INTELIGÊNCIA ARTIFICIAL E ÉTICA: Um caminho em construção

1 INTRODUÇÃO

Nas últimas décadas, o ritmo de transformação das tecnologias digitais remodelou práticas sociais, produtivas e institucionais. Ao mesmo tempo em que esses sistemas ampliam capacidades analíticas e operacionais, emergem questões éticas que exigem exame cuidadoso e fundamentado, conforme já discutido por autores que analisam a inserção responsável de tecnologias no âmbito social (SICHMAN, 2021).

Diversas pesquisas têm alertado para a exacerbação de vieses e preconceitos por sistemas inteligentes, uma vez que a inteligência da máquina depende diretamente da qualidade e do teor dos dados aos quais é submetida (GARCIA, 2020). Além disso, a complexidade inerente a muitos algoritmos de aprendizado profundo (*Deep Learning*) resulta em uma "caixa-preta" que impede a compreensão de como as decisões são tomadas, gerando uma falsa aura de imparcialidade devido ao seu embasamento matemático (LOPES; MENDES, 2023).

Nesse cenário, torna-se indispensável refletir não apenas sobre as potencialidades técnicas da Inteligência Artificial (IA), mas também sobre os limites éticos que orientam sua produção e aplicação. De acordo com Srour (2018), a ética científica constitui um corpo de conhecimentos que permite observar, descrever e explicar fatos morais, emitindo juízos de realidade e identificando padrões que possibilitam antecipar eventos. Essa perspectiva é particularmente relevante no campo da IA, onde a tomada de decisões algorítmicas pode gerar consequências amplas e imprevistas, exigindo um olhar que ultrapasse o senso comum e se fundamente em critérios científicos de avaliação moral.

Entre as inovações tecnológicas emergentes, a Inteligência Artificial destaca-se por alcançar de modo disruptivo as dimensões econômicas, sociais e ambientais e reconfigurar a relação humano-máquina, suscitando desafios como transparência algorítmica, vieses em modelos de aprendizado e governança de dados (COECKELBERGH, 2023). A amplitude desses impactos evidencia que a discussão sobre IA não pode restringir-se à dimensão técnica. É preciso também considerar os dilemas éticos que ela levanta. Srour (2018) propõe que a ética científica funciona como um instrumento de análise rigorosa desses dilemas, ao confrontar racionalidades universalistas (voltadas ao bem comum) e particularistas (orientadas a interesses excludentes). Aplicada à IA, a ética possibilita avaliar até que ponto o desenvolvimento tecnológico favorece a reprodução social das coletividades humanas ou, ao contrário, amplia desigualdades e exclusões.

Diante deste panorama, esse estudo tem como objetivo discutir as questões éticas associadas à aplicação de tecnologias de IA.

2 FUNDAMENTAÇÃO TEÓRICA

Conforme apontado por Cozman, Plonski e Neri (2021), a IA carece de uma definição consensual, sendo um conceito multifacetado que se transformou ao longo do tempo. A IA pode ser definida como o estudo de computadores que realizam tarefas que, no momento, são melhor executadas por humanos, ou "sistemas que pensam como humanos", "agem como humanos", "pensam racionalmente" ou "agem racionalmente".

Tendo em vista a rápida evolução e implementação da IA no cotidiano social, torna-se importante articular princípios e salvaguardas que mantenham a tecnologia alinhada a valores humanos, segurança e bem-estar, como proposto no debate sobre Inteligência Artificial Centrada no Ser Humano, sem perder de vista contribuições complementares da filosofia, da tecnologia e nos estudos críticos de ciência e sociedade, como em Shneiderman (2020).

Nesse contexto, é pertinente adotar também a perspectiva da ética científica. Para Srour (2018), a ética científica corresponde a um corpo de conhecimentos que observa, descreve e explica fatos morais, elabora conceitos, emite juízos de realidade e identifica padrões recorrentes, permitindo antecipar eventos com boa margem de probabilidade. Ao confrontar racionalidades universalistas, voltadas ao bem comum, e particularistas, de caráter excludente, esse enquadramento contribui para analisar como a Inteligência Artificial pode tanto reforçar práticas sociais inclusivas quanto acentuar exclusões e desigualdades.

