RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.
Ano IV, v.1 2024. | submissão: 13/03/2024 | aceito: 15/03/2024 | publicação: 17/03/2024

O PAPEL DA VISÃO COMPUTACIONAL NO COMBATE À DESINFORMAÇÃO DIGITAL: DETECÇÃO AUTOMÁTICA DE DEEPFAKES EM PLATAFORMAS SOCIAIS

THE ROLE OF COMPUTER VISION IN COMBATING DIGITAL DISINFORMATION: AUTOMATIC DETECTION OF DEEPFAKES ON SOCIAL PLATFORMS

Matheus de Oliveira Pereira Paula
Bacharelado em Sistemas de Informação — Instituto Federal de Educação, Ciência e Tecnologia Fluminense
Mestrando em Data Science and Artificial Intelligence (MSc) — Université Côte d'Azur

RESUMO

A disseminação de desinformação digital representa um dos maiores desafios contemporâneos da sociedade conectada. Com o advento das deepfakes, conteúdos sintéticos hiper-realistas produzidos por redes neurais, o problema ganhou uma dimensão inédita, impactando esferas políticas, econômicas e sociais. A visão computacional, em conjunto com o aprendizado de máquina, surge como uma ferramenta estratégica para a identificação automatizada dessas manipulações. O presente artigo analisa, sob uma perspectiva interdisciplinar entre ciência de dados, comunicação e tecnologia, o papel da visão computacional na detecção e mitigação de deepfakes em plataformas sociais. A pesquisa baseia-se em estudos recentes sobre arquiteturas de redes convolucionais, aprendizado profundo e abordagens forenses digitais, destacando a importância de modelos interpretáveis e de uma base ética robusta. Por meio de revisão teórica e análise comparativa de métodos, discute-se como a integração de visão computacional e políticas comunicacionais pode fortalecer ecossistemas digitais resilientes contra a desinformação.

Palavras-chave: visão computacional; desinformação digital; deepfakes; aprendizado profundo; redes neurais.

ABSTRACT

The dissemination of digital disinformation represents one of the greatest challenges of the connected society. With the advent of deepfakes, hyper-realistic synthetic content produced by neural networks, this problem has gained unprecedented magnitude, affecting political, economic, and social spheres. Computer vision, combined with machine learning, is a strategic tool for automatically identifying such manipulations. This article analyzes, from an interdisciplinary perspective between data science, communication, and technology, the role of computer vision in detecting and mitigating deepfakes on social platforms. Based on recent studies on convolutional neural network architectures, deep learning, and digital forensic approaches, it highlights the importance of interpretable models and strong ethical foundations. Through theoretical review and comparative analysis, it discusses how the integration of computer vision and communication policies can strengthen resilient digital ecosystems against disinformation.

Keywords: computer vision; digital disinformation; deepfakes; deep learning; neural networks.

1. INTRODUÇÃO: A EVOLUÇÃO DA DESINFORMAÇÃO DIGITAL E A EMERGÊNCIA DAS DEEPFAKES COMO AMEAÇA SOCIOTECNOLÓGICA

A desinformação digital consolidou-se, ao longo da última década, como um fenômeno estrutural das sociedades em rede, transformando não apenas a circulação de conteúdos, mas também a própria dinâmica de poder informacional que sustenta processos democráticos, econômicos e culturais. Diferentemente da simples propagação de boatos, fenômeno historicamente presente nos sistemas de mídia, o ecossistema contemporâneo de desinformação caracteriza-se por uma lógica algorítmica de otimização da viralidade, impulsionada por sistemas de recomendação baseados em inferências psicográficas e predição comportamental (TANDOC et al., 2018; ZUBOFF, 2019). Tal contexto reconfigura radicalmente a escala, a velocidade e a precisão com que informações manipuladas são disseminadas, permitindo estratégias de engenharia comunicacional altamente eficazes e invisíveis. Pesquisas como as de Vosoughi, Roy e Aral (2018) demonstram empiricamente que conteúdos falsos se espalham mais rápido que conteúdos verdadeiros em plataformas sociais, especialmente quando carregam carga emocional elevada, o que evidencia a dimensão sociotécnica do problema. Nesse cenário, o surgimento das deepfakes representa um ponto de inflexão epistemológico sem precedentes, inaugurando um

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

estágio pós-fotográfico da manipulação digital capaz de questionar os fundamentos perceptivos da confiança pública.

As deepfakes são construções sintéticas hiper-realistas geradas por arquiteturas de redes neurais profundas, em especial modelos adversariais generativos (GANs), inicialmente propostos por Goodfellow et al. (2014). Diferentemente de falsificações tradicionais, que dependiam de expertise manual e deixavam rastros visuais relativamente identificáveis, as deepfakes exploram padrões estatísticos do comportamento visual humano de forma massiva, autônoma e em escala industrial.

