

Lauro Accioly

ENTRE A PREVENÇÃO E A DETECÇÃO

¹As deepfakes são compreendidas como conteúdos sintéticos, gerados por aplicações de inteligência artificial generativa, e agora popularizados entre muitos consumidores.² Em uma abordagem mais técnica, as deepfakes são materiais criados por Deep Learning (DL), geralmente se refere a materiais criados por uma *Deep Neural Network* (DNN), um subconjunto de aprendizado de máquina. A rede neural mais conhecida é a *Generative Adversarial Networks* (GANs), cujos sistemas operam com base na interação entre duas redes: o gerador, que cria conteúdos sintéticos – como vídeos ou imagens –, e o discriminador, que avalia esses conteúdos e tenta identificar falhas nas criações do gerador para fornecer feedback ao aperfeiçoamento contínuo do processo. A dinâmica de competição interativa entre as redes resulta em mídias cada vez mais realísticas e difíceis de detectar como falsas.³

Deepfakes para além da desinformação política

No atual cenário, os impactos das deepfakes transbordam problemas do nicho desinformativo-político e assumem, por exemplo, papel de ferramentas de golpes financeiros em tempo real. Em um caso concreto, deepfakes foram usadas em videoconferência com o diretor financeiro e executivos de uma empresa, induzindo um funcionário a transferir cerca de US\$ 25 milhões a contas fraudulentas.⁴ O uso antiético e ilícito representa um risco crescente de fraudes no ambiente corporativo, intensificado por softwares de geração de voz combinados a outros esquemas de engenharia social, já que vítimas tendem a confiar em vozes e endereços eletrônicos familiares. Estudos indicam perdas significativas nos golpes citados, variando de U\$ 243 mil a U\$ 35 milhões.⁵

¹ Doutorando no Programa de Pós-Graduação em Relações Internacionais interinstitucional San Tiago Dantas (Unesp/Unicamp/Puc-SP), mestre em Relações Internacionais na Universidade Estadual da Paraíba com período sanduíche na American University (Washington DC). Técnico em Redes de Computadores pela Escola Técnica do Estado de Pernambuco (ETE-Gravatá).

² ACCIOLY FILHO, Lauro. "O que sabemos do impacto das Deepfakes?", *Le Monde Diplomatique Brasil*, 04 Abr. 2025, disponível em <https://diplomatique.org.br/o-que-sabemos-do-impacto-das-deepfakes/>

³ WANG, Yuchao *et al.* Multi-dimensional prediction method based on Bi-LSTMC for ship roll. *Ocean Engineering*, v. 242, p. 110106, 2021; ZHONG, Li *et al.* Bridging the theoretical bound and deep algorithms for open set domain adaptation. *IEEE transactions on neural networks and learning systems*, v. 34, n. 8, p. 3859-3873, 2021.

⁴ Maiores detalhes do caso podem ser acessados em: <https://www.cnnbrasil.com.br/economia/negocios/golpistas-usam-deepfake-de-diretor-financeiro-e-roubam-us-25-milhoes/>.

⁵ DE RANCOURT-RAYMOND, Audrey; SMAILL, Nadia. The unethical use of deepfakes. *Journal of Financial Crime*, v. 30, n. 4, p. 1066-1077, 2023.

A velocidade e o alcance das mídias sociais ampliam esses riscos. No Brasil, deepfakes de personalidades como Marcos Mion foram usadas para divulgar falsas promoções em redes sociais, levando vítimas a sites fraudulentos para pagamentos via Pix. O caso gerou prejuízos a um número considerável de pessoas e reforça a multidimensionalidade das deepfakes.⁶

Figura 01. Deepfake do apresentador Marcos Mion



Imagem 1. Captura de tela do site da CNN Brasil.

Surge, assim, a necessidade de soluções jurídicas para proteger o direito à imagem diante de conteúdos sintéticos utilizados para obter vantagens ilícitas. Parte dos golpes se fundamenta na familiaridade de figuras públicas, mais suscetíveis a terem suas imagens ilegalmente exploradas, dado o nível de exposição e confiança depositado em rostos e vozes reconhecíveis. Deste modo, é preciso fortalecer os

⁶ Maiores detalhes do caso podem ser acessados em: <https://www.cnnbrasil.com.br/nacional/centro-oeste/go/integrante-de-grupo-que-fazia-deepfake-de-marcos-mion-e-presos-em-goias/>.

mecanismos de proteção contra exploração comercial não autorizada, especialmente quando deepfakes são usadas em publicidade ou endossos sem consentimento, envolvendo questões significativas de propriedade intelectual e uso indevido de imagem.⁷