Ao considerar a participação da IA em cadeias de ação, torna-se evidente que sua contribuição para resultados no mundo não autoriza tratá-la como sujeito moral. A agência e a responsabilidade distribuem-se entre equipes de desenvolvimento, organizações implementadoras e instâncias regulatórias, o que cria lacunas de *accountability* em contextos de alta complexidade. Danos podem ocorrer sem que seja trivial identificar responsáveis, o que reforça a necessidade de mecanismos de rastreabilidade e de prestação de contas ao longo de todo o ciclo de vida dos sistemas (COECKELBERGH, 2023). Em resposta, propostas de arquitetura orientadas por princípios de segurança, avaliação contínua de risco e governança explícita buscam reduzir a probabilidade de danos e viabilizar reparação quando eventos adversos acontecem (SHNEIDERMAN, 2020).

A ausência do conhecimento dos resultados gerados pela IA compromete a auditabilidade e como consequência, a confiança de usuários e reguladores diminui, sobretudo em domínios de alto impacto social como saúde, justiça e crédito (BURRELL, 2016; LIPTON, 2018; DORAN; SCHULZ; BESOLD, 2017). Métodos de explicabilidade *pós-hoc*, como o *Local Interpretable Model-agnostic Explanations* (LIME) e o *Shapley Additive Explanations* (SHAP), aproximam fronteiras de decisões complexas ou decompõem resultados em contribuições de variáveis, produzindo subsídios para escrutínio técnico, relatórios de auditoria e decisões informadas por pessoas responsáveis pela operação de sistemas críticos (RIBEIRO; SINGH; GUESTRIN, 2016; LUNDBERG *et al.*, 2020). A explicabilidade, contudo, não é um fim em si mesma. Ela é mais eficaz quando integrada a procedimentos institucionais de avaliação ética e a processos de governança de dados.

A partir deste contexto, emergem os desafios de viés e justiça algorítmica. Dados históricos, escolhas de modelagem e objetivos de otimização podem reproduzir e intensificar desigualdades. Escores de risco criminal e iniciativas de policiamento preditivo produziram assimetrias que afetaram grupos específicos (ANGWIN *et al.*, 2016; LUM; ISAAC, 2016; COECKELBERGH, 2023). Em consequência, *frameworks* de ética algorítmica defendem qualidade e representatividade dos dados, proporcionalidade de uso, avaliação de impacto e monitoramento contínuo. Persiste o entendimento de que medidas técnicas precisam ser conectadas a instâncias de revisão e participação social, capazes de prevenir injustiças e corrigir desvios quando identificados (SAMPAIO; SABBATINI; LIMONGI, 2024).

Um exemplo notório desse problema ocorreu em 2018, quando a *Amazon* investiu em um sistema inteligente para pré-selecionar currículos, que acabou por discriminar candidatos do sexo feminino. O sistema, treinado com dados históricos predominantemente masculinos, aprendeu a penalizar currículos que continham palavras como "mulher" e outras associações femininas. O conhecimento contido nas bases de dados pode estar datado, e decisões tomadas em um contexto histórico podem ser inaceitáveis em outro (GARCIA, 2020).

O estudo de Garcia (2020), também revelou um viés racial em sistemas inteligentes na área da saúde. Uma empresa de seguro saúde norte-americana, ao buscar reduzir custos, um algoritmo para identificar pacientes com doenças crônicas graves que deveriam receber tratamentos preventivos. O critério de "caso crítico" estava associado ao quanto o paciente usava o sistema de saúde, o que, por sua vez, refletia desigualdades arraigadas: pacientes negros, devido a barreiras de acesso e prescrição, utilizavam menos os serviços de saúde, sendo, portanto, menos propensos a serem incluídos na lista para tratamento preventivo. Este exemplo

demonstra que os vieses podem estar escondidos nos dados, tornando sua identificação uma tarefa complexa (GARCIA, 2020).

