitation and sexual abuse online.

However, AI technologies, such as Deepfakes, present unique challenges that necessitate revisions to existing legal frameworks. Several legislative measures have been enacted to tackle these concerns. For example, the US National Defense Authorization Act (NDAA) directs the Department of Homeland Security to investigate deepfake creation technology and potential strategies for detection and mitigation. In 2022, the US Committee on Homeland Security and Governmental Affairs released a report titled "Deepfake Task Force Act: Report of the Committee on Homeland Security and Governmental Affairs, United States Senate, to Accompany S. 2559," which led to the establishment of the Deepfake Task Force Act and the National Deepfake and Digital Provenance Task Force as the administering body. Additional bills have been introduced to address AI-related concerns, including the Malicious Deep Fake Prohibition Act of 2018, the Defending Each, and the DEEPFAKES Accountability Act. Another recent example is the ratification of amendments to the Digital Services Act by the European Parliament on the criminalization of deepfake technologies.

However, scholars have criticized existing legal frameworks for being overbroad. For example, the Malicious Deep Fake Prohibition Act in the United States has significant shortcomings. Of particular concern is how it allows nearly all deepfakes to evade prosecution as long as the intent behind their creation is *not explicitly harmful*.[10] According to Section 3805 of the Act, a deepfake is defined as "any audiovisual record created or altered in a manner that the record would falsely appear to a reasonable observer to be an authentic record of the actual speech or conduct of an individual" (MDFPA, 2018). This contrasts with the UK Revenge Pornography Guideline, which requires an intent to harm for charges related to image-based abuse, potentially allowing for various harms to occur without explicit consent as the primary defense when no evidence of consent withdrawal is presented.

Internet. Forskningsrapport, BI.

[10] Delfino, R. A. (2019). Pornographic deepfakes: The case for federal criminalization of revenge porn's next tragic act. Fordham L. Rev., 88, 887.

In light of these complexities and the evolving landscape of AI-facilitated CSAM, it is clear that the definitions, conceptualizations, measures, and existing policies and practices surrounding this issue demand a comprehensive examination. There is a critical need for research that gathers consensus among experts in the field of child sexual abuse and protection to inform more rigorous and child-centric policymaking.

Question 4: What are the challenges that exist in the use of these digital technologies, products or services, that inhibit the work of law enforcement across jurisdictions in their work to investigate, detect, remove child sexual abuse materials online and prosecute these crimes?

Defining online child sexual exploitation and abuse, especially AI-generated CSAM is complex, encompassing a range of behaviors, from solicitation to grooming, cybersex, and accessing, producing, or sharing abusive images.[11] In many legal contexts, including Canada, it encompasses various activities, including CSAM, self-generated materials (often distributed without consent), sexting, sextortion, grooming, luring, lived child sexual abuse streaming, and made-to-order content.[12] The ongoing lag among policymakers in adapting regulations to address the multifaceted nature and forms of AI-generated CSAM poses significant challenges to law enforcement across jurisdictions in their work to define, investigate, and detect CSAM online and prosecute these crimes.

While we are gradually becoming aware of the diverse ways AI can be used to generate and distribute CSAM, little is known about the prevalence and characteristics of child sexual exploitation and abuse facilitated by AI. Moreover, gaps persist in our knowledge of how guardians, educators, and law enforcement officers address, track, and remove CSAM, or of their existing literacies and technological skills. It is highly likely that law enforcements in most jurisdictions lack the technological skills required to investigate, detect, and effectively remove AI-generated CSAM.

Question 7: In the case of generative Artificial Intelligence and end-to-end encryption, what are the challenges and recommended mitigation measures, including the application of advanced technology needed by technology companies, online service providers and law enforcement to prevent by blocking the sharing and removal of CSAM?

