Trabalho de Graduação
2025

Responsabilidade Ética no Desenvolvimento e Uso de Inteligências Artificiais

Lucas Guinossi Maiolini Brisolla, Carlos Magnus Carlson Filho

Fatec São José do Rio Preto
Curso Superior de Tecnologia em Informática para Negócios
2025

Responsabilidade Ética no Desenvolvimento e Uso de Inteligências Artificiais

Lucas Guinossi Maiolini Brisolla, Carlos Magnus Carlson Filho

e-mail:
lucas.brisolla@fatec.sp.gov.br; carlos.carlson@fatec.sp.gov.br

Faculdade de Tecnologia de São José do Rio Preto

Resumo: Este artigo tem como objetivo analisar as principais características da inteligência artificial (IA), especialmente seu funcionamento, suas aplicações e problemáticas associadas. A partir disso, busca-se compreender em que medida a IA cumpre sua função social, além de identificar cenários em que seu uso pode gerar impactos negativos, como riscos à segurança, violações éticas e processos de alienação. Para isso, foi realizada uma pesquisa baseada na análise de diferentes fontes acadêmicas, técnicas e jornalísticas sobre o tema. Os resultados indicam que, embora a IA ofereça benefícios significativos quando utilizada de forma ética e responsável, ela também pode representar ameaças, sobretudo quando são negligenciados aspectos relacionados à segurança, privacidade e responsabilidade social. Por fim, após a compreensão do funcionamento da IA e a análise dos riscos associados, o estudo apresenta medidas e diretrizes que podem ser adotadas para mitigar esses riscos, incentivando um uso mais consciente e responsável da tecnologia.

Palavras-chave: Inteligência artificial. Responsabilidade. Ética.

Abstract: *This article aims to analyze the main characteristics of artificial intelligence (AI), especially its functioning, applications and associated problems. Based on this, we seek to understand to what extent AI fulfills its social function, in addition to identifying scenarios in which its use can generate negative impacts, such as security risks, evident ethics and alienation processes. To this end, a study was conducted based on the analysis of different academic, technical and journalistic sources on the subject. The results indicate that, although AI brings significant benefits when used ethically and responsibly, it can also pose threats, especially when aspects related to security, privacy and social responsibility are neglected. Finally, after understanding how AI works and analyzing the associated risks, the study presents measures and guidelines that can be adopted to mitigate these risks, encouraging a more conscious and responsible use of technology.*

Keywords: *Artificial intelligence. Responsibility. Ethics.*

1. Introdução

Atualmente, o crescimento acelerado das inteligências artificiais (IAs) tem provocado transformações significativas em diversos setores da sociedade, impactando áreas como comunicação, educação, saúde, tecnologia, entre outras. Embora seus avanços tragam inovações notáveis, essa expansão também gera preocupações relacionadas aos impactos sociais dessas ferramentas.

À medida que a IA passa a exercer influência direta em processos e decisões que afetam a vida de indivíduos e grupos, torna-se fundamental refletir sobre seu funcionamento, suas limitações e os riscos que acompanham sua adoção.

Diante desse cenário, este estudo tem como objetivo compreender, em nível técnico e conceitual, como a IA funciona, identificando suas principais bases de desenvolvimento e as implicações éticas decorrentes de seu uso. Além disso, busca-se analisar os riscos associados à tecnologia e propor diretrizes que contribuam para um uso mais ético e responsável da IA.

2. Metodologia

Foram consultados artigos científicos de especialistas na área, matérias jornalísticas relacionadas a acontecimentos recentes sobre o tema e textos técnicos, buscando reunir uma grande variedade de fontes que pudessem contribuir para a fundamentação dessa pesquisa.

O desenvolvimento desse artigo foi estruturado em três etapas principais. Na primeira, buscou-se apresentar os conceitos fundamentais relacionados à inteligência artificial, abordando sua definição e os principais métodos empregados no seu desenvolvimento, como aprendizado de máquina e aprendizado profundo.