Diante dessas múltiplas dimensões, a ilustração abaixo fornece uma melhor compreensão da necessidade de uma atuação coordenada para mitigar essas externalidades negativas:

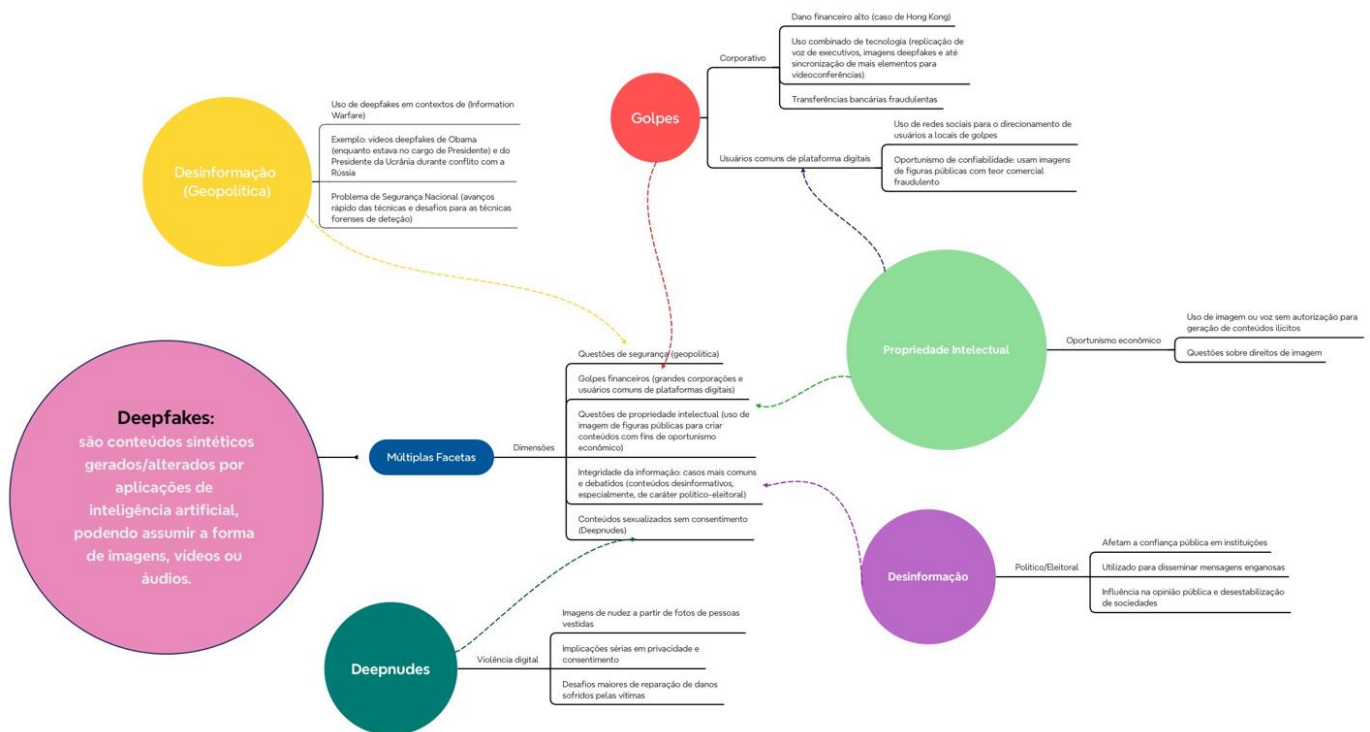


Figura 2. Mapa Mental das múltiplas facetas das Deepfakes, elaborado próprio autor no Xmind AI.

A ilustração permite observar como a questão das deepfakes está inserida numa miríade de questões complexas que requerem mais do que ações pontuais. Todavia, perante os abusos associados ao uso de deepfakes, diferentes mecanismos estão sendo desenvolvidos para autenticar e detectar conteúdos manipulados por aplicações de inteligência artificial. Essas abordagens operam em momentos diferentes do processo de credibilizar o conteúdo digital e com finalidades distintas.

Figura 03. Diferenças das medidas de autenticar e detectar deepfakes

⁷ ARNWINE, Danielle A. When Deepfakes Make Celebrities a Dime a Dozen Can the Right of Publicity Save Their Worth?. **Journal of Technology Law & Policy**, v. 29, n. 1, p. 5, 2025.

