Princípios Éticos na IA: O Impacto da Governança Ética na Redução de Vieses

Vanessa G. L. Pessoa[1]

[1]Centro de Informática– Universidade Federal da Paraíba (UFPB) João Pessoa– PB– Brazil

vanessalima@cc.ci.ufpb.br

Abstract. *With the increasing use of artificial intelligence, algorithmic biases—capable of replicating and amplifying existing societal prejudices—have become a significant challenge. This article explores how different organizations are addressing these issues by adopting ethical principles, including responsibility, transparency and explainability, data protection, fairness and equity, autonomy, human oversight, safety, and sustainability. A case study involving five companies illustrates differences in their approaches but also reveals convergence around fundamental values. The research highlights the importance of ethical governance, affirming that its effectiveness depends not only on the consistent application of these principles but also on the existence of clear regulations.*

Resumo. *Com o crescente uso de inteligência artificial, os vieses algorítmicos que podem replicar e amplificar preconceitos existentes na sociedade tornaram-se um desafio significativo. Este artigo explora como diferentes organizações estão abordando esses problemas através da adoção de princípios éticos, incluindo responsabilidade, transparência, explicabilidade, proteção de dados, justiça e equidade, autonomia, supervisão humana, segurança e sustentabilidade. Um estudo de caso envolvendo cinco empresas ilustra as diferenças em suas abordagens, mas também revela uma convergência em torno dos valores fundamentais. A pesquisa destaca a importância da governança ética ao afirmar que sua eficácia depende tanto da aplicação consistente desses princípios quanto à existência clara de regulamentações.*

1. Introdução

O crescimento da Inteligência Artificial é inegável e está cada vez mais presente em nosso cotidiano. Contudo, a disseminação desses sistemas levanta questões sobre a nossa capacidade de nos adaptarmos adequadamente ao seu uso. Em consequência do rápido avanço tecnológico, não antecipamos todas as implicações éticas e sociais decorrentes deste impacto da IA na sociedade.

O aprendizado das máquinas pode ser compreendido como a capacidade dos sistemas de aprender e adaptar-se à sua programação original [Carvalho, 2020]. Essa característica cria uma distinção entre as atividades do programador e as decisões automáticas tomadas pelo algoritmo, que em alguns casos, pode até mesmo gerar seu próprio código; isso torna mais difícil aos desenvolvedores explicarem todo o processo envolvido nessas escolhas [Requião and Costa, 2022]. Esse cenário levanta duas questões importantes: a opacidade que impede de entender com clareza como os critérios utilizados pelos algoritmos para a tomada de decisões; além da qualidade dos dados utilizados, que pode perpetuar os vieses já presentes na sociedade. Um exemplo

notável é o viés racial em algoritmos que avaliam as necessidades médicas com base nos custos dos cuidados, o que desfavorece pacientes negros que historicamente têm menos acesso à saúde [Obermeyer et al 2019]. Esses desafios destacam a importância de uma governança ética robusta para mitigar esses vieses discriminatórios e promover equidade nos sistemas inteligentes.

Este artigo está estruturado da seguinte maneira: a seção 2 oferece uma revisão da literatura sobre classificações em Inteligência Artificial, destacando o viés algorítmico e as implicações éticas do uso dessa tecnologia. Na seção 3, é discutida a questão ética e sua relevância na mitigação de vieses. Em seguida, na seção 4, apresentamos a metodologia adotada nesta pesquisa para possibilitar que um estudo de caso seja realizado na seção 5; este estuda cinco modelos de governança ética e analisa suas implementações dos princípios éticos. A discussão dos resultados ocorre na seção 6. Por fim, a seção 7 traz considerações finais e perspectivas futuras referentes à governança ética no desenvolvimento dos sistemas IA.

2. Referencial Teórico

Nesta seção, abordamos dois conceitos fundamentais para compreender a governança ética na inteligência artificial: a própria Inteligência Artificial e os preconceitos algorítmicos. Primeiro, são apresentados os conceitos e capacidades da inteligência artificial, diferenciando entre IA fraca e forte. Em seguida, discutimos o viés algorítmico como um fator que pode comprometer a equidade desses sistemas. Também analisamos como esses vieses podem surgir e ressaltamos a importância de uma taxonomia estruturada para entender suas origens. Esses tópicos estabelecem uma base teórica essencial para analisar as estruturas de governança ética nas seções seguintes.

2.1. Inteligência Artificial

Inteligência Artificial é uma área da computação que tem como objetivo simular a capacidade humana de resolver problemas [IBM]. Envolve diversos campos incluindo estatística, análise de dados, redes neurais e engenharia de hardware. Todos esses aspectos convergem para criar sistemas e máquinas capazes de raciocinar, aprender e agir como humanos - ou até mesmo processar grandes informações em escalas além do alcance humano comum [Cloud].

2.1.1. Capacidade da Inteligência Artificial

A capacidade da Inteligência Artificial pode ser diferenciada através de termos “IA fraca” ou “IA forte” [IBM]. A primeira, também conhecida como IA estreita, tem habilidade limitada com base em seu desenvolvimento e treinamento para executar conjuntos restritos. Enquanto a segunda é descrita como um sistema capaz de raciocinar e pensar na mesma medida que os seres humanos [Perez et al. 2016]. O principal objetivo no campo da IA é alcançar o que nos referimos como IA forte [Groumps 2023], ao longo da história houve vários momentos de grandes expectativas que iam além da capacidade computacional da época e da nossa compreensão sobre o funcionamento cerebral humano para possibilitar replicá-lo fielmente nesses sistemas [Perez et al. 2016].

Sistemas totalmente automatizados, ou seja, aquelas capazes de tomar decisões complexas sem qualquer intervenção humana, ainda estão apenas no campo teórico e amplamente discutidos em debates públicos - algo que motiva organizações a exagerarem as capacidades de suas tecnologias inteligentes [Moertl and Ebinger 2024; Bertoncini and Serafim 2023]. No entanto, a falta de entendimento sobre os limites da capacidade da IA pode resultar na dificuldade de distinguir qual é a responsabilidade do sistema e quando passa ser do operador humano. Em geral, os sistemas de inteligência artificial são destinados a auxiliar em vez de substituir totalmente as atividades do operador humano [Moertl and Ebinger 2024].

2.2. Viés Algorítmico

Ao tratar de vieses algorítmicos, é importante ter clareza sobre duas definições fundamentais: o conceito de algoritmo e o significado do termo viés. O viés é a tendência de ser contra ou a favor de indivíduos ou grupos com base em suas identidades sociais. [Kordzadeh and Ghasemaghaei 2021]. Já um algoritmo consiste em uma série ordenada de instruções que devem ser seguidas para se alcançar determinados resultados.