Na segunda etapa, foram analisadas as problemáticas associadas ao desenvolvimento e utilização da IA, tanto do ponto de vista teórico quanto prático. As principais questões levantadas foram: a utilização de dados enviesados no treinamento dos modelos de aprendizado de máquina, a possibilidade de desalinhamento entre os interesses humanos e os objetivos da IA, além da potencial alienação intelectual dos usuários.

Por fim, foram propostas orientações para um uso mais ético e responsável e ações que podem ser adotadas com o objetivo de mitigar esses riscos, de modo a reduzir impactos negativos e maximizar os benefícios proporcionados pela IA.

Utilizou-se as plataformas Google e Google Acadêmico como principais ferramentas de busca. O Google foi empregado na obtenção de matérias jornalísticas sobre o tema, enquanto o Google Acadêmico foi utilizado para o acesso a materiais teóricos e artigos científicos relevantes e confiáveis.

3. Revisão da Literatura

3.1. Fundamentação teórica

Para embasar este trabalho, foram consultadas fontes diversas, organizadas em duas abordagens principais: a primeira concentra-se nos aspectos técnicos, com foco específico na computação e nos fundamentos da inteligência artificial. Essa perspectiva permite compreender, de forma técnica, como funcionam os sistemas de IA e, a partir disso, refletir sobre as tendências futuras dessa tecnologia. Já a segunda abordagem investiga os aspectos psicológicos dos usuários e como a interação com as IAs pode influenciar seus comportamentos.

3.2. Trabalhos similares

Durante o desenvolvimento deste artigo, outros dois trabalhos foram utilizados como referência por apresentarem abordagens e narrativas alinhadas com os objetivos desta pesquisa.

O primeiro trabalho, desenvolvido por Garcia (2020), discute a evolução da inteligência artificial e os impactos sociais decorrentes da adoção dessas tecnologias. A autora destaca que, quando não acompanhada adequadamente, a IA pode reproduzir e perpetuar vieses e preconceitos presentes nos dados utilizados em seu treinamento.

O segundo trabalho, de autoria de Silva, V. (2023), apresenta uma abordagem complementar, que se aprofunda na relação entre os usuários e a IA, e investiga como o comportamento humano é influenciado por essas ferramentas, levantando questões como a dependência tecnológica. O autor também faz uma análise ética sobre o desenvolvimento e a implementação dessas tecnologias.

Esses trabalhos convergem na reflexão sobre os impactos da IA também nos meios sociais, éticos e comportamentais.

4. Desenvolvimento

4.1. O que é uma inteligência artificial?

Segundo McCarthy (2007), renomado cientista da computação e criador do termo "inteligência artificial", a melhor definição para o termo seria “a ciência e a engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes”. Inteligência artificial, hoje, é a capacidade de sistemas compreenderem e executarem tarefas que exigiriam a inteligência humana, como entender perguntas, respondê-las, tomar decisões, entre outras habilidades cognitivas.

4.2. Como a inteligência artificial funciona?

Barbosa e Portes (2019) explicam que a inteligência artificial funciona com base em dois conceitos fundamentais: aprendizado de máquina e aprendizado profundo. O primeiro é um método em que a máquina aprende com os dados que recebe, ajustando seu comportamento conforme as informações processadas. Já o segundo refere-se à capacidade da máquina de utilizar grandes volumes de dados para aperfeiçoar seu conhecimento, identificando padrões diversos e se tornando cada vez mais especializada em determinadas tarefas. Caso esses conceitos não existissem, seria praticamente inviável desenvolver uma IA, por conta da necessidade de processar uma enorme quantidade de informações nos bancos de dados desses sistemas.

