Este documento foi traduzido por inteligência artificial. Por favor, considere possíveis erros de tradução.

Envio da Human Rights Watch
ao Comitê das Nações Unidas sobre os Direitos da Criança
Resenha do Brasil
99º de Trabalho Pré-Sessional
Agosto de 2024

Escrevemos em antecipação à 99ª pré-sessão do Comitê dos Direitos da Criança e sua análise do Brasil. Este envio é uma atualização de nossos focos em nossa pesquisa recente sobre a raspagem e o uso indevido de fotos pessoais de crianças brasileiras para criar ferramentas de IA sem seu conhecimento ou consentimento.

Fotos pessoais de crianças brasileiras usadas indevidamente para alimentar ferramentas de IA (artigos 12, 16 e 34

Em junho de 2024, a Human Rights Watch informou que havia descoberto a coleta e o uso de fotos pessoais de crianças brasileiras para criar poderosas ferramentas de IA sem o conhecimento ou o consentimento das crianças ou de suas famílias.[1] Essas fotos são retiradas da Web e transformadas em um grande conjunto de dados que as empresas usam para treinar suas ferramentas de IA. Por sua vez, outros usam essas ferramentas para criar deepfakes maliciosos que colocam ainda mais crianças em risco de exploração e danos.

Uma análise feita pela Human Rights Watch constatou que o LAION-5B, um conjunto de dados usado para treinar ferramentas populares de IA e criado a partir da raspagem da maior parte da Internet, contém links para fotos identificáveis de crianças brasileiras. Os nomes de algumas crianças estão listados na legenda que as acompanha ou no URL onde a imagem está armazenada. Em muitos casos, suas identidades são facilmente rastreáveis, incluindo informações sobre quando e onde a criança estava no momento em que a foto foi tirada.

Uma dessas fotos mostra uma menina de 2 anos de idade, com os lábios entreabertos de admiração enquanto toca os dedos minúsculos de sua irmã recém-nascida. A legenda e as informações incorporadas na foto revelam não apenas os nomes das duas crianças, mas também o nome e a

[1] "Brasil: Children's Personal Photos Misused to Power AI Tools," comunicado à imprensa da Human Rights Watch, 10 de junho de 2024, https://www.hrw.org/news/2024/06/10/brazil-childrens-personal-photos-misused-power-ai-tools.

localização exata do hospital em Santa Catarina onde o bebê nasceu há nove anos em uma tarde de inverno.
A Human Rights Watch encontrou 170 fotos de crianças de pelo menos 10 estados: Alagoas, Bahia, Ceará, Mato Grosso do Sul, Minas Gerais, Paraná, Rio de Janeiro, Rio Grande do Sul, Santa Catarina e São Paulo. É provável que essa seja uma subcontagem significativa da quantidade total de dados pessoais de crianças existentes no LAION-5B, já que a Human Rights Watch analisou menos de 0,0001% dos 5,85 bilhões de imagens e legendas contidas no conjunto de dados.

As fotos analisadas pela Human Rights Watch abrangem toda a infância. Elas capturam momentos íntimos de bebês nascendo nas mãos enluvadas de médicos, crianças pequenas soprando velas em seu bolo de aniversário ou dançando de cueca em casa, estudantes fazendo uma apresentação na escola e adolescentes posando para fotos no carnaval da escola.

Muitas dessas fotos foram originalmente vistas por poucas pessoas e, anteriormente, tinham um certo grau de privacidade. Não parece ser possível encontrá-las por meio de uma pesquisa on-line. Algumas dessas fotos foram publicadas por crianças, seus pais ou familiares em blogs pessoais e sites de compartilhamento de fotos e vídeos. Algumas foram carregadas anos ou até mesmo uma década antes da criação do LAION-5B.

Depois que seus dados são coletados e inseridos em sistemas de IA, essas crianças enfrentam mais ameaças à sua privacidade devido a falhas na tecnologia. Os modelos de IA, inclusive os treinados no LAION-5B, são famosos por vazar informações privadas; eles podem reproduzir cópias idênticas do material em que foram treinados, inclusive registros médicos e fotos de pessoas reais.[2] As barreiras de proteção estabelecidas por algumas empresas para evitar o vazamento de dados confidenciais foram quebradas várias vezes.[3]

Esses riscos à privacidade abrem caminho para outros danos. O treinamento em fotos de crianças reais permite que os modelos de IA criem clones convincentes de qualquer criança, com base em um punhado de fotos ou até mesmo em uma única imagem do site .[4] Agentes mal-intencionados usaram ferramentas de IA treinadas pelo LAION para gerar imagens explícitas de crianças usando fotos inócuas,

