INTERFACE TECNOLÓGICA

SEGURANÇA EM SISTEMAS DE INTELIGÊNCIA ARTIFICIAL E MACHINE LEARNING ANÁLISE DE ATAQUES ADVERSARIAIS E ESTRATÉGIAS DE DEFESA

SECURITY IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING SYSTEMS ANALYSIS OF ADVERSARIAL ATTACKS AND DEFENSE STRATEGIES

Igor Luigi Fracarolli – igor.fracarolli@fatec.sp.gov.br
Faculdade de Tecnologia de Taquaritinga (Fatec) – Taquaritinga – São Paulo – Brasil

Robson Eduardo Galloppi – robson.galloppi@fatec.sp.gov.br
Faculdade de Tecnologia de Taquaritinga (Fatec) – Taquaritinga – São Paulo – Brasil

DOI: 10.31510/infa.v22i1.2240
Data de submissão: 10/04/2025
Data do aceite: 26/06/2025
Data da publicação: 30/06/2025

RESUMO

A Inteligência Artificial (IA) e o Machine Learning (ML) têm desempenhado papéis fundamentais na transformação digital, sendo amplamente utilizados em setores como saúde, finanças e transporte. Contudo, essas tecnologias apresentam vulnerabilidades significativas, especialmente em relação a ataques adversariais, que manipulam dados de entrada para comprometer a eficácia e segurança dos modelos. Além disso, o uso de dados enviesados para o treinamento de algoritmos pode perpetuar preconceitos e desigualdades, trazendo à tona preocupações éticas e sociais. Este trabalho tem como objetivo analisar os impactos de ataques adversariais em sistemas de IA e ML, bem como propor estratégias de defesa para garantir a confiabilidade e a segurança dessas tecnologias. Paralelamente, busca-se explorar os desafios éticos relacionados ao uso de dados enviesados, sugerindo diretrizes que promovam maior equidade e responsabilidade. A pesquisa adota uma metodologia bibliográfica, baseada na análise de artigos científicos, relatórios técnicos e estudos de casos. Essa abordagem permite identificar vulnerabilidades, exemplificar ataques adversariais e avaliar as soluções existentes para mitigar seus efeitos. Este trabalho baseia-se em autores que exploram o uso de inteligência artificial em diversas áreas, como linguagem natural (VASWANI et al., 2017; DEVLIN et al., 2019), reconhecimento de fala (RAVANELLI et al., 2020), agricultura (SILVA et al., 2018) e veículos autônomos (FENG et al., 2021), a vulnerabilidade a ataques adversariais (HOSPEDALES et al., 2020). Destacam-se as contribuições dos Transformers e do aprendizado profundo (OTTER et al., 2020). Os resultados indicam que medidas como a utilização de algoritmos robustos, a diversificação dos dados de treinamento e a implementação de testes adversariais contínuos são eficazes na proteção de sistemas de IA e ML. Além disso, destaca-se a necessidade de regulamentações que assegurem a transparência e a ética no desenvolvimento dessas tecnologias. Conclui-se que, para garantir a segurança e a equidade, é

essencial combinar estratégias técnicas com diretrizes éticas, promovendo a adoção responsável de IA e ML.

Palavras-chave: Inteligência Artificial. Machine Learning. Segurança. Ataques adversariais. Estratégias de defesa.

