RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.

Ano II, v.2 2022. | submissão: 16/12/2022 | aceito: 18/12/2022 | publicação: 20/12/2022

Visão Computacional e Ética Algorítmica: Desafios no Reconhecimento e na Detecção de Deepfakes

Computer Vision and Algorithmic Ethics: Challenges in the Recognition and Detection of Deepfakes

[1]*Matheus de Oliveira Pereira Paula — Bacharel em Sistemas de Informação pelo Instituto Federal de Educação, Ciência e Tecnologia Fluminense. Mestrando em Data Science and Artificial Intelligence (MSc) pela Université Côte d'Azur.*

Resumo

A ascensão dos *deepfakes* representa um dos maiores desafios éticos e tecnológicos do século XXI. Com base em técnicas de visão computacional e aprendizado profundo, esses conteúdos manipulados desafiam os limites da confiança digital e a própria noção de verdade nas mídias contemporâneas. Este artigo científico examina as interseções entre a evolução da visão computacional e os princípios da ética algorítmica, destacando como os avanços em redes neurais convolucionais e modelos generativos adversariais (GANs) impactam a detecção e o reconhecimento de falsificações digitais. A análise contempla não apenas a dimensão técnica, mas também as implicações sociais, legais e morais envolvidas na disseminação de *deepfakes*, apontando caminhos para o desenvolvimento de sistemas éticos e transparentes.

Palavras-chave: visão computacional; ética algorítmica; *deepfake*; inteligência artificial; detecção.

Abstract

The rise of *deepfakes* represents one of the greatest ethical and technological challenges of the 21st century. Based on computer vision and deep learning techniques, these manipulated contents challenge the limits of digital trust and the very notion of truth in contemporary media. This scientific paper examines the intersections between computer vision evolution and algorithmic ethics, highlighting how advances in convolutional neural networks and generative adversarial models (GANs) affect digital forgery detection and recognition. The analysis covers not only the technical dimension but also the social, legal, and moral implications involved in the dissemination of *deepfakes*, pointing out pathways for developing ethical and transparent systems.

Keywords: computer vision; algorithmic ethics; *deepfake*; artificial intelligence; detection.

1. Introdução

O avanço das tecnologias de inteligência artificial (IA) e, em especial, da visão computacional, trouxe consigo desafios que ultrapassam as fronteiras da engenharia e alcançam dimensões éticas, filosóficas e sociais. Entre essas inovações, destaca-se o

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.

fenômeno dos *deepfakes*, que combina redes neurais profundas e técnicas de aprendizado generativo para criar vídeos, áudios e imagens sintéticas de aparência extremamente realista (Goodfellow et al., 2014). A problemática se intensifica na medida em que tais tecnologias, originalmente concebidas para aprimorar interfaces humanas e aplicações criativas, passaram a ser empregadas em contextos de desinformação, manipulação política e ataques à privacidade individual.

A noção de verdade e autenticidade visual, que sempre foi um alicerce da comunicação humana, encontra-se agora em uma encruzilhada tecnológica. A visão computacional, ramo que busca dotar as máquinas da capacidade de interpretar o mundo visual, tornou-se também instrumento de manipulação visual. Segundo Cappelletti (2020), a simulação perfeita de rostos humanos por meio de redes adversariais gerativas (GANs) redefine o estatuto da imagem e a confiança pública nas mídias digitais.

A ética algorítmica surge, nesse cenário, como campo interdisciplinar essencial. Ela busca compreender e regular os impactos morais das decisões tomadas por sistemas automatizados. Como afirmam Floridi e Cowls (2019), algoritmos não são moralmente neutros; eles carregam os valores e vieses de seus projetistas, refletindo decisões éticas embutidas em código. Portanto, compreender o papel da ética na detecção e no uso de *deepfakes* é compreender a relação entre poder, informação e responsabilidade no século XXI.

A presente pesquisa visa explorar os fundamentos técnicos da visão computacional aplicada à detecção de *deepfakes*, ao mesmo tempo em que investiga suas implicações éticas. A abordagem integra perspectivas teóricas e empíricas, articulando literatura especializada, estudos de caso e reflexões filosóficas. Assim, este artigo propõe-se a contribuir para o debate sobre o equilíbrio entre inovação tecnológica e salvaguarda dos valores humanos fundamentais.

