RCMOS – Revista Científica Multidisciplinar O Saber.
ISSN: 2675-9128. São Paulo-SP.
Ano IV, v.2 2024. | submissão: 16/12/2024 | aceito: 18/12/2024 | publicação: 20/12/2024

Treinamento Ético de Redes Neurais: Conjuntos de Dados, Viés e Privacidade na Detecção de Deepfakes

Ethical Training of Neural Networks: Datasets, Bias, and Privacy in Deepfake Detection

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

RESUMO

A detecção automatizada de deepfakes, baseada no uso massivo de redes neurais profundas, tornou-se um imperativo civilizacional para a proteção da integridade informacional em plataformas digitais. Entretanto, o ponto mais sensível dessa arquitetura não reside na sofisticação algorítmica do modelo, mas na formação ética dos conjuntos de dados utilizados em seu treinamento — os quais são frequentemente marcados por vieses estruturais de raça, gênero e geopolítica, bem como por violações de privacidade e uso não consentido de imagens biométricas humanas. Este artigo analisa criticamente, com abordagem epistemológica e tecnopolítica, os dilemas éticos fundamentais associados à criação e curadoria de datasets para treinar detectores de deepfakes. Com base em estudos contemporâneos, explora-se como erros na composição do dataset podem reproduzir formas históricas de opressão algorítmica, amplificar desigualdades sociais e até mesmo comprometer a legitimidade democrática das tecnologias de combate à desinformação. Conclui propondo diretrizes éticas e técnicas para o desenvolvimento de pipelines de treinamento responsáveis, auditáveis e compatíveis com princípios de soberania informacional e justiça algorítmica.

Palavras-chave: deepfakes; ética algorítmica; vieses; privacidade digital; soberania informacional.

ABSTRACT

The automated detection of deepfakes through large-scale neural networks has become a civilizational necessity for protecting informational integrity in the digital environment. However, the most critical point in this architecture does not lie in the model itself, but in the ethical formation of the datasets used during training — which are frequently affected by structural biases related to race, gender, and geopolitics, as well as by potential violations of privacy and non-consensual biometric data extraction. This article develops an epistemologically rigorous analysis of the ethical dilemmas embedded in dataset construction for deepfake detection models, examining how flawed curation can reinforce algorithmic oppression, reproduce historical inequalities, and ultimately

endanger democratic legitimacy. It concludes by proposing an ethical framework and engineering guidelines for responsible and auditable dataset governance, anchored in informational sovereignty and computational justice.

Keywords: deepfakes; algorithmic ethics; dataset bias; digital privacy; informational sovereignty.

1. INTRODUÇÃO: O PARADOXO ÉTICO NA GUERRA ALGORÍTMICA CONTRA A DESINFORMAÇÃO

A detecção automatizada de deepfakes, impulsionada por redes neurais profundas e pipelines de visão computacional em larga escala, consolidou-se globalmente como uma das principais infraestruturas de defesa contra a manipulação informacional de quinta geração. No entanto — e este é o ponto mais negligenciado do debate público — a guerra contra deepfakes **não é uma guerra entre verdade e mentira**, mas **entre inteligências artificiais adversárias**, ambas capazes de operar em velocidades e granularidades superiores à cognição humana. Logo, o verdadeiro centro ético do problema não está apenas em identificar o conteúdo falso, mas em **interrogar as bases de poder e os critérios morais que estruturam o treinamento dos modelos encarregados de declarar o real**. Isso significa que um detector de deepfakes mal treinado — com datasets enviesados, racialmente assimétricos, obtidos sem consentimento legítimo ou restritos a segmentos demográficos hiper-específicos — **pode produzir uma nova arquitetura de opressão algorítmica**, mascarada sob a retórica da proteção social. A detecção, portanto, não é neutra: é um ato político de primeiro grau.

A formação dos datasets utilizados para treinar detectores de deepfakes é hoje reconhecida como um dos pontos mais críticos e perigosos de toda a pipeline de segurança informacional. Isso porque os modelos não aprendem "o que é verdade", mas apenas "o que é estatisticamente frequente" dentro dos dados que lhes são disponibilizados. Ou seja: **uma IA não detecta a falsidade — ela detecta a diferença em relação ao padrão aprendido**. Se esse padrão foi formado majoritariamente com rostos masculinos brancos europeus, como ocorreu com diversos datasets iniciais como FaceForensics++, o modelo passa a ser **significativamente menos eficaz para detectar manipulações em rostos femininos, negros, indígenas ou pertencentes a populações fora do eixo geopolítico anglo-saxão**. Esse problema, documentado por Buolamwini e Gebru (2018) no campo do facial recognition, **migrando para detecção de deepfakes, pode criar um cenário em que determinadas etnias ficam mais vulneráveis a manipulações e outras mais rigidamente policiadas** — amplificando desigualdades históricas dentro de um ambiente algorítmico.