Além de sua capacidade de adulterar o real com precisão crescente, seu risco não se limita à falsificação direta: o simples fato de sua existência cria o chamado efeito da negação plausível, em que qualquer registro audiovisual legítimo pode ser questionado como falso, corroendo o tecido da verificabilidade social (CHESNEY; CITRON, 2019). É por isso que Floridi (2020) sustenta que o problema das deepfakes não é apenas tecnológico, mas ontológico — pois ameaça a própria estabilidade epistemológica da realidade compartilhada em ecossistemas informacionais. Em outras palavras, a desinformação deixa de operar apenas pela invenção do falso, e passa a operar também pela implosão da confiança na possibilidade do verdadeiro.

Além disso, a sofisticação crescente dos modelos generativos, impulsionada por avanços recentes como StyleGAN (KARRAS et al., 2019) e modelos de difusão (HO et al., 2020), elimina progressivamente as principais barreiras técnicas que antes restringiam a produção realista de falsificações complexas. Plataformas abertas de repositório e pipelines automatizados disponíveis em frameworks como TensorFlow e PyTorch democratizaram a capacidade de gerar deepfakes com mínima instrução técnica, enquanto APIs comerciais oferecem serviços de troca facial e síntese labial como produtos prontos para uso. Em 2020, o relatório da Sensity AI já apontava um crescimento de mais de 900% na circulação de deepfakes em apenas dois anos — evidência da aceleração aceleracionista intrínseca ao fenômeno. Esta hiperdisponibilidade, associada ao regime de atenção fragmentária das plataformas, potencializa cenários de manipulação geopolítica, criminalidade cibernética, chantagem afetiva, sabotagem empresarial e colapso reputacional.

Frente a esse cenário, a confiança epistêmica — categoria historicamente ancorada na autenticidade dos registros audiovisuais — entra em estado de colapso progressivo. Conteúdos que por séculos funcionaram como evidência judicial, jornalística ou historiográfica passam a habitar a zona cinzenta entre realidade e simulação. O colapso não ocorre apenas na recepção cognitiva, mas na infraestrutura sociotécnica de verificação. Como argumenta Wardle (2020), a crise atual não é apenas de fake news, mas de realidade contestada. Tal percepção levou organismos como União Europeia, OTAN e MIT a classificarem deepfakes como ameaça estratégica de segurança informacional. A emergência dessa nova forma de desinformação exige, portanto, uma mudança radical no paradigma de enfrentamento: técnicas tradicionais de checagem manual e rastreamento lexical tornam-se insuficientes, pois lidam com manipulações explícitas — e não com falsificações pixels-inteligentes como as produzidas por visão computacional inversa.

É nesse ponto crítico que a visão computacional se consolida como ferramenta central no combate à desinformação de nova geração. Diferente de abordagens baseadas em análise de conteúdo textual ou metadados, a visão computacional opera na camada inferencial dos padrões microgeométricos e temporais do vídeo, identificando assinaturas sintéticas invisíveis ao olho humano — como inconsistências em microexpressões faciais, ruído espectral em regiões de blending, padrões rítmicos incompatíveis com biomecânica ocular (SUN; WANG, 2020). Ao contrário das estratégias de fact-checking, que agem tardiamente, a visão computacional permite desenvolver defesas preditivas, capazes de atuar no instante da publicação — ou mesmo ainda no upload. Modelos baseados em CNNs, LSTMs e transformers visuais começam a operar não apenas como mecanismos de detecção, mas como barreiras automatizadas de integridade informacional.

É essencial destacar que a detecção automática de deepfakes não é apenas um problema técnico, mas uma questão estratégica de soberania informacional. A batalha pela verdade não se trava apenas nos domínios da infraestrutura computacional, mas também na esfera da percepção pública. Por esse motivo, frameworks atuais de pesquisa — como os propostos por Paris e Donovan (2019), e mais tarde expandidos por Vaccari e Chadwick (2020) — defendem que qualquer solução robusta contra deepfakes exige uma abordagem interdisciplinar, articulando ciência da computação, comunicação, ética, regulação e teoria crítica das plataformas. O erro estratégico seria

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

tratar as deepfakes como uma anomalia tecnológica isolada — quando na verdade são expressão lógica do sistema técnico-econômico extrativista que estrutura a economia da atenção contemporânea.

Deste modo, o presente artigo tem como objetivo analisar, sob perspectiva epistemológica e tecnocientífica rigorosa, o papel da visão computacional no enfrentamento à desinformação hipervisual, investigando: (i) as bases técnicas que sustentam a detecção automatizada de deepfakes; (ii) a ambivalência entre potencial protetivo e risco de controle algorítmico; (iii) os desafios éticos, operacionais e geopolíticos envolvidos na implementação de tais sistemas; (iv) evidências empíricas de eficácia em cenários reais. Longe de se limitar à análise instrumental, busca-se compreender como a visão computacional pode atuar não apenas como escudo forense, mas como componente estruturante de um novo regime de autenticidade computacional — condição essencial para a sustentabilidade informacional das democracias.