[2] Carlini et al., "Extracting Training Data from Diffusion Models", 30 de janeiro de 2023, https://doi.org/10.48550/arXiv.2301.13188 (acessado em 9 de julho de 2024); Benj Edwards, "Artist finds private medical record photos in popular AI training data set", *Ars Technica*, 21 de setembro de 2022, https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/ (acessado em 9 de julho de 2024).
[3] Carlini et al., "Extracting Training Data from Diffusion Models" (Extração de dados de treinamento de modelos de difusão); Nasr et al, "Extracting Training Data from ChatGPT," 28 de novembro de 2023, https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html (acessado em 9 de julho de 2024); Mehul Srivastava e Cristina Criddle, "Nvidia's AI software tricked into leaking data," *Financial Times,* 9 de junho de 2023, https://www.ft.com/content/5aceb7a6-9d5a-4f1f-af3d-1ef0129b0934 (acessado em 9 de julho de 2024); Matt Burgess, "OpenAI's Custom Chatbots Are Leaking Their Secrets," *WIRED*, 29 de novembro de 2023, https://www.wired.com/story/openai-custom-chatbots-gpts-prompt-injection-attacks/ (acessado em 9 de julho de 2024).
[4] Benj Edwards, "AI image generation tech can now create life-wrecking deepfakes with ease", *Ars Technica*, 9 de dezembro de 2022, https://arstechnica.com/information-technology/2022/12/thanks-to-ai-its-probably-time-to-take-your-photos-off-the-internet/ (acessado em 9 de julho de 2024); Ibid., "Microsoft's VASA-1 can deepfake a person with one photo and one audio track", *Ars Technica*, 19 de abril de 2024, https://arstechnica.com/information-technology/2024/04/microsofts-vasa-1-can-deepfake-a-person-with-one-photo-and-one-audio-track/ (acessado em 9 de julho de 2024).

bem como imagens explícitas de crianças sobreviventes cujas imagens de abuso sexual foram coletadas no LAION-5B.[5]

Da mesma forma, a presença de crianças brasileiras no LAION-5B contribui para a capacidade dos modelos de IA treinados nesse conjunto de dados de produzir imagens realistas de crianças brasileiras. Isso amplia substancialmente o risco existente que as crianças enfrentam de que alguém roube sua imagem de fotos ou vídeos publicados on-line e use a IA para manipulá-las a dizer ou fazer coisas que elas nunca disseram ou fizeram.

Pelo menos 85 meninas de Alagoas, Minas Gerais, Pernambuco, Rio de Janeiro, Rio Grande do Sul e São Paulo relataram assédio por parte de seus colegas de classe, que usaram ferramentas de IA para criar deepfakes sexualmente explícitos das meninas com base em fotos tiradas de seus perfis de mídia social e, em seguida, circularam as imagens falsas on-line.

A mídia fabricada sempre existiu, mas exigia tempo, recursos e conhecimento para ser criada e era, em grande parte, irrealista. As ferramentas de IA atuais criam resultados realistas em segundos, geralmente são gratuitas e fáceis de usar, arriscando a proliferação de deepfakes não consensuais que poderiam recircular on-line para sempre e causar danos duradouros.

A LAION, organização alemã sem fins lucrativos que gerencia o LAION-5B, confirmou em 1º de junho que o conjunto de dados continha as fotos pessoais das crianças encontradas pela Human Rights Watch e se comprometeu a removê-las, dizendo que enviaria à Human Rights Watch a confirmação da remoção assim que ela fosse concluída. Até 16 de agosto, a empresa não havia confirmado a remoção dos dados das crianças de seu de dados. A LAION também contestou que os modelos de IA treinados no LAION-5B pudessem reproduzir dados pessoais literalmente. Ela disse: "Pedimos à HRW que entre em contato com os indivíduos ou seus responsáveis para incentivar a remoção do conteúdo de domínios públicos, o que ajudará a impedir sua recirculação."

Legisladores no Brasil propuseram a proibição do uso não consensual de IA para gerar imagens sexualmente explícitas de pessoas, inclusive crianças.[6] Esses esforços são urgentes e importantes, mas abordam apenas um sintoma do problema mais profundo de que os dados pessoais de crianças permanecem em grande parte desprotegidos contra o uso indevido. Da forma como está redigida, a lei de proteção de dados do Brasil - a Lei Geral de Proteção de Dados Pessoais - não oferece proteção suficiente para as crianças.

O governo deve reforçar a lei de proteção de dados adotando salvaguardas adicionais e abrangentes para a privacidade dos dados das crianças.

---

[5] Emanuel Maiberg, "a16z Funded AI Platform Generated Images That "Could Be Categorized as Child Pornography," Leaked Documents Show", *404 Media*, 5 de dezembro de 2023, https://www.404media.co/a16z-funded-ai-platform-generated-images-that-could-be-categorized-as-child-pornography-leaked-documents-show/ (acessado em 9 de julho de 2024); David Thiel, "Identifying and Eliminating CSAM in Generative ML Training Data and Models", Stanford Internet Observatory, 23 de dezembro de 2023, https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-23.pdf (acessado em 9 de julho de 2024).