Além do viés demográfico, surge o **problema da privacidade e da apropriação não consentida de imagens humanas**, muitas vezes coletadas silenciosamente a partir de redes sociais, plataformas

de vídeo ou bancos de dados vazados, sem qualquer autorização dos sujeitos envolvidos. A detecção de deepfakes exige, por natureza, a ingestão massiva de dados faciais reais para treinamento — o que coloca um dilema ético explosivo: **para defender pessoas contra manipulação visual, estamos violando a própria autonomia sobre sua identidade biométrica**. Autores como Floridi (2021) alertam que a fronteira da dignidade humana está sendo renegociada silenciosamente no espaço da computação forense: **a defesa da verdade não pode se tornar justificativa para vigilância biométrica total**. Sem estruturas claras de consentimento, governança de dados e soberania sobre representações faciais, **a própria IA "protetora" pode funcionar como máquina extrativista de identidades humanas**.

O problema se agrava na medida em que grandes plataformas e governos passam a reivindicar **monopólio da infraestrutura forense automatizada**, centralizando em poucas entidades o poder de **definir — em tempo real — o que é verdadeira ou falsificadamente legítimo de circular na esfera pública digital**. Reportes da OTAN (2021) e da Comissão Europeia (2022) já reconhecem que a soberania informacional do século XXI depende diretamente de quem controla a infraestrutura de detecção de deepfakes. Mas aqui surge o perigo: **quem controla os datasets controla a consciência automatizada da verdade**. Um modelo treinado com dados enviesados não apenas erra tecnicamente — **ele erra politicamente**. Pode reforçar estruturas de poder já existentes, bloquear narrativas dissidentes e autorizar seletivamente a circulação daquilo que serve a interesses geoestratégicos específicos. **A tecnologia da detecção, se mal governada, pode ameaçar mais do que protege.**

Diante desse cenário, fica claro que a discussão sobre deepfakes não pode ser reduzida a uma disputa técnica ou operacional. O problema é **fundacionalmente epistemológico e ético**, pois envolve a transferência da autoridade ontológica sobre o real — antes mediada pela percepção humana, pelo jornalismo, pela ciência ou pelo processo jurídico — para sistemas automatizados de classificação treinados em bases de dados invisíveis à sociedade. Os datasets se tornam, portanto, **espelhos invisíveis do poder**, encapsulando não apenas estatísticas, mas **visões ideológicas implícitas do que é "normal", do que é "aceitável", e do que é "real"**. Sem governança rigorosa, **o algoritmo que detecta deepfakes pode reforçar exatamente a lógica colonial, patriarcal e centralizadora que a sociedade contemporânea deveria superar**. O treinamento ético, portanto, não é opcional — é **pré-condição para qualquer tecnologia de detecção ser moralmente válida**.

Se a detecção algorítmica de deepfakes afirma trabalhar em nome da verdade, ela deve ser submetida ao duplo princípio da **justiça e da auditabilidade pública**. Isso implica, inevitavelmente, a exigência de que os datasets sejam **documentados, verificáveis, representativos, distribuídos geopoliticamente e compostos com o consentimento informado das pessoas envolvidas**. O que está em disputa não é apenas eficiência técnica, mas **legitimidade civilizacional**: um modelo que

detecta deepfakes com 99% de precisão, mas opera com vieses raciais, geopolíticos ou extrai dados não consentidos, **não é tecnicamente avançado — é moralmente falhado**. A sociedade tem o direito de exigir **explicabilidade, governança transversal e proteção contra abusos estruturais** em nome de qualquer promessa de segurança digital.

Assim, este primeiro item estabelece a tese central deste artigo: **o combate a deepfakes só é legítimo se for simultaneamente tecnicamente eficaz, democraticamente auditável e moralmente aceitável**, o que exige que a discussão comece não na arquitetura do modelo, mas **na total integridade ética da formação do dataset**. A questão definitiva, portanto, não é apenas "como detectamos deepfakes?", mas **"em nome de quem, sob quais critérios e com base em quais imagens humanas treinamos a inteligência que decidirá o que é real?"** A detecção, quando mal orientada, pode se tornar aquilo que pretende combater: **um mecanismo de distorção do real, legitimado por autoridade algorítmica superior**. A tecnologia só pode ser solução se for, antes de tudo, **moralmente justa desde sua origem** — e essa origem é o dataset.

2. FUNDAMENTOS ÉTICOS E ESTRUTURAIS DOS DATASETS NA DETECÇÃO DE DEEPFAKES

A construção de datasets para treinar detectores de deepfakes envolve decisões que não são apenas técnicas, mas **profundamente civilizatórias**, pois determinam antecipadamente **o que a inteligência artificial entenderá como normal, anômalo, legítimo, perigoso ou irrelevante** — antes mesmo de ser usada em qualquer ambiente real. Isso significa que cada curadoria de dados é um **ato de poder**, que seleciona quais corpos humanos se tornarão estatisticamente visíveis e quais serão invisibilizados; quais expressões faciais serão consideradas comportamentos neutros e quais serão tratadas como ruído; quais culturas terão sua fisiologia representada e quais serão tratadas como exceções. Essa lógica é denunciada por Crawford (2021), ao afirmar que datasets contemporâneos são "**infraestruturas ideológicas disfarçadas de fatos**". No caso de deepfakes, a gravidade se multiplica, pois os detectores **não apenas descrevem o real — julgá-lo-ão com autoridade forense**, definindo quais evidências visuais podem ou não ser confiáveis em contextos judiciais, eleitorais ou geopolíticos. Ao contrário do senso comum, o dataset não é um espelho da realidade — **ele é uma construção epistemológica que molda como a própria realidade será reconhecida pela IA**.