Os vieses humanos são reproduzidos pelos algoritmos, já que estes são criados e treinados por humanos [Simões-Gomes, Roberto and Mendonça 2020]. O uso da Inteligência Artificial pode reforçar os problemas sociais existentes, causando diferenciações, colocações preferenciais ou exclusões capazes de afetar o tratamento igualitário entre as pessoas [Requião and Costa 2022].

Apesar de o termo viés geralmente ter uma conotação negativa, nem todos os vieses são indesejáveis [Hellstrom, Dignum e Bensch 2020]. O impacto que um viés terá depende do contexto em que é aplicado e dos fatores contextuais envolvidos [Falzepour and Danks 2021]. Por exemplo, se uma empresa deseja aumentar a diversidade contratando pessoas de grupos minoritários, ela pode utilizar um sistema de IA projetado para favorecer currículos dessas minorias [Hellstrom, Dignum and Bensch 2020].

No caso dos algoritmos de previsão utilizados para identificar e ajudar pacientes com necessidades complexas de saúde, pode haver a introdução de viés racial quando estes empregam o custo dos cuidados médicos como parâmetro para avaliar as necessidades dos pacientes. Esse critério se mostra injusto pelo fato de que os pacientes negros tendem a receber menos assistência médica do que os brancos devido às desigualdades sistêmicas no acesso à saúde [Obermeyer et al., 2019].

Diante dessas questões, a equidade em sistemas que utilizam IA é de extrema importância, especialmente quando esses sistemas tomam decisões cruciais que impactam a vida das pessoas, como contratações ou cuidados médicos. Um sistema de IA só pode ser considerado justo se conseguir prever seu desempenho com precisão [Rieskamp et al., 2023] e sem perpetuar desigualdade estruturais.

2.2.1 Taxonomía de Viés Algorítmico

A taxonomia de vieses algorítmicos é uma sistemática classificação dos diversos tipos de viés existentes. Esta classificação ajuda a identificar e categorizar as diferentes formas pelas quais o viés pode influenciar os resultados da Inteligência Artificial

[Hellstrom, Dignum e Bensch 2020]. A agregação desses vieses em categorias facilita na compreensão da sua origem, seus efeitos, permitindo propor medidas para mitigá-los.

Os tipos de vieses não são mutuamente exclusivos e podem se relacionar em diferentes fases da construção e implantação dos sistemas de IA [Chaudhary 2024; Ferrara 2024]. A Tabela 1 apresenta os tipos de vieses que podem ocorrer com maior frequência em determinadas etapas do ecossistema de um sistema de Inteligência Artificial.

Tabela 1. Visão Geral dos Vieses em Sistemas de Inteligência Artificial

Etapa	**Tipo de Viés**	**Descrição**	**Referências selecionadas**
Especificação do problema	Viés de enquadramento	O problema é definido de forma que limita as possíveis soluções ou alternativas	Hellstrom, Dignum and Bensch 2020; Nazer et al. 2023
	Viés de interesse próprio	A definição do problema reflete interesses específicos organizações ignorando o benefício de todos os envolvidos	Hellstrom, Dignum and Bensch 2020
Coleta de Dados	Viés de representação	Os dados de entrada não são representativos para a população relevante	Giffen, Herhausen and Fahse 2022; Ferrara 2024
	Viés de rótulo	Quando os rótulos dos dados durante o processo de treinamento de um modelo são inconsistentes e tendenciosos	Giffen, Herhausen and Fahse 2022; Cozman and Kaufman 2022; Ferrara 2024
	Viés social	Os dados refletem viés existentes na sociedade	Giffen, Herhausen and Fahse 2022; Kordzadeh and Ghasemaghai 2021
Modelagem e validação	Viés de algorítmico	Considerações técnicas inadequadas ou enviesadas durante a modelagem	Giffen, Herhausen and Fahse 2022; Cozman and Kaufman 2022
	Viés de avaliação	São utilizadas população de	Cozman and Kaufman

		teste não representativa ou métricas de desempenho inadequadas durante a avaliação do modelo	2022; Ferrara 2024
Implantação	Viés de implantação	Modelo é utilizado em contexto diferente para qual foi criado	Giffen, Herhausen and Fahse 2022; Nazer et al. 2023
	Viés de feedback	O resultado de um sistema é utilizado como entrada para o mesmo sistema ou para futuros processos de decisão. Decisões anteriores, que podem ter sido enviesadas, reforçam ainda mais esse viés em interações subsequentes	Giffen, Herhausen and Fahse 2022; Chaudhary 2024
	Viés de interpretação	Pode ocorrer quando o indivíduo interpreta uma saída com base em seus preconceitos pessoais	Chaudhary 2024
	Viés de opacidade dos resultados	Quando os resultados de um modelo não são facilmente compreensíveis e acessíveis para os usuários ou partes interessadas	Chaudhary 2024; Giffen, Herhausen and Fahse 2022;

A primeira fase de especificação do problema é uma etapa crucial no desenvolvimento de projetos de IA, pois envolve a definição dos objetivos gerais, requisitos e limitações do problema [Mitchell et al. 2021 apud Fazelpour and Danks 2021]. Introduzimos vieses no sistema de IA tanto ao definirmos os objetivos quanto às variáveis-alvo usadas para medir se alcançamos o objetivo proposto.

Ao iniciar a análise dos dados, diversos tipos de vieses podem ser encontrados. Os próprios dados refletem os vieses existentes na população e sua manipulação pode gerar viés de representação ou rótulo. Essa etapa é crucial para o modelo aprendido e algumas vezes as estratégias conscientemente escolhidas visam alterar os desequilíbrios sociais [Hellstrom, Dignum and Bensch 2020];

Durante a fase de modelagem e validação, é necessário levar em conta considerações técnicas sobre quais recursos incluir ou excluir, pois isso pode introduzir viés discriminatório no modelo. Além disso, os critérios de avaliação geralmente envolvem avaliar o desempenho do modelo em relação a certas métricas de sucesso [Fazelpour and Danks 2021], envolvendo julgamentos de valor que dependem de objetivos e valores organizacionais.

3. Governança ética

Conforme os sistemas de IA assumam tarefas cognitivas com dimensões sociais, anteriormente realizadas exclusivamente por humanos, é imperativo que incorporem os requisitos sociais [Bostrom and Yudkowsky 2011]. O uso crescente da IA em processos decisórios com impacto significativo na vida das pessoas exige o desenvolvimento de estruturas de governança baseadas em princípios éticos sólidos.

