

Riscos Éticos da Inteligência Artificial: privacidade, viés algorítmico e governança

Gustavo Lunkes¹, Luan C. Muhlbeier¹, Gustavo M. Lima¹, Ayrán Dhroel Bade¹,
Tales U. W. F. Rosário², Julio C. E. Lauermann³, Jefferson dos Santos Vorpagel⁴

¹Curso de Engenharia de Software — Faculdade Donaduzzi

²Curso de Análise e Desenvolvimento de Sistemas — Faculdade Donaduzzi

³Curso de Inteligência Artificial — Faculdade Donaduzzi

⁴Faculdade Donaduzzi — Toledo, PR

{gustavolunkes1, luancarlosmuhlbeier, talesuriel, ayranbade,
juliolauermann16}@gmail.com, {gustavomarkeslima}@[outlook.com](mailto:gustavomarkeslima@outlook.com),
{jeffersonvorpagel}@hotmail.com

Abstract. Artificial Intelligence has evolved rapidly, bringing significant advances in productivity, automation, and decision-making. However, this technological advancement also introduces critical challenges such as algorithmic bias, digital surveillance, automated disinformation, cybersecurity threats, and concentration of power in large corporations. This paper discusses the “dark side” of AI, the risks and ethical dilemmas arising from its adoption, with emphasis on privacy, social bias, automation of malicious tasks, and the limits of human control over increasingly autonomous systems. The discussion is grounded in real cases and current developments, demonstrating that responsible and ethical use of AI is not optional, but imperative for society.

Keywords. Artificial Intelligence, algorithmic bias, digital surveillance, AI ethics, AI governance.

Resumo. A Inteligência Artificial evoluiu rapidamente, trazendo avanços expressivos em produtividade, automação e tomada de decisões. No entanto, esse avanço introduz desafios críticos como viés algorítmico, vigilância digital, automação de desinformação, ameaças à cibersegurança e concentração de poder em grandes corporações. Este trabalho discute o “lado sombrio” da IA, os riscos e dilemas éticos decorrentes de sua adoção, com ênfase em privacidade, preconceito social, automação de tarefas maliciosas e os limites do controle humano sobre sistemas cada vez mais autônomos. A discussão é fundamentada em casos reais e desenvolvimentos atuais, demonstrando que o uso responsável e ético da IA não é opcional, mas imperativo para a sociedade.

Palavras-chave. Inteligência Artificial, viés algorítmico, vigilância digital, ética em IA e governança de IA.

1. Introdução

Nos últimos anos, a Inteligência Artificial (IA) deixou de ser um conceito futurista e passou a integrar o cotidiano das pessoas, das empresas e dos governos. Algoritmos de recomendação, assistentes virtuais, sistemas de triagem médica e ferramentas de análise de crédito são exemplos de como a IA permeia decisões antes exclusivamente humanas. Diante dessa presença crescente da IA em decisões sociais, econômicas e institucionais, até que ponto esses sistemas podem ser considerados confiáveis, transparentes e justos?

Este trabalho apresenta um panorama crítico, se valendo de uma metodologia de revisão bibliográfica dos principais riscos associados à adoção da IA, estruturado em cinco dimensões: privacidade e vigilância, viés algorítmico, automação de atividades maliciosas, impactos sobre o mercado de trabalho e os limites éticos do controle humano.

2. Objetivo

Analisar os principais riscos éticos associados ao uso da Inteligência Artificial, com foco em privacidade, viés algorítmico, automação de usos maliciosos, impactos no mercado de trabalho e limites do controle humano.

3. Referencial Teórico

A discussão sobre os riscos da IA não é nova. Desde as primeiras reflexões de Norbert Wiener sobre a cibernética [Wiener 1948], pesquisadores alertam para os perigos de sistemas autônomos que escapam ao controle humano. Kate Crawford [Crawford 2021] e Safiya Noble [Noble 2018] documentaram sistematicamente como sistemas de IA reproduzem e amplificam preconceitos sociais presentes nos dados de treinamento.

