Deepfake no Caso Taylor Swift: a inteligência artificial generativa e o cerceamento da segurança das mulheres no espaço digital[1]

João Pedro Felix Ortiz CAMARGO[2]
Universidade Federal de Goiás, Goiânia, GO

RESUMO

O avanço da Inteligência Artificial e todas as suas ferramentas generativas trazem consigo a construção de um espaço hiper-real, sem o posicionamento de parâmetros éticos. O presente texto busca discutir os desafios e perigos que o avanço do *deepfake* traz para a segurança das mulheres no espaço digital. Utilizaremos aqui o caso referente a cantora estadunidense Taylor Swift, ao qual a tecnologia generativa foi utilizada para criação de imagens de teor sexual, ampliando a discussão política e social sobre o tema. Para contribuir com a discussão, recorremos a Han para debater o espaço digital e a violência, e autores diversos que contribuíram com análises compatíveis.

PALAVRAS-CHAVE: Inteligência Artificial; *deepfake;* generativa; violência; digital.

INTRODUÇÃO

A tecnologia *deepfake*, que envolve principalmente a criação de vídeos, imagens e áudios falsos com realidade e velocidade ampliada usando a inteligência artificial generativa, tem ganhado significativa atenção nos últimos anos (Mirsky e Lee, 2021). Essa tecnologia têm levantado preocupações quanto ao potencial uso indevido, pois, no que Han (2019) chama de espaço hiper-real, contextualizando o digital no contemporâneo, essa tecnologia pode desafiar a legitimidade e autenticidade de vídeos e áudios como evidências definitivas de eventos.

Por se tratar de uma tecnologia que se baseia em bancos de dados disponíveis em suas redes, o avanço do consumo e produção no espaço digital retroalimenta o potencial dessa ferramenta. A tecnologia possui maiores chances de alcançarem pessoas ricas, famosas e publicizadas (Seow *et al.*, 2022) Essa disposição das informações tem

[1] Trabalho apresentado ao Grupo de Trabalho Comunicação e Inteligência Artificial, evento integrante da programação do 27º Congresso de Ciências da Comunicação na Região Sudeste, realizado de 30 de maio a 1º de junho de 2024.
[2] Mestrando no Programa de Pós-Graduação em Comunicação (PPGCOM) da Faculdade de Informação e Comunicação (FIC) da Universidade Federal de Goiás (UFG), email: joaopedrofelix@discente.ufg.br

facilitado ainda mais a criação de notícias falsas e rumores realistas que podem ter um impacto sério na sociedade.

O presente trabalho usa como o caso da cantora estadunidense Taylor Swift, em um caso recente de ataques virtuais pela rede X, anteriormente, o Twitter. O caso aconteceu em janeiro de 2024, com o compartilhamento de imagens explicitamente sexuais por usuários anônimos. Apesar de não ser o primeiro caso que evidencia o uso da tecnologia de maneira criminosa e violenta contra a integridade feminina, o caso se destaca pela repercussão tomada, alcançando discussões no cenário político do país e a mobilização digital, na mesma rede social do ato, com a *hashtag "#ProtectTaylorSwift".*

As proporções que ferramentas generativas podem tomar, além da violação do direito à privacidade e de imagem, podem ir de ameaças e chantagens a danos psicológicos e pânico em massa pela população (Seow *et al.*, 2022). As imagens, segundo o jornal *The New York Times,* foram visualizadas mais de 47 milhões de meses. Na mesma matéria construída, o texto indica a chegada de um "tsunami" de imagens geradas por Inteligência Artificial e de caráter sexual.

Nesse sentido, o Han diz que a Inteligência Artificial é capaz de "influenciar nosso comportamento num nível que fica embaixo do limiar da consciência" (2019, p.23). Portanto, o cerceamento das mulheres no digital avança cada vez mais, em uma velocidade incontrolável. E, esse avanço, pode moldar a forma de consumir conteúdos tanto no digital quanto fora dele.

De acordo com Akhtar (2023), estima-se que cerca de 1,8 bilhão de imagens e vídeos sejam carregados diariamente no digital, desde redes sociais até sites de profissionais. Ele observa que uma parcela significativa, entre 40% e 50%, dessas mídias parece ser manipulada, seja para propósitos benignos, como retoques em imagens para capas de revistas, ou para fins adversos, como campanhas de propaganda ou desinformação.

Os dados trazidos pelo estudo de Akthar (2023) nos permite destacar com maior especificidade o problema da manipulação de imagens e vídeos de rostos humanos, destacando sua ameaça à integridade das informações na Internet e aos sistemas de reconhecimento facial, dada a importância central dos rostos nas interações humanas e na identificação biométrica de pessoas. O autor enfatiza que as manipulações plausíveis

em amostras de rostos podem comprometer criticamente a confiança nas comunicações digitais e nas aplicações de segurança, como a aplicação da lei.

No espaço digital, especialmente em redes sociais que possuem como base de consumo, a visualidade, a produção de conteúdos na contemporaneidade são acelerados e desordenados, conforme discute Han (2019). Esse fenômeno estrutura o que o autor chama de "supercomunicação" (Han, 2017), uma forma de violência neuronal e que resulta na desvalorização da linguagem como discurso e como ferramenta comunicacional.

Portanto, consideramos que esse excesso de informação é algo natural do presente tempo. Podemos identificar ruídos e resíduos de dados, imagens e conteúdos ao qual somos bombardeados, seja em plataformas de interações sociais, como o citado X, seja em plataformas de *streaming*. O ponto, portanto, é que nos vemos cercados pelo domínio de algoritmos e do consumo proporcionado pelos mesmos.