A governança ética da inteligência artificial é estruturada em quatro componentes inter-relacionados: teorias éticas, estruturas de governança, princípios éticos e diretrizes éticas. Esses elementos trabalham conjuntamente para assegurar o desenvolvimento responsável e a utilização adequada da IA. As teorias éticas procuram identificar o que torna uma ação mais ou menos ética [Stahl 2018], fornecendo assim uma base sólida para criar as estruturas de governança no campo da IA [Nassar and Kamal 2021; Bertoncini and Serafim 2023]. Essas estruturas funcionam como mecanismos práticos que convertem ideias abstratas das teorias em modelos organizacionais concretos. Dentro dessas estruturas, os princípios éticos atuam como metas que orientam a governança em termos como responsabilidade, segurança, justiça e transparência. Por fim, as diretrizes éticas oferecem instruções específicas às organizações, garantindo assim a implementação dos princípios e o alinhamento da inteligência artificial com os valores sociais e éticos.

3.1. Teorias Éticas na Governança de IA

A seguir, as teorias éticas mais influentes, juntamente com suas potenciais aplicações no contexto da governança ética em IA:

1. Utilitarismo: O utilitarismo promove a maximização do bem-estar, incentivando os responsáveis pelas decisões a analisar os efeitos de suas escolhas [Nassar and Kamal 2021]. No âmbito da inteligência artificial, esse conceito sugere que é fundamental garantir o desenvolvimento de sistemas destinados a proporcionar mais benefícios do que prejuízos, considerando as consequências tanto no curto quanto no longo prazo [Agarwal 2023].
2. Deontologia: A ética deontológica foca no que é intrinsecamente certo e errado nas ações, destacando o dever moral do agente que a executa [Stahl 2021]. Quando aplicada à IA, ressalta a importância para os desenvolvedores e usuários de seguirem regras e princípios éticos [Nassar and Kamal 2021].
3. Ética das virtudes: Essa abordagem ética foca no caráter e nas virtudes de indivíduos e instituições [Agarwal 2023]. No contexto da IA, isso significa promover uma cultura de integridade e responsabilidade, valorizando princípios como justiça, honestidade e transparência entre desenvolvedores, usuários e empresas [Nassar and Kamal 2021].
4. Contratualismo: O contratualismo fundamenta-se na ideia de que as normas morais são originárias de acordos sociais. Quando se trata da inteligência artificial, isso significa que os sistemas devem ser desenvolvidos para ganhar aceitação por todas as pessoas afetadas por sua utilização [Nassar and Kamal 2021].

5. Consequencialismo: As teorias consequencialistas focam nos resultados das ações realizadas [Stahl 2021]. No contexto da inteligência artificial, essas teorias orientam os decisores a considerar tanto as consequências negativas quanto positivas de suas decisões.

Essas teorias podem ser utilizadas de forma conjunta para orientar a governança ética da IA, promovendo o equilíbrio entre os princípios e abordando diversas preocupações relacionadas ao desenvolvimento e uso de sistemas inteligentes. Os valores éticos e as exigências legais devem ser incorporados como critérios nas decisões dos sistemas de IA, visando fornecer resultados mais transparentes e responsáveis [Bertoncini and Serafim 2023].

3.2. Princípios da Governança Ética

A governança ética deve definir e implementar procedimentos e padrões que orientem o desenvolvimento, uso e gerenciamento dos sistemas de inteligência artificial. Seu objetivo é minimizar os riscos associados à tecnologia, como a introdução de vieses discriminatórios, e promover seu uso em benefício do bem comum [Sigfrids et al., 2022]. Estruturas de governança ética têm o papel crucial de traduzir princípios éticos já conhecidos em objetivos compreensíveis [Lundgren 2023] que visam mitigar diretamente os vieses algorítmicos.

Os princípios mencionados a seguir promovem a confiança, garantem a aceitação social e reforçam o compromisso das entidades com o desenvolvimento de sistemas de IA que estejam alinhados aos valores humanos e ao bem-estar social [Agarwal 2023].

1. Responsabilidade:

   O princípio da responsabilidade se preocupa principalmente com a questão de quem responsabilizar pelas ações e decisões dos sistemas de IA, em quais aspectos e em que medida [Cheng and Liu 2022; Robles and Mallison 2023]. Em casos como o de um sistema de IA utilizado para diagnóstico médicos, a questão é: se o sistema errar na classificação de uma doença e o médico confiar cegamente no resultado, repassando o diagnóstico sem maior avaliação, quem deve ser responsabilizado? Quem desenvolveu e implementou o sistema inteligente ou o médico que o utilizou sem avaliação crítica? Esse é um dilema ético frequente em discussões sobre responsabilidade em IA [Stahl 2021], onde se busca um equilíbrio entre responsabilidade técnica e a responsabilidade profissional dos usuários [De Almeida, Dos Santos and Farias 2021]. Além disso, os usuários esperam pelo bom funcionamento dos sistemas [Camilleri 2023] e as falhas podem prejudicar essa confiabilidade.

2. Transparência e explicabilidade

   A falta de transparência pode dificultar a capacidade de responsabilizar os sistemas de IA por suas ações [Agarwal 2023]. Compreender o funcionamento dos sistemas de IA é pode ser um desafio, pois muitas vezes eles são vistos como uma "caixa-preta" [Perez et al. 2018; Bertoncini e Serafim 2023], dificultando ainda mais a tradução dos conceitos algorítmicos e decisões em termos que sejam claros para os usuários [Camilleri 2023]. Contudo, ter governança

transparente na área da IA não implica necessariamente que as pessoas precisam entender o código. Em vez disso, significa oferecer oportunidades para diálogo e expressão das preocupações relacionadas ao uso desses sistemas [Robles and Mallison 2023].

3. Privacidade e proteção de dados

   O princípio da privacidade e proteção de dados diz respeito ao direito de controlar quem pode acessar suas informações pessoais. O uso sem autorização ou consentimento dos dados coletados constitui uma violação da privacidade [Camilleri 2023, Kumar 2024]. Um grande problema no manuseio desses dados é que muitas vezes os usuários desconhecem que estão fornecendo essas informações e não sabem como elas serão utilizadas. Portanto, a transparência nas práticas de tratamento de dados é fundamental; deve-se informar aos usuários sobre a coleta, processamento e utilização dessas informações, além de estabelecer mecanismos para permitir o acesso, modificação e exclusão dos próprios dados pelos usuários [Agarwal 2023].