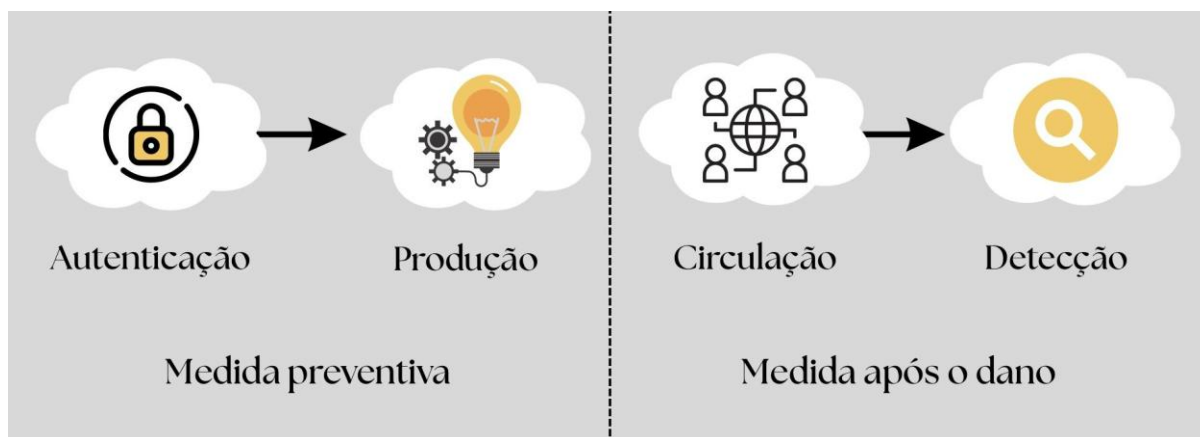


Figura 3. Elaborado pelo próprio autor no Canva

A **detecção** refere-se à identificação de conteúdos manipulados ou sintéticos que já estão em circulação. É um processo que depende de técnicas que analisam o conteúdo em busca de inconsistências visuais, temporais ou estatísticas — muitas vezes imperceptíveis ao olho humano, mas identificáveis por meio de redes neurais convolucionais (CNNs) ou outros algoritmos de aprendizado de máquina.⁸

Por outro lado, **autenticação** é um processo proativo de validação de conteúdos digitais, consiste em incorporar informações verificáveis — como marcas d'água digitais ou assinaturas criptográficas, permitindo uma verificação robusta e imediata de sua autenticidade, funcionando como uma camada preventiva contra manipulações e falsificações.⁹

As propostas de autenticação em evidência são: marcas d'água digitais (*watermarking*) e o uso da tecnologia blockchain.

O *watermarking* é apresentado por alguns como um mecanismo de etiquetagem anticlonagem, tornando possível verificar se a imagem marcada foi adulterada com base na presença ou ausência da etiqueta. O processo ocorre pela inserção de uma mensagem secreta criptografada (marca d'água) diretamente nos pixels da imagem, utilizando uma arquitetura de rede neural do tipo codificador-decodificador. Ela é imperceptível a olho nu e foi desenhada para ser resistente a modificações benignas — como compressão, redimensionamento, desfoque ou aplicação de filtros — mas sensível a alterações maliciosas, como manipulações

⁸ DAGAR, Deepak; VISHWAKARMA, Dinesh Kumar. A literature review and perspectives in deepfakes: generation, detection, and applications. **International journal of multimedia information retrieval**, v. 11, n. 3, p. 219-289, 2022.

⁹ AMERINI, Irene *et al.* Deepfake media forensics: Status and future challenges. **Journal of Imaging**, v. 11, n. 3, p. 73, 2025.

faciais baseadas em técnicas de troca de identidade, alteração de atributos ou *reenactment* — feitas com modelos GAN (ex: *StarGAN*, *AttGAN*), técnicas de *FaceSwap* ou modelos de difusão como o *Stable Diffusion*. Isso configura a natureza semi-frágil da watermarking: robusta o suficiente para sobreviver a edições triviais, mas vulnerável a distorções profundas associadas a deepfakes.¹⁰

