REVISTA OWL (*OWL Journal*)
www.revistaowl.com.br – ISSN: 2965-2634

DESINFORMAÇÃO DIGITAL E INTELIGÊNCIA ARTIFICIAL: DESAFIOS ÉTICOS E EPISTEMOLÓGICOS PARA A VERIFICAÇÃO AUTOMATIZADA DE INFORMAÇÕES

DIGITAL DISINFORMATION AND ARTIFICIAL INTELLIGENCE: ETHICAL AND EPISTEMOLOGICAL CHALLENGES FOR AUTOMATED INFORMATION VERIFICATION

DESINFORMACIÓN DIGITAL E INTELIGENCIA ARTIFICIAL: DESAFÍOS ÉTICOS Y EPISTEMOLÓGICOS PARA LA VERIFICACIÓN AUTOMATIZADA DE INFORMACIÓN

DOI: 10.5281/zenodo.20708180

Mislene Dalila da Silva[1]
Adriana Cristina Omena dos Santos[2]
Lucas Nunes Fonseca Costa[3]

1 Doutora em Engenharia Elétrica pela Universidade Federal de Uberlândia (UFU). Mestre em Tecnologia, Comunicação e Educação (UFU). Professora e pesquisadora do Centro Universitário de Patos de Minas – UNIPAM, Patos de Minas, Minas Gerais – Brasil. E-mail: mislenedalila@gmail.com. Lattes: http://lattes.cnpq.br/5593440909044581. ORCID: https://orcid.org/0000-0002-5657-4754
2 Doutora em Ciências da Comunicação pela Universidade de São Paulo (USP). Mestre em Ciências da Comunicação (USP). Professora e pesquisadora da Universidade Federal de Uberlândia – UFU, Uberlândia, Minas Gerais – Brasil. E-mail: adriomena@gmail.com. Lattes: http://lattes.cnpq.br/151543372591481. ORCID: https://orcid.org/0000-0002-8863-6219
3 Graduando em Sistemas de Informação pelo Centro Universitário de Patos de Minas – UNIPAM, Patos de Minas, Minas Gerais – Brasil. E-mail: lucasnfc@unipam.edu.br. Lattes: https://lattes.cnpq.br/8975660426439528. ORCID: https://orcid.org/0009-0002-5475-7538

REVISTA OWL (*OWL Journal*)
www.revistaowl.com.br – ISSN: 2965-2634

RESUMO

Este artigo analisa os desafios éticos e epistemológicos impostos pela proliferação da desinformação em ambientes digitais e pelo uso da Inteligência Artificial (IA) como mecanismo de verificação automatizada de informações. A pesquisa, de natureza qualitativa e exploratória, combina revisão de literatura com o desenvolvimento e a análise de uma plataforma de verificação de conteúdo, Veridix, baseada no modelo de linguagem Gemini. Os resultados indicam que, embora a IA apresente potencial significativo para a detecção de inconsistências informacionais, seu uso suscita questões éticas fundamentais relacionadas ao viés algorítmico, à opacidade dos modelos e à responsabilidade epistêmica dos sistemas automatizados. Argumenta-se que a integração responsável da IA no combate à desinformação requer marcos éticos robustos, transparência metodológica e letramento informacional crítico por parte dos usuários.

Palavras-chave: desinformação; inteligência artificial; ética algorítmica.

ABSTRACT

This article examines the ethical and epistemological challenges posed by the proliferation of *disinformation* in digital environments and by the use of Artificial Intelligence (AI) as a mechanism for automated information verification. The research, qualitative and exploratory in nature, combines a literature review with the development and analysis of a content verification platform, Veridix, based on the Gemini language model. Results indicate that, although AI presents significant potential for detecting informational inconsistencies, its use raises fundamental ethical questions related to algorithmic bias, model opacity, and the epistemic responsibility of automated systems. It is argued that the responsible integration of AI in combating disinformation requires robust ethical frameworks, methodological transparency, and critical information literacy on the part of users.

Keywords: *disinformation*; artificial intelligence; algorithmic ethics.