Esse posicionamento do digital transformam atualmente a forma de estar e consumir. Han (2019) diz que o digital contemporâneo se vê diante de uma forma de controle elaborada por ele como regime de informação. Esse regime é formado pelo controle e domínio, “no qual informações e seu processamento por algoritmos e inteligência artificial determinam decisivamente processos sociais e políticos” (Han, 2022, p. 7).

No caso de figuras publicizadas como a cantora Taylor Swift, o potencial de ferramentas generativas são violentas e cerceantes, não apenas para sua reputação e no que tange a moral e segurança de imagem, mas também, para sua integridade física e emocional. Podemos entender, como uma violação do arbítrio e da identidade, que a disseminação de *deepfakes* é um ato de violência característico da contemporaneidade. E, conforme afirma Arendt (1994), “a prática da violência como toda ação, transforma o mundo, mas a transformação provável é em um mundo mais violento.”. A aplicabilidade da afirmação de Arendt com os conceitos de Han nos permitem, então, categorizar que no digital, o uso de *deepfakes* é uma forma de violência.

Dessa forma, Sodré (2009, p. 45) diz “é cada vez mais difícil separar o imaginário do real ou o verdadeiro do falso”. Essa afirmação foi feita trabalhando, especificamente, o jornalismo e as narrativas criadas, porém, se aplica fielmente ao tema aqui tratado. Han (2022, p. 81) afirma que hoje, vivemos em uma nova forma de

niilismo, novamente contextualizado ao digital, com a perda da credibilidade factual e do compromisso com a verdade.

Com esse cenário de um "universo desfactuado", Han (2022, p.81) afirma que, juntamente ao regime da informação e a superprodução, "Passam a circular, então, informações totalmente desacopladas da realidade, formando um espaço hiper-real." Essas informações então criam narrativas, como a vista no Caso Taylor Swift, que apesar da consolidação que o conteúdo era falso, seu compartilhamento alcançou uma massa em milhões.

Casos como esse constroem narrativas, e segundo Han (2019, p. 113-14), essas narrativas não se destinam a argumentar, mas sim a buscar agradar e entusiasmar, o que constitui a base de sua alta efetividade. Ele argumenta que as formas de entretenimento narrativas presentes nas mídias de massa desempenham um papel crucial e perigoso na solidificação do novo comportamento da sociedade no digital, ao familiarizar as normas morais e enraizá-las nas inclinações, no cotidiano e na aceitação do "é assim que é", que, segundo o autor, não requer nenhum questionamento ou reflexão adicional.

Em suma, de acordo com Akhtar (2023), podemos observar na literatura que a maioria dos *frameworks* existentes para detecção de *deepfakes* apresentam uma queda significativa de desempenho ao serem testados com manipulações ou conjuntos de dados não utilizados durante o treinamento. Portanto, a detecção de imagens e vídeos ainda representam um grande desafio para garantir a credibilidade do conteúdo no digital.

O presente trabalho busca, ao fim, ampliar o desenvolvimento do tema no cenário nacional, vista a literatura acadêmica do tema no Brasil ainda se encontrar com baixo volumo, exigindo o uso de autores estrangeiros para trabalhar sobre a nova modalidade generativa. Ao fim, buscamos convergir autores que trabalham questões sociais com estudos localizados da área, para apresentar um dos novos desafios éticos e profissionais da comunicação. Esperamos, ao fim, poder expandir a discussão sobre a segurança das mulheres no digital e garantir a ampliação de vozes que permitam que o contexto político e social possam elaborar legislações e normativas em relação ao uso das I.A's.

De tal modo, a pressão social por parte da sociedade, contribuída pelo impacto ainda presente de profissionais da comunicação, pode permitir até mesmo considerações

por parte das empresas responsáveis pelo gerenciamento de banco de dados e algoritmos em torno de seus acessos e permissões. Vivemos em um mundo contemporâneo, onde ainda enfrentamos problemas arcaicos, portanto, assim como existe uma adaptabilidade por parte da tecnologia generativa, o combate contra a desinformação deve acompanhar o mesmo passo, ou concretizarmos a chamada morte da verdade.

REFERÊNCIAS

ARENDT, Hannah. **Sobre a Violência.** Rio de Janeiro: Relume-Dumará, 1994.

AKHTAR, Zahid. Deepfakes Generation and Detection: A Short Survey. J. **Imaging, v.** 9(1), 8, 2023. Disponível em: https://doi.org/10.3390/jimaging9010018

CONGER; Kate; YOON, John. Explicit Deepfake Images of Taylor Swift Elude Safeguards and Swamp Social Media. **The *New York Times*,** 26 jan. 2024. Disponível em:<https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html>

HAN, Byung-Chul. **Topologia da violência**. Rio de Janeiro: Vozes, 2017

HAN, Byung-Chul. **Infocracia**. Rio de Janeiro: Vozes, 2022.

HAN, Byung-Chul. **Bom Entretenimento.** Rio de Janeiro: Vozes, 2017.

MIRSKY, Yisroel; Lee, WENKE. The Creation and Detection of Deepfakes: A Survey. **ACM Computing Surveys,** v. 54, p 1-41, 2021. Disponível em: <https://dl.acm.org/doi/10.1145/3425780>

SEOW, Jia Wen; LIM, Mei Kuan; PHAN, Raphaël C.W.; LIU, Joseph K. (2022). A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. **Neurocomputing**, V. 513, p. 351-371. Disponível em <https://doi.org/10.1016/j.neucom.2022.09.135.>

SODRÉ, Muniz. **A narração do fato.** Rio de Janeiro: Vozes, 2009.