Within the realm of AI, deepfakes emerge as a prominent concern, encompassing images or recordings that undergo convincing alteration and manipulation, often portraying individuals engaging in actions or utterances they never actually did. These manipulations thrive in the pornography industry, where women's faces are superimposed onto others' bodies to create video

[11] De Santisteban, P., and M. Gámez-Guadix. 2018. Prevalence and risk factors among minors for online sexual solicitations and interactions with adults. The Journal of Sex Research. 55 (7): 939-950.
[12] Public Safety Canada. 2022. Child Sexual Exploitation on the Internet. Countering Crime.

illusions, resulting in non-consensual sexual image abuse and other harm. Researchers have identified three primary categories of image-based sexual abuse: non-consensual recording, the capture of nude or sexual imagery without consent; non-consensual distribution, sharing, posting, or dissemination of such content, whether online or offline; and sextortion, involving threats to distribute nude or sexual images, often conveyed in person, through cell phones, email, apps, or internet platforms.[13]

In AI, we must also confront the creation and manipulation of data, utilized to generate digital clones for explicit sexual abuse purposes and consequences through the application of deepfake technology. The sheer volume of CSAM that can be generated and distributed using AI tools, a number that is growing exponentially every year, far exceeds the existing capacities, resources, and abilities of law enforcement organizations, NGOs, platforms, moderators and tech companies to respond to, investigate, and address. The use of AI tools to detect and remove CSAM is promising in this regard, but also contains multiple uncertainties and unknowns. This ever-evolving landscape underscores the urgency of understanding and addressing the multifaceted challenges posed by AI-facilitated sexual abuse.

Question 8: Are there any examples of proactive measures taken to facilitate consultation and participation with a broad range of stakeholders, including children and child-rights organisations, for informing policy and legislation, setting technical standards and implementing processes to eradicate child sexual abuse and exploitation online?

The **National Participatory Action Research for Children's Safety: Artificial Intelligence-generated Child Sexual Exploitation and Abuse** project aims to bridge existing gaps in regulation and technology design by producing empirical evidence on research questions that align with three primary objectives:

1) To build consensus among experts in the field of child sexual abuse regarding problem identification (e.g., definition, impacted communities such as sexual minority children and children with autism); policy strategies with particular consideration of the issues relating to data privacy and ethics; and challenges related to AI-generated child sexual abuse materials (e.g., transparency and biases in AI algorithms).
2) To gather public opinions, experience, and sentiments concerning AI-generated CSAM to guide care for children potentially impacted by these materials.
3) To develop robust safeguarding responses, provide policy recommendations, and propose intervention strategies for practitioners to enhance child protection practices against AI-generated abuse. While our project is focused on Canada, a similar multi-sectoral approach would facilitate consultation and participation with a broad range of stakeholders for informing policy and legislation in other countries or internationally, as well as for setting technical standards and implementing effective processes to eradicate CSAM, child sexual abuse and exploitation in the global digital environment.

[13] Henry, N., & Flynn, A. (2019). Image-based sexual abuse: Online distribution channels and illicit communities of support. Violence against women, 25(16), 1932-1955.

Our collaboration is centred on consultation and the participation of a range of stakeholders and it aims to inform policy and legislation, design norms, and trust and safety procedures by identifying robust technological, regulatory, and social service solutions to AI-generated child sexual abuse materials. Our research asks: What is the extent of AI-generated CSAM and to what degree do experts concur on policy strategies, encompassing elements like problem identification (e.g., definitions, affected communities such as sexual minority children and children with autism), considerations of data privacy and ethics, and challenges related to AI generated CSAM (e.g., transparency and biases in AI algorithms)? Mobilizing Dr. Xue's expertise in responsible uses of big data and computational approaches, this project explores the creation of machine learning models capable of automatically identifying components related to the agenda-setting of AI-generated CSAM. Lastly, our project will identify technical measures and safety protocols for fostering collaboration among academia, practitioners, and industry stakeholders in addressing AI-generated CSAM, ensuring the protection of children, and addressing issues of child safety and ethical considerations.

We are eager to share the results and findings of this research as it becomes available. As we are committed to actively contributing to a global response to the UN's call for States and all stakeholders to work together to eradicate child sexual abuse and exploitation online, and to protect children and children's rights in the digital environment and in the field of AI.

Submitted by,

Dr. Jia Xue
Director, Artificial Intelligence for Justice Lab (AIJ)
Assistant Professor, Factor-Inwentash Faculty of Social Work
and the Faculty of Information
University of Toronto

Dr. Rhonda McEwen
President, Victoria University in the University of Toronto
Canada Research Chair in Tactile Interfaces, Communication and Cognition
Professor, Institute of Communication, Culture, Information and Technology, Faculty of Information, Department of Computer Science, and Institute of the History and Philosophy of Science and Technology
University of Toronto

Dr. Sara M. Grimes
Bell University Labs (BUL) Chair in Human-Computer Interaction
Director, Kids Play Tech Lab
Professor, Faculty of Information
University of Toronto