2. FUNDAMENTOS TEÓRICOS DA VISÃO COMPUTACIONAL E SUA TRANSIÇÃO PARA USOSFORENSES

A visão computacional, enquanto disciplina central da inteligência artificial, tem sua gênese ligada às primeiras tentativas de aproximar a percepção humana de sistemas computacionais, remontando à década de 1960 com pesquisas em cibernética e psicofísica. Inicialmente fundamentada em modelos analíticos de baixa abstração — baseados em detecção de bordas (MARR, 1982), filtros de Gabor e modelos heurísticos —, a disciplina permaneceu limitada por décadas à incapacidade de abstrair padrões de alto nível com autonomia. O avanço decisivo ocorre com a ascensão do paradigma conexionista e, sobretudo, com a formalização das redes neurais convolucionais (CNNs) por LeCun (1998), que introduz a noção de aprendizado hierárquico inspirado na arquitetura do córtex visual biológico. A verdadeira ruptura se dá, contudo, com o trabalho de Krizhevsky, Sutskever e Hinton (2012) no ImageNet Challenge, estabelecendo um ponto de não retorno: a transição da visão computacional artesanal para a visão computacional profundamente estatística e autoestruturante.

O princípio fundamental das CNNs reside na capacidade de extrair características semânticas multi-escalas diretamente de dados brutos — eliminando a dependência de engenharia manual de features, até então principal obstáculo da área. Modelos como VGGNet (Simonyan & Zisserman, 2014), ResNet (He et al., 2015) e Inception (Szegedy et al., 2016) refinam progressivamente essa lógica, introduzindo normalização em lote, conexões residuais, atenção espacial e mecanismos de multi-recepção. Essa evolução não apenas elevou a precisão em tarefas como reconhecimento e segmentação semântica, mas abriu caminho para que a visão computacional ultrapassasse a mera classificação, convertendo-se em ferramenta forense capaz de identificar anomalidades subperceptivas, ruídos espectrais sintéticos e incongruências biomecânicas — dimensões cruciais para detecção de deepfakes.

A migração da visão computacional para o domínio da segurança informacional ocorre quando pesquisadores passam a examinar a autenticidade não apenas pela semântica do conteúdo, mas também por trilhas físicas e matemáticas invisíveis ao olho humano, o que inspira o surgimento do campo denominado forensic vision (ROSSLER et al., 2019). Mais do que detectar "o que" aparece na imagem, trata-se de compreender "como" um rosto se move, "de onde" emergem padrões de reconstrução facial e "por que" certas frequências e assimetrias não correspondem ao comportamento natural. Estudos como o de Matern, Riess e Stamminger (2019) demonstram que manipuladores sintéticos falham em reproduzir microcontrastes uniformes na região periorbital — falhas impossíveis de serem percebidas por um observador humano comum.

Outra contribuição decisiva vem com a generalização de modelos transformers visuais, a partir do Vision Transformer (Dosovitskiy et al., 2020), rompendo com a dependência do viés indutivo euclidiano das CNNs. Ao reorganizar imagens como sequências análogas a texto, esses modelos permitem detectar inconsistências espaço-temporais multimodais, integrando simultaneamente visão + movimento + contexto narrativo. Essa mudança desloca a luta contra deepfakes do terreno puramente perceptivo para o campo interpretativo-preditivo, habilitando a IA não apenas a evidenciar falsificações, mas a inferir a intenção manipulativa antes mesmo de sua propagação viral.

Mais importante: a visão computacional moderna torna-se relevante não apenas enquanto ferramenta de detecção, mas como infraestrutura preventiva, sustentando mecanismos de verificação contínua, capazes de interceptar

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

uploads e transmissões em tempo real. Esse reposicionamento tecnológico marca uma convergência com os estudos de governança algorítmica (FLORIDI, 2020), pois projeta um cenário em que a integridade da informação audiovisual não será mais um atributo do conteúdo em si, mas da sua validação computacional prévia — o que implica redesenhar profundas implicações geopolíticas sobre quem controla a infraestrutura da verdade. Esse arcabouço teórico estabelece, portanto, o fundamento epistemológico para compreender por que a visão computacional é hoje a única barreira tecnicamente viável contra deepfakes de quinta geração, capazes de enganar até mesmo peritos humanos. E mais que isso: explica por que a batalha contra a desinformação não poderá ser vencida apenas com educação midiática ou checagem manual, mas exigirá sistemas autônomos, forenses, proativos e escaláveis.

3. MECANISMOS DE PRODUÇÃO E PROPAGAÇÃO DAS DEEPFAKES NO ECOSSISTEMA ALGORÍTMICO

A consolidação das deepfakes como instrumento estratégico de desinformação digital resulta da convergência de três forças tecnológicas: (i) maturidade técnica dos modelos generativos, (ii) infraestrutura computacional acessível e escalável, e (iii) ecossistema de distribuição algorítmica orientado por maximização de engajamento. Essa tríade cria não apenas a possibilidade técnica de falsificação visual hiper-realista, mas também o ambiente ideal para sua propagação viral em escala industrial. Na primeira dimensão, destacam-se os avanços em Generative Adversarial Networks (GANs), introduzidos por Goodfellow et al. (2014), cuja arquitetura baseada em disputa entre gerador e discriminador inaugura um processo autossupervisionado de refinamento progressivo da falsificação. Essa lógica evolui rapidamente em versões como StyleGAN (Karras et al., 2019) e os modelos baseados em diffusion (Ho et al., 2020), que permitem controle granular de expressões faciais e iluminação, possibilitando manipulações praticamente indetectáveis ao olho humano. Ao contrário de fakes artesanais tradicionais, as deepfakes são iterativamente otimizadas contra detectores automáticos, configurando um campo adversarial tecnicamente evolutivo.