No campo da privacidade, Shoshana Zuboff [Zuboff 2019] introduziu o conceito de “capitalismo de vigilância”, descrevendo como grandes empresas de tecnologia utilizam dados comportamentais para prever e influenciar decisões humanas sem pleno consentimento dos usuários. Na área de cibersegurança, estudos recentes demonstram que ferramentas baseadas em IA são capazes de automatizar a identificação de vulnerabilidades em sistemas computacionais, acelerando tanto a defesa quanto o ataque digital [Anderson et al. 2019]. O debate sobre IA e mercado de trabalho é polarizado entre visões catastrofistas e otimistas, com evidências empíricas indicando impactos diferenciados conforme o setor e o nível de qualificação dos trabalhadores [Acemoglu e Restrepo 2019].

4. Metodologia

Este trabalho adota uma abordagem qualitativo-descritiva de revisão bibliográfica sistemática e análise documental. A opção pelo caráter qualitativo se justifica pelo objetivo de compreender e interpretar fenômenos sociotécnicos complexos — como viés algorítmico e vigilância digital — a partir de múltiplas perspectivas teóricas. Os dados quantitativos citados ao longo do texto (como percentuais de pesquisas e estimativas de automação) são utilizados de forma ilustrativa, como evidências

empíricas que contextualizam a análise qualitativa, e não como objeto de tratamento estatístico próprio.

A revisão bibliográfica foi conduzida entre março e abril de 2026, utilizando as bases de dados Google Scholar, Scopus e ACM Digital Library. As palavras-chave empregadas nas buscas foram: "artificial intelligence ethics", "algorithmic bias", "AI surveillance", "AI governance", "AI labor market" e seus equivalentes em português. Foram adotados os seguintes critérios de seleção: (i) publicações a partir de 2018, priorizando produções dos últimos cinco anos; (ii) obras seminais anteriores a esse período, quando reconhecidas como referências consolidadas na área (como Wiener [1948]); e (iii) relevância direta para os cinco eixos temáticos definidos. Ao todo, foram analisados 27 trabalhos, entre artigos científicos, livros acadêmicos, relatórios de institutos de pesquisa (como McKinsey e Pew Research Center) e documentação técnica de empresas de IA. Os materiais foram organizados por eixo temático e interpretados mediante análise de conteúdo, buscando identificar convergências, divergências e lacunas na literatura.

A pesquisa foi estruturada em cinco eixos temáticos: (1) privacidade e vigilância digital; (2) viés algorítmico e discriminação; (3) automação de atividades maliciosas; (4) impactos sobre o emprego; e (5) limites do controle humano sobre sistemas autônomos. Para cada eixo, foram identificados casos reais, posicionamentos de especialistas e, quando disponíveis, dados empíricos quantitativos de suporte. A análise buscou mapear tanto as convergências entre os autores — especialmente quanto à urgência da regulação e à insuficiência dos mecanismos de controle atuais — quanto as divergências, notadamente no que se refere ao grau de impacto da IA sobre o mercado de trabalho.

5. Resultados

5.1. Privacidade e Vigilância

A coleta massiva de dados é o combustível dos sistemas de IA modernos. Algoritmos de recomendação processam dados comportamentais em tempo real para prever preferências e influenciar decisões de consumo. Uma pesquisa do Pew Research Center com 5.101 adultos norte-americanos revelou que 67% afirmam entender pouco ou nada sobre o que as empresas fazem com seus dados pessoais, e 73% acreditam ter pouco ou nenhum controle sobre essas informações [Pew Research Center 2023].

Tecnologias de reconhecimento facial, implementadas por governos e empresas, levantam questões sérias sobre vigilância em massa. Em países como a China, esses sistemas são utilizados para monitorar o comportamento público em escala sem precedentes. No Ocidente, o debate regulatório ainda é incipiente, com legislações como o GDPR europeu representando avanços, mas ainda insuficientes diante do ritmo de evolução tecnológica.

A análise da literatura revela convergência entre Zuboff [2019] e os dados do Pew Research Center [2023] quanto à assimetria estrutural entre empresas e usuários: enquanto organizações acumulam capacidade crescente de rastrear e prever comportamentos, os indivíduos permanecem amplamente desinformados sobre o alcance dessa coleta. Em perspectiva crítica, os autores deste trabalho avaliam que a abordagem regulatória atual é predominantemente reativa — legislações surgem após consolidação de práticas problemáticas — o que aponta para a necessidade de marcos normativos prospectivos que antecipem, e não apenas respondam, aos riscos da vigilância algorítmica.

