Proteção da Dignidade Feminina na Era dos Deepfakes e GANs

Luis Eduardo Rasch*
Universidade Federal de Pelotas
Pelotas, Brasil
lerasch@inf.ufpel.edu.br

William de Almeida Pavinato*
Universidade Federal de Pelotas
Pelotas, Brasil
wapavinato@inf.ufpel.edu.br

Larissa V. da Silva Mendes*
Universidade Federal de Pelotas
Pelotas, Brasil
larissa.mendes@ufpel.edu.br

Tatiana Aires Tavares*
Universidade Federal de Pelotas
Pelotas, Brasil
tatiana@inf.ufpel.edu.br

Micael Fontoura Mendes*
Universidade Federal de Pelotas
Pelotas, Brasil
micael.mendes@ufpel.edu.br

Manoela Viera de Moura*
Universidade Federal de Pelotas
Pelotas, Brasil
mvmoura@inf.ufpel.edu.br

Karen Satie Ono*
Universidade Federal de Pelotas
Pelotas, Brasil
ks.ono@inf.ufpel.edu.br

ABSTRACT

The rapid advancement of Generative Adversarial Networks (GANs), a subset of artificial intelligence (AI), has enabled the creation of highly realistic synthetic media known as deepfakes. While this technology offers innovative applications, it also presents significant legal and ethical challenges, especially in the creation of non-consensual pornographic content and so-called revenge pornography, endangering factual and legal safety, particularly for women. In this context, this article examines the legal and ethical implications of deepfakes, focusing on existing regulatory frameworks and proposing measures to mitigate their misuse as a means to advance the United Nations Sustainable Development Goal (SDG) number 5: achieving gender equality and the empowerment of all women and girls. Accordingly, we analyze methods for detecting GAN-generated images and ongoing legal initiatives, which reveal the need for a balanced approach combining technical solutions, legal accountability, and public awareness. Our findings highlight the urgency of regulatory updates and technological innovations to protect individual rights, democratic integrity, and women's freedom in the era of AI-generated content.

KEYWORDS

deepfake, artificial inteligence, gender equality, GANs, generative adversarial networks, non-consensual pornography, ODS

1 INTRODUÇÃO

Nas últimas décadas a área de Inteligência Artificial (IA) vem tendo sua evolução impulsionada pelos avanços no poder computacional, pela abundância de dados e, sobretudo, pelo desenvolvimento de métodos de *machine learning* e *deep learning* [17] [19].

Avanços recentes na área incluem modelos generativos, especialmente os *Large Language Models (LLMs)* e as *Generative Adversarial Networks (GANs)*, enquanto as primeiras revolucionaram o processamento de linguagem natural e a geração multimodal de texto, as *GANs* transformaram a criação de imagens e vídeos sintéticos de alta fidelidade em múltiplas áreas [18].

Contudo, é importante ressaltar que, apesar de seus avanços, o rápido desenvolvimento das tecnologias de inteligência artificial também ampliou desafios éticos e sociais, como a proliferação de deepfakes, a manipulação de informações visuais e a crescente dificuldade de distinguir conteúdos autênticos de materiais sintetizados. Diversos paradigmas de geração de imagens têm sido propostos nos últimos anos, incluindo modelos baseados em difusão, autoencoders variacionais e redes adversariais generativas.

O presente trabalho realiza um recorte analítico focado especificamente nas GANs, sem a pretensão de abranger todo o espectro de técnicas contemporâneas de geração de imagens sintéticas. Como consequência do avanço dessas técnicas de geração hiperrealista, observa-se também o agravamento de problemas éticos relevantes, especialmente aqueles que afetam corpos femininos, como a pornografia não consensual e a pornografia de vingança [22, 23].

Nesse contexto, outros pontos críticos dizem respeito aos vieses presentes nos dados de treinamento, que podem ser aprendidos e reproduzidos pelos modelos, perpetuando desigualdades e padrões discriminatórios, e à opacidade de seus processos internos, que reduz a transparência e dificulta auditorias independentes sobre o funcionamento e as decisões geradas por esses sistemas [9].

Ademais, no campo jurídico, observa-se que embora ainda inexista tipificação penal específica para *deepfakes* no ordenamento brasileiro, sua produção e circulação podem se enquadrar no **art. 218-C do Código Penal** [5], dispositivo destinado à proteção da intimidade sexual e que também abrange conteúdos fabricados artificialmente.