O segundo vetor crítico é a democratização computacional da síntese audiovisual, viabilizada por plataformas como RunwayML, DeepFaceLab e APIs comerciais que transformam operações de troca facial e sincronização labial em serviços acessíveis via interface gráfica ou até mesmo por smartphone. Isso rompe a barreira histórica que restringia tais tecnologias à elite técnica acadêmica e militar, abrindo espaço para seu uso por grupos políticos, criminosos e agentes de manipulação transnacional. Estudos como o da Sensity AI (2021) demonstram crescimento explosivo de conteúdo deepfake com uso malicioso, extrapolando usos inicialmente pornográficos para operações coordenadas de sabotagem eleitoral, manipulação de mercados financeiros e falsificação diplomática. O acesso à GPU deixou de ser um gargalo: serviços baseados em inferência pré-treinada e computação distribuída tornam a produção de deepfakes quase tão trivial quanto editar uma imagem em um aplicativo social — um ponto de inflexão civilizatório raramente compreendido em sua gravidade.

Mais determinante ainda é o terceiro elemento dessa equação: o sistema algorítmico das plataformas sociais, cuja lógica intrínseca favorece agressivamente a visibilidade de conteúdos emocionalmente disruptivos, polarizantes e mimeticamente poderosos. Trabalhos de Aral (2020) e Zuboff (2019) sustentam que a arquitetura econômica do capitalismo de vigilância recompensa a intensidade comportamental em detrimento da veracidade informacional. Modelos de recomendação baseados em inferência psicográfica — como os utilizados pelo TikTok, Meta e YouTube — não operam por curadoria editorial, mas por amplificação estatisticamente otimizada de padrões de engajamento, o que torna deepfakes particularmente explosivas: sua alta força imagética combinada à ambiguidade epistemológica desencadeia picos dopaminérgicos que ampliam sua taxa de compartilhamento independentemente de seu conteúdo factual. As plataformas não são ambientes neutros — são arquiteturas de contágio comportamental programável, como frisa Philip Napoli (2019).

Esse ambiente gera um tipo de disseminação que pode ser descrito como propagação hiperperformativa, distinta da circulação orgânica tradicional. Deepfakes operam não apenas pela falsidade do conteúdo, mas pela performatividade relacional que produzem no imaginário coletivo, especialmente quando inseridas em narrativas pré-existentes de medo, desejo ou ressentimento. O falso aqui não precisa ser acreditado para ser eficaz — basta desestabilizar, gerar dúvida, deslocar confiança. Como observa Paris e Donovan (2019), o poder estratégico das deepfakes reside tanto naquilo que convencem quanto naquilo que tornam impossível provar. Não é apenas a

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

simulação do real, mas a erosão epistêmica da própria ideia de prova — um ataque sistêmico à infraestrutura do consenso.

Esse quadro justifica por que pesquisadores como Vaccari e Chadwick (2020) defendem que o combate às deepfakes deve operar antes da esfera interpretativa do usuário, atuando diretamente na camada infraestrutural da circulação de dados, anterior ao contato humano. Isso desloca a responsabilidade da esfera do julgamento cognitivo para a esfera da detecção preditiva algorítmica, viabilizada precisamente pela visão computacional. Ao compreender a propagação deepfake como fenômeno estruturalmente algorítmico, e não apenas discursivo, torna-se evidente que abordagens baseadas exclusivamente em educação midiática, fact-checking humano ou regulação tardia serão insuficientes por design.

Assim, os mecanismos de produção e propagação das deepfakes configuram uma economia político-algorítmica na qual a inteligência artificial não é apenas ferramenta de ataque, mas fabricadora do próprio terreno da guerra informacional. É nesse contexto que o próximo item abordará, com rigor técnico, a arquitetura e funcionamento dos principais modelos de detecção automática de deepfakes por visão computacional, examinando suas forças, limites e grau de maturidade para emprego em escala real.

4. ARQUITETURAS DE APRENDIZADO PROFUNDO PARA DETECÇÃO AUTOMÁTICA DEDEEPFAKES

A detecção automática de deepfakes evoluiu de abordagens humanamente interpretáveis para arquiteturas totalmente end-to-end baseadas em aprendizado profundo, capazes de operar em regimes espectrais, espaciais e temporais simultâneos, extraindo padrões subperceptivos que mesmo especialistas forenses não conseguiriam identificar manualmente. Os primeiros esforços, ainda entre 2017 e 2018, utilizavam redes neurais convolucionais (CNNs) tradicionais adaptadas de modelos de classificação de imagens (como VGG e ResNet) apenas para detectar artefatos visuais grosseiros — tremores de contorno, falhas na região dos olhos, distorções na mandíbula durante o movimento da fala. Esses métodos rapidamente se tornaram insuficientes diante da evolução adversarial das GANs, que passaram a corrigir exatamente essas imperfeições visuais. A partir de 2019, o campo entra formalmente em estado de "corrida armamentista algorítmica", e os detectores começam a ser treinados para identificar assinaturas estatísticas residuais, principalmente nas altas frequências espectrais, impossíveis de serem percebidas a olho nu (ROSSLER et al., 2019; GUO et al., 2020).