[6] Projeto de Lei no PL 5342/2023, disponível em https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2401172 (acessado em 19 de agosto de 2024), e Projeto de Lei no PL 5394/2023, disponível em https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2402162 (acessado em 19 de agosto de 2024).

Em 2 de julho, a Autoridade Nacional de Proteção de Dados emitiu uma proibição preliminar do uso pela Meta dos dados pessoais de seus usuários baseados no Brasil para treinar seus sistemas de IA.[7] A decisão sem precedentes do governo segue a pesquisa da Human Rights Watch, conforme descrito acima, e incluiu dois argumentos que refletem as recomendações da Human Rights Watch: o primeiro é a importância de proteger a privacidade dos dados de crianças, dado o risco de danos e exploração que resulta do fato de seus dados serem raspados e usados por sistemas de IA. O segundo se concentra na limitação da finalidade e que as expectativas de privacidade das pessoas quando compartilham seus dados pessoais on-line devem ser respeitadas.

A Human Rights Watch recomenda que o Comitê:

- Parabenize o Brasil por sua proibição preliminar à Meta e pergunte que medidas foram tomadas para avaliar o cumprimento da decisão pela Meta.
- Pergunte ao Brasil se ele facilitará a reparação de crianças cuja privacidade foi violada por meio de raspagem não consensual e uso indevido de suas fotos pessoais.
- Pergunte ao Brasil se ele tomará medidas para evitar a futura extração não consensual e o uso indevido de dados pessoais de crianças.

A Human Rights Watch incentiva o Comitê a solicitar ao governo do Brasil que

- Adotar e aplicar leis para proteger os direitos das crianças on-line, incluindo a privacidade de seus dados
- Incorporar proteções de privacidade de dados para crianças em sua futura política nacional para proteger os direitos de crianças e adolescentes no ambiente digital, cujo processo de elaboração estava originalmente programado para ser concluído em julho de 2024 e parece ter sido adiado para outubro.[8]
- Incorporar proteções de privacidade de dados para crianças em suas regulamentações de IA propostas e no plano nacional de IA[9] , de modo que os direitos das crianças sejam respeitados, protegidos e promovidos durante todo o desenvolvimento e uso da IA. O governo deve tomar cuidado especial para proteger a privacidade das crianças com relação à IA, já que a natureza do desenvolvimento e do uso da tecnologia não permite que as crianças e seus responsáveis consintam de forma significativa sobre como a privacidade dos dados das crianças é tratada. Essas proteções devem:

[7] Hye Jung Han, "Brazil Prevents Meta from Using People to Power Its AI," comentário, Human Rights Watch dispatch, 3 de julho de 2024,
[8] Guilherme Seto et al, "Silvio Almeida e Moraes discutem proteção de crianças na internet," ("Silvio Almeida e Moraes discutem proteção de crianças na internet") , *Folha de São Paulo*, 20 de junho de 2024, https://www1.folha.uol.com.br/colunas/painel/2024/06/silvio-almeida-e-moraes-discutem-protecao-de-criancas-na-internet.shtml (acessado em 26 de julho de 2024); Conselho Nacional dos Direitos da Criança e do Adolescente, "Resolução No. 245, de 5 de abril de 2024, Dispõe sobre os direitos das crianças e adolescentes em ambiente digital" ("Resolução nº 245, de 5 de abril de 2024, Dispõe sobre os direitos das crianças e adolescentes em ambiente digital"), 16 de abril de 2024, https://www.gov.br/participamaisbrasil/blob/baixar/48630 (acessado em 7 de agosto de 2024).
[9] Ver Ministério da Ciência, Tecnologia e Inovação, "Plano brasileiro de IA terá supercomputador e investimento de R$ 23 bilhões em quatro anos," 30 de julho de 2024, atualizado em 12 de agosto de 2024, https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/07/plano-brasileiro-de-ia-tera-supercomputador-e-investimento-de-r-23-bilhoes-em-quatro-anos (acessado em 19 de agosto de 2024). O plano proposto não se refere à proteção dos direitos humanos e das crianças no desenvolvimento e uso planejados pelo governo da IA. A Human Rights Watch observa que o Brasil atuou como um dos principais patrocinadores da Resolução A/78/L.49 da Assembleia Geral das Nações Unidas, que "enfatiza que os direitos humanos e as liberdades fundamentais devem ser respeitados, protegidos e promovidos durante todo o ciclo de vida dos sistemas de inteligência artificial....". Consulte Assembleia Geral da ONU, Resolução A/78/L.49 (2024), disponível em
(acessado em 9 de julho de 2024), parágrafo 5.

- Proibir a coleta de dados pessoais de crianças em sistemas de IA, considerando os riscos de privacidade envolvidos e o potencial de novas formas de uso indevido à medida que a tecnologia evolui.
- Proibir a reprodução digital não consensual ou a manipulação de imagens de crianças.
- Fornecer àqueles que sofrem danos devido ao desenvolvimento e ao uso da IA mecanismos para buscar justiça e recursos significativos.