RESUMEN

Este artículo analiza los desafíos éticos y epistemológicos impuestos por la proliferación de la desinformación en entornos digitales y por el uso de la Inteligencia Artificial (IA) como mecanismo de verificación automatizada de información. La investigación, de carácter cualitativo y exploratorio, combina revisión de literatura con el desarrollo y el análisis de una plataforma de verificación de contenido , Veridix, basada en el modelo de lenguaje Gemini. Los resultados indican que, aunque la IA presenta un potencial significativo para detectar inconsistencias informacionales, su uso plantea preguntas éticas fundamentales relacionadas con el sesgo algorítmico, la opacidad de los modelos y la responsabilidad epistémica de los sistemas automatizados. Se argumenta que la integración responsable de la IA en la lucha contra la desinformación requiere marcos éticos sólidos, transparencia metodológica y alfabetización informacional crítica por parte de los usuarios.

Palabras clave: desinformación; inteligencia artificial; ética algorítmica.

1. INTRODUÇÃO

A proliferação de conteúdos falsos ou distorcidos em ambientes digitais constitui um dos fenômenos mais desafiadores da contemporaneidade para a produção, a circulação e a legitimação do conhecimento. A desinformação, compreendida aqui como a disseminação intencional de informações inverídicas ou enganosas, não representa apenas um problema comunicacional, mas uma ameaça concreta à credibilidade da ciência e à integridade dos processos democráticos (WARDLE; DERAKHSHAN, 2017). Esse cenário é amplificado pela velocidade e pelo alcance das plataformas digitais, que transformam rumores em narrativas de massa em questão de minutos.

Nesse contexto, a Inteligência Artificial (IA) emerge simultaneamente como parte do problema e como possível solução. Por um lado, modelos generativos de linguagem ampliam a capacidade de produzir textos sintéticos convincentes e de difundir desinformação em escala industrial (HARARI, 2024; SANTAELLA; KAUFMAN, 2024). Por outro, algoritmos de classificação e verificação automatizada prometem identificar inconsistências informacionais com rapidez e precisão superiores à capacidade humana individual. Essa ambivalência coloca a IA no centro de um debate ético urgente: como garantir que ferramentas de verificação sejam transparentes, confiáveis e eticamente responsáveis?

O presente artigo parte desse problema para investigar dois eixos inter-relacionados: (a) os desafios éticos e epistemológicos da verificação automatizada de informações em ambientes digitais; e (b) as possibilidades e limitações da IA na construção de soluções para o combate à desinformação. Para tanto, analisa-se o desenvolvimento e os resultados preliminares da plataforma Veridix, um sistema web de verificação de conteúdo baseado em IA, como estudo de caso concreto que materializa essas tensões.

A pergunta central que orienta a investigação é: quais são os limites éticos e epistemológicos do uso de sistemas baseados em Inteligência Artificial para a verificação de

informações em conteúdos digitais, e como esses limites se manifestam na prática de desenvolvimento de uma plataforma como o Veridix?

Este trabalho insere-se no campo das pesquisas em humanidades digitais e ética em pesquisa, dialogando com os debates recentes sobre integridade científica, desinformação e regulação da IA (MAINARDES, 2022; ROSSETTI; ANGELUCI, 2021). Sua contribuição reside na articulação entre a dimensão técnica do desenvolvimento de ferramentas de verificação e a perspectiva ética e epistemológica que deve orientar seu uso responsável.

2.REFERENCIAL TEÓRICO

2.1 Desinformação, Pós-Verdade e Credibilidade Científica

A desinformação contemporânea não pode ser reduzida ao fenômeno coloquialmente denominado "*fake news*". Como aponta Waisbord (2018), o termo é conceitualmente impreciso e politicamente instrumentalizado: ao agrupar sob uma mesma expressão fenômenos distintos, como erros jornalísticos involuntários, sátira política e propaganda deliberada, o conceito obscurece mais do que ilumina as dinâmicas da desinformação contemporânea. Wardle e Derakhshan (2017) avançam nessa crítica ao propor uma tipologia mais precisa, que distingue *misinformation* (informação falsa disseminada sem intenção de dano), *disinformation* (informação falsa disseminada com intenção deliberada de enganar) e *malinformation* (informação verdadeira usada para causar dano). Essa distinção é epistemicamente relevante porque implica responsabilidades éticas distintas para produtores, plataformas e verificadores de conteúdo.

