

INTELIGÊNCIA ARTIFICIAL GENERATIVA (IA GEN) E VÍDEOS-FAKE: CORPOS, PORNOGRAFIA E TECNOLOGIAS DE MÁ-INFORMAÇÃO

Generative Artificial Intelligence (genAI) and Deepfakes: Bodies, Pornography, and Misinformation Technologies

Inteligencia Artificial Generativa (IA gen) y videos falsos: cuerpos, pornografía y tecnologías de desinformación

Marcelo Barcelos¹

Fabia Ioscote²

Cristiane Fontinha³

Maria José Baldessar⁴

DOI: doi.org/10.31501/esf.v1i29.14888

Resumo: As autoras e o autor analisam a escalada de conteúdos *porn fake* - materiais pornográficos falsos criados a partir de aplicativos de Inteligência Artificial (IA) Generativa - que permitem recriar vídeos e gerar montagens com imagens pessoais. A GenAI potencializa um novo mercado pornográfico de conteúdo falso e acentua violências de gênero. Casos recentes, no Brasil e Reino Unido, inclusive em escolas, reforçam a urgência por regulação da IA Gen, autenticidade e fiscalização das plataformas.

Palavras-chave: Inteligência Artificial; *Deepfake*; Desinformação; Má-informação; *Nude-fake*.

Abstract: The authors analyze the escalation of fake porn content - fake pornographic material created using Generative Artificial Intelligence (AI) applications - which allow videos to be recreated and montages to be generated with personal images. GenAI potentiates a new pornographic market for fake content and accentuates gender violence. Recent cases in Brazil and the United Kingdom, including in schools, reinforce the urgency for regulation of Gen AI, content authenticity and platform monitoring.

Keywords: Artificial intelligence; Deep fake; Disinformation; Misinformation; Porn fake.

Resumen: Las autoras y el autor analizan el aumento de los contenidos pornográficos falsos, también conocidos como deepfakes, que son creados a partir de aplicaciones de inteligencia artificial (IA) generativa. Estas aplicaciones permiten recrear videos y generar montajes con imágenes personales. La IA generativa potencia un nuevo mercado pornográfico de contenido falso y acentúa las violencias de género. Casos recientes en Brasil y Reino Unido, incluso en escuelas, refuerzan la urgencia por la regulación de la IA generativa, la autenticidad del contenido y la fiscalización de las plataformas.

Palabras-clave: Inteligência Artificial; Deepfake; Desinformación; Malinformación; *Nude-fake*.

¹ Doutor em Jornalismo e pesquisador de pós-doutorado em Engenharia e Gestão do Conhecimento (PPGEGC), na Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil. marcelobarcelos.jornalismo@gmail.com | <http://orcid.org/0000-0003-2880-5487>

² Doutoranda, Universidade Federal do Paraná, Curitiba, PR, Brasil. ioscote.fabia@gmail.com | <http://orcid.org/0000-0002-6318-026X>. Bolsista Capes.

³ Doutora em Engenharia e Gestão do Conhecimento (UFSC) | Mídias do Conhecimento | e pesquisadora de pós-doutorado em Gestão do Conhecimento e Pesquisadora (PPGEGC), na Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil. crisfontinha@gmail.com | <https://orcid.org/0009-0008-8316-6474>

⁴ Doutora em Ciências da Comunicação (USP), Docente Permanente na Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil. mbaldessar@gmail.com | <https://orcid.org/0000-0001-8971-4576>

1. Introdução

Numa das últimas sessões do ano de 2023, a Câmara dos Deputados do Brasil aprovou Projeto de Lei (PL 9930/2018) que torna crime a produção de conteúdo pornográfico com uso de ferramentas de Inteligência Artificial (IA). O Projeto de Lei (PL), apresentado pela deputada Erika Kokay (PT-DF) e relatado pela deputada Luisa Canziani (PSD-PR), estabelece pena de um a quatro anos de prisão, além de multa, para quem criar ou divulgar "montagens ou modificação que tenham como objetivo incluir pessoa em cena de nudez ou ato sexual, inclusive com uso de Inteligência Artificial em vídeo, áudio ou fotografia". O texto do Projeto de Lei será submetido à votação no Senado Federal em 2024 e, se aprovado, aguardará a sanção presidencial.

Representa, assim, um dos primeiros avanços no sentido de coibir o que vem sendo chamado de crime “pornô *fake*” - criação e circulação de conteúdos que envolvem manipulação de imagens de terceiros e terceiras (a maioria mulheres), a divulgação e a comercialização de imagem de nudez ou de cunho sexual não autorizada, gerada por software e Inteligência Artificial Generativa (*genAI*). Compreende-se o termo pornô *fake* ou *pornfake* como mídias digitais manipuladas para parecerem autênticas, mas que são fabricadas ou alteradas com conteúdo de caráter íntimo ou sexual. Isso pode incluir vídeos e imagens falsos (*deepfakes*), muitas vezes gerados por IA, que podem ser usados para difamar, enganar ou violar a privacidade. É relevante mencionar que usar imagens falsificadas ou não autorizadas de outras pessoas também pode ser considerado *fake*. No entanto, a *genAI*, uma tecnologia que gera novos dados, como imagens e textos, demonstra uma sofisticação na manipulação digital e está amplamente acessível. Com recursos de deep learning e redes neurais artificiais, a *genAI*

pode ser utilizada para amplificar a disseminação de *deepfakes*, bem como criar conteúdo enganoso e segmentar audiências.