5.2. Viés Algorítmico

O viés algorítmico ocorre quando sistemas treinados com dados históricos reproduzem desigualdades já existentes. Em áreas como crédito, contratação e justiça, isso pode gerar decisões automatizadas injustas e difíceis de contestar [Crawford 2021; Noble 2018].

Um caso paradigmático é o sistema COMPAS, utilizado no judiciário norte-americano para prever a probabilidade de reincidência criminal. Investigações jornalísticas demonstraram que o algoritmo atribuía pontuações de risco mais altas a réus negros do que a réus brancos com perfis criminais similares. Esse tipo de viés, quando incorporado a decisões institucionais, tem o potencial de institucionalizar discriminações de forma opaca e difícil de contestar.

Crawford [2021] e Noble [2018] convergem ao apontar que o viés algorítmico não é um defeito técnico isolado, mas reflexo de estruturas sociais históricas embutidas nos dados de treinamento. A divergência entre ambas reside na ênfase: Crawford foca nas dimensões de poder e infraestrutura que sustentam esses sistemas, enquanto Noble concentra sua análise nos efeitos discriminatórios diretos sobre grupos marginalizados, especialmente mulheres negras. Na avaliação dos autores deste trabalho, ambas as perspectivas são complementares e necessárias: não basta corrigir dados tendenciosos se os incentivos econômicos e as relações de poder que os produziram permanecerem intactos. Isso implica que soluções técnicas isoladas, como re-balanceamento de datasets, são insuficientes sem acompanhamento de mudanças institucionais e regulatórias.

5.3. Automação de Atividades Maliciosas

A mesma capacidade de automação que torna a IA útil para tarefas legítimas pode ser explorada para fins maliciosos. Ferramentas de geração de texto e imagem podem ser utilizadas para criar desinformação em escala, falsificar documentos, produzir deep fakes convincentes e automatizar ataques de phishing altamente personalizados.

No campo da cibersegurança, modelos de IA treinados para identificar vulnerabilidades em código representam avanço significativo tanto para a defesa quanto para o ataque digital. Ferramentas de geração automatizada de malware, phishing

personalizado e exploração de falhas em sistemas foram documentadas em estudos de segurança ofensiva [Anderson et al. 2019]. Esse duplo potencial, ao mesmo tempo defensivo e ofensivo, ilustra um dilema central: quanto mais poderosa é a ferramenta, maior o risco de uso indevido.

A literatura neste eixo converge em torno da chamada "corrida armamentista" tecnológica: avanços em defesa e ataque digital ocorrem de forma simétrica e acelerada, sem que mecanismos de governança consigam acompanhar o ritmo. Os autores deste trabalho identificam uma lacuna importante na produção acadêmica analisada: há maior ênfase nos riscos do que em arquiteturas concretas de contenção. Entende-se que iniciativas como o rastreamento de proveniência de conteúdo gerado por IA (conforme proposto pelo C2PA — Coalition for Content Provenance and Authenticity) representam uma linha promissora, mas que ainda carecem de adoção em escala e respaldo regulatório internacional.

5.4. Impactos sobre o Mercado de Trabalho

A automação impulsionada pela IA tem o potencial de transformar profundamente o mercado de trabalho. Estudos do McKinsey Global Institute [McKinsey 2023] estimam que entre 60% e 70% das horas de trabalho na economia global têm potencial técnico de automação, com impactos desproporcionais sobre trabalhadores de baixa e média qualificação. Sem políticas públicas adequadas de requalificação profissional, a IA pode aprofundar desigualdades existentes.

Este é o eixo de maior divergência na literatura revisada. Acemoglu e Restrepo [2019] apresentam evidências empíricas de que a automação destrói postos de trabalho em ritmo mais rápido do que cria novas funções, especialmente em setores de manufatura e serviços rotineiros. Em contraposição, perspectivas mais otimistas — como as de relatórios do Fórum Econômico Mundial — argumentam que tecnologias disruptivas historicamente geram mais empregos do que eliminam, desde que acompanhadas de investimento em educação e requalificação. Na avaliação dos autores, a polarização entre catastrofismo e otimismo tecnológico dificulta a formulação de políticas públicas equilibradas. O consenso possível reside no caráter desigual da transição: independentemente do saldo líquido de empregos, os custos recaem desproporcionalmente sobre trabalhadores com menor escolaridade e menor acesso a programas de requalificação — o que torna a dimensão distributiva do problema tão relevante quanto a quantitativa.