No contexto brasileiro, a disseminação de desinformação tem afetado de modo documentado o campo científico, com narrativas contrárias a vacinas, ao aquecimento global e a tratamentos médicos consolidados (MASSARANI et al., 2021). Esse fenômeno compromete o que Mainardes (2022) denomina "dimensão ética estruturante" da produção do conhecimento: a responsabilidade de pesquisadores e instituições com a honestidade

intelectual e com a transparência metodológica não se limita ao âmbito acadêmico, mas estende-se à responsabilidade social com o uso e a circulação pública do conhecimento produzido.

A noção de pós-verdade, in*corpora*da ao *Oxford Dictionary* em 2016, *design*a um ambiente epistêmico em que apelos emocionais e crenças pessoais tendem a exercer mais influência sobre a opinião pública do que fatos objetivos. As consequências democráticas desse fenômeno são significativas: Maranhão et al. (2026) demonstram que as fake news moldam percepções, distorcem a realidade e fragilizam instituições ao influenciar negativamente a opinião pública, colocando em risco os fundamentos da democracia deliberativa. Esse ambiente cria condições propícias para que a desinformação prospere, pois fragiliza os critérios compartilhados de avaliação da confiabilidade informacional.

2.2 Inteligência Artificial, Viés Algorítmico e Ética

A IA generativa, especialmente os grandes modelos de linguagem (*LLMs*), representa um salto qualitativo na capacidade de produzir textos sintéticos indistinguíveis de textos humanos. Harari (2024) argumenta que essa capacidade transforma profundamente as redes de informação humanas, pois, pela primeira vez na história, agentes não humanos são capazes de produzir narrativas culturalmente coerentes em escala industrial. Essa mudança exige novas abordagens éticas e regulatórias que transcendem os marcos legais existentes.

Santaella e Kaufman (2024) propõem que a IA generativa representa uma "quarta ferida narcísica" para a humanidade: após Copérnico (não somos o centro do universo), Darwin (não somos a espécie privilegiada), Freud (não somos senhores de nossa própria mente) e, agora, a IA (não somos os únicos produtores de linguagem e cultura). Essa provocação filosófica tem consequências práticas para o debate sobre autoria, originalidade e integridade na produção de conhecimento.

O viés algorítmico constitui um dos desafios éticos mais relevantes para sistemas de verificação automatizada. Rossetti e Angeluci (2021) demonstram que algoritmos treinados

em *corpora* enviesados tendem a reproduzir e a amplificar preconceitos presentes nos dados de treinamento. No contexto da verificação de desinformação, isso significa que sistemas de IA podem sistematicamente classificar como falso conteúdo produzido por grupos sub-representados nos dados de treinamento, enquanto validam narrativas dominantes independentemente de sua veracidade. A dimensão regulatória desse problema é igualmente relevante: como demonstra Dorneles (2025), o combate à desinformação exige estratégias múltiplas que articulam regulação responsável de plataformas, promoção de alfabetização digital e incentivo à transparência, sem comprometer a liberdade de expressão. Esse risco é agravado pela opacidade dos modelos, a chamada "caixa-preta" algorítmica, que dificulta a auditabilidade e a responsabilização.

O'Neil (2020) descreve como algoritmos de classificação podem tornar-se instrumentos de destruição em massa quando seus critérios de avaliação não são transparentes, contestáveis e sujeitos a revisão democrática. Aplicada ao campo da verificação de conteúdo, essa perspectiva alerta para o risco de que plataformas de *fact-checking* automatizado concentrem poder epistêmico em entidades privadas, sem os freios e contrapesos necessários para garantir imparcialidade e equidade.

2.3 IA como Ferramenta de Combate à Desinformação: possibilidades e limites

A literatura recente sobre o uso de IA para detecção de desinformação revela um campo em rápido desenvolvimento, marcado por resultados promissores e por limitações estruturais significativas. Técnicas de processamento de linguagem natural (PLN), classificação de texto e análise de redes têm demonstrado capacidade de identificar padrões linguísticos associados à desinformação com precisão superior à de avaliadores humanos em experimentos controlados (ZHOU; ZAFARANI, 2020).