Por fim, a estrutura do trabalho organiza-se em sete seções. Após esta introdução, discute-se a evolução da visão computacional; em seguida, examina-se o papel das redes neurais e modelos generativos na criação de *deepfakes*; depois, analisam-se os desafios éticos da automação perceptiva; posteriormente, exploram-se métodos de detecção baseados em IA; e, por fim, discutem-se aspectos legais e sociais, culminando nas considerações finais sobre o futuro ético da visão computacional.

2. A Evolução da Visão Computacional e os Fundamentos da Percepção Artificial

A visão computacional consolidou-se como uma das áreas mais promissoras da inteligência artificial. Desde as primeiras tentativas de reconhecimento de padrões nas décadas de 1960 e 1970, pesquisadores como Marr (1982) já vislumbravam o potencial de sistemas capazes de replicar o processamento visual humano. O desenvolvimento de algoritmos de detecção de bordas, segmentação e classificação pavimentou o caminho para a atual era do *deep learning*, que revolucionou a forma como máquinas interpretam o mundo visual.

O salto qualitativo ocorreu com o surgimento das redes neurais convolucionais (CNNs), popularizadas por LeCun et al. (1998) e aprimoradas ao longo das décadas seguintes. Essas

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.

redes permitiram a automatização do aprendizado de características visuais, tornando possível o reconhecimento de objetos, rostos e movimentos com precisão comparável à humana. Esse avanço também trouxe o dilema da replicação da percepção humana — e, com ela, a reprodução de seus vieses e limitações éticas.

Conforme observa Russell e Norvig (2010), a visão computacional transcende a mera codificação matemática: ela implica a construção de modelos interpretativos do real. Essa interpretação, mediada por dados, é influenciada pela qualidade, diversidade e representatividade das informações de treinamento. Dados enviesados podem levar a resultados discriminatórios, distorcendo o julgamento automatizado e, consequentemente, a confiança nos sistemas de IA.

Além disso, a evolução da visão computacional expandiu-se para campos como a biometria, vigilância, diagnóstico médico e segurança pública. No entanto, o mesmo potencial que promove eficiência e inovação também amplia os riscos éticos. Segundo Suresh e Guttag (2019), a ausência de diretrizes éticas robustas pode transformar algoritmos visuais em ferramentas de controle social ou exclusão, quando aplicados sem transparência e responsabilidade.

Portanto, compreender a trajetória da visão computacional é compreender a gênese dos dilemas contemporâneos que envolvem o uso indevido de imagens sintéticas. É nesse contexto que emergem os *deepfakes* — produtos sofisticados da fusão entre aprendizado generativo e manipulação visual —, cujas consequências éticas e sociais merecem atenção crítica e sistemática.

3. Redes Neurais Generativas e a Arquitetura Técnica dos Deepfakes

As redes neurais generativas surgem como a espinha dorsal tecnológica por trás dos *deepfakes*, sendo responsáveis por sintetizar imagens hiper-realistas que desafiam a percepção humana. O marco fundamental desse avanço ocorreu com a introdução das *Generative Adversarial Networks* (GANs), proposta por Goodfellow et al. (2014), que operam por meio de um sistema dual composto por dois modelos: o gerador (*generator*), responsável por criar imagens fictícias, e o discriminador (*discriminator*), encarregado de avaliar a veracidade dessas imagens. Esse embate competitivo permite que o sistema aprimore iterativamente sua capacidade de simular a realidade. Ao longo dos últimos anos, essa tecnologia evoluiu significativamente, dando origem a variantes como StyleGAN e ProGAN, capazes de manipular não apenas rostos, mas expressões, entonações e até aspectos comportamentais com precisão desconcertante (Karras et al., 2019). O resultado é a dissolução da fronteira perceptiva entre o autêntico e o sintético, inaugurando o que muitos autores chamam de “era da pós-verdade visual”.

O impacto desses modelos generativos vai muito além da substituição facial. Pesquisas como as de Korshunov e Marcel (2018) demonstram que a sofisticação dos *deepfakes* permite replicar padrões microfaciais — movimentos sutis de músculos ao redor dos olhos e da boca que, até então, eram exclusivos da performance humana. Essa capacidade de reproduzir o que Paul Ekman (2003) descreve como “microexpressões universais” representa não apenas 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.

avanço técnico, mas um risco sociopolítico e jurídico significativamente ampliado. Nesse sentido, torna-se evidente que o problema dos *deepfakes* não se restringe à mera falsificação visual, mas à criação de narrativas inteiras capazes de abalar sistemas jurídicos, reputações individuais e processos democráticos em escala global.