Nesse sentido, a interpretação de que o conteúdo gerado por *GANs* aplica-se na letra da lei brasileira busca assegurar segurança jurídica, reduzir assimetrias de gênero e contribuir para o alcance do **5° Objetivo de Desenvolvimento Sustentável da ONU**, que reflete a necessidade da igualdade de gênero e do empoderamento feminino [3][12][20][24].

Contudo, a interpretação mencionada não é o suficiente, considerando que em países como a Espanha e o Reino Unido [1] já aprovaram normas específicas para conteúdo sexual proveniente de

*Autores que contribuíram igualmente para este trabalho.

deepfakes, dessa forma evitando as lacunas hermenêuticas que são absolutamente negativas para a liberdade e igualdade femininas.

Assim, o panorama contemporâneo da Inteligência Artificial mostra-se ambivalente: enquanto impulsiona inovações disruptivas e cria novas possibilidades em diversos setores, também impõe riscos e incertezas que exigem regulação clara, mecanismos efetivos de responsabilização e uma abordagem interdisciplinar alinhada a valores éticos e sociais.

2 RISCOS JURÍDICOS E SOCIAIS

Os modelos de geração audiovisual, ao permitirem a produção cada vez mais realista de imagens e vídeos manipulados, apresentam riscos diretos à integridade informacional, à segurança e à reputação de indivíduos. Dentre os usos mais preocupantes das *GANs* está a criação de pornografia não consensual, prática na qual rostos de vítimas são inseridos em cenas explícitas sem autorização, e ainda, a pornografia de vingança facilitada por IAs, a distribuição pública de imagens ou vídeos de teor sexual ou íntimo de uma pessoa sem o seu consentimento, geralmente praticada por ex-parceiros, na intenção de humilhar e diminuir o outro [2][3]. Desse modo, é essencial compreender que a produção e disseminação de *deepfakes* pornográficas constitui violação direta dos direitos da personalidade, especialmente da imagem, privacidade e honra, configurando dano moral presumido [1].

2.1 Casos Reais

Evidentemente, o uso de *deepfakes* para fins pornográficos não consensuais é um exemplo concreto de violência digital, atingindo tanto figuras públicas quanto pessoas comuns. Casos envolvendo celebridades como Taylor Swift, que teve imagens falsas amplamente disseminadas em plataformas como *X* e *Instagram*, e Scarlett Johansson, que manifestou publicamente frustração diante da facilidade de circulação desse conteúdo, ilustram como essas ferramentas podem se tornar instrumentos de controle e disciplinamento de corpos femininos, independentemente de notoriedade midiática [9][21].

Dessa forma, esses casos demonstram o atual paradigma onde a combinação de ferramentas avançadas de sintetização visual, plataformas digitais de grande alcance e uma falha na detecção e na regulamentação legal de conteúdos imagéticos gerados por IA afeta as mulheres, que se tornam vítimas de violência digital sem possibilidade de obtenção de justiça, proteção e reparação adequadas.

Seguramente, a opacidade dos sistemas de geração de imagem reforça esse problema: a falta de rastreabilidade e a dificuldade de auditoria ampliam a margem para abusos, dificultando a atribuição de responsabilidade [9]. Assim, é perceptível que as lacunas no âmbito jurídico e normativo quanto à responsabilização civil, propriedade intelectual de conteúdo sintético e governança algorítmica, reforçam a urgência de uma estrutura legal mais robusta [4][12].

Essas fragilidades, ao mesmo tempo, tornam-se ainda mais evidentes quando as vítimas não possuem visibilidade pública para impulsionar denúncias. Nesse sentido, o caso da australiana Hannah Grundy exemplifica essa realidade: em 2022, Andrew Hayler, ex-colega de trabalho, administrou um site denominado "The Destruction of Hannah", no qual publicou mais de 600 *deepfakes* de caráter sexual da vítima, acompanhadas de ameaças explícitas [8], mas mesmo diante da gravidade dos fatos, a polícia australiana relatou dificuldades investigativas, afirmando que crimes envolvendo IA demandam tempo e recursos extensos, o que atrasou e limitou a resposta do Estado.

2.2 Desafios Atuais e Consequências

Em suma, os episódios apresentados são uma mínima parcela da realidade, mas já demonstram como os atuais sistemas de rastreamento, identificação de autoria e controle de disseminação de *deepfakes* permanecem rudimentares. Desse modo, percebe-se que a insuficiência institucional contribui para a perpetuação de danos irreparáveis, especialmente para vítimas sem visibilidade e sem respaldo jurídico adequado [3][9][12][20].