Os temas de justiça se articulam às discussões de privacidade e vigilância. A expansão de reconhecimento facial, de rastreamento digital e de processamento massivo de dados pessoais impõe exigências de consentimento, finalidade e proporcionalidade. Abordagens de Inteligência Artificial Centrada no Ser Humano propõem transparência, contestabilidade e devido processo legal aplicado a sistemas algorítmicos, além de mecanismos de auditoria e reparação. Sem desenho institucional robusto e práticas efetivas de auditabilidade, tais propostas tendem a permanecer isoladas e pouco articuladas a políticas públicas e regulações setoriais. Ao mesmo tempo, a crítica pública ao modelo de monetização de dados comportamentais ganhou visibilidade, trazendo para o centro do debate a autonomia informacional do cidadão e a necessidade de controles externos sobre ecossistemas de dados (THE SOCIAL DILEMMA, 2020).

Em paralelo, a IA produz reconfigurações socioeconômicas relevantes. Estimativas variam conforme o método e o setor, mas convergem quanto ao aumento da substituição de tarefas rotineiras e previsíveis, com reforço de atividades que exigem julgamento contextual, interação social e criatividade. Habilidades complementares à IA, como pensamento crítico, colaboração e empatia, ganham centralidade, ao passo que respostas de política, como atualização de sistemas de proteção social, iniciativas de renda mínima e investimentos educacionais orientados a competências não rotineiras, são discutidas para favorecer transições justas e reduzir assimetrias de adoção tecnológica (COECKELBERGH, 2023).

A dimensão ambiental também impõe outra camada de responsabilidade. O treinamento e a operação de modelos de grande porte demandam significativa capacidade computacional e consumo energético, com impactos associados à pegada de carbono. Estudos apontam emissões relevantes ao longo de fases de treinamento e inferência e chamam atenção para ordens de grandeza que, em cenários específicos, se aproximam de atividades intensivas em energia. Recomenda-se explicitar indicadores de custo ambiental, estimular pesquisa em eficiência de modelos e de *hardware* e adotar métricas que relacionem desempenho a consumo energético (SCHWARTZ *et al.,* 2020).

As tensões apresentadas demonstram que os dilemas da Inteligência Artificial não podem ser reduzidos a uma oposição simplista entre bem e mal. Como observa Srour (2018), dilemas éticos frequentemente implicam escolhas entre dois bens (bem preferencial e bem preterido) ou entre dois males (mal menor e mal maior). Essa perspectiva amplia o horizonte analítico ao mostrar que, diante da IA, não se trata apenas de adotar princípios abstratos, mas de construir critérios científicos que orientem escolhas éticas complexas em contextos de incerteza tecnológica

3 METODOLOGIA

O método desse trabalho caracteriza-se como bibliográfico e qualitativo. A pesquisa bibliográfica é realizada com base em documentos, principalmente livros e artigos. Gil (2017, p. 152) ressalta que a pesquisa bibliográfica possibilita “ao investigador a cobertura de uma gama de fenômenos muito mais ampla do que aquela que poderia pesquisar diretamente”.

A abordagem qualitativa possibilita relevar aspectos da realidade que não podem ser explicados e medidos quantitativamente, ou seja, tem como foco a compreensão de temas a fim de exemplificar a dinâmica das relações sociais. A pesquisa qualitativa, trabalha com o universo de significados, motivos, aspirações, crenças, valores e atitudes (MINAYO, DESLANDES, GOMES; 2001).

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

A consolidação dos achados teóricos indica que a responsabilidade por danos e benefícios decorrentes de sistemas de Inteligência Artificial é uma propriedade sociotécnica distribuída. Não se trata de atribuir agência moral à máquina, mas de mapear como decisões de projeto, escolhas de dados, modelos de negócio e marcos regulatórios se combinam para produzir efeitos no mundo. Esse enquadramento desloca a análise de culpas individualizadas para arranjos institucionais com rastreabilidade e prestação de contas ao longo de todo o ciclo de vida dos sistemas, da concepção à desativação, e demanda governança explícita, avaliação contínua de riscos e supervisão humana significativa (COECKELBERGH, 2020; SHNEIDERMAN, 2020).

Nesse contexto, a explicabilidade revela-se condição necessária, porém não suficiente, para a confiança social. Procedimentos *pós-hoc* como LIME e SHAP auxiliam a compreender localmente por que um resultado foi produzido e a quantificar a contribuição das variáveis. Entretanto, quando desvinculados de processos organizacionais de revisão ética e de governança de dados, esses métodos tendem a gerar explicações formais de utilidade limitada para a responsabilização prática, sobretudo em domínios de alto impacto social como saúde, justiça e crédito (BURRELL, 2016; LIPTON, 2018; DORAN; SCHULZ; BESOLD, 2017; RIBEIRO; SINGH; GUESTRIN, 2016; LUNDBERG *et al.*, 2020).