ABSTRACT

Artificial Intelligence (AI) and Machine Learning (ML) have played fundamental roles in digital transformation, being widely used in sectors such as healthcare, finance, and transportation. However, these technologies present significant vulnerabilities, especially in relation to adversarial attacks, which manipulate input data to compromise the effectiveness and security of models. In addition, the use of biased data to train algorithms can perpetuate prejudices and inequalities, raising ethical and social concerns. This work aims to analyze the impacts of adversarial attacks on AI and ML systems, as well as propose defense strategies to ensure the reliability and security of these technologies. At the same time, we seek to explore the ethical challenges related to the use of biased data, suggesting guidelines that promote greater equity and responsibility. The research adopts a bibliographic methodology, based on the analysis of scientific articles, technical reports, and case studies. This approach allows identifying vulnerabilities, exemplifying adversarial attacks, and evaluating existing solutions to mitigate their effects. This work is based on authors who explore the use of artificial intelligence in several areas, such as natural language (VASWANI et al., 2017; DEVLIN et al., 2019), speech recognition (RAVANELLI et al., 2020), agriculture (SILVA et al., 2018) and autonomous vehicles (FENG et al., 2021), vulnerability to adversarial attacks (HOSPEDALES et al., 2020). The contributions of Transformers and deep learning (OTTER et al., 2020) stand out. The results indicate that measures such as the use of robust algorithms, the diversification of training data and the implementation of continuous adversarial testing are effective in protecting AI and ML systems. In addition, the need for regulations that ensure transparency and ethics in the development of these technologies is highlighted. It is concluded that, to ensure security and fairness, it is essential to combine technical strategies with ethical guidelines, promoting the responsible adoption of AI and ML.

Keywords: Artificial Intelligence. Machine Learning. Security. Adversarial attacks. Defense strategies.

1. INTRODUÇÃO

A Inteligência Artificial (IA) e o Machine Learning (ML) emergem como tecnologias centrais na transformação digital, impactando amplamente diversos setores econômicos e sociais. A IA é definida pela capacidade de sistemas computacionais de realizar tarefas complexas que normalmente exigiriam inteligência humana, como o reconhecimento de padrões e a tomada de decisões. Já o ML, um campo específico dentro da IA, é responsável por habilitar o aprendizado automatizado, permitindo que algoritmos aprimorem continuamente seu desempenho sem necessidade de programação explícita para cada função. Essas tecnologias

têm sido amplamente aplicadas em áreas como saúde, finanças, transporte e educação, mostrando sua relevância ao oferecer diagnósticos mais precisos, detecção de fraudes e personalização de serviços.

O crescimento do uso de IA e ML está diretamente relacionado à disponibilidade de grandes volumes de dados, ao avanço na capacidade de processamento computacional e ao desenvolvimento de modelos algorítmicos mais sofisticados (TORFI et al., 2020). Na saúde, essas inovações possibilitam tratamentos mais personalizados e diagnósticos antecipados. No setor financeiro, promovem maior segurança por meio da detecção automatizada de fraudes e da previsão de tendências de mercado. Apesar dos benefícios evidentes, o uso crescente dessas tecnologias traz desafios críticos, especialmente relacionados à segurança, como os ataques adversariais, e à perpetuação de desigualdades sociais, geradas por vieses presentes nos dados utilizados para o treinamento dos modelos.

Um dos principais desafios enfrentados por sistemas de IA e ML é a vulnerabilidade a ataques adversariais, que se caracterizam pela manipulação de entradas de dados com o objetivo de enganar os modelos e comprometer sua eficácia (HOSPEDALES et al., 2020). Esses ataques podem impactar diretamente a confiabilidade das aplicações, colocando em risco sua funcionalidade e segurança. Além disso, a questão dos vieses nos dados utilizados para o treinamento dos algoritmos suscita preocupações éticas significativas, uma vez que podem reforçar preconceitos existentes e ampliar desigualdades. Essas questões demandam uma abordagem crítica e a formulação de estratégias que assegurem tanto a integridade quanto a equidade no uso dessas tecnologias.

Diante disso, a presente pesquisa busca compreender como sistemas de IA e ML podem ser protegidos contra ataques adversariais e como minimizar os efeitos negativos do uso de dados enviesados. A análise desses aspectos se faz necessária para garantir que os benefícios dessas tecnologias sejam usufruídos de maneira ampla e justa, sem comprometer a segurança ou a equidade.

O objetivo geral deste trabalho é analisar os impactos de ataques adversariais em sistemas de IA e ML, além de propor estratégias de defesa que assegurem sua confiabilidade. Especificamente, busca-se investigar as principais vulnerabilidades desses sistemas, identificar exemplos concretos de ataques e seus efeitos, propor medidas eficazes de defesa e abordar os desafios éticos relacionados ao uso de dados enviesados, sugerindo diretrizes que promovam maior equidade.