4. Justiça e equidade

   O princípio da justiça em inteligência artificial busca corrigir os vieses algorítmicos, que podem ser introduzidos de maneira intencional ou não [Camilleri 2023]. Para isso, é necessário um sistema de governança capaz de fornecer métodos práticos para o desenvolvimento dos algoritmos a fim de evitar discriminação e, ao mesmo tempo, respeitar valores humanos [Robles and Mallison 2023; Agarwal 2023]. Isso requer um projeto ético e deliberado com constante envolvimento das partes interessadas. A participação contínua dessas partes - como desenvolvedores, usuários e reguladores - é essencial para assegurar que os sistemas de IA reflitam e promovam tais valores [Kumar 2024].

5. Liberdade e autonomia

   O princípio de liberdade e autonomia está relacionado à autodeterminação através de meios democráticos, ao direito de estabelecer e desenvolver relações com outras pessoas, bem como à liberdade para retirar consentimento ou utilizar uma plataforma ou tecnologia da preferência própria [Jobin, Ienca and Vayena 2019]. As inteligências artificiais influenciam o ambiente das possíveis ações ao fornecerem ou ocultar informações; tal comportamento pode diminuir a autonomia individual comprometendo a liberdade de escolha [Stahl 2021].

6. Supervisão humana

   Os sistemas de inteligência artificial devem ser desenvolvidos para complementar a autonomia humana, assegurando que suas decisões sejam transparentes e fáceis de entender. Isso permite que as pessoas mantenham o controle sobre suas ações e escolhas [Kutz et al 2023]. Além disso, é importante que as estruturas de governança integrem processos de revisão ética e avaliação de impacto [Agarwal 2023], estabelecendo mecanismos adequados para supervisão humana em situações críticas, com o objetivo de aumentar a segurança e responsabilidade.

7. Segurança e robustez

   O princípio de segurança e robustez enfatiza que as estruturas de governança em inteligência artificial devem contemplar planos de contingência para lidar com eventuais falhas, visando minimizar ou evitar danos não intencionais [Kutz et al. 2023]. Como os sistemas de IA aprendem por meio das interações com seus usuários, eles estão vulneráveis a ataques por parte de agentes mal-intencionados, o que transforma a questão da segurança em uma preocupação persistente [Camilleri 2023]. Além disso, esses sistemas podem ser usados para descobrir e explorar fraquezas existentes, aumentando assim os riscos relacionados à segurança [Krafft et al. 2020 apud Stahl, 2021]. Assim sendo, assegurar que a IA seja robusta e resiliente é crucial para mitigar tais riscos e proteger tanto os usuários quanto às informações processadas por estes sistemas.

8. Sustentabilidade

   Com o uso crescente de sistemas de IA, aumentam as preocupações sobre seus impactos ambientais. O princípio da sustentabilidade ressalta a importância da proteção do meio ambiente, dos ecossistemas e da biodiversidade [Jobim, Ienca, and Vayena 2019]. Embora a inteligência artificial possa ajudar na redução do consumo energético ao simplificar processos e melhorar a eficiência, sua implementação exige uma quantidade significativa de recursos tanto no desenvolvimento quanto na operação contínua [Stahl, 2021]. Novos produtos e serviços baseados em IA podem acarretar impactos ambientais negativos ao intensificarem questões como alto consumo energético e geração de resíduos eletrônicos . Portanto, é vital que a governança relacionada à IA adote estratégias voltadas para promover a sustentabilidade reduzindo danos ao ambiente.

Os princípios éticos, quando acompanhados de diretrizes específicas e objetivas, desempenham um papel crucial na identificação de vieses nos dados e nos processos de desenvolvimento dos sistemas de inteligência artificial. Ao serem aplicados desde as primeiras etapas de desenvolvimento, esses princípios fornecem não apenas orientações para o uso responsável da IA, mas também abordam questões sociais, éticas e relacionadas aos direitos humanos [Stahl 2021]. Com essa abordagem, as diretrizes éticas ajudam a equilibrar os benefícios tecnológicos da IA enquanto mitigam os riscos associados a decisões enviesadas e discriminatórias.

3.3. Níveis de aplicação da Governança Ética

Os mecanismos de governança da IA podem ser divididos em dois tipos de ética digital: rígida e branda. A ética rígida abrange leis formais e outras normas sancionadas, desenvolvidas através de processos legislativos oficiais [Sigfrids 2022], com o objetivo de garantir a conformidade e impor sanções em caso de violação. Em contraste, a ética branda envolve a adoção voluntária de princípios éticos, recomendações e diretrizes éticas, sem imposição legal obrigatória [Floridi 2018; Jobin, Ienca and Vayena 2019].

Existem, portanto, três níveis de aplicação de governança: Estado (política e legislação), organizações e mecanismos de orientação e integração com a sociedade [Auld et al. 2022]. O nível de aplicação do Estado, se refere à prática e implementação

de políticas iniciadas por autoridades. Um exemplo disso é a Lei de Serviços Digitais na União Europeia. No Brasil, um caso relevante dessa atuação é representado pela Lei Geral de Proteção de Dados (LGPD). O principal desafio nesse nível reside no rápido desenvolvimento tecnológico em larga escala das tecnologias de IA, criando uma brecha entre as diretrizes éticas estabelecidas e as leis existentes [Pöhler, Diepold and Wallach 2024].

O nível de aplicação de organizações refere-se às diretrizes éticas implementadas dentro de empresas e instituições. Essas diretrizes estabelecem premissas sobre como o sistema de IA deve operar em relação às partes interessadas, promovendo comportamento desejáveis além dos requisitos da lei, como os códigos de conduta [Birkstedt et al. 2022]. No entanto, as principais empresas tecnológicas exercem uma influência significativa na definição de questões de IA e na formulação pertinentes a esta área; essa influência pode acentuar desequilíbrios de poder e desigualdades sociais [Taeihagh 2021].

A aplicação de orientação e integração com a sociedade, no contexto de governança ética, refere-se ao envolvimento e desenvolvimento de sistemas de IA que priorizem o bem-estar social e a dignidade humana. Isso inclui não apenas a regulamentação e supervisão, mas também a garantia de que os sistemas de IA respeitem os direitos humanos [Singh 2021]. Além disso, existe uma ênfase na transparência dos sistemas de IA; entretanto muitos indivíduos carecem do conhecimento técnico necessário para supervisionar as decisões técnicas destes sistemas. Ademais, os sistemas baseados em inteligência artificial são inerentemente imprevisíveis devido à sua dependência dos dados fornecidos: pequenas alterações nas entradas podem resultar em diferenças significativas nos resultados.