Os resultados dos estudos citados acima também reforçam que justiça algorítmica não se resume à correção estatística de modelos. As métricas de equidade embutem escolhas normativas e relações de compromisso entre erros de classificação em diferentes grupos, o que exige deliberação pública e critérios transparentes de aceitabilidade. Os casos relatados na literatura, mostram como vieses estruturais podem ser internalizados pelo modelo e permanecer invisíveis se as variáveis de decisão não forem transparentes (GARCIA, 2020). A implicação prática é acoplar técnicas de mitigação a processos participativos de definição de objetivos, políticas de dados com representatividade e revisão periódica de impactos, a fim de prevenir injustiças e corrigir distorções quando identificada (ANGWIN *et al.*, 2016; LUM; ISAAC, 2016; COECKELBERGH, 2020).

As tensões entre justiça e privacidade agravam-se em ecossistemas de dados intensivos. Abordagens de Inteligência Artificial Centrada no Ser Humano propõem salvaguardas que se mostram mais efetivas quando integradas à minimização de dados, anonimização adequada e controles externos independentes (SHNEIDERMAN, 2020).

Do ponto de vista socioeconômico, observa-se reconfiguração do trabalho com substituição de tarefas rotineiras e ascensão de atividades que combinam julgamento contextual, interação social e criatividade. Listas de ocupações supostamente ameaçadas são instrumentos descritivos limitados, pois desconsideram fatores institucionais como organização do processo produtivo, relações laborais e políticas de qualificação. Implicações práticas incluem programas de requalificação orientados a competências não rotineiras, atualização de proteções sociais e pactos setoriais de adoção responsável para favorecer transições justas (COECKELBERGH, 2020).

A dimensão ambiental adiciona uma camada decisiva de responsabilidade. O treinamento e a operação de modelos de grande porte demandam energia e recursos computacionais expressivos. A ética de eficiência ecológica proposta na literatura recomenda explicitar orçamentos de treino e inferência, adotar métricas de desempenho por unidade de energia, incentivar pesquisa em modelos e *hardware* mais eficientes e priorizar contratação de energia de menor intensidade de carbono, além de estratégias como reaproveitamento de calor em *data centers* (SCHWARTZ *et al*., 2020).

À luz da ética científica, que distingue entre racionalidades universalistas e particularistas, é possível prover critérios para escolhas éticas em contextos de incerteza tecnológica, elevando a probabilidade de benefícios sociais líquidos e de alinhamento às metas

do Desenvolvimento Sustentável (SROUR, 2018; COECKELBERGH, 2020; SHNEIDERMAN, 2020;).

5 CONCLUSÃO

Os resultados discutidos indicam que a ética em Inteligência Artificial deve ser entendida como problema sociotécnico, que demanda soluções combinando desenho técnico, arranjos institucionais e controle social informado. A evidência reunida mostra que a responsabilidade por danos e benefícios não reside em agentes isolados, mas é distribuída ao longo do ciclo de vida dos sistemas. Esse enquadramento exige mecanismos robustos de rastreabilidade, prestação de contas e possibilidade real de contestação, ancorados em governança explícita e supervisão humana significativa, com atenção a contextos de alto impacto social.

A discussão também reforça que a explicabilidade é condição necessária para a confiança, porém insuficiente quando dissociada de processos organizacionais e regulatórios. Abordagens *pós-hoc* como LIME e SHAP são úteis, se articuladas a escolhas de modelagem transparentes quando viáveis, documentação rigorosa de dados e modelos e instâncias formais de revisão ética. No campo da justiça algorítmica, os achados destacam que não há métrica universal e as decisões sobre equidade envolvem escolhas normativas, devendo ser tratadas por meio de processos participativos, políticas de dados com representatividade e monitoramento contínuo de impactos.

Quanto à privacidade e vigilância, a consolidação de ecossistemas intensivos em dados impõe salvaguardas de finalidade, proporcionalidade e consentimento informado, além de transparência e devido processo aplicado a sistemas algorítmicos. No plano socioeconômico, a substituição de tarefas rotineiras e a reconfiguração ocupacional apontam para políticas de qualificação continuada e proteção social, de modo a favorecer transições justas. Por fim, a dimensão ambiental recomenda metas explícitas de eficiência energética e de emissões, bem como métricas de desempenho por unidade de energia, de forma a internalizar custos ambientais no processo decisório.