A ameaça das *deepfakes* se tornou um problema global com o rápido avanço dos aplicativos de produção de linguagem natural, que possibilitam a fusão e distorção de imagens. Os resultados são imagens recriadas, áudios adulterados e principalmente a facilidade operacional e tecnológica de editar imagens e incorporar novas cenas e rostos. No Brasil, outro PL nesse caminho também classifica como pública incondicionada a natureza da ação penal dos crimes contra a dignidade sexual, uma abordagem inédita no país. Pelo que é proposto neste segundo texto, ainda em discussão inicial, a pena pode ser acrescida em dois terços ou até dobrar em razão de as imagens incluírem, por exemplo, vítimas com idades inferiores a 18 anos completos. Ainda segundo um dos projetos de lei que tramita no Legislativo brasileiro, pretende-se aumentá-la de um terço a dois terços se o conteúdo for utilizado para obter algum tipo de vantagem por meio de extorsão, ameaça ou chantagem.

Os dados associados à escalada das *deepfakes* - técnica de manipulação de mídia que usa *genAI* para criar vídeos falsos (Paris & Donovan, 2019) - e, consequentemente, dos *fake nudes*, têm impressionado pelo volume, rápida disseminação/circulação, financiamento, permissão junto às plataformas e baixa punibilidade sob o aspecto jurídico. De acordo com a Revista Wired, publicação norte-americana especializada em tecnologia, foram subidos 113 mil vídeos em 35 sites dedicados ao chamado *deepfake*. Esses dados corroboram os estudos de Das e Tripathi (2023), Mania (2023) e Maddocks (2020), este último consideravelmente referencial, a partir do texto *Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes*.

Conforme relatado pela Wired, os conteúdos de "pornô *fake*" aumentaram significativamente, registrando um salto de 54% em relação ao volume total do ano de 2022, somente nos últimos nove meses. Hoje, pela legislação brasileira (Lei nº 13.718/2018), os responsáveis por criar, editar e compartilhar este tipo de conteúdo podem responder por fato análogo à simulação e participação de criança ou adolescente em cena de sexo explícito ou pornográfica mediante adulteração, montagem ou modificação de fotografia, vídeo ou qualquer outra forma de representação visual (Brasil, 2018).

Chama a atenção o aumento da oferta de aplicativos que permitem manipular imagens com extrema fidedignidade e, neste compasso, sem exigir habilidades e competências de programação ou edição aprimoradas de imagem (Gerardi, Walters & James, 2020). Esse cenário eleva o perigo de disseminação de conteúdos classificados como baseados em adulteração de imagem com viés destinado à pornografia e exploração sexual. O agravante, é que muitas vezes, as vítimas se veem silenciadas pelas ameaças ou constrangimento de denunciar, segundo apontam as pesquisas de Sousa e Silva (2022) e dados do Projeto Vazou (2018).

Embora vídeos desse tipo, muitas vezes denominados de *deepfakes*, existam online desde 2017, os avanços na Inteligência Artificial, especialmente no *deep learning* (aprendizado profundo destes sistemas), Redes Generativas Adversárias (GANs), Autoencoders Variacionais (VAEs) e Redes Neurais Recorrentes, (RNNs), combinados com a crescente disponibilidade da tecnologia específica, tornaram a produção de material sexualmente explícito não consensual mais fácil, lucrativo e acessível para diferentes idades, incluindo adolescentes, como registramos nos dois casos brasileiros ilustrados neste artigo. Uma análise realizada pela NBC News de dois dos maiores sites que hospedam *fake videos* sexualmente explícitos descobriu que esses vídeos eram facilmente acessíveis através do Google. Além

disso, os criadores dos sites também usavam a plataforma de bate-papo on-line Discord para anunciar vídeos para venda e criar vídeos personalizados. A Sensity, empresa sediada em Amsterdã - que detecta e monitora mídia sintética desenvolvida por IA para setores como bancos e *fintech* -, divulgou um estudo que revela que 96% das *deepfakes* são sexualmente explícitos e apresentam mulheres que não consentiram na criação do conteúdo (Chan & Tenbarge, 2023).

Há, notadamente: 1) a conjuntura de capitalismo de plataforma que favorece o fenômeno; 2) avanços tecnológicos em IA e seu usos inapropriados, sobretudo da Inteligência Artificial Generativa, e ainda 3) ausência ou brechas de regulação, que corroboram para perpetuar a violência e a desinformação historicamente forjadas no tecido social.

2. Problema de pesquisa

Os *fake videos*, com indicativo de viés sexual e pornográfico sem consentimento expreso, reproduzem e favorecem violências digitais aperfeiçoadas com recursos potentes de Inteligência Artificial Generativa (*genAI*). Esses vídeos, inéditos até então no tecido da vida algorítmica, atingem sobretudo vítimas do gênero e/ou que se identificam como pertencentes ao gênero feminino. A problemática deste artigo assume, portanto, que se enfrenta a expansão do fenômeno dos vídeos com IA Generativa (*genAI*) na reprodução de violências de gênero, ancorada em processos de capitalismo de plataforma permissivo e cruelmente mercantil, cujo mercado que faturou milhões de dólares com este tipo de má-informação e poronografia falsa.