Diante disso, compreende-se que, além de impactos sociais profundos, os *deepfakes* impõem desafios éticos e jurídicos significativos, com a ausência de regulamentação específica, somada ao ritmo acelerado do desenvolvimento tecnológico, cria-se um ambiente de lacunas legais que dificulta a responsabilização de autores e o controle do uso dessas ferramentas.

Atualmente, o Brasil tem Projetos de Lei no sentido de corrigir algumas dessas falhas em trâmite: o **PL 3821/2024**, que aguarda apreciação pelo Senado Federal, tem como intuito principal tipificar o crime de manipulação digital de imagens por inteligência artificial, e ainda, agravar a pena em casos de crimes contra mulheres e candidaturas em período eleitoral, alterando o Código Penal e a Lei das Eleições [4]. Também há o **PL 370/2024** que aguarda sanção e tem como objetivo estabelecer causa de aumento de pena no crime de violência psicológica contra a mulher quando praticado com o uso de inteligência artificial ou de qualquer outro recurso tecnológico que altere imagem ou som da vítima [5].

Assim, é compreensível que como catalisam inovações e transformações positivas, as IAs também operam como vetores de risco, exigindo respostas regulatórias consistentes, fiscalização ativa e abordagens multidisciplinares orientadas por valores humanos e éticos.

3 DETECÇÃO DE DEEPFAKES

Inicialmente, nos estágios primários do desenvolvimento das *GANs*, as imagens sintéticas apresentavam imperfeições e incoerências visuais notórias, o que facilitava a identificação. Entretanto, com o avanço contínuo dos modelos e a evolução progressiva da qualidade multimídia dos resultados, causado pelo uso massivo destas ferramentas pela população e consequente treinamento dos modelos generativos, o realismo visual das imagens aumentou drasticamente, dificultando e até mesmo impossibilitando a distinção entre um conteúdo real e um gerado artificialmente. Assim, a necessidade de aplicação de técnicas robustas de detecção emergiu paralelamente ao aperfeiçoamento dessas tecnologias.

Um estudo publicado na *Royal Society Open Science* investigou a capacidade humana de reconhecer rostos sintéticos gerados por IA, comparando indivíduos sem treinamento prévio e um grupo instruído posteriormente, designado Super-Recognizers (SRs) [14]. Os resultados revelaram grande dificuldade na identificação de rostos artificiais em ambos os cenários experimentais realizados sem treinamento inicial.

Figure 1: Exemplo de dois olhos gerados por *GANs* [13]

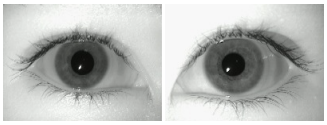

Figure 2: Exemplo de dois olhos de faces humanas reais [2]

Na Tarefa de Escolha Forçada, em que apenas uma imagem era apresentada por vez, as taxas de detecção foram baixas: 41% para SRs, 31% para o grupo controle *Prolific* e 30% para o grupo controle *Database*. No segundo experimento, com duas imagens para comparação, considerado menos suscetível a vieses decisionais, o desempenho permaneceu insatisfatório, com sensibilidade média de 0.182 para SRs, 0.317 para o grupo *Prolific* e 0.307 para o grupo *Database*. Embora os SRs tenham exibido desempenho ligeiramente superior, sua precisão não diferiu estatisticamente do acaso, enquanto os grupos de controle ficaram abaixo dele, indicando forte vulnerabilidade ao hiper-realismo das imagens.

3.1 Métodos Computacionais de Detecção

No entanto, a automatização da detecção pode superar as limitações perceptivas humanas, como mostra o estudo *"Eyes Tell All"*. A análise apresenta evidências de que *GANs* não preservam adequadamente padrões fisiológicos oculares, resultando em assimetrias e falhas estruturais. Com base nisso, a pesquisa propôs um método de identificação fundamentado na análise pupilar, que alcançou um desempenho robusto: AUC = 0.91, na distinção entre rostos reais e sintéticos [15].

Apesar de *CNNs* e outros classificadores obterem bons resultados, muitos enfrentam problemas de generalização e interpretabilidade. Desse modo, com o objetivo de mitigar essas limitações, a pesquisa *"Robust Attentive Deep Neural Network for Detecting GAN-Generated Faces"* [16] introduziu uma arquitetura neural capaz de localizar automaticamente inconsistências oculares entre pares de imagens. O modelo utiliza uma função de perda combinada, entropia cruzada somada a um relaxamento *ROC-AUC* pela estatística *WMW*, permitindo aprendizado eficiente mesmo em bases desbalanceadas e demonstrando desempenho superior em cenários adversariais.