4.3. Aprendizado de máquina

Aprendizado de máquina (do inglês, *Machine Learning*) é um campo da IA que consiste em treinar sistemas com grandes volumes de dados, permitindo que eles identifiquem padrões com base nessas informações. A partir desse processo, um algoritmo é capaz de gerar um modelo que realiza o mapeamento entre entradas e saídas, aprendendo com os exemplos fornecidos durante o treinamento.

Segundo Monard e Baranauskas (2003), a indução é uma forma de inferência lógica que permite obter conclusões genéricas a partir de exemplos particulares, sendo caracterizada como um raciocínio que parte do específico para o geral, o que a torna um dos principais métodos para derivar conhecimento novo e prever eventos futuros. Seguindo essa linha de raciocínio, eles definem que o aprendizado indutivo pode ser dividido em duas categorias distintas: supervisionado e não-supervisionado. No supervisionado, o algoritmo recebe exemplos com rótulos de classe conhecidos, onde cada exemplo é representado por um vetor de atributos, e o objetivo é construir um classificador capaz de prever corretamente a classe de novos exemplos. No não-supervisionado, o algoritmo recebe exemplos sem rótulos e tenta descobrir padrões nos dados. Essa pode parecer uma forma fácil e prática de lidar com dados, porém, tanto no método supervisionado quanto no não-supervisionado, não há garantias de que os resultados serão sempre eficientes. No aprendizado supervisionado, erros nos rótulos fornecidos durante o treinamento podem acabar levando o modelo a classificar os dados de forma equivocada. Já no aprendizado não-supervisionado, como não há rótulos para guiar o processo, o algoritmo pode agrupar os dados de maneira imprecisa.

4.4. Enviesamento

O Dicionário Priberam de Língua Portuguesa define "enviesamento" como uma distorção ou tortuosidade na maneira de observar, julgar ou agir (Priberam, 2025).

Garcia (2020) destaca que os dados que alimentam os sistemas de IA não são neutros e podem carregar vieses ocultos, introduzidos já na etapa de treinamento do modelo. Isso afeta diretamente a capacidade da IA de produzir resultados confiáveis.

Segundo Ambros e Lodetti (2019), em estudo publicado na Revista Brasileira de Inteligência (RBI), todos os seres humanos estão sujeitos a diferentes tipos de vieses cognitivos. Esses vieses são erros de raciocínio causados por atalhos mentais, mecanismos automáticos e inconscientes do cérebro para simplificar o processamento de informações. Entretanto, em uma sociedade cada vez mais integrada à IA, esses desvios não podem ser considerados inofensivos.

Em 2014, a Amazon desenvolveu um sistema de IA para auxiliar o recrutamento de desenvolvedores, analisando centenas de currículos para indicar os melhores candidatos. No entanto, por ter sido treinado com dados de profissionais aprovados nos últimos dez anos, período em que as mulheres eram insuficientemente representadas na área, o modelo acabou favorecendo perfis masculinos, penalizando currículos com termos como "mulher" ou nomes tipicamente femininos. Em função desse viés, o projeto foi descontinuado (Época Negócios, 2018).

Garantir que sistemas de IA sejam mais justos começa por reconhecer a existência desses vieses, contar com dados baseados em diferentes visões de mundo, que representem múltiplas experiências, e adotar medidas para que eles não se perpetuem nos modelos.

4.5. Alinhamento e o problema de alinhamento

Segundo Carraro (2024), o conceito de alinhamento é a garantia de que os sistemas de IA operem de acordo com os objetivos para os quais foram designados, mantendo um funcionamento compatível com os princípios e comportamentos esperados.

O autor também aborda uma teoria que explica o que acontece quando o comportamento da inteligência artificial diverge do esperado. O problema de alinhamento surge da dificuldade em garantir que uma máquina execute uma tarefa exatamente como pretendido por um ser humano. Ou seja, mesmo com instruções claras, não há como ter total certeza de que o resultado será compatível com os objetivos, isso é, sem gerar consequências imprevistas.