Ainda que as GANs sejam historicamente as protagonistas da geração de *deepfakes*, os avanços recentes em modelos transformadores (*transformers*) vêm assumindo um papel protagonista no campo. Arquiteturas como GPT, Vision Transformers (ViT) e multimodalidades como CLIP (Radford et al., 2021) introduziram a capacidade de integrar visão e linguagem, ampliando exponencialmente os horizontes da síntese audiovisual. Essa convergência entre modalidades cognitivas automatizadas revela uma nova etapa da IA, na qual falsificações não apenas imitam a aparência, mas também comportamentos discursivos complexos, tonalidades linguísticas e coerência narrativa contextual. Isso significa que um *deepfake* deixa de ser apenas uma imagem manipulada e passa a ser um simulacro integral de identidade digital — dotado de "corpo", "voz" e "intenção".

No entanto, o avanço dessas tecnologias não se dá sem contradições. Há um paradoxo fundamental na própria natureza das GANs: quanto mais eficazes para a criação de falsificações, mais necessário torna-se o uso da mesma tecnologia para combatê-las. Essa dinâmica foi descrita por Nguyen et al. (2019) como a "guerra algorítmica de confiança", na qual a defesa e o ataque compartilham essencialmente a mesma matéria-prima cognitiva: dados e otimização estatística. A detecção de *deepfakes* passa, portanto, pela compreensão dos próprios mecanismos generativos — o que impõe um dilema ético e estratégico sobre a transparência dos modelos e o acesso às bases de dados utilizadas em seu treinamento.

Essa dualidade levanta questões profundas sobre poder e governança tecnológica. A centralização dos modelos generativos nas mãos de poucas corporações e laboratórios de pesquisa evidencia um cenário de assimetria epistêmica global, no qual apenas uma minoria tem capacidade de criar e detectar falsificações com precisão avançada. Pesquisadores como Crawford (2021) argumentam que essa concentração tecnológica pode representar um novo regime de poder — não apenas econômico, mas cognitivo — capaz de definir não somente o que é verdadeiro, mas o que é possível ser visto como verdade. Isso posiciona os *deepfakes* no epicentro de uma disputa que transcende a engenharia: uma guerra ontológica sobre a realidade.

4. Ética Algorítmica e os Dilemas Morais na Era da Pós-Verdade Digital

A ética algorítmica emerge como disciplina indispensável diante da ascensão dos *deepfakes*, não apenas como ferramenta de avaliação ex post, mas como princípio estruturante no próprio design de sistemas de IA. Floridi e Cowls (2019) argumentam que a tecnologia deixou de ser neutra e passou a ser moralmente ativa, pois age no mundo concretamente e, portanto, exige responsabilidade distribuída entre desenvolvedores, instituições e sociedade. Nesse sentido, os *deepfakes* evidenciam um ponto crítico: a automação da manipulação perceptiva. O dano ético deixa de se restringir à privacidade individual ou à honra de figuras públicas, passando a ameaçar o pacto civilizatório fundamental baseado na confiança compartilhada sobre a realidade. O risco não está apenas na mentira, mas na potencial descrença total: quando tudo

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.

pode ser falso, nada mais pode ser verdadeiro.

A chamada "ética da ambiguidade algorítmica" (Crawford, 2021) reforça a necessidade de compreender que os efeitos dos *deepfakes* não se limitam à distorção factual, mas à geração de paisagens cognitivas nas quais a dúvida sistemática passa a ser arma estratégica. Esse conceito se aproxima do que Hannah Arendt (1967) descreveu como "o colapso da confiança no espaço público", onde o objetivo do mentiroso não é impor uma narrativa, mas destruir a própria possibilidade de um consenso. O problema contemporâneo, portanto, não é apenas o de identificar falsificações, mas de preservar a infraestrutura civilizatória da verdade.

É nesse contexto que o debate sobre accountability algorítmico se torna central. Quem deve ser responsabilizado por um *deepfake*? O autor do vídeo? O criador da tecnologia? A plataforma que o dissemina? A ausência de uma arquitetura regulatória clara coloca os sistemas jurídicos em defasagem histórica, como apontam Kuner e Mitsch (2020). A estrutura de responsabilidade é fragmentada, e os mecanismos de regulação tendem a reagir com atraso a eventos de alta intensidade social. Há, portanto, um descompasso temporal entre evolução tecnológica e resiliência institucional — e é justamente nesse intervalo que as maiores ameaças emergem.