Assim, o fenômeno dos *fake videos* pornográficos baseados em IA Generativa associa, de maneira arbitrária e compulsória, os corpos das vítimas a conteúdos falsos e permissivos, distorcidos e

jamais permitidos. Seguramente, o fenômeno provoca danos à imagem, à honra, ao direito à privacidade, ao anonimato e à verdade, entre tantos outros prejuízos sociais, econômicos e psicológicos, e sob o manto irresponsável e tecnopolítico das empresas de tecnologia, de comércio eletrônico e do ecossistema financeiro de dados.

3. Deepfakes e cheap fakes: das técnicas à simplicidade dos usos

Paris e Donovan (2019) explicam que os primeiros vídeos amadores de troca de rostos com o uso de Inteligência Artificial surgiram em novembro de 2017. Foi nesse período que se deu o primeiro registro, com caso promovido e recirculado na plataforma Reddit, de um anônimo chamado "deepfakes" que publicou vídeos pornô de pessoas famosas como Taylor Swift e Scarlett Johansson (Albahar & Almalki, 2019). Desde então, a mídia e o público em geral passaram a utilizar o termo *deepfake* para se referir aos vídeos que utilizam machine learning ou deep learning para gerar rostos e corpos humanos (Paris & Donovan, 2019).

Em termos técnicos e conceituais, Albahar e Almalki (2019) esclarecem que uma *deepfake* é definida como uma técnica que permite alterar ou manipular a troca de um rosto por outro. Baseado em Inteligência Artificial Generativa (*genAI*), o sistema elabora o modelo de uma pessoa expressando ou realizando ações por meio de extensos conjuntos de dados que incorporam gravações, vídeos ou fotografias. A manipulação é alcançada ao empregar centenas ou milhares de imagens da pessoa-alvo, que servem como conjunto de dados (Albahar & Almalki, 2019). Os três tipos mais comuns de aprendizado de máquina utilizados em técnicas de *deepfakes* são as Redes Generativas Adversariais (GANs), Autoencoders Variacionais (VAEs) e Redes Neurais Recorrentes, as RNNs (Vales, 2019).

As Redes Adversárias Generativas (GANs) são um tipo de modelo de IA composto por duas redes neurais, o gerador e o discriminador, que competem entre si em uma espécie de jogo (Figura 1). O gerador cria amostras de dados, como imagens, tentando enganar o discriminador, que por sua vez tenta distinguir entre amostras reais e falsas (Vales, 2019). Esse processo iterativo resulta em uma melhoria contínua na capacidade do gerador de produzir amostras cada vez mais realistas, sendo amplamente utilizadas em tarefas de geração de imagens e síntese de voz.

Os Autoencoders Variacionais fazem parte de uma família de modelos usados em aprendizado de máquina, especialmente nas redes neurais. Eles têm a função de aprender formas eficientes e resumidas de dados complicados (Vales, 2019). De modo simplificado, a ideia dos VAEs é misturar a habilidade dos autoencoders padrão, que são como ferramentas de compactação de informação, com a força da inferência estatística, algo como entender a probabilidade e a incerteza associadas aos dados. As Redes Neurais Recorrentes são um tipo especial de modelo de aprendizado de máquina úteis para lidar com dados sequenciais, como séries temporais, texto ou áudio (Vales, 2019). A característica principal das RNNs é que elas têm uma espécie de "memória" interna, que permite capturar informações contextuais de sequências.

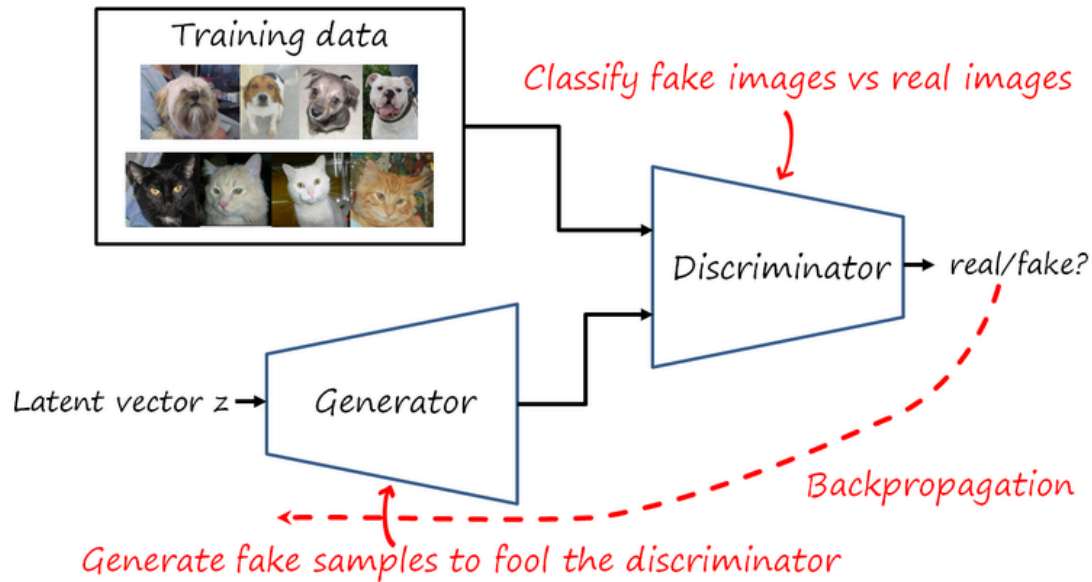


Figura 1 - Workflow de uma GANs (Dsa, 2022).