Recomenda-se, portanto, ampliar investigações sobre a efetividade de explicações na tomada de decisão humana, os efeitos distributivos de intervenções algorítmicas e a mensuração padronizada da pegada ambiental da IA.

REFERÊNCIAS

ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L. **Machine bias.** ProPublica, 23 maio 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 07 mai. 2025.

BURRELL, J. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. **Big Data & Society**, v. 3, n. 1, p. 1-12, 2016. DOI: 10.1177/2053951715622512

COECKELBERGH, M. **Ética na Inteligência Artificial**. Tradução de Clarisse de Souza et al. São Paulo/Rio de Janeiro: Ubu/Editora PUC-Rio, 2023.

COZMAN, F. G.; PLONSKI, G. A.; NERI, H. (orgs.). **Inteligência Artificial**: Avanços e Tendências. São Paulo: Instituto de Estudos Avançados, 2021. DOI: 10.11606/9786587773131.

DORAN, D.; SCHULZ, S.; BESOLD, T. R. What does explainable AI really mean? A new conceptualization of perspectives. **CEUR Workshop Proceedings**, v. 2071, p. 1–8, 2018. Disponível em: https://ceur-ws.org/Vol-2071/CExAIIA_2017_paper_2.pdf.

GARCIA, A. C. B. Ética e inteligência artificial. **Computação Brasil**, [S. l.], p. 14–22, nov. 2020.

GIL, A. C. **Como elaborar projetos de pesquisa**. 6 ed. São Paulo: Atlas, 2017.

LIPTON, Z. C. The mythos of model interpretability. **Queue**, v. 16, n. 3, p. 31-57, 2018. DOI: 10.1145/3236386.3241340.

LOPES, C. de M. N.; MENDES, J. C. Ética e inteligência artificial: desafios e melhores práticas. **Revista da UFMG**, Belo Horizonte, v. 30, fluxo contínuo, e47673, 2023. DOI: 10.35699/2965-6931.2023.47673.

LUNDBERG, S. M.; ERION, G.; CHEN, H. *et al.* From local explanations to global understanding with explainable AI for trees. **Nature Machine Intelligence**, v. 2, p. 56-67, 2020. DOI: 10.1038/s42256-019-0138-9.

LUM, K.; ISAAC, W. To predict and serve? **Significance**, v. 13, n. 5, p. 14-19, 2016. DOI: 10.1111/j.1740-9713.2016.00960.

MINAYO, M. C. S.; DESLANDES, S. F.; GOMES, R. **Pesquisa social**: teoria, método e criatividade. Petrópolis: Vozes, 2016.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. 2016. DOI: 10.1145/2939672.2939778.

SAMPAIO, R. C.; SABBATINI, M.; LIMONGI, R. **Diretrizes para o uso ético e responsável da Inteligência Artificial Generativa:** um guia prático para pesquisadores. São Paulo: Editora Intercom, 2024.

SCHWARTZ, R. *et al.* Green AI. **Communications of the ACM**, v. 63, n. 12, p. 54 63, 2020. DOI: 10.1145/3381831.

SHNEIDERMAN, B. Human-centered artificial intelligence: reliable, safe & trustworthy. **International Journal of Human-Computer Interaction**, v. 36, n. 6, p. 495-504, 2020. DOI: 10.1080/10447318.2020.1741118.

SICHMAN, J. S. Inteligência Artificial e sociedade: avanços e riscos. **Estudos Avançados**, v. 35, n. 101, p. 37-49, 2021. DOI: 10.1590/s0103-4014.2021.35101.004. Disponível em: https://www.scielo.br/j/ea/a/c4sqqrthGMS3ngdBhGWtKhh/. Acesso em: 8 jan. 2025.

SROUR, R. H. **Ética empresarial**. 5. ed. Rio de Janeiro: Elsevier, 2018. 266 p.

THE SOCIAL DILEMMA. Direção: Jeff Orlowski. Estados Unidos: Netflix, 2020. Documentário. 94 min.