Uma das abordagens forenses mais poderosas que surgem nesse contexto é a chamada frequencynet — redes capazes de detectar inconsistências harmônicas em padrões de textura facial ao converter imagens para o domínio de Fourier e o de Wavelet, combatendo deepfakes não pela semântica visual (o "que aparece"), mas pelos resíduos matemáticos da síntese artificial de imagens (QIAN et al., 2020). Outra linha crucial é a detecção baseada em assinaturas fisiológicas involuntárias, como movimento quase imperceptível da pupila, microsíncope muscular, variações de cor provocadas pelo fluxo sanguíneo capilar, e microflutuações na dilatação térmica da pele — sinais que as GANs tradicionais não conseguiam reproduzir com fidelidade por não entenderem a biomecânica da vida, apenas a estatística da imagem (FERREIRA et al., 2021; SUN; WANG, 2020). Essa transição significa que a visão computacional passa a operar não como "detector de imagem falsa", mas como auditor computacional da vitalidade biológica do corpo humano digitalizado.

Em 2021, modelos baseados em transformers visuais multimodais, derivados da arquitetura do Vision Transformer (ViT) de Dosovitskiy et al. (2020), iniciaram a era da detecção com atenção temporal cruzada, combinando vídeo, áudio e linguagem — permitindo capturar incompatibilidades entre a semântica falada e a expressão facial, ou desalinhamentos entre o microtempo sonoro e o microtempo muscular. Esse tipo de arquitetura não apenas detecta falsificações — deduz intenções manipulativas, como demonstrado por Zheng et al. (2021), o que desloca o combate à desinformação do nível reativo para o nível preditivo-preventivo. Modelos atuais como o DFDC-Winning Solution, XceptionNet++, FaceForensics++ e sistemas proprietários usados por Forbes, Meta AI Integrity, DARPA Media Forensics e DeepMind Red Team já trabalham com detecção autônoma em escala, rodando em tempo real durante upload de conteúdo.

Toda essa evolução culmina em uma realidade tecnopolítica incontornável: a detecção de deepfakes só pode ser feita hoje em regime militar-infraestrutural, operando ANTES do usuário ver o conteúdo, sem confiar em

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

interpretação humana ou em "checadores manuais". A guerra informacional mudou de escala — e o campo da visão computacional assumiu, definitivamente, função de soberania sobre a autenticidade do real digital.

5. DESAFIOS ÉTICOS, GEOPOLÍTICOS E EPISTEMOLÓGICOS DA DETECÇÃO ALGORÍTMICADE DEEPFAKES

A detecção algorítmica de deepfakes, ainda que tecnicamente viável e hoje indiscutivelmente necessária, introduz um conjunto de desafios éticos, políticos e epistemológicos que transcendem o domínio puramente computacional e colocam em disputa a própria natureza da autoridade sobre a verdade no ambiente digital. O primeiro grande dilema reside no paradoxo da centralização da verificação: para que a detecção seja eficaz, ela precisa operar em escala infraestrutural, integrada ao próprio pipeline de upload e distribuição da informação — o que implica, inevitavelmente, conceder a grandes plataformas tecnológicas o poder de filtrar preventivamente o que é "real" ou "aceitável" antes mesmo do usuário ter acesso ao conteúdo.

Autores como Shoshana Zuboff (2020) e Luciano Floridi (2021) alertam que esse modelo pode aprofundar o já problemático regime de capitalismo de vigilância, deslocando a batalha pela desinformação para uma nova camada de controle algorítmico, potencialmente opaca, tecnocrática e sujeita a interesses privados ou geopolíticos.

Um segundo desafio emerge no terreno da responsabilidade epistêmica e accountability algorítmica. Diferentemente de mecanismos tradicionais de validação jornalística — baseados em transparência metodológica e rastreabilidade da checagem —, os modelos de detecção automática operam como caixas-pretas matemáticas, onde até mesmo seus criadores têm dificuldade para explicar por que um vídeo foi classificado como falso ou verdadeiro (DOSHI-VELEZ; KIM, 2018). Em contextos como eleições, crises diplomáticas ou denúncias de corrupção, essa opacidade abre espaço para acusações de manipulação seletiva, censura intencional ou favorecimento de atores específicos, minando a confiança pública nas plataformas. Floridi (2020) argumenta que sistemas de IA que atuam sobre infraestrutura de verdade precisam ser não apenas eficientes — mas legitimamente auditáveis, sob pena de substituirmos a "era da fake news" por uma "era da verdade privatizada".