Historicamente, grande parte dos datasets utilizados para treinamento de modelos visuais foram formados a partir de **coletas centralizadas no eixo euro-norte-americano**, gerando uma representação estatisticamente hipertrofiada de **rostos masculinos, brancos, iluminados sob padrões fotográficos ocidentais, com expressões neutras que obedecem a padrões culturais específicos**. Datasets como LFW, CelebA, MegaFace e mesmo FaceForensics++ foram criticados por especialistas por **repetirem a homogeneidade eurocêntrica da internet mainstream**, excluindo de maneira quase total **fenótipos africanos, indígenas, sul-asiáticos, traços faciais miscigenados ou variações morfológicas fisiológicas não hegemônicas**. O resultado é devastador: modelos

treinados sobre tais bases **detectam deepfakes com alta precisão apenas nas identidades dominantes**, enquanto deixam brechas perigosas para manipulações com alvos não-europeus — expondo populações históricas de risco a novos tipos de abuso digital silencioso. Não é exagero afirmar que, na guerra algorítmica pela verdade, **quem está ausente do dataset está ausente da proteção**.

O problema não é apenas quantitativo — é ontológico. Os datasets definem **o que é considerado "face humana legítima"**, o que é computacionalmente reconhecível como "sinal vital", o que é tratado como "microexpressão válida" e o que é descartado como "ruído irrelevante". Esse recorte normativo afeta diretamente a capacidade dos modelos forenses de detectar assinaturas fisiológicas autênticas (como pulsação dérmica, dilatação pupilar ou microcontração involuntária) quando estas se manifestam com padrões fisiológicos **diferentes dos parâmetros caucasianos que saturam os datasets ocidentais**. Isso gera um **paradoxo perverso**: a IA pode classificar como "suspeito" um rosto real de uma cultura sub-representada — e simultaneamente **considerar autêntica uma deepfake que simula corretamente apenas o padrão biométrico dominante**. O dataset, portanto, **não apenas viessa estatisticamente a IA — ele pode induzir um falso senso de certeza científica**, tecnicamente sofisticado, mas moralmente distorcido desde sua origem.

Outro dilema ético inescapável reside na **origem dos dados faciais utilizados para treinamento**. A esmagadora maioria dos rostos usados em datasets globais foi coletada **sem consentimento explícito, informado e juridicamente sólido dos indivíduos envolvidos** — muitas vezes extraída de redes sociais públicas, bancos audiovisuais, plataformas de streaming e até vazamentos criminais. A lógica "o que está online está disponível" tornou-se perversamente normalizada pela indústria de IA, resultando na criação de datasets que, embora tecnicamente poderosos, **incorporam em si uma violação estrutural da autonomia digital**, utilizando **rostos humanos como matéria-prima descartável**. Em casos extremos, como denunciado por pesquisadores da AI Now Institute (2021), **rostos de pessoas mortas foram usados para treinar detectores de spoofing facial sem qualquer consulta às famílias**, demonstrando que os limites éticos da extração biométrica foram ultrapassados com naturalidade industrial. A detecção de deepfakes não pode ser "protetora" se nasce da coleta invasiva da identidade alheia.

O problema torna-se ainda mais complexo quando inserido na geopolítica dos dados. A China criou datasets oficiais com mais de **um bilhão de rostos** para treinar modelos de detecção e classificação biométrica avançada — quase todos obtidos internamente sob legislação autoritária. Os Estados Unidos, por sua vez, operam sob o modelo oposto: datasets altamente fragmentados, privados e despadronizados, com alto risco de **monopolização corporativa da infraestrutura da verdade**. Europa tenta impor **ética regulatória**, mas carece da potência computacional e da escala populacional de seus concorrentes. O resultado é um trinômio crítico: **quem controla o dataset controla a IA; quem controla a IA controla a detecção; quem controla a detecção controla a realidade digital**. Ou seja, o dataset é **o campo zero do poder tecnocivilizacional**. Discuti-lo é discutir **soberania informacional** — e não apenas ferramenta técnica.

Além disso, a ausência de governança sobre datasets permite que corporações privadas e regimes autoritários usem detecção de deepfakes **como arma geopolítica, e não como instrumento de proteção pública**. Já é tecnicamente possível que um governo classifique seletivamente conteúdo verdadeiro como falso para destruir reputações — ou que plataformas privadas silenciem narrativas políticas sob o pretexto de "segurança algorítmica". Quando o dataset é oculto da sociedade, **a democracia deixa de ter poder sobre a definição de "verdade visual válida"**. Eis o perigo supremo: um mundo no qual **a percepção do real é mediada por modelos que não podem ser auditados, contestados ou substituídos**. A batalha ética começa, portanto, **não no código-fonte, mas na curadoria radical da matéria-prima humana usada para fabricá-lo**.

Portanto, os datasets de treinamento de detectores de deepfake representam o **lugar de maior risco e de maior poder no ecossistema da segurança informacional contemporânea**. Controlá-los exige **ética aplicada, ciência da informação crítica, filosofia da tecnologia e arcabouço jurídico internacional robusto**. O presente artigo defende que **não há tecnologia legítima de proteção contra deepfakes que não seja acompanhada de uma governança ética explícita da coleta, curadoria e auditoria de datasets**. Sem isso, qualquer modelo — por mais tecnicamente impressionante — será **um simulacro de proteção construído sobre violação silenciosa e viés estrutural inconfessável**.

