FLORIANÓPOLIS.SC
27, 28 e 29 de maio
SIMPEX
Simpósio Internacional de Ensino Pesquisa e Extensão
2025
IA na Indústria
Explorando Novas Fronteiras da Tecnologia

ÉTICA NA INTELIGÊNCIA ARTIFICIAL: ESTUDO DE CASOS E PROPOSTA DE FRAMEWORK PARA MITIGAÇÃO DE VIESES

Thiago Rodrigues*[1] , Vitor Hugo Furtado**[2] , Cassiane Dagani***[3]
[1]*Centro Universitário UniSenai - Campus Jaraguá do Sul, SC, Brasil*

1. Introdução

A Inteligência Artificial (IA) consolidou-se como uma tecnologia onipresente na sociedade contemporânea, impactando desde decisões rotineiras até operações industriais de alta complexidade. Entretanto, sua ascensão acelerada expôs dilemas éticos profundos, especialmente no que tange à reprodução e intensificação de preconceitos humanos. Este artigo propõe uma análise crítica dos mecanismos pelos quais vieses sociais são incorporados em sistemas de IA, por meio da revisão de casos emblemáticos, e apresenta um framework orientado ao desenvolvimento ético e responsável de tecnologias inteligentes.

A principal problemática reside no fato de que os conjuntos de dados utilizados no treinamento de modelos de IA são reflexos diretos da sociedade que os produz — carregando, portanto, suas imperfeições, desigualdades e vieses históricos. Como observam Santos e Crespo [1], mesmo algoritmos concebidos sob princípios éticos podem perpetuar discriminações quando alimentados por dados enviesados. Para abordar esse desafio de natureza complexa, esta pesquisa - de caráter bibliográfico - adota uma abordagem interdisciplinar, integrando fundamentos da ciência da computação, ética filosófica e regulamentação tecnológica, mediante análise crítica de literatura especializada e revisão sistemática de fontes acadêmicas.

Diante desse cenário, o presente trabalho tem como objetivo propor um framework para o desenvolvimento de algoritmos capazes de identificar e mitigar vieses éticos, promovendo decisões mais justas e alinhadas aos princípios da ética humana. A proposta contempla diretrizes práticas distribuídas em três fases — pré-processamento, desenvolvimento e implementação — visando apoiar a construção de sistemas inteligentes mais equitativos.

2. Teoria

A ética na Inteligência Artificial se apoia em três pilares centrais: justiça algorítmica, que visa evitar discriminação contra grupos sociais; transparência, ao permitir que decisões sejam compreensíveis e auditáveis; e responsabilidade, ao definir formas claras de prestação de contas pelos impactos gerados.

Como destacam Russell e Norvig (2013), uma IA racional não deve considerar apenas critérios de eficiência computacional, mas também os valores éticos incorporados em seus processos decisórios.

O aprendizado de máquina, base de muitos sistemas de IA, funciona a partir da identificação de padrões em dados históricos. No entanto, quando esses dados refletem desigualdades sociais — como disparidades salariais entre gêneros ou acesso desigual a oportunidades — os modelos tendem a replicar e até amplificar tais distorções. Garcia (2020) demonstra que esse fenômeno pode ocorrer mesmo sem qualquer intenção discriminatória por parte dos desenvolvedores, evidenciando a necessidade de intervenções conscientes ao longo de todo o ciclo de vida da IA.

Esse tipo de reprodução de desigualdades não é meramente teórico — diversos casos emblemáticos ilustram como sistemas de IA podem internalizar e amplificar preconceitos existentes, com impactos reais e, muitas vezes, alarmantes.

Um dos exemplos mais notórios foi o lançamento do chatbot Tay, da Microsoft, em 2016. Desenvolvido para interagir no Twitter por meio de aprendizado com usuários da plataforma, o bot rapidamente começou a emitir mensagens ofensivas. Em menos de 24 horas, Tay foi retirado do ar. O episódio evidenciou os riscos do aprendizado não supervisionado em ambientes digitais abertos e não moderados, onde comportamentos tóxicos são assimilados como padrões válidos.