Contudo, pesquisadores apontam que a generalização desses resultados para ambientes reais é problemática por múltiplas razões: (a) a desinformação evolui mais rapidamente do que os modelos conseguem acompanhar; (b) contextos culturais e linguísticos específicos

exigem modelos localmente treinados; (c) a fronteira entre opinião legítima e desinformação é epistemicamente contestada; e (d) sistemas automatizados podem ser deliberadamente contornados por atores mal-intencionados (ZHOU; ZAFARANI, 2020; ROBL FILHO; MARRAFON; MEDON, 2022).

Robl Filho, Marrafon e Medon (2022) analisam especificamente o fenômeno das *deepfakes* [4]e das redes automatizadas de desinformação, argumentando que a IA utilizada para desinformar representa uma ameaça ao mercado livre de ideias e aos fundamentos da democracia deliberativa. Essa análise jurídica e filosófica ilumina a dimensão política da desinformação automatizada e reforça a necessidade de marcos regulatórios adequados. No Brasil, o Marco Civil da Internet (Lei 12.965/2014) representa o principal instrumento normativo vigente para a governança do ambiente digital, embora não contemple especificamente a regulação de sistemas de IA. No âmbito internacional, o AI Act europeu (Regulamento UE 2024/1689) inaugura um modelo de regulação baseado em risco que tem orientado debates regulatórios em diversas jurisdições (MONARI; OTT, 2025).

No campo educacional, a questão da desinformação articula-se ao debate sobre letramento midiático e informacional. A Estratégia Brasileira de Educação Midiática (SECOM, 2023) reconhece que a capacidade de avaliar criticamente fontes informacionais é uma competência cidadã fundamental na era digital, competência que ferramentas de verificação automatizada podem apoiar, mas não substituir.

2.4 Ética em Pesquisa e Integridade Científica na Era Digital

Mainardes (2022, 2023) argumenta que a ética não é um apêndice do processo de pesquisa, mas seu elemento estruturante: ela orienta todas as etapas, desde a formulação do problema até a disseminação dos resultados. Essa perspectiva, que o autor denomina ético-

4 *Deepfakes* são conteúdos audiovisuais sintéticos gerados por sistemas de inteligência artificial - especialmente por redes generativas adversariais (GANs) - com o objetivo de simular, de forma convincente, a aparência, a voz ou as expressões de pessoas reais. O termo deriva da combinação de deep learning (aprendizado profundo) e fake (falso).

ontoepistemológica, é especialmente relevante no contexto do desenvolvimento de sistemas de IA, pois implica que as escolhas técnicas de um desenvolvedor (qual modelo usar, em quais dados treinar, como apresentar os resultados) são, ao mesmo tempo, escolhas éticas.

A ausência de regulação específica para sistemas de IA no Brasil constitui uma lacuna normativa relevante para o campo aqui investigado. A Lei 14.874/2024, que institui o Sistema Nacional de Ética em Pesquisas com Seres Humanos, foi concebida para contextos de pesquisa biomédica e social, não para o desenvolvimento de sistemas automatizados de classificação informacional. Sua mobilização analógica neste campo é possível, mas requer cautela: os dados que treinam modelos de IA frequentemente são gerados por usuários ou coletados de fontes consideradas públicas, sem que haja consentimento informado equivalente ao exigido em pesquisas com seres humanos. Essa distinção aponta para a necessidade de marcos éticos próprios para o desenvolvimento de IA, atualmente em construção no Brasil por meio de projetos de lei em tramitação no Congresso Nacional.

A questão da integridade científica na era da IA generativa acrescenta uma camada adicional de complexidade ao debate. Conforme aponta Mainardes (2024), o crescimento do uso de ferramentas de IA para auxiliar na produção de textos acadêmicos coloca em xeque noções estabelecidas de autoria, originalidade e responsabilidade intelectual. Esse debate é constitutivo do dossiê em tela e fornece o horizonte ético dentro do qual o desenvolvimento de ferramentas como o Veridix deve ser compreendido.

3. METODOLOGIA

A presente pesquisa adota uma abordagem qualitativa de natureza exploratória e descritiva, articulando revisão de literatura com o desenvolvimento e a análise de um estudo de caso instrumental (STAKE, 1995): a plataforma Veridix. O estudo de caso instrumental é adequado quando o caso concreto serve para iluminar uma questão mais ampla: neste trabalho, os desafios éticos e epistemológicos da verificação automatizada de informações.