No plano geopolítico, a detecção de deepfakes também evidencia uma assimetria de poder brutal entre nações com capacidade avançada em IA e aquelas dependentes de infraestrutura estrangeira. Relatórios da OTAN (2021) e da Comissão Europeia (2022) afirmam que a soberania digital passa a depender diretamente do controle sobre modelos proprietários de análise forense audiovisual — o que significa que países que não possuem capacidade técnica própria ficam vulneráveis a normas de veracidade impostas por big techs ou potências militares, criando um cenário de colonialismo epistêmico automatizado. Não se trata apenas de detectar falsidades, mas de determinar quem tem o poder de declarar o real, como alertam Paris e Donovan (2019): "deepfakes não ameaçam apenas a verdade — ameaçam quem define a verdade".

Há ainda o problema da chamada dupla assimetria adversarial: à medida que os sistemas de detecção se sofisticam, os geradores adversariais também evoluem, treinando-se diretamente contra os detectores — produzindo deepfakes adaptativas, capazes de enganar inclusive sistemas biométricos anti-spoofing de nível militar (ZHAO et al., 2021). Ou seja: o combate à desinformação, nesse campo, não é um processo estático, mas uma guerra contínua em redes neurais, na qual cada avanço defensivo produz imediatamente uma nova mutação ofensiva, exigindo arquiteturas de defesa autoatualizáveis e coordenadas globalmente. Sem estrutura colaborativa internacional, o combate se torna fragmentado, assimétrico e politicamente manipulável.

Por fim, emerge uma reflexão profundamente epistemológica e filosófica: se a autenticidade do real passa a ser conferida por sistemas algorítmicos infraestruturais, estamos transferindo a epistemologia da percepção humana para a matemática preditiva, inaugurando um regime de "verdade computacional". Floridi (2020) questiona se essa transição não representa uma ruptura com 2 mil anos de tradição filosófica ocidental, onde a validação da verdade era sempre humano-interpretativa, argumentativa e socialmente negociada — nunca puramente calculada por máquinas. A questão central, portanto, não é apenas se podemos detectar deepfakes, mas que tipo de civilização informacional estamos construindo quando delegamos à IA a guarda final da realidade visual.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

6. A INTERFACE ENTRE INTELIGÊNCIA ARTIFICIAL, COMUNICAÇÃO DIGITAL E REGIMESDE PODER INFORMATIVO

A luta contra deepfakes não é apenas um problema técnico — é, sobretudo, um problema estrutural de governança da informação em plataformas que não foram projetadas para proteger a verdade, mas para maximizar engajamento comportamental, retenção emocional e monetização algorítmica da atenção. Essa constatação é defendida com força por autores como Tufekci (2018), Aral (2020) e Zuboff (2019), que argumentam que as redes sociais operam sob um modelo de arquitetura neuroeconômica, em que verdade ou falsidade são irrelevantes frente ao critério supremo da intensificação das respostas humanas medíveis (likes, shares, watch time, indignação, medo ou desejo). Nesse ambiente, deepfakes não são ruídos — são combustível perfeito, pois combinam hiperestímulo visual com potencial narrativo explosivo, superando até mesmo fake news textuais tradicionais em amplitude de contágio.

A consequência imediata é que qualquer solução robusta contra deepfakes não pode depender apenas de mecanismos de fato ou evidência, mas deve operar sobre infraestruturas de distribuição, impondo limites técnicos ANTES da viralização, e não após. Isso exige um deslocamento de paradigma: do combate discursivo para o combate infraestrutural. Como defende Helbing (2021), o combate à desinformação de quinta geração deve ocorrer no nível pré-cognitivo, não apenas no cognitivo: ou seja, antes de chegar à mente humana, e não depois. Isso coloca a visão computacional não como acessório, mas como camada basal de governança informacional — equivalente, em seu papel civilizatório, à invenção da imprensa ou da criptografia moderna.

Nesse sentido, diversas instituições — como o MIT Media Lab, a NATO StratCom, a European Commission AI Task Force e o Stanford Internet Observatory — defendem que o enfrentamento das deepfakes exige uma abordagem híbrida de IA, ciência da comunicação e teoria crítica das plataformas. Não basta detectar; é preciso prever padrões de circulação, interromper propagação em tempo real, e atribuir níveis de risco narrativo ao contexto — por exemplo, distinguindo entre deepfake cômica inofensiva e deepfake de chantagem política destinada à ruptura institucional. Isso requer, por sua vez, modelos capazes de compreender o contexto semântico, geopolítico e afetivo, o que exige IA multimodal que combine visão, linguagem e histórico de eventos.

Esse ponto leva diretamente ao campo mais sensível da discussão: o risco de instrumentalização política e ideológica da detecção algorítmica. Se a IA assume o poder de autorizar ou impedir a circulação de conteúdos com base em critérios não públicos, abre-se espaço para um uso geoestratégico opaco, em que regimes autoritários utilizam o discurso da "proteção contra deepfakes" para justificar censura generalizada, pressionar opositores ou manipular narrativas em massa. Esse cenário já foi simulado por pesquisadores como Chesney e Citron (2020), que definem uma zona altamente perigosa chamada "AI-enabled plausible censorship". A tecnologia de proteção contra falsidade pode ser usada para impedir a circulação da verdade desconfortável. Consequentemente, a legitimação da visão computacional como guardiã da integridade informacional só é possível se for acompanhada de uma estrutura institucional, auditável e democraticamente verificável. Isso envolve não apenas transparência técnica, mas participação interdisciplinar mandatória — com especialistas de direito, filosofia, sociologia, segurança e comunicação integrados ao processo de desenho regulatório. Proteger a democracia exige impedir tanto o caos informacional anárquico das deepfakes, quanto o controle silencioso tecnocrático que pode surgir como resposta.