A eficácia de cada abordagem varia conforme o contexto social e o nível de aplicação. As regulamentações formais são essenciais para garantir segurança e proteger os direitos fundamentais, mas sozinhas podem não ser suficientes para guiar a inteligência artificial de maneira ética e benéfica [Sigfrids 2022]. Nesse sentido, práticas éticas brandas podem atuar como um complemento ao promover boas práticas voluntárias que vão além das exigências legais [Jobin, Ienca and Vayena 2019]. Contudo, essas diretrizes éticas podem falhar quando empresas ou governos atuam em contextos onde os direitos humanos são negligenciados [Floridi 2018]. Assim sendo, é necessário encontrar um equilíbrio entre as duas abordagens com base nas particularidades específicas de cada situação.

Figura 1. Interseção entre ética digital rígida e branda com os níveis de aplicação de governança em IA

4. Metodologia

Este estudo utiliza uma abordagem qualitativa, centrada na análise da problemática do viés algorítmico e da proposta da governança ética em sistemas inteligentes como forma de mitigar os vieses discriminatórios. O método qualitativo possibilita explorar fenômenos complexos e subjetivos ligados às implicações éticas e sociais [Cresswell 2021].

O método envolve a análise dos regulamentos organizacionais de empresas globais no que tange às estruturas de governança ética da IA. As organizações selecionadas para o estudo são Microsoft, Google, Accenture, Salesforce e IBM. O objetivo é examinar e contrastar os princípios éticos relacionados à estrutura de governança ética da inteligência artificial dessas empresas com aqueles discutidos neste artigo. A metodologia segue as etapas abaixo:

1. Seleção das empresas: Serão selecionadas empresas que possuam políticas públicas documentadas sobre governança de IA.
2. Coleta de dados: A coleta de dados será realizada através da análise de relatórios de governança e políticas de IA das empresas. O foco será nas seções que tratam dos princípios éticos.
3. Análise das diretrizes éticas: A análise será baseada nos oitos princípios de governança ética em IA discutidos no artigo: responsabilidade, transparência e explicabilidade, privacidade e proteção de dados, justiça e equidade, liberdade e

autonomia, supervisão humana, segurança e robustez e sustentabilidade. Para cada princípio, será avaliada a presença ou ausência de diretrizes específicas nas políticas das empresas, bem como o grau de detalhamento dessas diretrizes.

4. Discussão dos resultados: A discussão buscará compreender em que medida as diretrizes organizacionais refletem os princípios de governança ética da IA mencionados.

5. Estudos de casos

Para entender como diversas organizações tratam a governança ética na inteligência artificial, foi realizado um estudo de caso com cinco empresas: Microsoft, Google, Accenture, Salesforce e IBM. Essas companhias foram selecionadas devido à sua importância no campo tecnológico e no desenvolvimento da inteligência artificial.

Cada estudo de caso examinará a estrutura de governança adotada pela empresa, destacando os princípios éticos que estão em alinhamento com os oito princípios apresentados neste artigo: responsabilidade, transparência e explicabilidade, privacidade e proteção de dados, justiça e equidade, liberdade e autonomia, supervisão humana, segurança e robustez, e sustentabilidade.

5.1. Estudo de Caso: Microsoft

A Microsoft é uma corporação global de tecnologia, reconhecida pelo desenvolvimento de plataformas e ferramentas digitais, com foco crescente em soluções impulsionadas por IA. A empresa emprega mais de 200 mil colaboradores em todo o mundo [Microsoft],estando presente no Brasil há 35 anos [Microsoft Mais Brasil - Relatório de Impacto].

A empresa divulgou suas diretrizes internas [Microsoft 2022] para projetar inteligência artificial de forma responsável, estabelecendo princípios como metas para orientar seu desenvolvimento e uso internos. No entanto, os princípios de liberdade e autonomia, supervisão humana e sustentabilidade não foram incluídos nas diretrizes éticas da Microsoft. A seguir, apresentamos uma comparação entre os princípios contemplados pela governança ética da Microsoft e aqueles discutidos neste artigo:

1. Responsabilidade

   Os sistemas de IA da Microsoft passam por Avaliações de Impacto durante todo o ciclo de desenvolvimento, reforçando o monitoramento contínuo de seus efeitos. Trata a responsabilidade não apenas em termos de impacto social e organizacional, mas também integra dois outros princípios da governança ética: supervisão e controle humanos, além da privacidade e proteção dos dados, garantindo que as soluções sejam apropriadas e eficazes para seus propósitos estabelecidos

2. Transparência e explicabilidade

   Os sistemas são desenvolvidos para oferecer suporte à explicabilidade às partes interessadas, garantindo uma comunicação clara sobre seus recursos e limitações. Além disso, asseguram que os usuários sejam informados quando

estão interagindo com um sistema de IA ou conteúdos gerados por algoritmos que imitam a sua autenticidade.

3. Justiça e equidade

   A Microsoft aborda esse princípio dividindo-o em dois aspectos: justiça e inclusão. A Microsoft projeta suas IAs para minimizar disparidades nos resultados entre diferentes grupos demográficos e evitar a reprodução de estereótipos. Na inclusão, há um foco na aderência aos padrões de acessibilidade previamente estabelecidos pelos Padrões da Microsoft.

4. Segurança e robustez

   A Microsoft chama esse princípio de confiabilidade e prevenção de riscos. Os sistemas de IA da empresa são projetados para minimizar o tempo necessário para corrigir falhas conhecidas ou previsíveis, além de estarem constantemente monitorados dentro dos intervalos operacionais seguros.

5. Privacidade e proteção dos dados

   Neste princípio, a proteção dos dados é abordada com foco específico nos padrões de conformidade internos na empresa: Padrão de Privacidade da Microsoft e Política de Segurança da Microsoft.

Além de incorporar esses princípios em suas diretrizes, a governança ética permite que a Microsoft colabore com outras empresas em iniciativas como o AI for Good. Esta plataforma contínua promove debates e soluções práticas alinhadas aos Objetivos de Desenvolvimento Sustentável da ONU [International Telecommunication Union, 2024]. A AI for Good Foundation está estabelecendo padrões para boas práticas na implementação empresarial, focando áreas como viés e equidade, privacidade dos dados, democratização do acesso à tecnologia e governança responsável [Ethics - AI for Good Foundation]. Essas áreas convergem no objetivo comum de promover equidade e ampliar o acesso seguro aos dados.

5.2. Estudo de Caso: Google

A Google é uma empresa multinacional amplamente reconhecida por seus serviços online e desenvolvimento de software. A companhia emprega aproximadamente 300 mil pessoas [Google] e tem sua sede brasileira situada em São Paulo. Em seu site oficial [Google], a empresa reafirma o compromisso com o desenvolvimento responsável da inteligência artificial, destacando princípios fundamentais que orientam suas práticas. Entre esses princípios estão justiça e equidade, segurança e robustez, responsabilidade, privacidade e proteção dos dados; no entanto, ficam ausentes os aspectos de transparência e explicabilidade, liberdade e autonomia, supervisão humana e sustentabilidade.