A revisão de literatura foi conduzida nas bases Scopus, Web of Science e SciELO, utilizando os descritores "*disinformation*", "*fake news* detection", "algorithmic ethics", "artificial intelligence ethics" e "information verification", com recorte temporal de 2017 a 2025. A busca inicial resultou em 312 registros. Após a remoção de duplicatas e a aplicação dos critérios de inclusão (estudos em português, inglês ou espanhol; publicados entre 2017 e 2025; com foco em desinformação, ética algorítmica ou verificação de conteúdo), foram selecionados 42 textos para leitura na íntegra. Destes, 16 foram incorporados ao referencial teórico do artigo, com base na relevância direta para os eixos investigados. Foram priorizados artigos com alto índice de citação e trabalhos de pesquisadores brasileiros que dialogam diretamente com o campo das humanidades.

O desenvolvimento do Veridix seguiu metodologia ágil baseada no *framework Scrum*, com *sprints* quinzenais documentadas. A escolha do modelo Gemini (Google) como motor de análise foi precedida por testes comparativos com modelos locais e outras interfaces de programação de aplicativos (APIs) disponíveis. Os critérios de avaliação incluíram: (a) precisão na detecção de inconsistências informacionais; (b) tempo de resposta; (c) transparência nos critérios de classificação; e (d) comportamento frente a conteúdos ambíguos.

A análise dos resultados combinou avaliação técnica de desempenho com análise crítica das implicações éticas das escolhas de *design*, em consonância com a perspectiva de que decisões técnicas são sempre, também, decisões éticas (MAINARDES, 2022; ROSSETTI; ANGELUCI, 2021).

Ética em Pesquisa: Este estudo não envolveu coleta de dados com seres humanos identificáveis. A plataforma Veridix não armazena conteúdos submetidos pelos usuários de forma vinculada a dados pessoais. Os princípios éticos que orientaram a pesquisa incluem transparência metodológica, responsabilidade com o impacto social da ferramenta desenvolvida e honestidade intelectual quanto às limitações do sistema. Conforme

preconizado pela ANPEd (2021), adota-se aqui a autodeclaração de princípios e procedimentos éticos como mecanismo de responsabilização.

4. RESULTADOS E DISCUSSÃO

4.1 O Veridix como Estudo de Caso: descrição da plataforma

O Veridix é uma plataforma web de análise de informações em conteúdos digitais, desenvolvida pela equipe de pesquisadores responsável por este estudo e acessível pelo endereço https://veridix.xlucazzz.dev/. O sistema permite que usuários submetam textos, URLs ou documentos para análise, recebendo como resposta: (a) um percentual estimado de confiabilidade informacional; (b) um resumo interpretativo do conteúdo analisado; e (c) a categorização temática do material. A arquitetura técnica combina Node.js no *backend*, React no *frontend* e MySQL para gestão de dados, tendo o modelo Gemini (Google) como núcleo da análise semântica.

A disponibilização pública da plataforma, de acesso gratuito, responde a uma demanda social identificada na literatura: ferramentas de verificação de conteúdo existentes são, em sua maioria, direcionadas a jornalistas e pesquisadores especializados, com interfaces pouco acessíveis ao cidadão comum (MASSARANI et al., 2021). O Veridix busca democratizar o acesso à verificação crítica de conteúdo, aspecto que, como se discutirá adiante, gera seus próprios desafios éticos.

4.2 Desafios Éticos Identificados no Desenvolvimento

O processo de desenvolvimento do Veridix revelou, em diferentes momentos, tensões éticas que merecem problematização à luz do referencial teórico mobilizado.

4.2.1 O problema da "alucinação" e da desinformação gerada pela própria IA

Durante os testes iniciais com uma versão anterior do modelo Gemini, o sistema produziu análises factualmente incorretas sobre os conteúdos submetidos à verificação; ou seja, a ferramenta destinada a combater a desinformação produziu desinformação. Esse fenômeno, denominado na literatura técnica de "alucinação" dos modelos de linguagem, tem implicações éticas profundas: um sistema que apresenta resultados com falsa confiança a usuários não especializados pode consolidar crenças equivocadas com aparência de autoridade algorítmica.