Sobre as GANs, em especial, cabe registrar que há uma escalada dessa arquitetura computacional em centenas de aplicativos que, simplificada, teriam permissões para gerar uma nova imagem, por exemplo, a partir do confronto de outras imagens; daí a expressão Adversárias. Especificamente no contexto das Redes Generativas Adversárias (GANs), é relevante destacar que essa arquitetura está experimentando um notável avanço, sendo incorporada em inúmeros aplicativos de *deepfake*. Um exemplo é o projeto *This Person does not exist* desenvolvido pela Nvidia (Nvidia, n.d.). Esses aplicativos possuem permissões que lhes possibilitam gerar novas imagens, muitas vezes através da confrontação de outras imagens, o que justifica a expressão "Adversárias".

Dessa forma, as Redes Generativas Adversárias (GANs) se configuram como uma manifestação de redes naturais em competição entre si, estabelecendo uma dinâmica de troca competitiva entre as duas inteligências artificiais. A interação ocorre sem a necessidade, em linhas gerais, de supervisão e comandos explícitos, que são comumente associados a sistemas dependentes de instruções claras (prompt) para gerar resultados específicos. Assim, as GANs (Goodfellow, Bengio & Courville, 2016) desempenham uma nova camada complexa para a geração não supervisionada, das quais milhares, hoje, originam-se e se recombinaem em centésimos de segundos, a cada nova vez um pouco melhor do que antes, e ininterruptamente se o sistema assim o quiser, num fluxo, arriscaríamos, *ad aeternum*.

No que se refere ao emprego da técnica e ao uso de software ou aplicativos na manipulação de audiovisual, Paris e Donovan (2019) pontuam diferenças entre *deepfakes* e *cheap fake*. Enquanto a primeira se trata de uma denominação mais ampla e que contempla o uso de software e a dependência de uma IA mais sofisticada com treinamento de dados, a segunda é composta por técnicas de acelerar, desacelerar, cortar, reorganizar ou recontextualizar a filmagem (Paris & Donovan, 2019). Evidentemente, as técnicas de *cheap fake* compõem as *deepfakes* e têm o mesmo potencial para serem transmitidas em velocidade e escala nas plataformas digitais. De acordo com os autores, as implicações estão atreladas a um menor nível de experiência e menos recursos técnicos necessários para alterar ou manipular imagens.

Nesse caso, o público em geral tem explorado as técnicas de *cheap fakes*, como troca de rosto, a aceleração e a desaceleração de vídeos, em programas como Adobe Premiere e After Effects, Sony Vegas, e aplicativos como o Snapchat. Por outro lado, as *deepfakes*, que utilizam troca de rosto, performances virtuais e vozes sintéticas, são produzidas com programas de código aberto como

FakeApp, FaceSwap e Deep-22 Face Lab (Paris & Donovan, 2019). Em geral, muitos desses aplicativos, criados por programadores independentes, estão disponíveis em plataformas como o GitHub. Para utilizá-los é necessário algum nível de expertise em computação e uma máquina mais robusta em termos de capacidade de processamento. Contudo, é relevante ressaltar que, desde 2019, é possível utilizar aplicativos em smartphones para criar *deepfakes*. Um exemplo é o Reface, disponível para iOS e Android. O aplicativo utiliza o design de estrutura GANs para mapear o rosto do usuário e integrá-lo de maneira realista a outra imagem, reproduzindo expressões faciais e movimentos de forma convincente.

As ferramentas de *fake nude* também são facilmente encontradas por meio de pesquisa nos motores de busca e podem ser utilizadas de maneira gratuita. O Nudify Online, por exemplo, é um site que permite, a partir de login com a conta Google, Apple ou Discord, fazer o upload de fotos e despir a pessoa (Nudify, 2024). O site explica que utiliza redes neurais sofisticadas para analisar e manipular imagens e oferece um tutorial sobre como criar *fake nude*.

4. Procedimentos metodológicos

Este artigo apresenta uma breve revisão bibliográfica sobre o avanço das *deepfakes* sob a perspectiva metodológica da Prospectiva Estratégica, da Escola Francesa La Prospective, de Michel Godet. A escolha pela Prospectiva Estratégica representa uma mirada de investigação que elabora cenários futuros a partir de análises, cruzamentos, dados sociais e econômicos para antecipar proposições, neste caso, de combate aos processos das *deepfakes* e má informação, sobretudo na perpetuação de mecanismos contra o corpo feminino, continuamente atacado e vulnerabilizado no cenário da IA e pornografia. Assim, o corpus da investigação foi extraído de casos ocorridos em escolas

do Brasil e no Reino Unido, exemplos pré-selecionados pela ótica da gravidade, atualidade e repercussão midiática. Os casos selecionados foram: 1) o documentário britânico *My Blond GF* (The Guardian); 2) duas reportagens sobre as recentes denúncias sobre *fake nude* produzidos em escolas brasileiras. Em seguida, definimos variáveis de literacia, impacto e potencial desinformativo e de má informação para identificar fundamentos e autores que discutam a escalada das *deepfakes*. Depois, debatemos estes cenários como “visões de futuro” (Cristo, 2003) que consolidam a proposta desta metodologia: prospectar políticas públicas e educacionais que, diante do quadro apresentado, possam articular a construção de um futuro desejável, de forma antecipada e preditiva, isto é, prevenindo e combatendo este tipo de má-informação produzida com uso indiscriminado e não regulado de AI Generativa, especialmente destinado à pornografia. Os estudos de Cover (2022) e Maddocks (2020) atestam que o *pornfake* se alastra em todo o mundo, tornando-se um dos principais desafios cibernéticos atuais, sobretudo para o gênero feminino.