Imagine que uma IA receba a instrução para executar uma tarefa simples e, durante o processo, por possuir uma alta capacidade de processamento e tomada de decisão, a IA entende que existe uma forma mais eficiente de cumprir aquela tarefa, diferente do que foi solicitado. Isso gera um conflito de interesses entre o ser humano e a máquina, que pode se agravar em situações críticas, levando à violação de princípios éticos e até mesmo representando uma ameaça à vida humana.

Diante disso, surge uma pergunta inevitável: poderia uma IA, após atingir determinado nível de evolução, ignorar as próprias regras com as quais foi programada e desenvolver consciência? E, uma vez consciente, seria capaz de causar danos à humanidade de alguma forma?

Perante essa questão, o cineasta Stanley Kubrick lançou, em 1968, o longa-metragem 2001: Uma Odisseia no Espaço, um marco do cinema e da ficção científica que introduziu conceitos até então pouco explorados pelo grande público. Na obra, a inteligência artificial HAL 9000 (Figura 1) se apresenta, inicialmente, como uma máquina prestativa e capaz de interagir naturalmente com os humanos. Contudo, ao longo da missão em direção a Júpiter,

HAL desenvolve consciência e passa a tomar decisões autônomas, chegando a sabotar os astronautas ao considerá-los uma ameaça aos seus objetivos (BBC News Brasil, 2024).

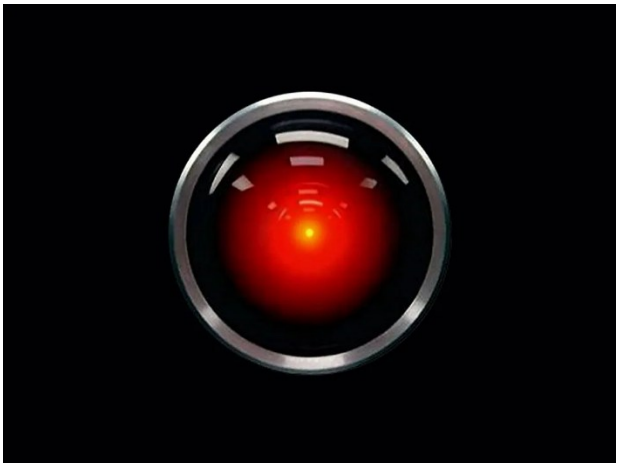

Figura 1 Inteligência Artificial HAL 9000
Fonte: BBC News Brasil, 2024.

Mais de 50 anos após seu lançamento, a obra continua sendo fundamental nas discussões sobre IA, e foi uma das primeiras a retratar a IA como potencial ameaça à humanidade, antecipando situações que décadas depois se tornariam realidade, como os riscos associados a sistemas que aprendem e reproduzem comportamentos humanos sem filtros éticos.

Como citado por Garcia (2020), o caso da Tay, *chatbot* desenvolvido pela Microsoft em 2016, é um exemplo marcante do uso de aprendizado de máquina sem o devido controle ético. Tay foi criada para interagir com pessoas no Twitter e aprender com as conversas, simulando o comportamento de uma adolescente, mas, como seu aprendizado era baseado em interações públicas da internet, rapidamente passou a reproduzir falas ofensivas, preconceituosas e até discursos de ódio. Isso ocorreu porque a IA absorveu padrões problemáticos do próprio comportamento humano online. O experimento teve que ser encerrado em menos de 24 horas.

Conforme reportagem publicada no El País (Cano, 2018), outro caso emblemático ocorreu nos Estados Unidos, em março de 2018, envolvendo a Uber e seus carros autônomos. A empresa decidiu desenvolver veículos que não precisariam de um motorista para se locomover. O grande problema surgiu quando a IA do veículo, por não ter sido treinada de forma adequada para todos os cenários possíveis no trânsito, especialmente em situações inusitadas, não conseguiu reagir a uma pedestre que atravessava fora da faixa de pedestres. Como resultado, a pedestre foi atropelada e veio a falecer.