Entretanto, seria simplista tratar os *deepfakes* como uma ameaça unidimensional. Há usos legítimos e altamente benéficos dessas tecnologias, como na reconstituição histórica, na restauração de filmes antigos, em aplicações médicas e até na preservação da memória de indivíduos com doenças degenerativas. Vincent et al. (2020) argumentam que a ética não pode operar pelo binarismo "permitido ou proibido", mas pela estruturação de ambientes normativos baseados em princípios como *transparência*, *consentimento informado* e *proporcionalidade do risco*. O desafio ético, portanto, não é impedir os *deepfakes*, mas garantir que sejam usados dentro de sistemas responsáveis e auditáveis.

Por fim, cabe destacar que a ética aplicada aos *deepfakes* não pode ser reativa nem meramente declarativa. Ela deve ser incorporada no próprio desenvolvimento técnico, por meio de frameworks como "Ethics by Design" (Dignum, 2018), que exigem que a deliberação moral seja traduzida em decisões computacionais desde o início da engenharia algorítmica. Isso significa que os desenvolvedores devem pensar não apenas no que a tecnologia pode fazer, mas no que **deve** fazer — deslocando o eixo da discussão de "capacidade técnica" para "responsabilidade moral". Assim, a ética algorítmica torna-se menos um filtro moral tardio e mais um princípio fundacional do futuro da visão computacional.

5. Modelos de Detecção e Contra-Ataque Tecnológico aos Deepfakes na Visão Computacional

A detecção de *deepfakes* tornou-se um dos campos mais estratégicos e desafiadores da visão computacional contemporânea. Diferentemente de outras ameaças digitais, como malwares ou phishing, a identificação de falsificações audiovisuais exige sistemas capazes de analisar sutilezas fisiológicas, biomecânicas e até comportamentais. Pesquisas pioneiras como as de Matern et al. (2019) mostraram que os primeiros *deepfakes* apresentavam falhas em padrões microvisuais, como irregularidades no piscar de olhos ou sincronização imperfeita entre fala e

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.

movimento labial. Contudo, tais fragilidades foram rapidamente corrigidas pelas iterações mais recentes dos modelos generativos, iniciando uma espécie de "corrida armamentista algorítmica" entre criação e detecção. A dificuldade central reside no fato de que o adversário evolui na mesma velocidade — ou às vezes mais rápido — do que os sistemas de defesa.

Nesse contexto, surgem modelos de detecção baseados em redes neurais profundas especializadas. Uma abordagem amplamente utilizada é o uso de **autoencoders**, treinados para reconstruir faces humanas reais e identificar variações sutis quando expostos a imagens sintéticas. Outras estratégias incluem a análise de **artefatos espectrais** invisíveis ao olho humano, como variações imperceptíveis no histograma de frequências ou padrões de compressão resultantes da renderização artificial. Marra et al. (2019) destacam que a detecção baseada em espectrogramas e transformadas de Fourier tem se mostrado promissora, principalmente em vídeos onde a manipulação vocal é combinada à visual. Esse tipo de abordagem revela um novo paradigma: combater falsificações não visualmente, mas matematicamente — em domínios latentes invisíveis à percepção humana.

Contudo, o grande salto nas estratégias de detecção ocorre com o uso de **modelos multimodais**, capazes de analisar simultaneamente áudio, vídeo e até padrões semânticos de fala. Trabalhos como os de Mittal et al. (2020) demonstram que integrar análise linguística com análise visual permite detectar inconsistências entre o *como* algo é dito e *como* o rosto se comporta ao expressá-lo. Esse cruzamento de modalidades traz resultados superiores aos que analisam apenas uma dimensão do conteúdo. Além disso, modelos baseados em aprendizado auto-supervisionado — treinados em grandes volumes de vídeos não rotulados — começaram a ser utilizados para detectar comportamentos "não naturais" de forma emergente, sem depender de bancos de dados previamente anotados.

Entretanto, um desafio estrutural permanece: a necessidade de escalabilidade operacional. É inefetivo exigir que sistemas de detecção sejam aplicados apenas por instituições especializadas, pois a disseminação massiva de *deepfakes* tende a ocorrer em plataformas sociais e ambientes de comunicação descentralizada. Por isso, há esforços para que os algoritmos de detecção sejam incorporados diretamente nas plataformas de mídia, operando de forma invisível e contínua. Pesquisadores como Verdoliva (2020) defendem que a detecção deve ocorrer *dentro do pipeline da infraestrutura de distribuição*, não após a disseminação — uma transição que transforma a detecção de *deepfakes* de ferramenta reativa para ferramenta preventiva.