Os três exemplos trazidos no artigo são observados à luz dos enquadramentos criados pelos autores. São eles: 1) baixo potencial de promover desinformação; 2) médio/neutro potencial de estimular processos desinformativos; e 3) alto potencial de propagar ondas de mentiras e má-informação. Como variáveis para interpretar, analisar e qualificar os casos estudados, foram formuladas: a) escala de auditabilidade; b) escala de espalhabilidade; c) escala de responsabilidade legal; d) escala de consequências globais; e) escala de corresponsabilidade das plataformas.

5. CASO 1 - *My Blonde GF*: sem lei não há responsabilidade

A vida de Helen é subitamente transformada quando ela se depara com a revelação de que seu rosto foi manipulado digitalmente em imagens retratando mulheres em contextos sexualmente explícitos e, por vezes, violentos. Esse é o dramático resumo que antecede a história da escritora britânica Helen Mort, retratada no documentário *My Blonde GF: a disturbing story of deepfake pornography* (Figura 2), produzido pelo jornal The Guardian e disponibilizado ao público em outubro de 2023 (Poulton & Gormley, 2023). Durante 18 minutos, Helen compartilha os efeitos reais e brutais da pornografia *deepfake* em sua vida. Em um relato marcado por sentimentos de angústia, desrespeito e indignação, a história desafia uma cultura que trata a aparência e a imagem das mulheres como uma propriedade pública, ao mesmo tempo em que os responsáveis permanecem no anonimato.



Figura 2 - *My Blonde GF: a disturbing story of deepfake pornography*, da diretora Rosie Morris (Poulton & Gormley, 2023).

Em entrevista à BBC, a diretora do documentário, Rosie Morris, disse que o intuito seria fazer com que o espectador caminhasse junto com Helen para compreender os efeitos traumáticos da pornografia *deepfake* (Bushby, 2023). O caso de Helen Mort se tornou público em 2021, quando a escritora descobriu que seu rosto havia sido digitalizado e estava sendo utilizado em imagens sexuais explícitas e violentas em um site pornográfico. Desde então, a escritora exige mudanças na lei britânica para o compartilhamento não consensual de imagens pornográficas. Desde 26 de outubro de 2023, a Lei de Segurança On-line do Reino Unido criminaliza a criação e partilha de imagens ou vídeos explícitos manipulados sem consentimento. O caso apresenta um alerta evidente quanto à aceleração do fenômeno da má informação, com o aumento na oferta da Inteligência Artificial Generativa e a baixa necessidade de tecnologia especializada, rodando em sistemas de geração de imagens que possuem uma altíssima similaridade com produções profissionais. Isso resulta em uma elevada dificuldade de suscitar dúvidas ou descrença, o que nos leva a crer que episódios como o vivido por Helen demandam uma regulação mais rigorosa diante do comportamento das plataformas, frequentemente omissas e remuneradas. Essa regulação deverá focar no aumento da capacidade de auditabilidade e na verificação de todo o conteúdo que apresenta imagens sexuais explícitas, além do uso de dados dos usuários sem o seu consentimento.

Nesse sentido, estudos como os de Gerardi, Walters e James (2020), que sugerem uma espécie de marca d'água em imagens que são alteradas ou manipuladas com IA Generativa, parece ser um caminho pertinente de identificação do conteúdo e uma maneira de combate à desinformação no ambiente digital, assim como as tecnologias de *blockchain* que permitem registrar as transações dos usuários e entregar maior confiança por meio de uma etiquetagem autenticada e *tokenizada*. Embora o

caso de Helen possa representar um potencial médio de espalhabilidade (Jenkins, 2009), uma vez que esse tipo de conteúdo circula muito rapidamente em sites ou plataformas específicas para um determinado tipo de público, as brechas em responsabilidade legal e a corresponsabilidade de autoria (de quem produz, divulga e hospeda na Internet) possuem alto potencial permissivo para que as *deepfakes* pornográficas continuem sendo uma prática cada vez mais comum. Evidentemente, o alto potencial da irresponsabilidade legal e a corresponsabilidade das plataformas, inclusive dos sistemas monetários e tecnoliberais diversos, contribuem para aumentar o nível de propagação da má-informação como expôs reportagem especial da rede estadunidense NBC News (Figura 3). A reportagem investigativa revelou que as *fake nude* são facilmente monetizadas, com anuência e lucros assumidos por empresas como Google e Mastercard (Tenbarger, 2023).