A solução adotada, que consistiu em migrar para a versão mais recente do Gemini e refinar as instruções fornecidas ao sistema, reduziu significativamente a incidência do problema, mas não o eliminou. Cabe considerar que parte das imprecisões pode estar relacionada ao viés linguístico dos dados de treinamento: modelos de linguagem de grande escala são predominantemente treinados em inglês, o que pode comprometer sua capacidade de interpretar contextos culturais e linguísticos específicos do português brasileiro. Essa hipótese aponta para uma direção promissora em pesquisas futuras: o teste comparativo com modelos treinados prioritariamente em português, como o Maritaca AI, poderia contribuir para avaliar em que medida o desempenho do sistema é afetado pela língua dos dados de treinamento. Isso ilustra uma limitação estrutural que O'Neil (2020) já havia identificado: algoritmos que produzem julgamentos com aparente objetividade podem conferir legitimidade a erros sistêmicos de modo mais insidioso do que avaliadores humanos, justamente porque sua opacidade dificulta a contestação.

4.2.2 A opacidade dos critérios de verificação

O modelo Gemini, como todos os grandes modelos de linguagem proprietários, opera como uma "caixa-preta": não é possível auditar completamente os critérios pelos quais classifica um conteúdo como inverídico ou confiável. Isso implica que o Veridix, ao apresentar percentuais de confiabilidade informacional ao usuário, pode criar uma aparência

de objetividade mensurável para um processo que é, na prática, opaco e potencialmente enviesado.

Rossetti e Angeluci (2021) argumentam que a ética algorítmica exige não apenas que algoritmos produzam bons resultados, mas que seus critérios de avaliação sejam explicáveis, contestáveis e sujeitos a revisão. A interface do Veridix, em sua versão atual, apresenta o resultado da análise sem expor a cadeia de raciocínio do modelo, limitação que deve ser tratada em versões futuras, especialmente se a plataforma se destinar a contextos educacionais ou jornalísticos.

4.2.3 Responsabilidade epistêmica e letramento do usuário

A democratização do acesso à verificação automática de conteúdo traz consigo o risco do que se poderia denominar "delegação epistêmica irrefletida": usuários que transferem a um algoritmo a responsabilidade de avaliar a confiabilidade de uma informação, sem desenvolver as competências críticas necessárias para questionar os próprios resultados da ferramenta. Esse risco é especialmente relevante em contextos educacionais, nos quais ferramentas de verificação automatizada podem ser incorporadas ao currículo sem o acompanhamento de uma formação voltada ao pensamento crítico.

A Estratégia Brasileira de Educação Midiática (SECOM, 2023) aponta na direção correta ao enfatizar que o letramento midiático não se reduz ao uso de ferramentas, mas envolve a capacidade de compreender seus mecanismos, questionar seus critérios e reconhecer suas limitações. Ferramentas como o Veridix devem ser concebidas, portanto, não como substitutos do julgamento humano, mas como andaimes cognitivos que ampliam a capacidade crítica sem substituí-la.

4.3 Possibilidades e Contribuições da IA no Combate à Desinformação

A despeito dos desafios éticos identificados, os resultados preliminares do Veridix demonstram que sistemas baseados em IA podem oferecer contribuições relevantes para o ecossistema de verificação de conteúdo. O sistema demonstrou capacidade de: (a) identificar marcadores linguísticos de desinformação em textos curtos com rapidez superior à da análise humana; (b) detectar inconsistências factuais em relação a informações consolidadas presentes no *corpus* de treinamento do modelo; e (c) apresentar os resultados em linguagem acessível a usuários não especializados.

Esses resultados estão em consonância com a literatura que aponta o potencial dos modelos de linguagem de grande escala para tarefas de verificação factual (ZHOU; ZAFARANI, 2020), ao mesmo tempo em que confirmam as limitações já documentadas: o desempenho é sensível à qualidade e ao recorte temporal do *corpus* de treinamento, e a fronteira entre opinião e desinformação permanece epistemicamente problemática para sistemas automatizados.