3. VIÉS ALGORÍTMICO E DISPARIDADE SOCIOTÉCNICA NA DETECÇÃO DE DEEPFAKES

O viés algorítmico presente em detectores de deepfakes não é um efeito colateral isolado, mas a consequência direta da maneira como datasets são estruturados — e, sobretudo, da lógica histórica que define **quais corpos são considerados normativos e quais são tratados como exceções estatísticas**. Isto significa que a IA **não falha "por acidente" em determinados grupos**; ela falha porque **foi treinada para reproduzir uma estrutura pré-existente de visibilidade seletiva do mundo**. Estudos conduzidos por Buolamwini e Gebru (2018), inicialmente em sistemas de reconhecimento facial, demonstraram que modelos líderes de mercado atingiam **99% de acurácia para homens brancos**, mas **menos de 70% para mulheres negras**, evidenciando uma **hierarquização algorítmica da importância humana**. Essa assimetria já foi flagrada também em modelos de detecção de deepfakes, especialmente nos primeiros datasets públicos, nos quais **rostos caucasianos recebiam mais de 70% da representatividade** — e portanto, **eram os mais protegidos**. A tecnologia, assim, **não protege igualmente — protege primeiro quem é estatisticamente dominante**.

Esse fenômeno cria um paradoxo de injustiça estrutural: populações que historicamente foram mais expostas a violências visuais e representacionais — mulheres, pessoas negras, povos indígenas, grupos politicamente oprimidos — **são exatamente as mais vulneráveis no contexto das deepfakes**, seja por **exposição maior ao ataque**, seja por **detecção tardia ou ineficaz de manipulações** realizadas contra elas. Há casos documentados pelo European Digital Media Observatory (2021) em que **mulheres racializadas sofreram ataques pornográficos deepfake,**

mas os detectores corporativos classificaram o conteúdo como "natural", por pura ausência de exemplos similares no dataset de treinamento. Isso significa que o **risco real não é apenas técnico — é moral e estrutural**. Uma tecnologia vendida como "protetora" pode estar, de forma invisível, **operando como mecanismo de continuidade de violência algorítmica histórica**.

A assimetria também se manifesta em dimensão geopolítica. A maioria absoluta dos datasets utilizados para treinar detectores modernos é **ocidentalizada e monolíngue**, com baixíssima representação de contextos do **sul global**, como África, América Latina ou Sudeste Asiático. Isso gera o efeito descrito por Milan e Treré (2019) como **colonialismo de dados**, no qual "a verdade computacional é definida nos hemisférios hegemônicos e importada como padrão civilizatório global". Consequentemente, **os mesmos modelos têm alta performance ao detectar deepfakes em figuras públicas norte-americanas, mas falham gravemente em políticos africanos ou indígenas brasileiros**. Como consequência, **a segurança informacional torna-se profundamente assimétrica entre civilizações**: a IA detecta mais e melhor aquilo que interessa às potências de infraestruturas centrais. Isso comprova que **os datasets não são apenas conjuntos de imagens — são projeções estratégicas de poder global**.

Outro fenômeno crítico é o chamado **viés de confirmação retroalimentado por IA**. Modelos de detecção podem ser inclinados a **detectar com maior rigor manipulações associadas a ideologias, etnias ou posturas políticas previamente estigmatizadas** — amplificando sistemas de crença já presentes na sociedade. Ou seja: **a IA não é neutra — ela pode funcionar como espelho ampliado de preconceitos culturais invisíveis**, disfarçados de objetividade matemática. Em cenários extremos, isso poderia permitir o uso de detectores como **ferramentas de supressão seletiva de narrativas incômodas**, sob pretexto de "risco forense", "anomalia", "conteúdo suspeito". A "verdade automática" deixa de ser apenas tecnologia — ela se torna **infraestrutura de poder interpretativo**.

O combate ao viés, portanto, **não pode ser feito apenas por ajustes estatísticos tardios** — como "balanceamento matemático" ou "pesos compensatórios" nos modelos. A raiz do problema exige **reforma ética radical na origem do dataset**, garantindo **representatividade multirracial, diversidade fisiológica real, equilíbrio de gênero e inclusão deliberada de composições sociais sub-representadas**. A ética do treinamento, portanto, não é departamento opcional — é **pilar de soberania cognitiva da democracia**. Se a detecção automática será o filtro do real, então **o real precisa estar integralmente representado desde a formação do modelo**.