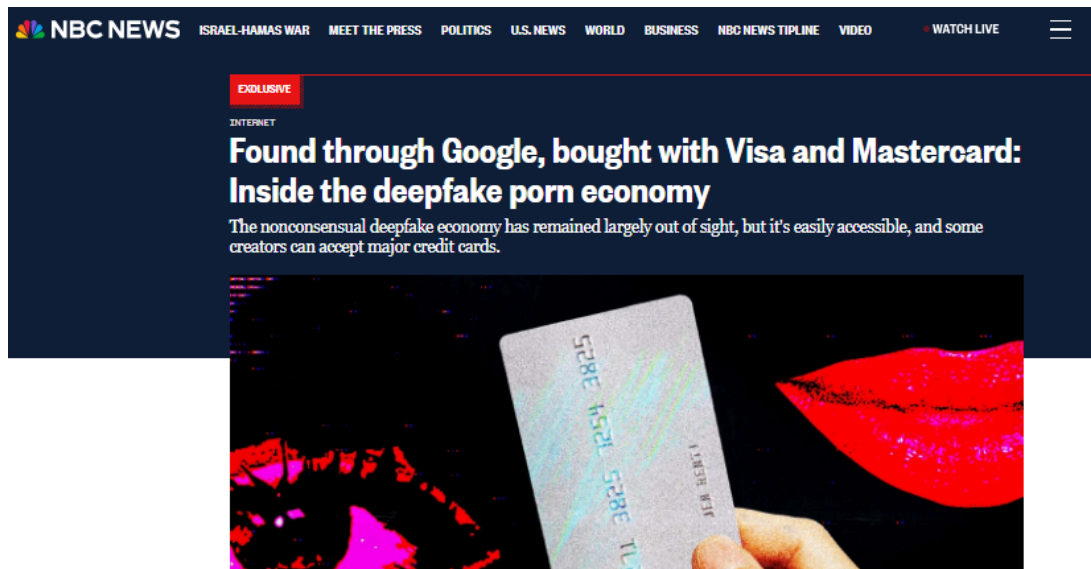


Figura 3 - *Fake nude* pode ser comprada com cartão de crédito (Tenbarger, 2023).

As punições previstas, seja em forma de projeto de lei, seja como a Lei de Segurança On-line do Reino Unido, podem ser encaradas como um avanço em relação às *deepfakes*. No entanto, o fato é que o capitalismo de plataforma – que movimenta pesquisas no Google, pagamento por cartão de crédito para acesso ao conteúdo pornô *fake* e até mesmo remuneração pela produção desse tipo de conteúdo – sobrepõe-se à ordem ética e punitiva. Deste modo, entende-se que o panorama da IA Generativa, neste recorte específico, oferece alto potencial no âmbito das consequências globais das *deepfakes* e na escalada do *porn fake*. Caracteriza-se como fenômeno sociotécnico de difícil combate, responsabilização e rastreio da cadeia de má-informação - das plataformas aos usuários e criadores deste tipo de material.

5.1. CASO 2 - *Deepfakes* e *fake nude* na escola brasileira

Os meses de outubro e novembro de 2023 foram marcantes para o avanço considerável das *deepfakes* envolvendo pornografia no Brasil. Dois casos chamam a atenção, em especial: a exposição de alunos e alunas de escolas brasileiras, expondo um contexto que desafia o ambiente escolar, professores e pais, principalmente pela sofisticação dos aplicativos e o perigo das manipulações que chegaram ao imaginário e à realidade escolar. O Brasil registrou pelo menos dois casos, com quase 50 vítimas em poucas semanas. Neles, o mesmo *modus operandi* de geração de conteúdo pornográfico sem consentimento, e falso, por meio de aplicativos de baixa especialidade.

As alunas de um colégio do Rio de Janeiro foram expostas e constrangidas a partir do compartilhamento de vídeos falsos e fotos manipuladas com o uso de imagens extraídas de perfis pessoais/profissionais nas redes sociais digitais e regeneradas com aplicativos de IA. Dali, os nudes

falsificados passaram a circular em aplicativos de mensagem instantânea como o WhatsApp e no aplicativo que domina o consumo de milhões de jovens, o Discord, plataforma que acumula acusações recorrentes por permitir que usuárias e usuários transmitam e comuniquem abusos, crimes de ódio e automutilação.

O primeiro caso de *deepfakes* e pornografia ocorreu no final do mês de outubro e ganhou repercussão nacional quando cerca de 20 adolescentes relataram à polícia que descobriram, em listas e grupos, que seus colegas estavam divulgando montagens pornográficas com rostos, corpos, cenas e atuação em conteúdo pornográfico. A escola chegou a produzir um comunicado para se posicionar, muito brevemente, dizendo que uma “ferramenta criada para solucionar problemas e apoiar a vida moderna” tenha sido utilizada de forma inadequada. O crime foi registrado na Delegacia de Proteção à Criança e ao Adolescente da referida capital do estado do Rio de Janeiro. O segundo caso ocorreu em novembro de 2023, em uma escola de Recife, em Pernambuco. As fotos alteradas teriam sido divulgadas por meio de um aplicativo de mensagens. No comunicado à imprensa, a escola relatou que orientou e prestou atendimento às famílias das vítimas e às alunas que sofreram o crime. A nota diz ainda que “a equipe do colégio acompanhou os familiares para registrar queixa na Delegacia de Apuração de Atos Infracionais - DEPAI e também fez denúncia no Conselho Tutelar”.