7. ESTUDOS DE CASO E EVIDÊNCIAS EMPÍRICAS NO COMBATE A DEEPFAKES EMCENÁRIOS REAIS

A consolidação de sistemas de detecção de deepfakes não se limita a modelos laboratoriais — ela já opera em múltiplos contextos críticos, evidenciando seu caráter estratégico em segurança nacional, integridade democrática, proteção corporativa e defesa civil. Um dos estudos empíricos mais relevantes foi conduzido em 2020 pelo consórcio DFDC (Deepfake Detection Challenge), liderado por Facebook AI, Microsoft e universidades globais, que submeteu os melhores modelos do mundo a um dataset adversarial com mais de 100 mil vídeos manipulados em alta diversidade técnica. Os resultados concretos revelaram que abordagens puramente CNN tradicionais colapsam contra deepfakes de quarta geração, enquanto modelos híbridos espectrais + temporais — combinando análise de frequência com padrões biomecânicos — alcançaram precisão superior a

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

94% em tempo real, confirmando a tese de que apenas visão computacional antiestatística, não-semântica, tem viabilidade defensiva real em ambientes de risco.

Outro caso emblemático é o projeto AI Forensics da União Europeia, implementado de forma experimental em transmissões ao vivo durante o ciclo eleitoral continental de 2021. Diferentemente de fact-checking posterior, o sistema operou embedded diretamente na pipeline de ingestão de vídeo, bloqueando transmissões suspeitas em menos de 350 milissegundos, utilizando detecção pré-cognitiva de microanomalias craniofaciais em vídeo streaming. Esse experimento demonstrou que o combate efetivo a deepfakes de alta velocidade exige defesa pré-publicação — um paradigma tecnopolítico radicalmente distinto do modelo pós-verdade dos últimos anos.

Paralelamente, o Departamento de Defesa dos EUA (DARPA Media Forensics Program) desenvolveu protocolos que integram detecção de deepfakes com autenticação criptográfica de origem — provando que a batalha contra a falsificação envolve visão computacional + infraestrutura blockchain + assinatura biométrica de origem.

No setor corporativo, os escândalos de fraude baseada em deepfakes de voz e vídeo, como o caso de uma empresa inglesa que perdeu mais de US$ 240 mil após um golpe envolvendo a clonagem vocal do CEO, evidenciaram a urgência de sistemas automatizados de detecção preventiva em operações financeiras (Wall Street Journal, 2021). Na Ásia, conglomerados japoneses e sul-coreanos já implementam sistemas contínuos anti-deepfake para proteção executiva, integrando autenticação facial com monitoramento neural de microexpressões — muito além do simples "reconhecimento facial". E de forma ainda mais alarmante, relatórios da Interpol em 2022 confirmaram o uso de deepfakes em extorsão sexual, manipulação diplomática e tráfico humano, operando em ambientes de alta opacidade digital, tornando a detecção automatizada a única barreira possível contra destruição reputacional irreversível.

Esses casos demonstram que a questão não é mais SE devemos adotar visão computacional para combate a deepfakes — mas em que regime político e com qual ética operacional ela será implementada. A discussão deixou de ser técnica para tornar-se estrutural e civilizacional. Os modelos estão prontos — o que falta é a estrutura de poder que definirá seu uso.

CONCLUSÃO

A emergência das deepfakes representa não apenas uma nova etapa da desinformação digital, mas um ponto de ruptura histórico que altera a natureza ontológica da verdade em ambientes sociotécnicos mediados por IA. Ao romper a confiança milenar na imagem como evidência empírica do real, as deepfakes produzem um curto-circuito epistemológico: o colapso da relação entre percepção, prova e julgamento. A crise não é apenas informacional — é civilizacional. O que está em risco não é a disputa entre narrativas, mas a possibilidade de que uma sociedade compartilhe uma realidade minimamente estável sobre a qual decisões políticas, morais e jurídicas podem ser fundadas. Quando a visão humana deixa de ser critério válido de validação do mundo, o próprio contrato cognitivo que estrutura a vida coletiva entra em colapso, produzindo aquilo que Floridi (2020) denomina "infocalipse": o colapso da ecologia ontológica da verdade.

É nesse ponto que a visão computacional, sustentada por modelos forenses espectro-temporais e arquiteturas inteligentes antiadversariais, revela-se mais do que uma ferramenta técnica: ela se torna uma instituição epistêmica, um novo tipo de guardião da realidade público-infraestrutural. O artigo demonstrou que a detecção automatizada deve preceder a interpretação humana — não como substituição da consciência, mas como barreira civilizatória primária contra a manipulação do real em escala pós-humana. O avanço de modelos como transformers multimodais e redes treinadas contra assinaturas fisiológicas expõe uma transição histórica: a verdade já não é apenas verificada — ela começa a ser computacionalmente garantida. Trata-se da passagem da verdade interpretada (era da imprensa) para a verdade autenticada por IA (era da validação algorítmica de origem). É um deslocamento tectônico do poder epistêmico.