4. PRIVACIDADE, CONSENTIMENTO E A EXPROPRIAÇÃO BIOMÉTRICA NO TREINAMENTO DE IA

Se há um ponto onde a detecção de deepfakes revela sua contradição civilizacional mais grave, ele está na fronteira entre **proteção informacional** e **violação massiva da privacidade humana**. Para treinar redes neurais capazes de reconhecer manipulações visuais hiper-realistas, é necessário alimentá-las com **quantidades colossais de dados faciais reais**, registrando expressões, padrões microvasculares, reflexos pupilares, geometria óssea e milhares de outras variações fisiológicas que

compõem a individualidade corporal. O problema ético nasce do fato de que **a imensa maioria desses dados é coletada sem qualquer consentimento explícito e informado**, sendo extraída silenciosamente de redes sociais, bancos audiovisuais públicos, leaks criminais e até plataformas jornalísticas — muitas vezes sob a lógica tácita de que "se está na internet, é tecnicamente disponível". Ou seja: **modelos treinados para "defender identidades" normalmente começam violando identidades.**

Esse fenômeno tem sido descrito por especialistas como **expropriação biométrica silenciosa**, uma nova forma de colonialismo digital na qual rostos humanos são transformados em matéria-prima computacional sem consulta, indenização ou governança. Diferente de imagens genéricas, rostos carregam não apenas a identidade de uma pessoa, mas sua **possibilidade existencial no espaço público** — sua capacidade de ser reconhecida ou confundida, celebrada ou destruída. Roubar um rosto não é copiar uma foto — **é sequestrar o direito à própria existência digital**. E quando o mesmo dataset é utilizado para treinar sistemas de vigilância e de detecção de deepfakes, **a fronteira entre proteção e controle se dissolve perigosamente. O que impede que um modelo "protetor" se torne, no mesmo segundo, um mecanismo de rastreio massivo ou censura prévia?**

A dimensão ética se agrava porque não estamos falando de privacidade no sentido trivial (gosto pessoal, preferências, localização), mas de **bioprivacidade** — ou seja, da posse legítima sobre sinais corporais impossíveis de serem substituídos. Senhas podem ser trocadas; rostos, não. Portanto, qualquer sistema que capture traços faciais, mesmo sob pretexto de segurança, **assume potência estrutural de dominação irreversível**, caso não seja regido por um protocolo jurídico e ético de soberania. Um ambiente onde empresas ou governos podem extrair, manipular e treinar modelos com dados faciais sem autorização **é um ambiente onde o corpo humano deixa de ser pessoa e torna-se recurso computacional**.

Nesse contexto, a ausência de **consentimento ativo, específico e auditável** deixa de ser falha legal — é violação ontológica. Acompanhando a tese de Luciano Floridi (2020), o corpo digitalizado torna-se **metafenômeno informacional de dignidade humana** — e qualquer extração não consensual de sua representação constitui **agressão metacognitiva**, mesmo que alegadamente "para protegê-lo". Afirmar que "a tecnologia de detecção é necessária para proteger a democracia" só é verdadeiro **se essa proteção não for construída sobre a violação silenciosa e irreversível do próprio cidadão**.

Por isso, este artigo sustenta com firmeza: **nenhum modelo de detecção de deepfakes pode ser eticamente validado se seu dataset base violar direitos fundamentais de identidade e autodeterminação informacional**. A guerra contra a manipulação da realidade não pode ser vencida à custa da destruição da soberania do indivíduo sobre seu próprio corpo digital. A **ética do consentimento e da governança dos dados biométricos** não é detalhe técnico — **é fronteira civilizacional definitiva**.

5. DIRETRIZES PARA TREINAMENTO ÉTICO E GOVERNANÇA RESPONSÁVEL DE DATASETS

Diante do risco de que detectores de deepfakes se tornem simultaneamente **armas de proteção e opressão**, a questão crucial passa a ser: **como projetar e treinar redes neurais com base em princípios éticos verificáveis, transparentes e auditáveis em escala global?** A resposta não pode residir em genéricos "ajustes técnicos", mas em **protocolos de governança radicalmente orientados por justiça informacional e soberania biométrica**. Isso implica que a curadoria de datasets não é um subproduto da engenharia — **é uma responsabilidade política de primeira ordem**, equiparável à organização de eleições, à autenticação documental e à proteção constitucional de direitos fundamentais. Logo, deve seguir **princípios não negociáveis**.

O primeiro desses princípios é o da **representatividade demográfica e fenotípica obrigatória**. Um dataset que concentre excessivamente rostos europeus, masculinos, adultos, normatizados por estética ocidental — **está condenado à injustiça técnica e política desde sua origem**. Portanto, modelos de detecção de deepfakes precisam ser treinados em bases **geograficamente distribuídas**, com ampla inclusão **afro-diaspórica, indígena, asiática, árabe, latina e diversa em gênero, idade e expressão identitária**. Não se trata de "diversidade decorativa", mas de **garantir que a IA reconheça o humano em todas as suas formas**, e não apenas no recorte dominante. A proteção algorítmica só será democrática se a **experiência humana for integralmente representada na sua fonte de treinamento.**

O segundo princípio é o do **consentimento biométrico expresso, granular e revogável**, sustentado por mecanismos auditáveis de rastreabilidade de origem. Isso significa que datasets devem ser construídos com **participação consciente dos indivíduos**, com **direito documental de retirada posterior**, e com modelos de uso que devem ser **legivelmente explicáveis a qualquer cidadão, não apenas a especialistas em IA**. A extração silenciosa de rostos de redes sociais públicas é **eticamente inválida e civilizacionalmente perigosa**, ainda que tecnicamente "eficiente". Um sistema que promete proteger o usuário **não pode começar violando sua autonomia irreversivelmente**.