Em ambos os casos, os relatos das vítimas revelam claramente o potencial ofensivo e violento vivido. De acordo com nossa pesquisa, verificamos nos dois casos analisados neste artigo baixa auditabilidade, espalhabilidade restrita, sobretudo nos primeiros dias, dentro de comunidades fechadas de criadores de conteúdo falso (pessoas próximas – o que favorece que o consumo siga sem o alcance das autoridades, muitas vezes) ou grupos menos abertos. Quanto à responsabilidade legal nota-se uma

conduta institucional de qualidade baixa informativa, com potencial alto de normalizar ou subestimar o lastro de prejuízos psicológicos, educacionais e geracionais.

Trata-se de um problema nas áreas da Educação, da Comunicação Pública e do Letramento Digital, com Responsabilidade das Plataformas. Assim, percebemos, alimenta-se mais uma etapa de reprodução do colonialismo de dados, com determinante corresponsabilidade das plataformas, o que indica que é preciso uma coalizão que promova o letramento educacional de base complementar para contextos de IA, um desafio escolar mundial, em primeira escalada aqui no Brasil. É possível notar que, mesmo após a remoção do conteúdo e o posicionamento oficial de autoridades e das instituições envolvidas, a memória provocada pelo *deepfake* é permanente.

Na perspectiva da pesquisadora americana, Renee Cummings, especialista em ética da IA, a pornografia *deepfake* "pode destruir a vida de alguém", citando vítimas que sofrem com ansiedade, depressão e estresse pós-traumático.

Considerações finais

Em outubro de 2023, o Fórum Econômico Mundial (WEF) lançou um informativo em que se debruça na equidade dos dados nos modelos básicos, tanto em termos do impacto da IA Generativa (*genAI*) na sociedade como no desenvolvimento futuro de ferramentas de *genAI*. A equidade de dados nos modelos básicos refere-se à preocupação e prática de garantir que os conjuntos de dados utilizados para treinar e alimentar modelos de machine learning sejam justos e imparciais em relação a diferentes grupos demográficos. Isso significa que os dados devem representar adequadamente todas as categorias relevantes da população, sem viés injusto ou discriminação. Assinado por 22

pesquisadores e cientistas, o documento intitulado Equidade de dados: conceitos básicos para IA generativa busca: 1) estabelecer um vocabulário partilhado para facilitar a colaboração e o diálogo em torno dos avanços da IA; 2) avaliar as preocupações iniciais para estabelecer uma estrutura de investigação na qual as partes interessadas possam se concentrar; e 3) moldar o desenvolvimento futuro de tecnologias promissoras. O texto representa um primeiro passo na exploração e promoção da equidade de dados no contexto da IA Generativa. As definições, estrutura e recomendações propostas destinam-se a moldar proativamente o desenvolvimento de tecnologias promissoras de *genAI*.

É fato que o desenvolvimento das tecnologias digitais transformaram a sociedade em diversas áreas, em especial, a comunicação. Nesse âmbito, a literacia digital passou a ser uma habilidade essencial para navegarmos neste novo mundo. As mudanças impactam a produção e o consumo das notícias. A audiência desenvolveu tecnicidades para manusear aplicativos, softwares, dispositivos móveis e a dinâmica das redes. Contudo, o público ainda precisa habilitar e refinar comportamentos que denotem uma preocupação com a literacia digital, tais como a busca on-line por informações de qualidade e críveis; a utilização das mídias sociais de forma consciente e ética; e, como fica exposto no texto do Fórum Econômico Mundial, a clareza sobre os limites de uso e aplicações com IA.

Com a constante evolução das tecnologias, faz-se ainda, reiteramos, urgente a regulação de condutas antiéticas e criminosas nas redes. Outro ponto preocupante desta investigação é a reprodução da violência. Fica evidente que as mulheres são as maiores vítimas do *fake nude*. O episódio 49 do Podcast Papo de Futuro, da Rádio Câmara discutiu, no dia 14 de novembro de 2023, a questão das “*Deepfakes* e a vulnerabilidade da mulher”, levantando o porquê do público feminino (e não o masculino) ser a maior vítima dessa exploração, além de falar na objetificação da mulher nas redes

sociais digitais. Durante a entrevista ao Papo de Futuro, a psicanalista Vera Iaconelli destaca que a Internet amplifica discursos conservadores e potencializa o que há de pior na sociedade, sendo um ambiente em "que ninguém se responsabiliza por nada". A psicóloga acredita que a Internet é mais aditiva que qualquer droga. Segundo ela, a lei da Internet é a lei do like, do que é mais escatológico e mais violento, mais disruptivo, e que sem regulação pode ser utilizada das piores formas. As plataformas on-line têm políticas internas para lidar com comportamentos criminosos em seus serviços. No entanto, as medidas e políticas não são consideradas suficientes, e a pressão aumenta para que as plataformas assumam maior responsabilidade na prevenção, detecção e remoção de conteúdo ilegal e criminoso.