5.5. Limites do Controle Humano

À medida que sistemas de IA ganham autonomia crescente, surgem questões fundamentais sobre responsabilidade e controle. Grandes modelos de linguagem podem produzir saídas imprevisíveis e difíceis de interpretar, mesmo por seus próprios desenvolvedores, fenômeno conhecido como “caixa-preta” algorítmica. Organizações como a Anthropic, a OpenAI e o DeepMind têm investido em pesquisas sobre

“alinhamento de IA”, o problema de garantir que sistemas de IA ajam de acordo com valores humanos, embora o campo ainda esteja em estágios iniciais [Crawford 2021].

A revisão da literatura permite identificar uma síntese transversal aos cinco eixos analisados: os riscos éticos da IA não são acidentais, mas estruturais — resultam de escolhas de design, de incentivos econômicos e de lacunas regulatórias. Crawford [2021] e Wiener [1948], separados por décadas, convergem na mesma advertência: sistemas técnicos poderosos sem supervisão adequada tendem a escapar ao controle humano de formas imprevisíveis. Os autores deste trabalho posicionam-se no sentido de que a solução não é frear o desenvolvimento da IA, mas sim condicionar seu avanço a exigências concretas de transparência, auditabilidade e responsabilização. Mecanismos como auditorias algorítmicas independentes, obrigatoriedade de explicabilidade em decisões de alto impacto e responsabilidade civil objetiva para desenvolvedores de sistemas autônomos representam passos institucionais necessários — e tecnicamente viáveis — nessa direção.

6. Conclusão

Os riscos da IA não estão apenas na tecnologia em si, mas nas escolhas feitas durante seu desenvolvimento, treinamento, aplicação e fiscalização. Vigilância em massa, discriminação algorítmica, automação de atividades maliciosas, desemprego estrutural e perda de controle humano são desafios que demandam ação consciente de pesquisadores, desenvolvedores, reguladores e da sociedade civil.

Conclui-se que a Inteligência Artificial oferece benefícios relevantes, mas sua adoção exige transparência, auditoria, regulação e governança responsável. Os riscos identificados neste trabalho indicam que o avanço tecnológico precisa ser acompanhado por critérios éticos e mecanismos de responsabilização. Iniciativas como o AI Act europeu e as diretrizes de IA responsável da UNESCO representam passos importantes nessa direção.

Referências

- Acemoglu, D. e Restrepo, P. (2019). “Automation and New Tasks: How Technology Displaces and Reinstates Labor”. *Journal of Economic Perspectives*, 33(2), 3-30. Disponível em: https://www.google.com.br/books/edition/Automation_and_New_Tasks/9pGH0AEA-CAAJ?hl=pt-BR
- Anderson, R., Barton, C. e Böhme, R. et al. (2019). “Measuring the Changing Cost of Cybercrime”. *Proceedings of the 18th Workshop on the Economics of Information Security (WEIS 2019)*. Boston College. Disponível em: https://weis2019.econinfosec.org/wp-content/uploads/sites/6/2019/05/WEIS_2019_paper_25.pdf

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press. Disponível em: https://www.google.com.br/books/edition/The_Atlas_of_AI/KfodEAAAQBAJ?hl=pt-BR&gbpv=0

McKinsey Global Institute (2023). “The Economic Potential of Generative AI: The Next Productivity Frontier”. McKinsey & Company. Disponível em: <https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier>

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press. Disponível em: https://www.google.com.br/books/edition/Algorithms_of_Oppression/g8OSDgAAQBAJ?hl=pt-BR&gbpv=0

Pew Research Center (2023). “How Americans View Data Privacy”. Pew Research Center. Disponível em: <https://www.pewresearch.org/internet/2023/10/18/how-americans-view-data-privacy/>. Acesso em: 22 abr. 2026.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press. Disponível em: https://www.google.com.br/books/edition/Cybernetics_Or_Control_and_Communication/NnM-uISyywAC?hl=pt-BR&gbpv=0&kptab=overview

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs. Disponível em: https://www.google.com.br/books/edition/The_Age_of_Surveillance_Capitalism/MJ5ADwAAQBAJ?hl=pt-BR&gbpv=0