O terceiro princípio é a adoção obrigatória do **protocolo de documentação e auditabilidade** — nos moldes do *Datasheets for Datasets* (Gebru et al., 2018) —, em que **cada dataset deve declarar explicitamente sua origem, composição demográfica, fontes, propósitos, limitações e riscos potenciais**. Esse modelo impede que conjuntos de dados se tornem **caixas-pretas geopolíticas**, invisíveis à sociedade, protegidas por cláusulas corporativas. A documentação não deve ser "padrão opcional" — **deve ser exigência legal e internacional**, como passaporte informacional obrigatório para qualquer sistema que interfira na percepção pública do real.

O quarto princípio é a **participação multidisciplinar antes do código**. A definição do que é "viés", "aceitável", "fraude", "normal" **não pode ser definida exclusivamente por engenheiros ou investidores corporativos**. É necessário que **filósofos, especialistas em direitos humanos, juristas, comunicólogos e representantes de grupos vulneráveis** estejam formalmente envolvidos **na fase de definição do dataset**, e não apenas nos relatórios finais. Detectores de deepfakes **não são ferramentas neutras — são infraestruturas de poder perceptivo**. Logo, devem ser

desenhados sob processos **que representem a pluralidade real da sociedade.**

O quinto princípio é o da **interoperabilidade ética global** — de modo que protocolos de detecção não se tornem **armas geopolíticas disfarçadas de proteção democrática**. Isso exige **padronização internacional**, com regras verificáveis que impeçam o uso unilateral da tecnologia por governos autoritários ou corporações monopolistas, bem como **a execução de auditorias independentes em nível civilizacional**. A verdade digital não pode ser propriedade de players únicos.

Esses princípios, juntos, estabelecem as bases para aquilo que este artigo defende como o **pilar estruturante para uma Inteligência Artificial eticamente legítima no combate à desinformação: a governança radical dos datasets como condição absoluta de justiça informacional**.

6. AUDITORIA, TRANSPARÊNCIA E SOBERANIA ALGORÍTMICA COMO EIXOS NÃO-NEGOCIÁVEIS

A criação de detectores automáticos de deepfakes representa um ponto de inflexão na história da governança informacional digital. Se antes a disputa pela verdade ocorria na esfera interpretativa — entre jornalistas, instituições, testemunhos e consenso público — agora ela migra para uma camada **infraestrutural e algorítmica**, onde decisões são tomadas em milissegundos, muitas vezes antes que qualquer ser humano tenha acesso ao conteúdo. Esse deslocamento é tão radical que transforma a detecção automatizada não apenas em uma ferramenta de segurança, mas em uma **tecnologia de soberania epistêmica**, capaz de determinar o que "chega a existir" na esfera pública e o que é interceptado antes da circulação. Por isso, a questão fundamental não é apenas "a IA detecta corretamente?", mas sim "**quem controla a IA que detecta — e sob quais regras visíveis e verificáveis?**". Sem auditoria obrigatória, **o combate às deepfakes pode se converter no maior experimento de censura algorítmica invisível da história.**

A auditabilidade precisa ser compreendida como **direito democrático e princípio civilizacional**, não como recurso opcional de honestidade corporativa. Isso implica tornar obrigatória a criação de **mecanismos independentes e internacionais de revisão contínua** da performance, do viés e dos potenciais abusos dos detectores de deepfakes — com participação **não apenas técnica, mas também filosófica, jurídica, jornalística e representativa de grupos vulneráveis à opressão digital**. Os algoritmos que operam como "guardas da verdade" não podem ser julgados apenas por sua acurácia matemática, mas por sua **conformidade ética com os direitos humanos fundamentais**, incluindo liberdade de expressão, integridade informacional, privacidade e o direito à refutação pública.

Não há transparência legítima sem **documentação integral dos datasets, das regras de decisão do modelo e das políticas de fallback** (isto é, o que acontece quando a IA está em dúvida ou detecta risco). Em outras palavras: **é tão importante saber por que um vídeo foi bloqueado quanto saber por que outro foi aprovado**. O perigo reside justamente na assimetria: um modelo pode deliberadamente ser treinado para "falhar para alguns, proteger outros". É por isso que pesquisadores como Brundage e Whittlestone (2020) defendem a criação de **sistemas de accountability algorítmico com logs invioláveis — auditáveis por organismos civis —**

semelhantes ao funcionamento das caixas-pretas da aviação. Sem isso, qualquer arquitetura de detecção pode ser programada para parecer justa — enquanto perpetua injustiças silenciosamente.

A questão se agrava quando se reconhece a dimensão geopolítica do problema. **Quem controla a detecção automática controla a arma mais poderosa da era informacional: o poder de autorizar a existência pública de uma verdade**. Se apenas EUA, China ou conglomerados privados do Vale do Silício controlarem os pipelines de validação visual do mundo, então países do Sul Global serão **reduzidos a consumidores passivos de epistemologias computacionais impostas**, incapazes de contestar falsos positivos, falsos negativos ou censura seletiva. Isso cria o risco de um **colonialismo algorítmico**, onde um vídeo real produzido no Brasil pode ser bloqueado automaticamente por um modelo treinado nos EUA que não reconhece padrões expressivos brasileiros como autênticos. **Sem soberania algorítmica, não há soberania nacional nem soberania informacional.**

Assim, este artigo defende o princípio da **soberania algorítmica distribuída** como imperativo estratégico. Isso significa que modelos de detecção devem ser **governados por consórcios públicos e multilaterais**, com direito explícito de auditoria por universidades, imprensa livre e organismos de direitos humanos. Detectores não devem operar como "oráculos proprietários", mas como **infraestruturas auditáveis de confiança pública**, com mecanismos de contestação institucionalizados, logs acessíveis sob critérios legais controlados e protocolos claros para correção e revisão. A IA não pode ter a palavra final sobre a realidade — **ela deve ser sujeita à contestação humana garantida por lei**.