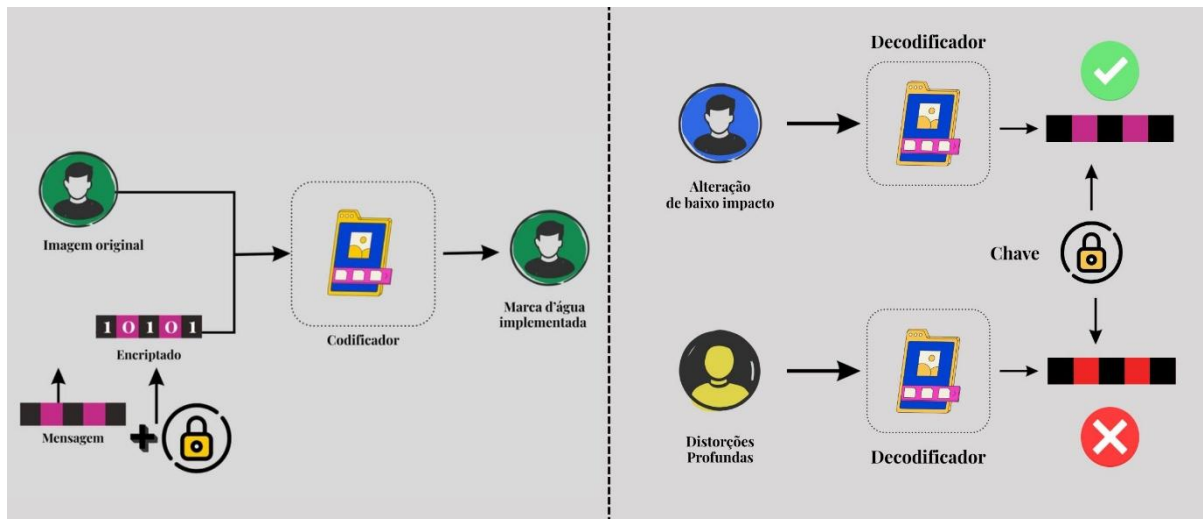


Figura 04. Ilustração simplificada do processo de watermarking, elaborado pelo autor.

Nesse cenário, mecanismos de autenticação, como a *watermarking*, não substituem a detecção, mas oferecem uma camada adicional de verificação ao identificar alterações diretamente na origem do conteúdo.

Por outro lado, a tecnologia blockchain, de forma simplificada, estaria atrelada a registrar vídeos, áudios e imagens quando são criados, por meio de contratos inteligentes (smart contracts) e com armazenamento descentralizado. O registro inclui informações como data, hora, local, dispositivo utilizado e identidade digital do criador — tudo isso criptografado e vinculado a um código único (hash) que representa aquele conteúdo específico. No entanto, a abordagem enfrenta limitações de altos custos operacionais e a necessidade de um ecossistema global padronizado, com a adesão de plataformas, governos, produtoras de conteúdo e usuários.¹¹

¹⁰ ZHAO, Yuan *et al.* Proactive deepfake defence via identity watermarking. In: **Proceedings of the IEEE/CVF winter conference on applications of computer vision**. 2023. p. 4602-4611; NADIMPALLI, Aakash Varma; RATTANI, Ajita. Social media authentication and combating deepfakes using semi-fragile invisible image watermarking. **Digital Threats: Research and Practice**, v. 5, n. 4, p. 1-30, 2024.

¹¹ DOUGHMAN, Scott. **How blockchain data storage can protect us from deepfakes**. World Economic Forum Website, Sep 22, 2023. Disponível em: <https://www.weforum.org/stories/2023/09/how-blockchain-can-protect-us-again-ai-threats/>; HEIDARI, Arash *et al.* A novel blockchain-based deepfake detection method using federated and deep learning models. **Cognitive Computation**, v. 16, n. 3, p. 1073-1091, 2024.

Quanto aos mecanismos de detecção, muitos estão progredindo com uso de *Deep Learning* por redes neurais (CNN) que recebem imagens e vídeos — camada de entrada — responsável por receber os dados, convertendo cada pixel em valores numéricos que representam cor e intensidade. Durante o treinamento — camada oculta — a rede vai errando e corrigindo até aprender como identificar esses padrões com precisão. Após treinada, a rede neural pode receber um vídeo ou imagem e “quebrar” esse material em informações matemáticas, analisando pixel por pixel ou quadro por quadro (*frame to frame*), comparando com o que aprendeu para decidir se aquilo é real ou falso — camada de saída. Em vídeos, além de olhar cada quadro, a rede também verifica se há inconsistências no movimento entre os quadros, como movimentos labiais que não combinam com o áudio ou expressões faciais que mudam de forma artificial. Já em áudios, a rede analisa padrões de frequência e pausas na fala para perceber se a voz foi gerada por inteligência artificial.¹²