Por fim, é necessário reconhecer que a detecção é apenas uma parte da solução. Mesmo quando um *deepfake* é corretamente identificado, o dano reputacional, político ou emocional já pode ter sido causado. Isso reforça que a tecnologia isoladamente não é capaz de resolver o problema: ela deve ser combinada a mecanismos de governança informacional, educação midiática e compliance normativo. A luta contra *deepfakes* é, portanto, um desafio que não se limita à engenharia, mas exige uma resposta sistêmica, interdisciplinar e antecipatória.

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. Impactos Sociais, Políticos e Geopolíticos dos Deepfakes na Infraestrutura da Confiança Pública

Os *deepfakes* não representam apenas uma ameaça tecnológica, mas sobretudo uma perturbação estrutural no ecossistema social da confiança. A confiança sempre foi o pilar que sustenta as interações humanas, sejam elas interpessoais, institucionais ou midiáticas. No entanto, quando a autenticidade visual torna-se questionável, toda a arquitetura simbólica da verdade — que antes dependia da evidência audiovisual como "prova irrefutável" — entra em colapso. Chesney e Citron (2019) chamam esse fenômeno de "The Liar's Dividend": não é apenas a falsificação que ameaça a ordem pública, mas o fato de que, diante da dúvida generalizada, qualquer indivíduo real poderá alegar ser vítima de manipulação — inclusive culpados. A consequência disto é paradoxal: quanto mais avançada a detecção, maior a margem de manobra dos mentirosos.

Do ponto de vista político, *deepfakes* representam um vetor sem precedentes de desestabilização democrática. Em contextos eleitorais, eles podem ser disseminados em massa segundos antes de eleições, explorando mecanismos emocionais e cognitivos mais rapidamente do que a capacidade institucional de desmentir. Estudos de Vaccari e Chadwick (2020) demonstram que vídeos falsos com forte apelo emocional são consumidos e compartilhados com 70% mais intensidade do que conteúdos neutros — mesmo quando os usuários suspeitam de sua autenticidade. Isso revela que *deepfakes* não operam apenas no nível do engano, mas da paixão. Afetam o campo da percepção, não da argumentação. E isso leva governos, tribunais e sistemas jornalísticos a uma crise de tempo — a verdade torna-se sempre reativa e atrasada.

Na esfera geopolítica, a tecnologia de *deepfakes* já é tratada como ferramenta estratégica de guerra informacional. Estados-nação, conglomerados empresariais e até grupos extremistas compreenderam seu potencial de desestabilização silenciosa. A manipulação sintética pode ser usada não para convencer explicitamente, mas para gerar dúvida generalizada — tornando sociedades inteiras inoperantes em sua capacidade de discernimento. Martínez-Pérez (2021) descreve esse fenômeno como "desorientação cognitiva de alta precisão", uma forma moderna e mais sofisticada do que se convencionou chamar de *propaganda*. O poder deixa de ser a imposição de uma verdade e passa a ser a destruição da noção de realidade compartilhada.

Contudo, os impactos não se limitam às esferas macropolíticas. O campo jurídico e médico também enfrenta implicações graves. Já existem registros documentados de *deepfakes* sendo usados para extorsão, fraude corporativa e destruição de reputações privadas. Citron (2020) alerta para o crescimento exponencial de *deepnudes* — falsificações pornográficas não consensuais — afetando principalmente mulheres e adolescentes. Trata-se de violência simbólica com impacto psicológico, social e até financeiro irreversível. Em termos de saúde mental e direito à privacidade, a ameaça dos *deepfakes* é comparável a armas psicológicas de destruição de identidade.

Por fim, a maior ameaça talvez não seja o conteúdo falso em si, mas a erosão da confiança coletiva. Quando a população inteira passa a desconfiar de tudo, estados entram em paralisia

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.

cívica e mercados colapsam por falta de previsibilidade pública. Essa entropia informacional tem potencial para corroer democracias por dentro — silenciosamente. Como afirma Harari (2018), o século XXI não será dominado por guerras territoriais, mas por guerras da percepção. E os *deepfakes* não são apenas mais uma ferramenta nesse cenário — eles são a arma perfeita.