Mais do que isso: a defesa real da sociedade não está apenas na IA que detecta deepfakes — **mas no direito da sociedade de verificar a IA que detecta deepfakes**. Em um contexto onde até a proteção pode se tornar arma, o verdadeiro poder está na **transparência radical, no controle distribuído e na eliminação do privilégio opaco sobre a definição do real**. Nenhuma democracia está segura se a verdade visual for terceirizada a máquinas inacessíveis e não contestáveis.

Portanto, auditoria, transparência e soberania não são complementos opcionais da detecção de deepfakes — **são as condições mínimas para que qualquer sistema assim seja sequer moralmente permitido existir**. **Sem governança auditável, toda tecnologia de defesa se converte em risco potencial de tirania infraestrutural.** Este é o ponto-limite: a guerra pela verdade não será vencida com algoritmos mais fortes — **mas com algoritmos mais justos, radicalmente auditáveis e politicamente subordinados ao princípio da dignidade humana.**

7. IMPLICAÇÕES SOCIAIS, POLÍTICAS E FUTUROS POSSÍVEIS PARA UMA DETECÇÃO ÉTICA DE DEEPFAKES

O enfrentamento ético da desinformação algorítmica exige compreender que a crise das deepfakes é apenas a manifestação mais visível de um problema mais profundo: **a substituição progressiva da mediação humana pela automação decisória nas estruturas de validação da verdade**.

Detectores automáticos não são neutros — são sistemas políticos travestidos de técnica — e o

modo como são desenhados, treinados e auditados **determinará se viveremos em uma democracia informacional ou em um regime de tecnocracia epistêmica**. A história mostra que cada nova infraestrutura de comunicação reconfigura o poder. Se o século XX foi marcado pela luta pelo controle da mídia, o século XXI será definido pela luta **pelo controle da inteligência que decide o que é informação legítima**. Nesse cenário, a ética dos datasets e das redes neurais não é um apêndice de governança — **é a própria arena onde o destino da liberdade cognitiva humana está sendo decidido.**

A sociedade contemporânea vive o paradoxo de depender de tecnologias que não compreende, mas das quais depende para manter sua própria segurança. Essa dependência cria o risco do que Floridi (2020) chama de "**heteronomia algorítmica**" — a delegação irreversível do julgamento moral para sistemas automáticos. Se a detecção de deepfakes for tratada como monopólio técnico, ela produzirá não apenas novas assimetrias de informação, mas **novas desigualdades ontológicas**, nas quais diferentes grupos humanos terão diferentes graus de acesso à confiança pública. É o retorno silencioso de uma **casta epistêmica digital**, formada por quem detém a infraestrutura da verificação. A defesa ética contra deepfakes, portanto, não pode limitar-se ao âmbito da IA — **deve reconfigurar a própria relação entre tecnologia, verdade e democracia.**

Nesse contexto, surge a necessidade de incorporar **educação algorítmica** como política pública. Cidadãos precisam compreender o que é uma rede neural, como se forma um dataset e quais são os limites do que uma IA pode decidir. Isso implica um novo tipo de alfabetização, não apenas digital, mas **epistêmica**, para que indivíduos possam participar conscientemente das decisões que moldam sua própria realidade informacional. Sem essa base, o poder algorítmico permanecerá concentrado nas mãos de poucos, e o discurso da "proteção contra deepfakes" servirá apenas para consolidar novas formas de controle. A **transparência técnica precisa ser acompanhada de participação cidadã** — ou a auditoria pública será uma ficção.

Além disso, a governança ética da detecção de deepfakes requer o estabelecimento de **instituições globais de regulação e certificação**, nos moldes da OMS ou da ONU, mas voltadas à integridade informacional. Tais organismos precisariam definir padrões universais de coleta ética de dados, protocolos de auditoria de modelos, sanções para uso indevido e mecanismos de cooperação transnacional para compartilhamento de conhecimento forense. Só assim seria possível **impedir a privatização da verdade** e evitar que a defesa informacional se converta em guerra fria algorítmica.

Essa infraestrutura de governança deve ser acompanhada por legislação que reconheça **a autenticidade computacional como direito humano emergente**, garantindo que toda pessoa tenha o direito de existir digitalmente sem ser falsificada, manipulada ou extraída sem consentimento.