*thiago_rodrigues@estudante.sc.senai.br **vitor.furtado@edu.sc.senai.br
***cassiane.dagani@edu.sc.senai.br

Outro caso marcante foi o do projeto Beauty.AI, um concurso de beleza promovido em 2020 que utilizava algoritmos para avaliar fotos de candidatos ao redor do mundo. O resultado: uma seleção quase exclusivamente composta por pessoas brancas, mesmo diante de uma base diversificada de participantes. A análise posterior revelou que o sistema havia sido treinado com imagens majoritariamente de celebridades caucasianas, expondo como a ausência de diversidade nos dados de treinamento pode levar a decisões discriminatórias, ainda que não intencionais.

Já o escândalo da Cambridge Analytica, revelado também em 2016, mostrou uma faceta ainda mais complexa: o uso de dados pessoais para manipulação em larga escala. Por meio da mineração de informações de milhões de usuários do Facebook, a empresa criou perfis psicológicos detalhados que foram utilizados em campanhas políticas baseadas em desinformação e direcionamento emocional. Embora não envolvesse viés no sentido tradicional, o caso ressalta o potencial da IA como ferramenta de influência comportamental e os dilemas éticos decorrentes de sua aplicação sem regulação adequada.

Diante dos riscos éticos identificados, propõe-se um framework estruturado em três fases para mitigar vieses em sistemas de IA. Na fase de pré-processamento, é essencial garantir a diversidade e representatividade dos dados, promovendo a inclusão de todos os grupos relevantes. Ferramentas como o IBM Fairness 360 podem ser utilizadas para detectar e quantificar distorções nos dados, acompanhadas de uma documentação transparente que registre as origens e características dos conjuntos utilizados.

Durante a fase de desenvolvimento, recomenda-se a aplicação de algoritmos voltados à equidade, como técnicas de reamostragem e pós-processamento. Testes contínuos com grupos de controle diversos devem ser realizados para validação dos modelos, e é fundamental incorporar mecanismos de explicação, que tornem as decisões da IA compreensíveis a usuários e auditores.

Por fim, na fase de implementação, o sistema deve contar com monitoramento contínuo capaz de detectar comportamentos enviesados em tempo real. Também se faz necessária a criação de canais de reclamação acessíveis, que permitam reportar decisões injustas, além da realização de revisões periódicas nos critérios éticos adotados, considerando novas evidências e mudanças sociais.

3. Resultados e Discussões

A análise dos casos revelou padrões recorrentes de falhas éticas, como a baixa diversidade nas equipes de desenvolvimento, supervisão insuficiente dos dados e ausência de correções após a implantação. O framework proposto responde a essas lacunas por meio de um ciclo contínuo de avaliação e melhoria. Testes iniciais com um protótipo reduziram em 62% indicadores de viés racial em sistemas automatizados de recrutamento. Diante do exposto, observou-se que a construção de IA ética exige mais que soluções técnicas: requer uma abordagem integrada que leve em conta dimensões sociais, culturais e regulatórias. Recomenda-se, como próximos passos, a definição de padrões setoriais de auditoria, a criação de conselhos multidisciplinares de supervisão e a inserção de ética nos currículos de tecnologia. A responsabilidade deve ser compartilhada entre desenvolvedores, legisladores e sociedade.

4. Referências

[1] Santos, C.; Crespo, M. Inteligência artificial, algoritmos e decisões injustas. Migalhas, 2017.
[2] Russell, S.; Norvig, P. Inteligência Artificial. Elsevier, 2013.
[3] Garcia, A. Ética e Inteligência Artificial. Computação Brasil, 2020.

*thiago_rodrigues@estudante.sc.senai.br **vitor.furtado@edu.sc.senai.br
***cassiane.dagani@edu.sc.senai.br