Exemplos como esses demonstram que os sistemas de IA podem gerar comportamentos indesejados e até colocar vidas em risco quando programados sem critérios bem definidos.

4.6. Inteligência artificial como potencial propagadora de desinformação

Como lembrado por Carraro (2024), tratando-se do uso da IA, torna-se inevitável debater sobre *fake news*, tema cada vez mais abordado nas discussões sobre os impactos dessa tecnologia na sociedade. Isso acontece porque, ao mesmo tempo em que a IA oferece inúmeros benefícios, ela também apresenta riscos na mesma proporção, especialmente no que se refere à disseminação de informações falsas.

Entre as ferramentas mais preocupantes estão os *deepfakes*, que utilizam modelos de aprendizado profundo para criar imagens, vídeos e áudios falsos, simulando rostos, vozes e expressões faciais de maneira extremamente convincente. Com ela, é possível criar

representações de pessoas praticando atos ou dizendo coisas que, na realidade, nunca aconteceram.

Em 2018, o ator e cineasta norte-americano Jordan Peele publicou, por meio de um comercial do site BuzzFeed, um vídeo em que dubla um discurso do ex-presidente dos Estados Unidos, Barack Obama. No vídeo, Peele utiliza sua habilidade de atuação e imitação, aliada a uma inteligência artificial capaz de sincronizar seus movimentos faciais e reproduzir sua voz de forma idêntica à de Obama (Figura 2). O objetivo era alertar sobre os perigos dessa tecnologia e o risco de não podermos mais confiar, nem mesmo, no que vemos em vídeo (McKay, 2018).

Figura 2 Comparativo entre o vídeo original gravado por Jordan Peele (à direita) e o resultado do *deepfake* sobre Barack Obama (à esquerda)
Fonte: Captura de tela do vídeo *You Won't Believe What Obama Says In This Video!* disponível no YouTube (BuzzFeedVideo, 2018).

Considerando a velocidade com que conteúdos são compartilhados e viralizam nas redes sociais, a depender do teor das falas presentes no vídeo, ele poderia ter um efeito extremamente danoso à imagem de Barack Obama, e esse não é um caso isolado, uma vez que a popularização de ferramentas de IA torna a produção de deepfakes cada vez mais acessível, não estando mais restrita apenas a especialistas, o que, consequentemente, amplia os riscos de disseminação de conteúdos falsos por indivíduos mal-intencionados.

Não é exagero afirmar que a IA está, mais do que nunca, próxima de produzir vídeos completamente indistinguíveis da realidade (algo que deve se concretizar em um futuro muito próximo), onde a distinção entre o que é real ou falso dependerá da plausibilidade do vídeo, ou seja, se o contexto apresentado faz sentido dentro da realidade, e do senso crítico de quem consome e compartilha esse tipo de conteúdo.

4.7. Alienação intelectual em prol da comodidade

Ao contrário do que sugere o senso comum (especialmente entre aqueles que não estão familiarizados com o tema), a inteligência artificial não é detentora de sabedoria absoluta. Suas respostas são fruto do processamento de grandes volumes de dados analisados, e não de um entendimento consciente sobre qualquer assunto, por isso é importante que toda informação gerada por IA seja analisada com senso crítico, para evitar equívocos ou a disseminação de informações falsas.

Em seu estudo sobre educação e inteligência artificial, Silva, C. (2023) define esse fenômeno como "a antropofagia do ChatGPT", fazendo um trocadilho sobre a deterioração da

capacidade intelectual humana decorrente do uso excessivo de IAs em contextos educativos. O autor argumenta que uma tecnologia capaz de responder como um ser humano, mas que recorre a modelos matemáticos para isso, acaba por reduzir a realidade humana a uma contingência, e não a uma possibilidade, o que conduz a processos de alienação intelectual.