4 CONSIDERAÇÕES FINAIS

Conforme exposto previamente, a disseminação de deepfakes representa um desafio substancial para os pilares da democracia, da verdade e da dignidade humana [10]. Embora o ordenamento jurídico brasileiro disponha de garantias constitucionais e civis relativas à imagem, à privacidade e à honra, o avanço acelerado das tecnologias de síntese visual tem superado a capacidade de resposta das estruturas legais tradicionais. A velocidade de criação e compartilhamento desses conteúdos torna insuficiente a aplicação isolada dos mecanismos já existentes, exigindo o desenvolvimento e a utilização de ferramentas mais ágeis e específicas.

Nesse cenário, torna-se urgente desenvolver medidas de prevenção, suporte às vítimas e meios eficazes de rastreamento para responsabilizar os autores. A LGPD reforça essa necessidade ao classificar imagem como dado pessoal sensível e estabelecer limites rígidos para tratamento ilícito [7]. No entanto, a velocidade de compartilhamento e disseminação de informações desafia o modelo reativo previsto pelo Marco Civil da Internet [6], evidenciando um déficit regulatório que contrasta com o ritmo de evolução tecnológica [11].

Concomitantemente, a pesquisa científica oferece caminhos promissores. Segundo Guo (2021), *GANs* podem falhar em reproduzir padrões fisiológicos com perfeição, o que abre espaço para métodos de detecção capazes de identificar imagens manipuladas enquanto marcos regulatórios mais robustos não se consolidam [16]. Assim, governança, tecnologia e educação digital devem atuar de forma conjunta: atualização da legislação, implementação de ferramentas automáticas de detecção de conteúdo manipulado artificialmente e conscientização da população acerca de crimes cibernéticos.

Finalmente, a disparidade entre a velocidade de geração de *deepfakes* e a lentidão da obtenção de respostas jurídicas demonstra que o desafio é contínuo e abrange dar suporte e proteção jurídica às vítimas, proteger a legitimidade das instituições democráticas e garantir, neste âmbito, o direito à dignidade humana.

REFERENCES

[1] Apelação cível 70083317115, 2019. Reconhece dano moral por deepfake.
[2] Ahmed. c-kks2r dataset. Roboflow Universe, 2024.
[3] Emily et al. Bender. On the dangers of stochastic parrots. *FAccT*, pages 610–623, 2021.
[4] Rishi et al. Bommasani. On the opportunities and risks of foundation models, 2021. arXiv:2108.07258.
[5] Brasil. Código penal brasileiro, 1940.
[6] Brasil. Marco civil da internet, 2014.
[7] Brasil. Lei geral de proteção de dados, 2018.
[8] BBC Brasil. Transformação de fotos em pornografia com ia, 2025.
[9] Kate Crawford. *Atlas of AI*. Yale University Press, 2021.
[10] Amanda Passos Ferreira and Carolina da Silva Leme. O fenômeno da deep fake no contexto eleitoral. *Boletim IBCCRIM*, 31(363):21–23, 2023.
[11] Luciano Floridi. *The Fourth Revolution*. Oxford University Press, 2018.
[12] Luciano et al. Floridi. Ai4people—an ethical framework for a good ai society. *Minds and Machines*, 28(4):689–707, 2018.
[13] Generated Photos. This person does not exist, 2019. Acessado em 24/07/2025.
[14] Katie et al. Gray. Training super-recognizers for ai face detection. *R. Soc. Open Sci.*, 2025.
[15] Hui et al. Guo. Eyes tell all. *arXiv:2109.00162*, 2021.
[16] Hui et al. Guo. Robust attentive neural network for gan-detection. *IEEE Access*, 10:32574–32583, 2022.
[17] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.
[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for gans. In *CVPR*, pages 4401–4410, 2019.
[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
[20] ONU Brasil. Ods 5 — igualdade de gênero.
[21] Globo Staff. Taylor swift e outras celebridades vítimas de deepfake, 2025.
[22] Daniel Story and Ryan Jenkins. Deepfake pornography and the ethics. *Philosophy Technology*, 36:1–22, 2023.
[23] Mustafa Kaan Tuysuz and Ahmet Kılıç. Legal and ethical considerations of deepfakes. *ISSLP*, 2(2):4–10, 2023.
[24] Agnes Venema. Deepfakes as a security issue: Why gender matters, 2025.