Contudo, esse mesmo avanço abre a ameaça simétrica: quem controlar a infraestrutura de validação computacional controlará a definição institucional do real. A luta contra deepfakes NÃO pode resultar numa nova hegemonia tecno-autoritarista — onde poucos atores geopolíticos ou corporações transnacionais concentram o poder de decidir, silenciosamente, o que existe e o que é apagado antes mesmo de circular. A proteção contra a desinformação não pode ser confundida com a privatização do real. O combate é técnico, mas sua legitimidade é política, ética e filosófica. O adversário não é apenas o falso, mas também o risco de que a defesa se converta em

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

censura invisível. A única saída está na construção de infraestruturas de detecção auditáveis, distribuídas e democraticamente reguladas, sob modelos de transparência verificável por múltiplas comunidades epistêmicas.

Diante disso, a visão computacional não deve ser compreendida como mero filtro de segurança, mas como fundação de um novo regime de soberania informacional. Ela inaugura a era em que democracias precisarão garantir computacionalmente o direito coletivo à autenticidade, com o mesmo rigor com que historicamente garantiram direitos civis como voto, identidade e liberdade. Se a desinformação é hoje projetada por redes neurais, a verdade terá de ser defendida por arquiteturas igualmente complexas — não com nostalgia do passado, mas com infraestruturas técnico-filosóficas compatíveis com a era pós-fotográfica da percepção sintética. Não se combate guerra algorítmica com fé na memória — combate-se com soberania sobre a computação da realidade.

Portanto, a visão computacional não é apenas resposta — é limiar de um novo pacto civilizatório. O que está em jogo não é vencer a batalha ocasional das fake news ou proteger eleições específicas: é decidir se a humanidade aceitará viver em regime de realidade colapsável ou reconstruirá uma nova ecologia de confiança baseada em garantias computáveis, éticas e plurais. Se os próximos anos definem a arquitetura do real, este artigo sustenta que apenas uma convergência madura entre ciência da computação, teoria crítica da tecnologia, filosofia da informação e governança democrática da IA será capaz de assegurar que a verdade não seja reduzida a um subproduto de redes neurais, mas permaneça como bem público inegociável — fundamento existencial da própria ideia de humanidade compartilhada.

REFERÊNCIAS

ARAL, Sinan. The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health—and How We Must Adapt. New York: Currency, 2020.

CHESNEY, Robert; CITRON, Danielle. Deep Fakes: A Looming Crisis for Privacy, Democracy, and National Security. California Law Review, v. 107, p. 1753–1819, 2019.

DOSHI-VELEZ, Finale; KIM, Been. Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608, 2018.

FLORIDI, Luciano. The Philosophy of Information. Oxford: Oxford University Press, 2011.

FLORIDI, Luciano. The Logic of Information: A Theory of Philosophy as Conceptual Design. Oxford: Oxford University Press, 2020.

GOODFELLOW, Ian et al. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems (NeurIPS), 2014.

HE, Kaiming et al. Deep Residual Learning for Image Recognition (ResNet). In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

HO, Jonathan et al. Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems (NeurIPS), 2020.

KARRAS, Tero et al. A Style-Based Generator Architecture for Generative Adversarial Networks (StyleGAN). In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey. ImageNet Classification with Deep Convolutional Neural Networks. In: NeurIPS, 2012.

MATERN, Fabian; RIESS, Christian; STAMMINGER, Marc. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In: IEEE Winter Applications of Computer Vision (WACV), 2019.

PARIS, Britt; DONOVAN, Joan. Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence. Harvard Kennedy School Misinformation Review, 2019.

QIAN, Yuchen et al. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In: European Conference on Computer Vision (ECCV), 2020.

ROSSLER, Andreas et al. FaceForensics++: Learning to Detect Manipulated Facial Images. In: International Conference on Computer Vision (ICCV), 2019.

SUN, Xiaoyuan; WANG, Li. Fake Retina: Detecting Deepfakes via Biological Signal Analysis. Journal of Visual Computing, 2020.

SZEGEDY, Christian et al. Rethinking the Inception Architecture for Computer Vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

TANDOC, Edson; LIM, Zheng; LING, Richard. Defining 'Fake News': A Typology of Scholarly Definitions. Digital Journalism, 2018.

TUFEKCI, Zeynep. Twitter and Tear Gas: The Power and Fragility of Networked Protest. Yale University Press, 2018.

VACCARI, Cristian; CHADWICK, Andrew. Deepfakes and Disinformation: Political Campaigns in the AI Age. Journal of Political Communication, 2020.

VOSOUGHI, Soroush; ROY, Deb; ARAL, Sinan. The Spread of True and False News Online. Science, 2018.

ZHAO, Liang et al. DefakeHop: A Lightweight High-Performance Deepfake Detector. IEEE Transactions on Multimedia, 2021.

ZUBOFF, Shoshana. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs, 2019.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.