A pesquisa será conduzida com base em um estudo bibliográfico, com foco em artigos científicos, relatórios técnicos e casos práticos relacionados ao tema. O estudo abordará tanto os aspectos técnicos dos ataques adversariais e das estratégias de defesa quanto as implicações éticas e regulatórias, visando contribuir para o desenvolvimento seguro e responsável de sistemas de IA e ML (VASWANI et al., 2017).

Este trabalho fundamenta-se em uma seleção de autores que abordam avanços significativos na aplicação de aprendizado profundo e inteligência artificial em diferentes contextos. Vaswani et al. (2017) introduzem o modelo Transformer, base para diversas arquiteturas subsequentes em processamento de linguagem natural (PLN), como o BERT proposto por Devlin et al. (2019), que aprimora a compreensão de linguagem ao utilizar treinamento bidirecional.

Hospedales et al., (2020) relata que um dos principais desafios enfrentados por sistemas de IA e ML é a vulnerabilidade a ataques adversariais, que se caracterizam pela manipulação de entradas de dados com o objetivo de enganar os modelos e comprometer sua eficácia.

Otter et al. (2020) oferecem uma revisão abrangente sobre o uso de deep learning em PLN, destacando sua evolução e aplicações práticas. No campo do reconhecimento de fala, Ravanelli et al. (2020) exploram o uso de aprendizado auto-supervisionado e multitarefa para tornar os sistemas mais robustos. Complementarmente, Silva et al. (2018) demonstram a aplicação de técnicas de IA na agricultura, com foco na detecção e contagem de plantas, enquanto Feng et al. (2021) aplicam esses conceitos ao desenvolvimento de testes inteligentes para veículos autônomos em ambientes realistas e adversos, ampliando o escopo de atuação da IA para além do campo textual.

2. FUNDAMENTAÇÃO TEÓRICA

Os setores que mais adotam Inteligência Artificial (IA) e Machine Learning (ML) têm registrado grandes avanços, especialmente entre 2023 e 2025, período em que os investimentos globais em IA ultrapassaram US$ 300 bilhões (IDC, 2025). Essas tecnologias otimizam processos, melhoram previsões e automatizam operações complexas.

No setor financeiro, IA e ML são amplamente usadas para análise preditiva, gestão de risco, automação bancária e detecção de fraudes. Segundo PwC (2024), 86% das instituições financeiras já aplicam IA, reduzindo custos operacionais em média 25% e aumentando a precisão nas decisões de investimento em até 30%. Algoritmos preveem oscilações de mercado

e aceleram a identificação de fraudes em até 60%, reforçando a segurança do sistema (Rangel Neto, 2025; Torfi et al., 2020).

No varejo, essas tecnologias personalizam ofertas, recomendam produtos e otimizam a cadeia de suprimentos. ML analisa dados de clientes para ajustar estratégias de marketing em tempo real (Hospedales et al., 2020) e melhora a logística com previsão de demanda e gestão de estoques, reduzindo custos e aumentando a satisfação (Vaswani et al., 2017).

Na saúde, IA e ML transformam diagnósticos, administração hospitalar e desenvolvimento de medicamentos. Sistemas de apoio à decisão clínica ajudam a identificar doenças rapidamente, com base em grandes volumes de dados (Silva et al., 2018; Strelkova, 2017).

Na manufatura, IA é usada para automação, manutenção preditiva e otimização da produção. Isso reduz falhas, diminui custos e melhora a qualidade, tornando o processo mais ágil e eficiente (Hospedales et al., 2020; Lecun et al., 2015).

No setor de transportes e logística, IA e ML otimizam rotas, preveem tráfego e gerenciam frotas, economizando combustível e tempo (Vaswani et al., 2017). Além disso, o desenvolvimento de veículos autônomos promete aumentar a segurança e eficiência, transformando a mobilidade e contribuindo para a sustentabilidade (Torfi et al., 2020).

Esses setores evidenciam como IA e ML estão impulsionando a eficiência, reduzindo custos e fortalecendo a competitividade num mercado cada vez mais tecnológico.

3. PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa adotou a revisão bibliográfica como método principal, com enfoque qualitativo e descritivo, visando investigar os ataques adversariais em sistemas de inteligência artificial e machine learning, bem como estratégias de defesa. A revisão bibliográfica é um recurso metodológico adequado para a análise crítica e sistemática de produções científicas já publicadas, permitindo a identificação de padrões e lacunas no conhecimento sobre o tema (Gil, 2019). A pesquisa qualitativa possibilita uma compreensão aprofundada dos fenômenos, sendo apropriada para estudos que buscam explorar aspectos complexos, como a segurança em sistemas de IA (Marconi; Lakatos, 2019).

O estudo foi conduzido em artigos publicados em um período de 10 anos, abrangendo publicações de 2015 a 2025, com o objetivo de mapear os avanços e desafios recentes no campo. A escolha desse intervalo temporal é justificada pela rápida evolução das tecnologias de inteligência artificial e a crescente relevância dos ataques adversariais no cenário atual (Gil,

2019). O local de estudo compreendeu o ambiente digital, utilizando-se bases de dados online como Google Acadêmico para a identificação de artigos e publicações científicas relevantes.

O público-alvo da pesquisa inclui pesquisadores, profissionais da área de ciência da computação, desenvolvedores de sistemas de IA e machine learning, além de especialistas em cibersegurança. A escolha desse público se deve à necessidade de entendimento dos riscos associados aos ataques adversariais e às práticas de defesa mais eficazes (Marconi; Lakatos, 2019).

Os descritores utilizados para a busca de dados foram "ataques adversariais", "segurança em machine learning", "defesas em IA", "inteligência artificial adversarial" e "robustez em machine learning". Esses termos permitiram uma busca focada nos aspectos centrais da pesquisa, otimizando a identificação de estudos relevantes (Gil, 2019).

Os critérios de inclusão abrangeram artigos publicados entre 2013 e 2023, em português e inglês, que tratassem diretamente de ataques adversariais e estratégias de defesa em IA e machine learning. Adotou-se como critério de exclusão trabalhos que não abordassem os aspectos de segurança em IA ou que se limitassem a áreas não relacionadas à cibersegurança ou machine learning (Marconi; Lakatos, 2019).

A análise dos dados foi feita de maneira descritiva e interpretativa, considerando as contribuições teóricas e práticas dos estudos selecionados. A técnica de análise de conteúdo foi empregada para identificar os principais padrões de ataque e as respostas de defesa, proporcionando uma compreensão ampla dos desafios e soluções (Marconi; Lakatos, 2019).

4. RESULTADOS E DISCUSSÃO

4.1 Preconceitos e desigualdades resultantes do uso de algoritmos de IA e ML, com foco nos dados utilizados para treiná-los

A segurança em sistemas computacionais ganha importância com o aumento das ameaças cibernéticas e o avanço de tecnologias como inteligência artificial (IA), big data e aprendizado de máquina. Sistemas modernos exigem métodos eficazes para proteger a confidencialidade, integridade e disponibilidade das informações. Técnicas de aprendizado profundo, usadas em modelos de linguagem natural, ajudam a identificar padrões anômalos que indicam possíveis ataques (OTTER et al., 2020).

Com a expansão de sistemas automatizados, como veículos autônomos, é fundamental testar a segurança em ambientes adversos. Feng et al. (2021) destacam a importância de testes realistas que avaliem a resistência dos algoritmos a condições variadas e hostis, exigindo sistemas resilientes e adaptativos.

Na proteção contra ataques cibernéticos, o aprendizado supervisionado tem apresentado bons resultados. Rangel Neto (2025) propõe um método que combina aprendizado de máquina e big data para detectar ataques DDoS em tempo real, antecipando ameaças e reforçando a segurança.

Arquiteturas como o BERT, criado por Devlin et al. (2019), são eficazes na análise semântica de textos, auxiliando na identificação de ameaças linguísticas, como tentativas de engenharia social, por meio de representações contextuais profundas.

Além do setor tecnológico, a segurança também é essencial em áreas como previdência complementar. Custódio (2021) mostra que o uso de algoritmos inteligentes melhora a gestão e protege dados sensíveis, evidenciando a abrangência da segurança digital.