A detecção ética de deepfakes, portanto, precisa ser compreendida como **campo interdisciplinar e transnacional**, que articula engenharia, comunicação, direito, filosofia e ética aplicada. É preciso reconhecer que cada dataset é também uma narrativa sobre o humano — e que treinar uma IA é, de certo modo, **escrever uma nova ontologia da humanidade**. Esse reconhecimento implica responsabilidade histórica: quem treina uma rede neural não está apenas ensinando uma máquina a

ver — está ensinando o mundo a reconhecer-se através dela. O futuro da verdade, da confiança e até da noção de identidade coletiva dependerá da capacidade de a sociedade **instituir princípios técnicos e morais equivalentes em importância aos direitos civis da modernidade.**

Por fim, a ética no treinamento de redes neurais não é uma opção idealista, mas a única estratégia de sobrevivência cognitiva da civilização informacional. O mundo das deepfakes mostra que **a tecnologia é capaz de dissolver as fronteiras entre o falso e o verdadeiro mais rápido do que nossa capacidade institucional de reagir**. A única defesa possível é antecipatória: criar infraestruturas de IA que incorporem a moralidade na sua própria arquitetura — o que Floridi chama de "ética por design". Sem isso, a humanidade viverá eternamente sob o regime da suspeita e do caos perceptivo. **A ética dos datasets é, portanto, o novo contrato social da era digital: quem controla a formação da IA controla a forma do real.**

CONCLUSÃO

A crise das deepfakes expõe, com violência inédita, o colapso das fronteiras tradicionais entre verdade, imagem e poder. O que antes era apenas uma disputa discursiva — travada entre jornalismo, política e ciência — hoje se converteu em uma guerra infraestrutural, onde a própria realidade é mediada por algoritmos treinados em dados que refletem, reproduzem e amplificam as desigualdades do mundo. O artigo demonstrou que o cerne dessa disputa não é tecnológico, mas ético: não se trata de inventar máquinas mais inteligentes, mas de **ensinar às máquinas o que significa agir com justiça epistemológica**. O treinamento ético de redes neurais, portanto, não é um luxo acadêmico, mas uma questão de sobrevivência civilizatória. A proteção contra deepfakes só será legítima quando a verdade deixar de ser monopólio técnico e voltar a ser um bem público compartilhado, sustentado por transparência, auditabilidade e consentimento.

As reflexões aqui desenvolvidas evidenciam que a noção de ética em IA precisa ser deslocada do campo das intenções para o campo das infraestruturas. A injustiça não nasce apenas nas decisões humanas sobre tecnologia — nasce no código, no dataset, naquilo que é invisível ao cidadão comum. Por isso, um sistema de detecção de deepfakes só será moralmente válido se for **tecnicamente preciso, socialmente justo e politicamente auditável**. Isso exige um novo contrato social entre ciência, Estado e sociedade civil, em que o acesso a datasets, algoritmos e critérios de decisão seja tratado como direito democrático, não como segredo industrial. A IA que detecta falsidades precisa ser também capaz de revelar a verdade sobre si mesma.

No horizonte geopolítico, este estudo reforça que o domínio sobre os datasets é o novo eixo de soberania das nações. Assim como no passado as potências controlavam território e energia, no século XXI o poder é definido pela **posse e pela ética dos dados**. Uma civilização que aceita datasets enviesados, opacos e extrativistas como base de sua infraestrutura de verdade aceita, na prática, viver sob hegemonia epistemológica. A soberania algorítmica é, portanto, a nova fronteira da independência informacional: sem ela, países inteiros serão apenas consumidores de modelos estrangeiros que decidem, de forma silenciosa, o que é autêntico e o que não existe. A ética do

treinamento é, nesse sentido, também **política de Estado e defesa nacional.**

Entretanto, a verdadeira batalha é interior à própria cultura contemporânea: a de reconectar tecnologia e responsabilidade moral. O avanço da IA revela que a humanidade só será capaz de viver com máquinas inteligentes se aprender a **reintroduzir valores humanos no centro da computação**. A ética dos datasets é uma forma de reeducação civilizacional — o esforço de lembrar que dados não são neutros, que imagens não são inocentes e que decisões automatizadas não estão acima da moral. O treinamento ético é o antídoto contra a indiferença algorítmica. Ele transforma o código em compromisso, o modelo em instituição e a técnica em extensão da dignidade humana.

Conclui-se, portanto, que a detecção ética de deepfakes deve ser entendida como um novo capítulo do Iluminismo — um Iluminismo computacional — no qual a razão é reprogramada com moralidade. Nenhum avanço técnico terá valor se não for acompanhado de reflexão ética e regulação democrática. O futuro da verdade depende de um gesto simples, porém revolucionário: **ensinar as máquinas a respeitar o humano que as criou**. Somente assim será possível transformar a inteligência artificial em aliada da verdade, e não em cúmplice do esquecimento.

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.

BRUNDAGE, Miles; WHITTLESTONE, Jess. *Governance of Artificial Intelligence: Ethics, Law and Policy*. Oxford Internet Institute, 2020.

BUOLAMWINI, Joy; GEBRU, Timnit. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. Proceedings of Machine Learning Research, 2018.

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

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

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

GEBRU, Timnit et al. *Datasheets for Datasets*. Proceedings of the Conference on Fairness, Accountability, and Transparency, 2018.

MILAN, Stefania; TRERÉ, Emiliano. *The Rise of Data Colonialism: Reclaiming Digital Sovereignty*. Social Media + Society, 2019.

NATO STRATCOM. *Deepfake Detection and Information Integrity Report*. Brussels, 2021.

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

ZUBOFF, Shoshana. *The Age of Surveillance Capitalism*. New York: PublicAffairs, 2019.