Do ponto de vista da pesquisa em humanidades, o Veridix é relevante não apenas como ferramenta, mas como objeto de estudo que materializa as tensões éticas do campo. Seu desenvolvimento revelou que a construção de sistemas de IA eticamente responsáveis não é um problema técnico que se resolve após a conclusão do projeto, mas um processo contínuo de reflexão que deve integrar todas as etapas do desenvolvimento, em consonância com a perspectiva ético-ontoepistemológica proposta por Mainardes (2022).

5.CONSIDERAÇÕES FINAIS

Este artigo buscou demonstrar que o combate à desinformação digital por meio de Inteligência Artificial é um campo atravessado por tensões éticas e epistemológicas que não podem ser resolvidas exclusivamente por avanços técnicos. A análise do desenvolvimento e dos resultados preliminares da plataforma Veridix permitiu identificar três desafios centrais: o problema da "alucinação" algorítmica e da desinformação produzida pelos próprios sistemas

de verificação; a opacidade dos critérios de classificação nos modelos de linguagem proprietários; e o risco de delegação epistêmica irrefletida por parte dos usuários.

Esses desafios não inviabilizam o uso de IA no combate à desinformação, mas delimitam as condições éticas de seu uso responsável. Articulando o referencial de Mainardes (2022, 2023) sobre ética em pesquisa com os debates contemporâneos sobre ética algorítmica (ROSSETTI; ANGELUCI, 2021; O'NEIL, 2020), argumenta-se que sistemas de verificação automatizada precisam ser: (a) transparentes quanto a seus critérios e limitações; (b) contestáveis pelos usuários afetados; (c) acompanhados de estratégias de letramento crítico; e (d) submetidos a marcos regulatórios democráticos, como o *AI Act* europeu e a Lei 14.874/2024 no Brasil.

Como limitação deste estudo, destaca-se que os resultados do Veridix foram avaliados apenas em testes funcionais controlados, sem validação sistemática com usuários em contextos reais de uso. Pesquisas futuras poderão investigar como diferentes perfis de usuários interpretam e utilizam os resultados da plataforma, bem como desenvolver métricas mais refinadas para avaliar o impacto de ferramentas de verificação automatizada sobre o comportamento informacional de populações específicas. Aponta-se, ainda, a necessidade de in*corpora*r etapas de curadoria e classificação humana ao processo de aprendizado do sistema: a supervisão especializada na rotulação de conteúdos permite corrigir vieses do modelo, ampliar a diversidade dos dados de treinamento e aprimorar progressivamente a precisão das análises, aproximando o desempenho automatizado de padrões editoriais e científicos de verificação.

A contribuição central deste trabalho reside na articulação, raramente realizada na literatura nacional, entre a dimensão técnica do desenvolvimento de sistemas de verificação de conteúdo e a perspectiva ética e epistemológica que deve orientar sua concepção, implementação e uso. Essa articulação é indispensável para que a IA se torne, de fato, uma aliada na construção de ambientes informacionais mais confiáveis e democráticos.

REFERÊNCIAS

BRASIL. Lei n.º 12.965, de 23 de abril de 2014. Institui o Marco Civil da Internet. Brasília: Presidência da República, 2014. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 12 mar. 2025.

BRASIL. Secretaria de Comunicação Social da Presidência da República. Estratégia Brasileira de Educação Midiática. Brasília: SECOM, 2023.

DORNELES, Elton Roberto. Entre likes e fake news: a liberdade de expressão em tempos de desinformação. Revista OWL (OWL Journal), Campina Grande, v. 3, n. 3, p. 1-18, 2025. Disponível em: https://www.revistaowl.com.br/index.php/owl/article/view/423. Acesso em: 15 fev. 2025.

HARARI, Yuval Noah. Nexus: uma breve história das redes de informação, da Idade da Pedra à inteligência artificial. São Paulo: Companhia das Letras, 2024.

MAINARDES, Jefferson. Ética em pesquisa: princípios, aspectos problemáticos e agenda de pesquisa. Arquivos Analíticos de Políticas Educativas, v. 30, n. 146, p. 3-21, 2022. Disponível em: https://doi.org/10.14507/epaa.30.7436. Acesso em: 10 jan. 2025.