Técnicas de autoaprendizado multitarefa, como as de Ravanelli et al. (2020), aumentam a robustez de sistemas de reconhecimento, tornando-os resistentes a ruídos e adulterações, fundamentais para operar em ambientes cibernéticos dinâmicos e ameaçadores.

O uso crescente de IA e aprendizado de máquina levanta preocupações sobre vieses e desigualdades, pois os dados de treinamento podem refletir preconceitos sociais. Zhang et al. (2020) alertam que modelos como o DialoGPT podem reproduzir esses vieses, tornando essencial a análise cuidadosa dos dados para evitar discriminações.

Os vieses em IA surgem porque os dados de treinamento são historicamente construídos a partir de ambientes marcados por desigualdades sociais e econômicas. Segundo Custódio (2021), os sistemas de IA empregados em entidades previdenciárias, por exemplo, podem reproduzir desigualdades de acesso a benefícios, se os dados de entrada forem desiguais. Isso é particularmente evidente em algoritmos usados para prever o risco de inadimplência de beneficiários, os quais podem penalizar grupos sociais que já enfrentam desvantagens estruturais, como minorias étnicas ou pessoas de baixa renda.

Outro fator a ser considerado é o viés presente na coleta de dados para IA e ML. Otter et al. (2020) destacam que o processo de coleta pode ser influenciado por uma série de fatores, como a subjetividade humana ou falhas no método de captura dos dados, o que gera distorções nos resultados produzidos pelos algoritmos. Esses vieses são transmitidos para os modelos, que aprendem com esses dados e acabam reforçando preconceitos que deveriam ser combatidos.

Um exemplo disso pode ser visto em sistemas de reconhecimento facial que, devido a desequilíbrios nos dados, apresentam maior taxa de erro para pessoas de pele escura.

Adicionalmente, algoritmos de IA e ML são frequentemente avaliados com base em métricas que não consideram adequadamente a distribuição equitativa dos erros. Ravanelli et al. (2020) apontam que, no campo do reconhecimento de fala, os modelos tendem a ser mais precisos para grupos de pessoas cujas características fonéticas são mais representadas no conjunto de dados de treinamento. Isso reforça a exclusão de grupos linguísticos minoritários, uma vez que suas vozes não são suficientemente consideradas no desenvolvimento de tais sistemas.

O desenvolvimento de tecnologias como o BERT, descrito por Devlin et al. (2019), ilustra bem o desafio de reduzir os vieses em IA. Embora o BERT tenha sido projetado para melhorar a compreensão da linguagem natural, ele ainda depende de grandes quantidades de dados de treinamento, os quais frequentemente carregam vieses linguísticos e culturais. Assim, mesmo ferramentas avançadas de processamento de linguagem natural podem perpetuar desigualdades, caso não sejam cuidadosamente desenhadas para mitigar esses problemas.

Outro exemplo de impacto dos dados enviesados é abordado por Feng et al. (2021), que estudam algoritmos aplicados a veículos autônomos. Esses sistemas podem apresentar um desempenho desigual dependendo do ambiente em que foram treinados, levando a situações perigosas em cenários sub-representados, como ruas de periferias ou áreas urbanas em países em desenvolvimento. A falta de dados diversificados para treinar os algoritmos torna essas tecnologias propensas a erros que afetam desproporcionalmente comunidades vulneráveis.

Por fim, o desafio de mitigar preconceitos em IA e ML está diretamente relacionado à melhoria dos processos de coleta e curadoria de dados. Zhang et al. (2020) enfatizam que, sem uma abordagem consciente para eliminar vieses dos conjuntos de dados, os algoritmos continuarão a reproduzir padrões discriminatórios. Assim, é necessário um esforço constante para garantir que os dados usados no treinamento sejam representativos e equilibrados, a fim de que os sistemas de IA possam operar de maneira justa e equitativa.

4.2 As atuais diretrizes e regulamentações existentes para a aplicação de IA e ML, com ênfase em transparência e responsabilidade

As diretrizes e regulamentações para a aplicação de Inteligência Artificial (IA) e Machine Learning (ML) têm evoluído rapidamente, com ênfase crescente em aspectos de transparência e responsabilidade. A transparência, no contexto da IA, refere-se à capacidade de

compreender e rastrear como os algoritmos chegam às suas decisões, especialmente em áreas sensíveis como diagnósticos médicos, conforme apontado por Lecun et al. (2015).