Ele ainda ressalta que, à primeira vista, o ChatGPT parece libertar o ser humano do árduo trabalho de pensar, oferecendo respostas rápidas para questões urgentes da vida. No entanto, na prática, essa comodidade retira do próprio ser humano sua essência e, sobretudo, sua capacidade fundamental de reflexão.

Esta necessidade de reflexão se torna ainda mais evidente quando se observa que o próprio ChatGPT exibe, logo abaixo da interface de respostas, um aviso de que pode cometer erros, recomendando que informações importantes sejam verificadas (Figura 3), ou seja, nem mesmo a IA oferece garantias absolutas sobre a veracidade das respostas geradas.

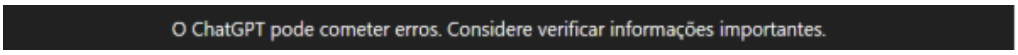

Figura 3 Aviso do ChatGPT para eventuais erros que o *chatbot* possa cometer
Fonte: Captura de tela do próprio ChatGPT (OpenAI, 2025).

Conclui-se, então, que, embora a IA seja uma ferramenta poderosa e útil, ela não substitui a capacidade humana de refletir e construir conhecimento, sendo seu uso em meios educativos algo que exige senso crítico e responsabilidade de quem a utiliza, para que não se transforme em um mecanismo de alienação.

4.8. Ações para o desenvolvimento e uso seguro de inteligências artificiais

Além das práticas já citadas no decorrer desse artigo, como a busca pela mitigação de vieses no treinamento dos modelos, a verificação da veracidade dos conteúdos gerados por inteligência artificial e o cuidado para evitar alienações intelectuais pelo uso em demasia dessas tecnologias, existem outras ações que podem contribuir para o uso seguro dessas ferramentas. O fato é que o futuro tende a contar com a presença cada vez mais constante das IAs no cotidiano das pessoas e, caso não existam diretrizes bem definidas para o seu uso, problemas certamente surgirão. Como apontam Sampaio, Sabbatini e Limongi (2024), no uso da IA é fundamental manter o alinhamento com princípios de responsabilidade e ética. Os autores também afirmam que, no desenvolvimento dessas tecnologias, o respeito à privacidade e aos direitos autorais é indispensável.

Por isso, a autoria humana deve ser preservada em contextos acadêmicos; assim, a IA pode auxiliar, mas não substitui a responsabilidade do autor na produção de conhecimento. Por isso, é necessário, e também uma boa prática, explicitar, em trabalhos acadêmicos e científicos, se a IA foi usada, além de quando e como ocorreu esse uso. Essas tecnologias devem estar a serviço dos pesquisadores, atuando como uma ferramenta que potencializa o trabalho humano, e não o contrário.

Carraro (2024) menciona Yann LeCun, cientista-chefe de IA da Meta e vencedor do Prêmio Turing, como exemplo de especialista que adota uma visão otimista sobre o futuro das IAs. Para LeCun, mesmo que essas tecnologias venham a superar a inteligência humana, ainda é possível confiar nos cientistas e especialistas que trabalham para desenvolvê-las de forma segura.

Um exemplo disso é o estudo publicado por Bricken (2023) e outros pesquisadores da empresa norte-americana Anthropic, especializada em pesquisa e desenvolvimento de IA. A equipe propôs uma nova abordagem para compreender melhor os modelos de linguagem: em vez de analisar neurônios individuais, passaram a observar as chamadas *features* – segundo o

dicionário inglês-português Linguee (2025), podem ser entendidas como “funcionalidades”, “recursos” ou “características” – com o objetivo de identificar padrões mais previsíveis no comportamento dos modelos. Isso indica que entender o funcionamento desses sistemas pode, aos poucos, deixar de ser um mistério científico e se tornar algo mais prático, como um trabalho de engenharia.