No entanto, esse tipo de mecanismo requer revisões que mitiguem **falsos positivos** (rotular vídeos reais como deepfakes) e **falsos negativos** (não identificar deepfakes em certos grupos). Especialmente, por essa questão se tratar de uma problemática de viés no desempenho da detecção de conteúdos deepfakes, em grande maioria, por causa de atributos como, faces de pessoas de pele mais escura, mulheres, idosos ou pessoas usando óculos e acessórios. Nestes casos, as taxas de erro foram consideravelmente maiores em comparação com outros grupos, visto que os bancos de dados usados para treinar os detectores são, em sua maioria, desbalanceados, concentrando amostras em grupos específicos, deixando outros grupos sub-representados. Com isso, os detectores aprendem padrões predominantes no conjunto de dados e falham ao encontrar padrões em grupos com menor representatividade.¹³

Medidas multissetoriais: de intervenção a mitigação dos danos

O avanço técnico por si só não é suficiente, intervenção multissetorial aponta um caminho mais profícuo, podendo mobilizar diversos atores na mitigação dos danos

¹²KAUR, Achhardeep et al. Deepfake video detection: challenges and opportunities. **Artificial Intelligence Review**, v. 57, n. 6, p. 159, 2024.

¹³XU, Ying et al. Analyzing fairness in deepfake detection with massively annotated databases. **IEEE Transactions on Technology and Society**, v. 5, n. 1, p. 93-106, 2024; PANDEY, Mayank et al. Detecting low-resolution deepfakes: an exploration of machine learning techniques. **Multimedia Tools and Applications**, v. 83, n. 25, p. 66283-66298, 2024.

gerados pelas externalidades negativas das deepfakes. O seu carácter transnacional demonstra como instrumentos legais isoladas não bastam para combater o uso ilícito de deepfakes. Um exemplo a não seguir é o *Take It Down Act* nos Estados Unidos, visto que isola aplicação de penalidades pra externalidade do uso indevido de deepfakes apenas para a chamada *deepnudes*, sem sequer conceituar os casos em que se aplica a remoção, possibilitando o potencial de denúncias abusivas devido sua ambiguidade conceitual.¹⁴As medidas precisam ser claras, tipificando usos ilícitos e estabelecendo mecanismos de reparação de danos, considerando que as deepfakes abrangem fraude financeira, chantagem, roubo de identidade e extorsão.

Em vista disso, os mecanismos de remoção de conteúdos deepfakes precisam equilibrar sinalização obrigatória de conteúdos manipulados e a remoção ágil em casos de violação extrema, como *deepnudes*, evitando medidas de censura abruptadas ou automatizadas que tragam mais problemas do que soluções. É fundamental também avançar em cooperação internacional para tentar padronizar tais medidas, assim como aplicar protocolos éticos no processo de detecção de deepfakes é indispensável, especialmente considerando as limitações técnicas e humanas na identificação desses conteúdos.¹⁵

Estudos destacam a literacia digital e as agências de fact-checking como ferramentas essenciais para um ambiente informacional seguro. Contudo, sua eficácia depende de confiança e engajamento social. Assim, é necessário promover a literacia digital não apenas como ferramenta de empoderamento, mas também como meio de conscientização sobre golpes financeiros e outros ilícitos envolvendo deepfakes.¹⁶

A conscientização social sobre o funcionamento, os riscos e os usos das deepfakes é vital para fortalecer a confiança e o interesse em aprimorar habilidades digitais. Para isso, é fundamental apresentar o tema de forma acessível, incentivando

¹⁴ACCIOLY FILHO, Lauro. "The Take It Down Act shows the fragmentation of policies addressing deepfakes and generative AI", **LSE United States Politics and Policy**, 10 Jul. 2025, disponível em <https://blogs.lse.ac.uk/usappblog/2025/07/10/the-take-it-down-act-shows-the-fragmentation-of-policies-addressing-deepfakes-and-generative-ai/>

¹⁵GHEDIRI, Karima. Countering the negative impacts of deepfake technology: Approaches for effective combat. **International journal of economic perspectives**, v. 18, n. 12, p. 2871-2890, 2024.

¹⁶MATLI, Walter. Extending the theory of information poverty to deepfake technology. **International Journal of Information Management Data Insights**, v. 4, n. 2, p. 100286, 2024; SULTANBAYEVA, Gulmira et al. Digital Literacy as a Tool for Identifying Fake News: A Comparative Analysis Using the Example of European and Kazakh Media. **Journal of Information Policy**, v. 15, 2024.

o aprendizado de medidas de precaução frente ao seu uso ilícito. No ambiente corporativo, a literacia digital também pode prevenir fraudes, sendo crucial capacitar funcionários sobre deepfakes e estimular uma cultura de verificação. Essa formação deve ser acompanhada de protocolos rigorosos de comunicados e transações financeiras.¹⁷

¹⁷LAKHERA, Girish et al. (Ed.). **Navigating the World of Deepfake Technology**. IGI Global, 2024.