A falta de explicabilidade em modelos de ML, como as redes neurais profundas, frequentemente leva a problemas de confiança e adoção, o que tem incentivado o desenvolvimento de técnicas de "caixa branca", que tornam os sistemas mais interpretáveis (TORFI et al., 2020). Hospedales et al. (2020) destacam que a transparência é crucial para garantir que os sistemas de IA sejam justos, imparciais e seguros, especialmente quando aplicados em contextos de alto impacto, como o diagnóstico médico e a automação de decisões governamentais.

A responsabilidade, por outro lado, trata da alocação de culpa ou crédito pelas ações tomadas por sistemas de IA. Em muitas situações, os sistemas de ML agem com base em grandes volumes de dados, e suas previsões podem ter consequências significativas. A responsabilidade recai tanto sobre os desenvolvedores dos algoritmos quanto sobre as organizações que os implementam (Vaswani et al., 2017).

Como mencionado por Albuquerque (2023), as aplicações clínicas que utilizam IA requerem estruturas de responsabilização rigorosas, considerando as implicações legais de diagnósticos errôneos e decisões automáticas. Para enfrentar esses desafios, regulamentos, como o Regulamento Geral sobre a Proteção de Dados (GDPR) da União Europeia, impõem a necessidade de que os sistemas de IA sejam explicáveis e que os indivíduos tenham o direito de contestar decisões automatizadas (Silva et al., 2018).

Além disso, Colodetti (2022) ressalta que a implementação responsável de IA deve seguir as recomendações éticas de organismos internacionais. Normas como as diretrizes da UNESCO para a IA ética e o projeto de lei da União Europeia sobre IA, de 2021, fornecem um quadro regulatório detalhado para assegurar que as aplicações de IA e ML não resultem em discriminação ou prejuízo a indivíduos ou grupos. Strelkova (2017) observa que a criação de IA ética exige que os algoritmos sejam testados quanto a vieses, e que as empresas devem ser responsáveis por remediar falhas identificadas. Essa abordagem regulatória busca mitigar os riscos e garantir a equidade em sistemas que afetam diretamente a vida das pessoas.

Outro ponto importante destacado por Lecun et al. (2015) é a necessidade de padrões para avaliação contínua de sistemas de IA após sua implementação. Isso significa que, além da transparência durante o desenvolvimento, deve haver monitoramento constante para garantir que os sistemas permaneçam confiáveis e que as falhas possam ser corrigidas rapidamente. Albuquerque (2023) salienta que, na área médica, esse acompanhamento contínuo é

fundamental, uma vez que a evolução dos dados utilizados em sistemas de IA pode alterar seus resultados ao longo do tempo.

Essas diretrizes, combinadas com um foco crescente em responsabilidade e transparência, são passos essenciais para o desenvolvimento de uma IA que não apenas seja tecnicamente avançada, mas também eticamente sólida e confiável.

5. CONCLUSÃO

O presente estudo faz uma análise detalhada da literatura científica recente, que mostra o rápido crescimento da aplicação de sistemas de Inteligência Artificial (IA) e Machine Learning (ML), junto com os riscos crescentes relacionados à segurança e à ética. Estudos de casos reais e experimentos demonstram que ataques adversariais podem manipular os resultados desses modelos, comprometendo decisões importantes em áreas sensíveis. Além disso, a identificação de vieses algorítmicos, muitas vezes causados por dados enviesados, destaca a necessidade de considerar os impactos sociais da IA. As propostas de defesa, como métodos para aumentar a robustez dos sistemas e regulamentações éticas, indicam que soluções eficazes exigem a colaboração entre ciência, engenharia, políticas públicas e filosofia. Assim, conclui-se que é urgente promover um desenvolvimento tecnológico responsável, seguro e justo.