1. Justiça e equidade

   O Google compromete-se a evitar impactos injustos sobre as pessoas, especialmente aqueles relacionados a características sensíveis como raça, etnia, renda, orientação sexual, capacidade e crença política ou religiosa.

2. Segurança e robustez

Os sistemas de IA do Google são projetados de acordo com as melhores práticas estabelecidas por pesquisa de segurança em IA e, em determinados casos, serão testados em ambientes restritos e monitorados em sua operação pós implantação.

3. Responsabilidade

   Esse princípio é discutido em conjunto com o da transparência e do controle humano. Os sistemas são projetados para permitir que os usuários forneçam feedback sobre seu funcionamento, além de oferecer explicações relevantes e compreensíveis sobre como as decisões foram tomadas. Isso promove que os humanos tenham a capacidade de supervisionar e intervir nas operações quando necessário.

4. Privacidade e proteção dos dados

   O Google aplica esse princípio oferecendo a oportunidade de notificação e consentimento, além de garantir transparência e controle adequados sobre os dados.

5.3. Estudo de Caso: Accenture

Accenture é uma empresa de consultoria com atuação global, presente em 49 países e contando com mais de 730 mil colaboradores [Accenture]. No seu site, a empresa divulga um estudo de caso intitulado "O projeto da Accenture para IA responsável", que descreve os princípios aplicados aos seus sistemas internos de inteligência artificial. Esses princípios são analisados comparativamente aos mencionados neste artigo, porém deixando de fora os relacionados à transparência e explicabilidade, liberdade e autonomia, além da segurança e robustez que não foram abordados pela Accenture.

1. Justiça e equidade

   A Accenture se desempenha em garantir que todos os seus modelos de sistemas de inteligência artificial tratem todos os grupos de forma equitativa, reconhecendo essa ação como necessária para a mitigação de vieses.

2. Transparencia e explicabilidade

   A empresa adota uma política de transparência e explicabilidade divulgando o uso da IA, onde todos possam entender e avaliar os seus resultados e processos de tomada de decisão

3. Responsabilidade

   O princípio de responsabilidade é aplicado através de estruturas de governança dentro de toda empresa, com funções, políticas e responsabilidades claras.

4. Privacidade e proteção dos dados

   A Accenture garante que a IA esteja em conformidade com as leis relevantes, que esteja protegida contra ataques cibernéticos e que os dados estejam protegidos com a proteção de dados.

5. Sustentabilidade

A empresa se compromete a mitigar qualquer impacto negativo que suas soluções possam ter no planeta, priorizando inovações que favoreçam a sustentabilidade.

5.4. Estudo de Caso: Salesforce

A Salesforce é uma empresa global focada em soluções de gerenciamento do relacionamento com clientes (CRM). Com sede em San Francisco, a organização conta com mais de 78 mil funcionários distribuídos por diversos países [Salesforce]. A Salesforce utiliza inteligência artificial para otimizar as áreas de marketing, vendas e atendimento ao cliente. A empresa mantém seu compromisso com o desenvolvimento ético e responsável dessas tecnologias. Em seu site oficial, a Salesforce divulgou seus princípios éticos que incluem privacidade e proteção dos dados, transparência e explicabilidade, liberdade e autonomia e sustentabilidade, mas não foram mencionados outros princípios abordados anteriormente neste artigo.

1. Privacidade e proteção dos dados

   Este princípio na Salesforce é chamado internamente de "segurança", este princípio abrange a proteção de informações pessoais e a redução de vazamentos e vieses. A empresa enfatiza a importância de evitar conteúdos tóxicos e discriminatórios em interações automatizadas.

2. Transparência e explicabilidade

   A Salesforce adota esse princípio, chamando-o de transparência e precisão. No que diz respeito à transparência, a empresa enfatiza a importância de respeitar a origem dos dados e deixar claro quando o conteúdo é gerado por IA. Em termos de precisão, destaca-se a necessidade de informar sempre que uma resposta não for totalmente clara, além da possibilidade de habilitar a verificação dos fatos quando viável.

3. Liberdade e autonomia

   A Salesforce adota uma postura centrada no ser humano, permitindo que os humanos permaneçam no controle, potencializando as capacidades humanas, garantindo práticas trabalhistas responsáveis.

4. Sustentabilidade

   O princípio da sustentabilidade é implementado através da otimização dos modelos de IA para garantir eficiência no uso de recursos, minimizando as emissões de carbono. Além disso, a Salesforce também se dedica à representatividade e alta qualidade dos dados empregados no treinamento desses modelos.

5.5. Estudo de Caso: IBM

A IBM é uma empresa que fornece tecnologia e serviços essenciais para resolver problemas empresariais. A organização emprega mais de 300 pessoas localizadas em diversos países [IBM]. A IBM está empenhada na promoção da IA responsável e publicou em seu site princípios éticos adotados no desenvolvimento dessa tecnologia, além de um documento detalhado com diretrizes ilustrando a abordagem utilizada para

cada item mencionado [IBM]. Entre esses princípios estão transparência e explicabilidade, justiça e equidade, segurança e robustez, privacidade e proteção dos dados; no entanto, ficam ausentes os aspectos de responsabilidade, liberdade e autonomia, supervisão humana e sustentabilidade.

1. Transparência e explicabilidade

   Esse princípio é discutido de maneira independente. A transparência é abordada para assegurar que todas as partes envolvidas estejam informadas sobre quais dados foram coletados, como são utilizados e armazenados, e quem tem acesso a eles. A explicabilidade assegura que os humanos sejam capazes de perceber e compreender o processo decisório da IA.

2. Segurança e robustez

   A IBM tem o compromisso de que toda equipe trabalhe em conjunto a fim de escolher componentes robustos que minimizem os riscos e permitam a confiança dos usuários nos resultados dos sistemas de IA.

3. Justiça e equidade

   No início, a IBM enfatiza a importância de identificar e enfrentar preconceitos para promover uma representação inclusiva. Isso é realizado por meio de pesquisas contínuas e da coleta responsável e representativa de dados que refletem uma população diversificada.

4. Privacidade e proteção dos dados

   Este princípio assegura a conformidade com as regulamentações de proteção de dados nos países onde os produtos da IBM são utilizados. Além disso, sugere medidas para preservar e fortalecer o controle dos usuários sobre seus próprios dados e como eles são usados.

Embora não mencione explicitamente a responsabilidade como um princípio, o documento destaca a importância de assumir responsabilidades pelos resultados do sistema de IA no mundo real. Enfatiza que cada indivíduo envolvido na criação da IA, em qualquer etapa do processo, tem a obrigação de considerar seu impacto, assim como as empresas dedicadas ao seu desenvolvimento.