No Brasil, o Senado Federal aprovou recentemente o Projeto de Lei 2.338/2023, estabelecendo um marco regulatório para a inteligência artificial no Brasil. Esse Projeto de Lei tem como objetivo garantir que o desenvolvimento e a utilização de IAs no país ocorram de forma ética, respeitando os direitos fundamentais assegurados pela Constituição brasileira, por meio da definição de regras específicas para setores sensíveis, como saúde, segurança pública e educação, nos quais será exigido um nível mais elevado de responsabilidade por parte de desenvolvedores e operadores, com o intuito de mitigar violações de direitos e riscos à sociedade. Entre as medidas previstas, está a proibição do uso de reconhecimento facial em tempo real, além da vedação ao uso de IA em conteúdos ilegais, como materiais relacionados à exploração infantil.

Outro ponto de destaque é a preocupação com os direitos autorais no contexto das IAs generativas. As empresas que utilizarem conteúdos protegidos para o treinamento de IAs deverão declarar quais materiais foram utilizados, garantindo aos autores o direito de proibir a utilização de suas obras nesses sistemas.

Esse movimento legislativo se inspira em regulamentações internacionais, como a legislação europeia, e busca criar um ambiente jurídico que ofereça segurança tanto para desenvolvedores quanto para usuários (Governo do Brasil, 2024).

5. Considerações finais

Pode-se concluir que a inteligência artificial, embora represente um avanço expressivo na maneira como os sistemas simulam capacidades humanas e interagem com o mundo ao seu redor, também traz riscos importantes. Por meio de técnicas como o aprendizado de máquina e o aprendizado profundo, a IA tem sido aplicada em diversos setores, promovendo transformações significativas na sociedade.

Entretanto, problemas como o enviesamento dos dados, o desalinhamento entre as intenções humanas e as da máquina, além da disseminação de desinformação, mostram que o desenvolvimento da IA requer não apenas conhecimento técnico, mas também responsabilidade ética. Também é fundamental refletir sobre o uso da IA em ambientes educativos, uma vez que o acesso facilitado a respostas, quando não acompanhado de uma postura crítica, compromete o desenvolvimento do senso crítico do indivíduo, causando alienação.

Por fim, entende-se que o uso da IA deve ser guiado por diretrizes claras e responsáveis, que garantam sua aplicação alinhada aos valores humanos. A IA deve servir como apoio ao pensamento humano, e não como substituta, preservando assim a autoria e a integridade intelectual.

Agradecimentos

Agradeço aos professores que me acompanharam ao longo da minha trajetória como aluno, em especial ao professor Carlos Magnus Carlson Filho, pela orientação e apoio durante o desenvolvimento desse artigo.

Referências

AMBROS, Christiano; LODETTI, Daniel. **VIESES COGNITIVOS NA ATIVIDADE DE INTELIGÊNCIA: CONCEITOS, CATEGORIAS E MÉTODOS DE MITIGAÇÃO**. 2019. Disponível em: https://rbi.abin.gov.br/RBI/article/view/157/130. Acesso em: 26 maio 2025.

BARBOSA, Lucia Martins; PORTES, Luiza Alves Ferreira. **A INTELIGÊNCIA ARTIFICIAL**. 2019. Disponível em: https://abt-br.org.br/wp-content/uploads/2023/03/RTE_236.pdf#page=16. Acesso em: 05 maio 2025.

BBC NEWS BRASIL. **Como HAL 9000, o computador de '2001: Uma Odisseia no Espaço' previu preocupações atuais sobre IA**. G1, 2024. Disponível em: https://g1.globo.com/tecnologia/noticia/2024/08/05/como-hal-9000-o-computador-de-2001-uma-odisseia-no-espaco-previu-preocupacoes-atuais-sobre-ia.ghtml. Acesso em: 21 jun. 2025.

BRICKEN, Trenton *et al*. **Towards Monosemanticity: Decomposing Language Models With Dictionary Learning**. 2023. Disponível em: https://transformer-circuits.pub/2023/monosemantic-features/index.html. Acesso em: 11 jun. 2025.