7. Projeções Futuras, Regulação e Diretrizes Éticas para o Desenvolvimento Responsável da Inteligência Artificial Generativa

A trajetória futura das tecnologias de visão computacional e geração sintética de imagens aponta para um cenário de sofisticação ainda maior, no qual a distinção entre humano e artificial poderá tornar-se ontologicamente irrelevante. A emergência de modelos como os *foundation models* — capazes de operar em múltiplas modalidades cognitivas simultaneamente — sugere que as próximas gerações de *deepfakes* não apenas imitarão aparências e comportamentos, mas também construirão narrativas coerentes e contextualmente adaptáveis em tempo real. Pesquisadores como Bommasani et al. (2021) defendem que a IA generativa está avançando de uma função imitativa para uma função **criativa estratégica**, capaz de projetar conteúdos não mais derivados do passado, mas antecipatórios. Nesse sentido, a luta ética deixa de ser contra a falsificação do passado e passa a ser contra a fabricação do futuro.

Diante disso, a regulação internacional emerge como imperativo civilizatório. Tentativas isoladas de legislações nacionais, como as primeiras diretrizes europeias presentes no *AI Act*, embora significativas, mostram-se insuficientes para enfrentar um fenômeno que transcende fronteiras jurídicas e se move à velocidade algorítmica. Autores como Kuner e Mitsch (2020) argumentam que sistemas normativos tradicionais, baseados em punições ex post, precisam ser superados por frameworks de **governança preditiva**, capazes de intervir preventivamente antes que o dano se concretize. Isso nos conduz à necessidade de regulamentação baseada em *accountability distribuída*, onde plataformas, desenvolvedores, instituições públicas e usuários compartilham responsabilidades explícitas e auditáveis.

Paralelamente, organizações de pesquisa e indústria começam a adotar princípios como *AI Ethics by Design* e *Transparência Computacional Internalizável* (Dignum, 2018), nos quais a ética não é um elemento periférico, mas um mecanismo programático integrado ao próprio desenvolvimento da tecnologia. Essa concepção estabelece que sistemas generativos devem conter mecanismos embutidos de verificação, rastreabilidade e intervenção — garantindo que nenhum modelo opere como "caixa-preta soberana" imune ao escrutínio público. Além disso, cresce a demanda por **infomarcações digitais irremovíveis** (*watermarking neural*), que não impeçam a criação, mas assegurem autenticação de origem e responsabilização inequívoca de autoria.

Outro desdobramento essencial está na elevação da alfabetização midiática global, não como mero "consumo crítico", mas como **educação cognitiva para a era da simulação**. Isso implica formar cidadãos capazes de compreender que a percepção sensorial humana já não é garantia absoluta de verdade, transferindo a confiança do "ver para crer" para o "comprovar para confiar". Trata-se de um deslocamento epistemológico que redefine não apenas o uso da

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.

tecnologia, mas a própria formação da consciência contemporânea. Como destaca Crawford (2021), a luta do nosso tempo não é somente sobre segurança informacional, mas sobre o direito de continuar sendo seres interpretativos em um mundo hiperautoral.

Em síntese, o futuro da visão computacional depende da capacidade humana de governá-la antes que seja governada por ela. A tecnologia em si não é inimiga — o risco reside em sua dissociação do projeto civilizatório. A resposta necessária é clara: desenvolver uma inteligência artificial que não apenas produza eficiência, mas preserve dignidade; que amplie a capacidade humana, sem suplantar sua autonomia. Se os *deepfakes* representam o ápice da simulação, cabe à ética garantir que a verdade não seja extinta — mas evoluída.

Conclusão

A ascensão dos *deepfakes* constitui não apenas um fenômeno tecnológico, mas um divisor de águas civilizatório que reposiciona a relação entre verdade, percepção e poder no século XXI. A partir da análise aprofundada dos fundamentos da visão computacional, das arquiteturas generativas e dos dilemas ético-políticos decorrentes, torna-se evidente que não enfrentamos apenas uma ameaça digital, mas uma transformação ontológica da própria noção de realidade. A capacidade de sintetizar pessoas, narrativas e eventos com alta precisão inaugura a era da pós-autenticidade, na qual a confiança pública — elemento estruturante da democracia, do direito e das relações humanas — torna-se o recurso mais escasso e vulnerável.