6. Discussão

Os estudos de casos apresentados demonstram que as empresas estão cada vez mais cientes das responsabilidades éticas associadas ao desenvolvimento e uso da inteligência artificial. O viés discriminatório pode surgir em várias etapas do processo de desenvolvimento dos sistemas de IA, levando as organizações a buscar a implementação de princípios éticos como forma de mitigar esses vieses. Embora cada empresa adote abordagens distintas, os princípios éticos compartilham o objetivo comum de promover um uso justo e seguro da inteligência artificial. A tabela 2 ilustra como a variedade nos princípios das diretrizes éticas indica que não há um padrão universalmente aceito para as estruturas de governança ética na IA; em vez disso, observa-se uma convergência em torno de princípios fundamentais que guiam essas práticas.

Tabela 2. Análise comparativa dos princípios éticos nas empresas

	Microsoft	Google	Accenture	Salesforce	IBM
Responsabilidade	X	X	X		
Transparência e explicabilidade	X			X	X
Privacidade e proteção de dados	X	X	X	X	X
Justiça e equidade	X	X	X		X
Liberdade e autonomia				X	
Supervisão humana					
Segurança e robustez	X	X			X
Sustentabilidade			X	X	

Observa-se que Microsoft, em seu documento, detalha minuciosamente cada etapa da implementação de seus princípios e indica quem é o responsável por garantir sua execução. Em contraste, a IBM e o Google descrevem os princípios com sugestões práticas para implementá-los sem definir um modelo ou atribuir responsabilidades específicas às pessoas envolvidas. Por sua vez, a Salesforce e a Accenture adotaram uma abordagem mais abstrata ao descrever seus princípios, tratando os princípios éticos como metas para alcançar inteligência artificial confiável, mas sem explicitar metodologias precisas para a sua implementação.

É importante destacar que, embora os nomes dos princípios variem entre as empresas, existe uma correspondência significativa entre as ideias apresentadas. Por exemplo, a Microsoft denomina equidade como inclusão e traduz robustez por confiabilidade e prevenção de riscos. Já o Salesforce denomina liberdade e autonomia como empoderamento. O Google opta por uma abordagem distinta, articulando seus princípios através de ações concretas, como "evitar criar ou reforçar preconceitos injustos", em vez de apenas classificá-los sob o rótulo de "justiça e equidade".

Outra observação relevante é que nenhuma organização adotou todos os oito princípios expostos no artigo; contudo, mesmo quando não são explicitamente aceitos como um princípio individual, eles acabam sendo abordados indiretamente dentro de outros ou nas orientações gerais dessas entidades de forma indireta ou transversal. A Microsoft e o Google discutem controle humano junto à autonomia e liberdade em seus capítulos sobre responsabilidade. Por outro lado, Salesforce e IBM não tratam diretamente responsabilidade como um princípio ético específico, mas ela é destacada ao longo do documento como resultado final alcançado pela aplicação dos demais

conceitos. Além disso, ainda que a Salesforce não possua formalmente um princípio dedicado à justiça e equidade, se ocupa do tema vieses sob o princípio de privacidade e proteção aos dados.

Por fim, esta análise reforça a relevância da governança ética para mitigar os vieses discriminatórios, consolidando o compromisso das organizações em desenvolver tecnologias de IA de forma responsável e confiável. Como discutido ao longo deste estudo, os vieses podem surgir em qualquer fase do desenvolvimento da IA, desde a coleta dos dados até a aplicação dos modelos. Por isso, estabelecer medidas de monitoramento contínuo em todas etapas é essencial para identificar esses vieses, responsabilizar os envolvidos e implementar as correções necessárias.

Ao abordar os vieses discriminatórios, as organizações reconhecem que o objetivo é minimizá-los e não eliminá-los completamente. Isso se deve à complexidade dos vieses, que podem variar de acordo com diferentes culturas e sociedades. Além disso, as empresas enfrentam limitações em relação ao uso final das ferramentas que oferecem; por exemplo, uma IA projetada para gerar imagens não tem controle sobre como suas criações serão utilizadas por terceiros. Esses desafios destacam a necessidade de legislações mais abrangentes e críticas, levando em consideração um amplo espectro de cenários para mitigar riscos além do alcance organizacional.

A diversidade entre as diretrizes dos princípios éticos não devem ser vistos como fraqueza. Pelo contrário, essas diretrizes internas refletem os contextos e prioridades específicas de cada organização. No entanto, esses esforços fragmentados evidenciam a urgência de uma estrutura globalmente coordenada. A lei de IA da União Europeia [União Europeia 2023], a ordem executiva da Casa Branca dos Estados Unidos [The White House 2023] e o apelo do secretário-geral da ONU por uma governança global da IA [ONU 2024] são exemplos significativos de iniciativas que buscam padronizar e alinhar as diretrizes. Tais esforços pressionam as organizações a se posicionarem com mais seriedade em prol de uma IA justa e confiável, alinhando as práticas às expectativas sociais e as leis emergentes.

Embora as empresas estejam empenhadas no desenvolvimento responsável, nosso acesso se limita às informações divulgadas em suas plataformas públicas. Essas informações são frequentemente muito vagas, o que impede uma compreensão clara sobre a aplicação e implementação dos princípios estabelecidos. Para alcançar essa clareza, seria necessário disponibilizar ferramentas que permitam monitorar e controlar esses sistemas de inteligência artificial. Atualmente, sem esse nível de controle disponível, as estruturas de governança dentro das empresas atuam como um gerenciamento de risco: não garantem a inexistência total de riscos — como a existência de vieses discriminatórios — mas buscam minimizar sua ocorrência e proporcionar soluções mais ágeis.

7. Conclusão

Os vieses discriminatórios nos algoritmos constituem um problema complexo e desafiador. Identificar se um viés presente em um sistema é discriminatório depende diretamente do contexto no qual ele está inserido. Além disso, o desenvolvimento de

sistemas de inteligência artificial envolve várias etapas, cada uma sujeita à introdução de diferentes tipos de vieses, conforme mostrado na Tabela 1 com a taxonomia dos vieses.

Dada a dificuldade de eliminar completamente esses vieses, esforços estão sendo direcionados tanto através de legislações formais quanto por meio de iniciativas privadas. O objetivo final é alcançar um equilíbrio entre regulamentações governamentais e práticas corporativas responsáveis através das diretrizes éticas. No entanto, enquanto essa convergência ainda está se desenvolvendo, empresas globais, como as analisadas neste estudo, desempenham um papel crucial ao incorporar princípios éticos nos sistemas de inteligência artificial. A incorporação desses princípios nas diretrizes de governança voltadas para IA promove maior aceitação social.