A pesquisa sobre segurança em IA e ML, especialmente no que diz respeito a ataques adversariais e estratégias de defesa, evidencia a complexidade e importância do tema atualmente. A expansão dessas tecnologias em setores cruciais como saúde, finanças e transporte traz avanços significativos, mas também desafios que demandam atenção imediata. Os ataques adversariais exploram vulnerabilidades dos modelos, afetando a integridade e confiabilidade das soluções, o que pode causar graves consequências para indivíduos e instituições.

Além disso, os vieses nos dados de treinamento perpetuam preconceitos e desigualdades, reforçando a necessidade de uma abordagem ética e responsável no desenvolvimento e uso dessas tecnologias. Embora tragam inovação e eficiência, sistemas de IA e ML podem ampliar injustiças sociais se não forem projetados e monitorados com transparência e responsabilidade.

As estratégias para defender os sistemas contra ataques adversariais e a implementação de diretrizes éticas que promovam a equidade são essenciais para o avanço seguro da tecnologia. Investir em métodos robustos para detectar e mitigar ataques é fundamental para

manter a confiança nessas soluções. Paralelamente, regulamentações que garantam responsabilidade e transparência no uso de dados são indispensáveis para equilibrar os benefícios e riscos dessas inovações.

Portanto, construir um ambiente tecnológico mais seguro e justo depende da cooperação entre profissionais de tecnologia, formuladores de políticas e especialistas em ética. Somente assim será possível maximizar o potencial da IA e do ML, reduzir riscos e garantir que seus benefícios sejam distribuídos de forma responsável e inclusiva. Esta pesquisa contribui para esse debate, oferecendo fundamentos teóricos e práticos que podem orientar futuras ações nesse campo em evolução.

REFERÊNCIAS

ALBUQUERQUE, Bárbara Beatriz Fernandes. **A revolução na prática clínica: O impacto da Inteligência Artificial (IA) nas aplicações radiológicas e diagnóstico médico**. 2023. Trabalho de Conclusão de Curso. Universidade Federal do Rio Grande do Norte.

COLODETTI, Pedro Vinicius Baptista. **Matemática aplicada à inteligência artificial**: a base fundamental do machine learning. 2022.

CUSTÓDIO, Elaine Cristina Pereira. **Aplicações de inteligência artificial em entidades fechadas de previdência complementar**. 2021.

DEVLIN, J. et al. **BERT**: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805, 2019.

FENG, S. et al. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. **Nat Commun**, v.12, p.748, 2021.

GIL, A. C. (2019). **Como elaborar projetos de pesquisa**. São Paulo: Atlas.

HOSPEDALES, T. et al. A. **Meta-Learning in Neural Networks**: A Survey. arXiv:2004.05439, 11 Abr. 2020.

LECUN, Y. et al. **Deep learning. Nature** v.521, p.436-44, 2015

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 8. ed. São Paulo: Atlas, 2019.

OTTER, D. W. et al. **A survey of the usages of deep learning for natural language processing**. IEEE Transactions on Neural Networks and Learning Systems, v.32, n.2, p.604-24, 2020.

RANGEL NETO, Digenaldo de Brito. **Detecção de ataques DDoS na camada de aplicação: um esquema com aprendizado de máquina e Big Data**. 2025. Dissertação de Mestrado.

RAVANELLI, M. et al. **Multi-task self-supervised learning for robust speech recognition**. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, 2020. p.6989-93.

SILVA, Giovanni Cimolin da et al. **Detecção e contagem de plantas utilizando técnicas de inteligência artificial e machine learning.** 2018.

STRELKOVA, O. **Three types of artificial intelligence**. 2017.

TORFI, A. et al. **Natural language processing advancements by deep learning:** A survey. arXiv preprint arXiv:2003.01200 (2020).

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan; KAISER, Łukasz; POLOSUKHIN, Illia. **Attention is all you need**. NeurIPS, [S. l.], 2017.

ZHANG, Yizhe; SUN, Siqi; GALLEY, Michel; CHEN, Yen-Chun; BROCKETT, Chris; GAO, Xiang; GAO, Jianfeng; LIU, Jingjing; DOLAN, Bill. **DialoGPT:** Large-Scale Generative Pre-training for Conversational Response Generation. ACL, system demonstration, [S. l.], p. 1-10, 2020.