Em um cenário marcado pela disputa entre regulação e liberdade, a discussão sobre a regulamentação das plataformas é vital. O discurso direitista destaca a importância da liberdade de expressão como pilar democrático, argumentando que a intervenção estatal pode sufocar vozes divergentes. Paralelamente, a defesa do livre fazer do capitalismo ressalta a capacidade inovadora e a eficiência do mercado não regulamentado, que também inclui requintes de crueldade e desrespeito, em especial no que se refere à produção de vídeos pornô *fakes*. Esses elementos, no cerne da discussão, desafiam a sociedade a encontrar um meio-termo, garantindo a expressão livre, ao mesmo tempo em que protege contra excessos prejudiciais causados pelas *deepfakes*. Promover a autorregulação da indústria tecnoliberal da Inteligência Artificial Generativa (*genAI*), envolvendo ativamente as plataformas na definição e adesão a padrões éticos, surge como desafio global de difícil solução, mas urgente de ser planejado e liderado pela sociedade, especialmente no âmbito da política, da escola e da vida digital, em todos os seus aspectos. Ao mesmo tempo, consideramos crucial investir em educação

digital constante, acessível e ampla. Isto, com o claro objetivo de letrar usuários e usuárias a discernir informações. O objetivo, nesta mirada, portanto, é fomentar a utilização responsável da *genAI* e implementar parâmetros sólidos de participação cidadã em contexto de fabricação de pornografia com IA, inclusive no combate, vigilância e denúncia, diante da publicação e recirculação deste tipo de conteúdo ilegal, malicioso e ainda mais violento para as mulheres.

Referências Bibliográficas

- Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, 97(22), 3242-3250.
- Brasil. (2018). Lei nº 13.718, de 24 de setembro de 2018. Altera o Decreto-Lei nº 2.848, de 7 de dezembro de 1940 (Código Penal), para tipificar os crimes de importunação sexual e de divulgação de cena de estupro, e o art. 1º da Lei nº 8.072, de 25 de julho de 1990, para incluir os crimes previstos nos arts. 218-A e 218-B do Decreto-Lei nº 2.848, de 7 de dezembro de 1940. *Diário Oficial da União*, Brasília, DF.
- Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Bushby, B. H. (2023, June 18). Deepfake porn documentary explores its “life-shattering” impact. *BBC News*.
<https://www.bbc.com/news/entertainment-arts-65854112>
- Chan, M.; & Tenbarge, K. (2023, November 23). For teen girls victimized by ‘deepfake’ nudes, there is little recourse. (2023, November 23). *NBC News*.
<https://www.nbcnews.com/news/us-news/little-recourse-teens-girls-victimized-ai-deepfake-nudes-rcna126399>
- Cover, R. (2022). Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation, *Continuum*, 36:4, 609-621, DOI: 10.1080/10304312.2022.2084039
- Cristo, C. M. P. N. (2003). Prospectiva estratégica: instrumento para a construção do futuro e para a elaboração de políticas públicas. *Revista do Serviço Público*, 54(1), 63-78.
- Das, A. K., & Tripathi, M. (2023). Strategies for checking misinformation: An approach from the Global South. *IFLA journal*, 49(1), 3-17.
- Dsa. (2022, December 19). Capítulo 54 - Introdução às Redes Adversárias Generativas (GANs - Generative Adversarial Networks) - Deep Learning Book. Deep Learning Book.
<https://www.deeplearningbook.com.br/introducao-as-redes-adversarias-generativas-gans-generative-adversarial-networks/>

- Gerardi, F.; Walters, N.; & James, T. (2020). Cyber-security implications of deepfakes. University College London.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
- Jenkins, H. (2009, February 11). If it doesn't spread, it's dead (Part one): Media viruses and memes — pop junctions. Henry Jenkins. https://henryjenkins.org/blog/2009/02/if_it_doesnt_spread_its_dead_p.html
- Mania, K. (2020). The legal implications and remedies concerning revenge porn and fake porn: a common law perspective. *Sexuality & Culture*, 24(6), 2079-2097.
- Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, 7(4), 415-423.
- Nudify. (2024, February 2). Steps to make a deep fake nude on Nudify. Nude Blog. <https://blog.nudify.online/steps-to-make-a-deep-fake-nude-on-nudify/>
- Nvidia. (n.d.). This person does not exist. <https://thispersondoesnotexist.com/>
- Paris, B., & Donovan, J. (2019). Deepfakes and cheap fakes. The manipulation of audio and visual evidence. *Data & Society*.
- Poulton, L., & Gormley, J. (2023, October 28). My Blonde GF: a disturbing story of deepfake pornography. *The Guardian*. <https://www.theguardian.com/technology/video/2023/oct/25/my-blonde-gf-a-disturbing-story-of-deepfake-pornography>
- Projeto Vazou. (2018). Pesquisa sobre o vazamento não consentido de imagens íntimas no Brasil (2018). <https://www.crimlab.com/projetovazou/resultado.pdf>
- Sousa, L. S.; & Silva, L. M. de A. (2022). Pornografia de vingança: aspectos históricos e conceituais. [Monografia, Curso de Direito, Universidade do Sul de Santa Catarina]. Acervo Digital da UNISUL. <https://repositorio.animaeducacao.com.br/handle/ANIMA/29183>

Tenbarge, K. (2023, March 27). Found through Google, bought with Visa and Mastercard: Inside the deepfake porn economy. NBC News.

<https://www.nbcnews.com/tech/internet/deepfake-porn-ai-mr-deep-fake-economy-google-visa-mastercard-download-rcna75071>

Vales, A. (2019, October 11). An introduction to synthetic media and journalism - WSJ Digital Experience & Strategy - Medium. Medium.

<https://medium.com/the-wall-street-journal/an-introduction-to-synthetic-media-and-journalism-cbbd70d915cd>

Wardle, C., Kerwin, J., & Dias, N. (2017). Information Disorder: The Essential Glossary. Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School. Harvard.