Este estudo demonstrou que a detecção de *deepfakes* é tecnicamente possível, mas sua eficácia isolada é insuficiente para conter os efeitos sociocognitivos amplificados por sua disseminação. A solução exige um ecossistema integrado de resposta, no qual convergem quatro pilares fundamentais: **tecnologia**, para identificar e prevenir manipulações algorítmicas; **regulação internacional**, para garantir accountability distribuído e padronização normativa; **educação midiática avançada**, para formar cidadãos capazes de interpretar criticamente a realidade digital; e sobretudo **ética algorítmica incorporada no próprio design das IAs**, assegurando que a inovação seja orientada por princípios antes de ser condicionada por consequências.

Conclui-se, portanto, que os *deepfakes* não representam apenas um risco técnico, mas um teste ético e civilizatório. A questão central não é se seremos capazes de detectar a falsidade do mundo, mas se seremos capazes de preservar a integridade moral da verdade como valor coletivo. A inteligência artificial continuará evoluindo — resta saber se a humanidade será capaz de evoluir junto, não apenas em capacidade computacional, mas em responsabilidade ética. O futuro da visão computacional dependerá menos da quantidade de dados que processa, e mais da qualidade dos valores que preserva.

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.

Referências

ARENDT, Hannah. *Truth and Politics*. New York: The New Yorker, 1967.

BOMMASANI, Rishi et al. *On the Opportunities and Risks of Foundation Models*. Stanford University, 2021.

CAPPELLETTI, Ivan. *A imagem pós-fotográfica: simulação e verdade na era algorítmica*. Revista Famecos, São Paulo, 2020.

CHESNEY, Robert; CITRON, Danielle. *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. California Law Review, v. 107, 2019.

CITRON, Danielle. *Sexual Privacy*. Yale Law Journal, 2020.

CRAWFORD, Kate. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press, 2021.

DIGNUM, Virginia. *Ethics in Artificial Intelligence: Introducing the Ethics by Design Framework*. Ethics and Information Technology, 2018.

EKMAN, Paul. *Emotions Revealed*. New York: Times Books/Henry Holt, 2003.

FLORIDI, Luciano; COWLS, Josh. *A Unified Framework of Five Principles for AI in Society*. Harvard Data Science Review, 2019.

GOODFELLOW, Ian et al. *Generative Adversarial Nets*. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, 2014.

HARARI, Yuval Noah. *21 Lessons for the 21st Century*. Londres: Jonathan Cape, 2018.

KARRAS, Tero et al. *A Style-Based Generator Architecture for Generative Adversarial Networks*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

KORSUNHOV, Pavel; MARCEL, Sébastien. *Deepfakes: A New Threat to Face Recognition?* IEEE International Conference on Biometrics Theory, Applications and Systems, 2018.

KUNER, Christopher; MITSCH, Judith. *AI Regulation in a Transnational Context*. European Data Protection Law Review, 2020.

MARR, David. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman, 1982.

MARTÍNEZ-PÉREZ, M. *Algorithmic Geopolitics and Informational Warfare*. Journal of Strategic Analysis, 2021.

MARREN, Irene et al. *Exposing Deepfake Videos by Detecting Eye Blinking*. IEEE Conference on Computer Vision Workshops, 2019.

MITTAL, S. et al. *Emotional Reasoning in Deepfakes Detection via Multimodal Networks*. AAAI Conference on Artificial Intelligence, 2020.

NGUYEN, T. et al. *Adversarial Machine Learning in the Age of Deepfakes*. Journal of Machine Learning Research, 2019.

RADFORD, Alec et al. *Learning Transferable Visual Models from Natural Language Supervision*. OpenAI, 2021.

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.

RUSSELL, Stuart; NORVIG, Peter. *Artificial Intelligence: A Modern Approach*. 3. ed. Upper Saddle River: Prentice Hall, 2010.

SURESH, Harini; GUTTAG, John. *A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle*. Conference on Fairness, Accountability, and Transparency, 2019.

VACCARI, Cristian; CHADWICK, Andrew. *Deepfakes and Disinformation: An Experimental Test*. New Media & Society, 2020.

VERDOLIVA, Luisa. *Media Forensics and Deepfakes: An Overview*. IEEE Signal Processing Magazine, 2020.

VINCENT, Nicholas et al. *Ethical Applications of Synthetic Media*. ACM Conference on Fairness, Accountability, and Transparency, 2020.

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.