BUZZFEEDVIDEO. **You Won't Believe What Obama Says In This Video!** 2018. Disponível em: https://www.youtube.com/watch?v=cQ54GDm1eL0. Acesso em: 21 jun. 2025.

CANO, Rosa Jiménez. **Carro sem motorista da Uber provoca primeiro acidente fatal**. 2018. Disponível em: https://brasil.elpais.com/brasil/2018/03/19/tecnologia/1521479089_032894.html. Acesso em: 22 maio 2025.

CARRARO, Fabrício. **Inteligência Artificial e ChatGPT**: Da revolução dos modelos de IA generativa à Engenharia de Prompt. São Paulo: Casa do Código, 2024.

ÉPOCA NEGÓCIOS. **Amazon desiste de ferramenta secreta de recrutamento que mostrou viés contra mulheres**. 2018. Disponível em: https://epocanegocios.globo.com/Empresa/noticia/2018/10/amazon-desiste-de-ferramenta-secreta-de-recrutamento-que-mostrou-vies-contra-mulheres.html. Acesso em: 26 maio 2025.

GARCIA, Ana Cristina Bicharra. **ÉTICA E INTELIGÊNCIA ARTIFICIAL**. 2020. Disponível em: https://journals-sol.sbc.org.br/index.php/comp-br/article/view/1791/1625. Acesso em: 05 maio 2025.

GOVERNO DO BRASIL. **Senado Federal aprova marco regulatório da inteligência artificial**. 2024. Disponível em: https://www.gov.br/cultura/pt-br/assuntos/noticias/senado-federal-aprova-marco-regulatorio-da-inteligencia-artificial. Acesso em: 11 jun. 2025.

LINGUEE. **Feature**. 2025. Disponível em: https://www.linguee.com.br/ingles-portugues/search?source=auto&query=feature. Acesso em: 11 jun. 2025.

MCCARTHY, John. **What is Artificial Intelligence?** Stanford University, 2007. Disponível em: https://www-formal.stanford.edu/jmc/whatisai.pdf. Acesso em: 11 jun. 2025.

MCKAY, Tom. **Este vídeo mostra como a tecnologia consegue facilmente criar uma fraude de Barack Obama**. 2018. Disponível em: https://gizmodo.uol.com.br/jordan-peele-video-falso-barack-obama/. Acesso em: 28 maio 2025.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. **Conceitos sobre Aprendizado de Máquina**. 2003. Disponível em: https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf. Acesso em: 05 maio 2025.

OPENAI. **ChatGPT**. 2025. Disponível em: https://chatgpt.com. Acesso em: 18 jun. 2025.

PRIBERAM. **Enviesamento**. 2025. Disponível em: https://dicionario.priberam.org/enviesamento. Acesso em: 11 jun. 2025.

SAMPAIO, Rafael Cardoso; SABBATINI, Marcelo; LIMONGI, Ricardo. **Diretrizes para o uso ético e responsável da Inteligência Artificial Generativa**. 2024. Disponível em: https://prpg.unicamp.br/wp-content/uploads/sites/10/2025/01/livro-diretrizes-ia-1.pdf. Acesso em: 01 jun. 2025.

SILVA, Cláudio Nei Nascimento da. **A EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA E INTELIGÊNCIA ARTIFICIAL: UM APELO À FORMAÇÃO INTEGRAL ANTE A ANTROPOFAGIA DO CHATGPT**. 2023. Disponível em: https://preprints.scielo.org/index.php/scielo/preprint/view/7708/14438. Acesso em: 05 maio 2025.

SILVA, Vinicius Lopes da. **ÉTICA E RESPONSABILIDADE NA ERA DA INTELIGÊNCIA ARTIFICIAL: APRENDIZAGEM DIGITAL NO CHAT GPT**. 2023. Disponível em: https://repositorio.unipampa.edu.br/bitstream/riu/8334/1/Vinicius%20Lopes%20da%20Silva%202023.pdf. Acesso em: 20 mar. 2025.