MAINARDES, Jefferson. Ética, integridade e cultura de integridade: reflexões a partir do contexto brasileiro. Horizontes, Itatiba, v. 41, n. 1, p. 1-23, 2023. Disponível em: https://doi.org/10.24933/horizontes.v41i1.1624. Acesso em: 10 jan. 2025.

MAINARDES, Jefferson. A ética na formação de pesquisadores/as na Pós-Graduação em Educação: uma revisão sistemática. Roteiro, v. 49, e34826, 2024. Disponível em: https://doi.org/10.18593/r.v49.34826. Acesso em: 10 jan. 2025.

MARANHÃO, F. T. S. et al. Democracia e desinformação no cenário digital: os impactos das fake news sob a perspectiva de Chomsky. Revista OWL (OWL Journal), Campina Grande, v. 4, n. 3, p. 1-29, 2026. Disponível em: https://doi.org/10.5281/zenodo.19336086. Acesso em: 20 abril. 2026.

MASSARANI, Luisa et al. Vacinas contra a COVID-19 e o combate à desinformação na cobertura da Folha de S. Paulo. Fronteiras: estudos midiáticos, v. 23, n. 2, 2021. Disponível em: https://doi.org/10.4013/fem.2021.232.03. Acesso em: 15 nov. 2024.

MORO, Catia; COUTINHO, Andréa Santos; PINHO, Glauco. Ética na pesquisa em Educação: desafios perante encaminhamentos sobrepostos à Plataforma Brasil. Práxis Educativa, v. 18, p. 1-17, 2023. Disponível em: https://doi.org/10.5212/PraxEduc.v.18.21835.079. Acesso em: 10 jan. 2025.

MONARI, Ana Carolina Soares; OTT, Clarissa. Regulação de inteligência artificial: perspectivas comparadas entre o AI Act europeu e os marcos normativos brasileiros. Revista de Direito e Tecnologia, v. 7, n. 1, p. 45-68, 2025.

O'NEIL, Cathy. Algoritmos de destruição em massa: como o big data aumenta a desigualdade e ameaça a democracia. Santo André: Rua do Sabão, 2020.

ROBL FILHO, Ilton Norberto; MARRAFON, Marco Aurélio; MEDON, Filipe. A inteligência artificial utilizada para desinformar ou a serviço da desinformação: como as deepfakes e as redes automatizadas abalam o mercado livre de ideias e a democracia constitucional e deliberativa. Economic Analysis of Law Review, v. 13, n. 3, p. 32-47, 2022. Disponível em: https://doi.org/10.31501/ealr.v13i3.12527. Acesso em: 15 nov. 2024.

ROSSETTI, Roseli; ANGELUCI, Alan. Ética Algorítmica: questões e desafios éticos do avanço tecnológico da sociedade da informação. Galáxia, São Paulo, n. 46, p. 1-14, 2021. Disponível em: https://doi.org/10.1590/1982-2553202150301. Acesso em: 10 jan. 2025.

SANTAELLA, Lucia; KAUFMAN, Dora. A inteligência artificial generativa como quarta ferida narcísica do humano. Matrizes, v. 18, n. 1, p. 37-53, 2024. Disponível em: https://doi.org/10.11606/issn.1982-8160.v18i1p37-53. Acesso em: 10 jan. 2025.

STAKE, Robert E. The art of case study research. Thousand Oaks: Sage, 1995.

WARDLE, Claire; DERAKHSHAN, Hossein. Information disorder: toward an interdisciplinary framework for research and policymaking. Strasbourg: Council of Europe, 2017.

WAISBORD, Silvio. Truth is what happens to news: on journalism, fake news, and post-truth. Journalism Studies, v. 19, n. 13, p. 1866-1878, 2018. Disponível em: https://doi.org/10.1080/1461670X.2018.1492881. Acesso em: 15 dez. 2024.

ZHOU, Xinyi; ZAFARANI, Reza. A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Computing Surveys, v. 53, n. 5, p. 1-40, 2020. Disponível em: https://doi.org/10.1145/3395046. Acesso em: 15 nov. 2024.

Recebido em: 25/04/2026
Aprovado em: 19/05/2026
Publicado em: 15/06/2026