Este artigo sugere que a incorporação de princípios éticos é uma estratégia para orientar o desenvolvimento responsável de sistemas de inteligência artificial. Esses princípios são baseados em teorias éticas já estabelecidas na filosofia. Para investigar como essa abordagem tem sido aplicada na prática, foi conduzido um estudo de caso envolvendo cinco empresas: Microsoft, Google, Accenture, Salesforce e IBM.

A análise comparativa dos estudos de caso revela que, apesar da falta de padronização na aplicação dos princípios éticos, as empresas estão atentas às questões éticas no desenvolvimento de sistemas inteligentes. Cada organização adotou estratégias e enfoques distintos, porém todas compartilham o objetivo comum de mitigar vieses e garantir maior transparência e controle nos processos de desenvolvimento. Este trabalho ressalta a importância da governança ética como elemento central para desenvolver inteligência artificial alinhada com valores humanos e capaz de reduzir vieses discriminatórias. Contudo, é fundamental reconhecer certas limitações do estudo: a dependência em fontes secundárias para coletar dados sobre os exemplos analisados e o rápido avanço do campo da IA podem rapidamente desatualizar algumas informações apresentadas aqui. Assim sendo, ressalta-se a necessidade futura por uma investigação mais aprofundada que analise detalhadamente o impacto da governança da IA e dos princípios éticos na prática real das organizações.

Referencias

ACCENTURE. **Blueprint for Responsible AI**. Disponível em: <https://www.accenture.com/us-en/case-studies/data-ai/blueprint-responsible-ai>. Acesso em: 20 set. 2024.

ACCENTURE. **Sobre a nossa empresa**. Disponível em: <https://www.accenture.com/br-pt/about/company-index#:~:text=Somos%20um%20time%20global&text=Cidades%20com%20instala%C3%A7%C3%B5es%20e%20opera%C3%A7%C3%B5es%20Accenture%20em%2049%20pa%C3%ADses.> . Acesso em: 19 out. 2024.

Ethics - AI for Good Foundation Disponível em: <https://ai4good.org/ethics/>. Acesso em: 19 out. 2024.

AULD, G. et al. Governing AI through ethical standards: learning from the experiences of others private governance initiatives. **Journal of European Public Policy**, v. 29, n. 11, p. 182201844, 22 ago. 2022.

AGARWAL, L. **Defining Organizational AI Governance and Ethics**. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4553185>. Acesso em: 20 set. 2024.

BERTONCINI, A. L. C.; SERAFIM, M. C. Ethical content in artificial intelligence systems: A demand explained in three critical points. **Frontiers in Psychology**, v.14, 30 mar. 2023.

BIRKSTEDT, T. et al. AI governance: themes, knowledge gaps and future agendas. **Internet Research**, v.33, n-7, p. 133-167, 27 jun. 2023.

BOSTROM, N.; YUDKOWSKY, E. **THE ETHICS OF ARTIFICIAL INTELLIGENCE**. [s.l: s.n]. Disponível em: <https://nickbostrom.com/ethics/artificial-intelligence.pdf>. Acesso em: 20 set. 2024.

CAMILLERI, M. A. Artificial intelligence governance: Ethical considerations and implications for social responsibility. **Expert Systems**, 18 jul. 2023.

CARVALHO, A. P. DE. Viés algorítmico e discriminação: possíveis soluções regulatórias para o Brasil. **lume.ufrgs.br**, 2020.

CHAUDHARY, A. K. Algorithmic Bias: An Integrative Review and Scope for Future Research. **Research Square (Research Square)**, 21 ago. 2024.

CHENG, L; KUY, X. From principles to practices: the intertextual interaction between AI ethical and legal discourses. **International journal of legal discourse**, v.8, n.1, p. 31-52, 1 abr. 2023.

CLOUD, G. **O que é Inteligência Artificial (IA)?** Disponível em : <https://cloud.google.com/learn/what-is-artificial-intelligence?hl=pt-br>. Acesso em: 10 ago, 2024.

COZMAN, F. G.; KAUFMAN, D. Viés no aprendizado de máquina em sistemas de inteligência artificial: a diversidade de origens e os caminhos de mitigação. **Revista USP**, n. 135, p. 195-210, 22 dez. 2022.

CRESWELL, John W.; CRESWELL, J. D. **Projeto de pesquisa: métodos qualitativo, quantitativo e misto.** 5th ed. Porto Alegre: Pens 2021. E-book. p.i. ISBN 9786581334192. Disponível em: <https://integrada.minhabiblioteca.com.br/reader/books/9786581334192/>. Acesso em : 15 ago. 2024.

DE ALMEIDA, P. G. R.; DOS SANTOS, C. D.; FARIAS, . S. Artificial Intelligence Regulation: a Framework for Governance. **Ethics and Information Technology**, v. 23, n. 3, p. 505-525, 21 abr. 2021.

FAZELPOUR, S.; DANKS, D. Algorithmic bias: Senses, sources, solutions. **Philosophy Compass**, v. 16, n.8, 12 jun. 2021.

FERRARA, E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. **Sci**, v.6, n.1, p.3, 26 dez. 2023.

FLORIDI, L. Soft Ethics and the Governance of the Digital. **Philosophy & Technology**, v. 31, n. 1, p. 1-8, 17 fev. 2018.

GIFFEN, V. B.; HERHAUSEN, D.; FAHSE, T. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. **Journal of Business Research**, v.144, n.1, p.93-106, maio 2022.

GOOGLE. **Google AI Principles**. Disponível em: <https://ai.google/responsibility/principles/>. Acesso em: 20 set. 2024.

GOOGLE. **AI principles: 2023 progress update**. [S.l.]: Google, 2023. Disponível em: <https://ai.google/static/documents/ai-principles-2023-progress-update.pdf>. Acesso em: 20 set. 2024.

GOOGLE. **Google | LinkedIn**. Disponível em: <https://www.linkedin.com/company/google/people/> . Acesso em: 19 out. 2024.

GROUMPOS, P. P. A Critical Historic Overview of Artificial Intelligence: Issues, Challenges, Opportunities, and Threats. **Artificial Intelligence and Applications**, v.1, n.4, p.197-213, 24 jul. 2023.

HELLSTRÖM, T.; DIGNUM, V.; BENSCH, S. Bias in Machine Learning – What is it Good for? **arXiv (Cornell University)**, 1 jan. 2020.

IBM. **Everyday Ethics for Artificial Intelligence A practical guide for designers & developers Introduction 6**. [s.l: s.n]. Disponível em <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>. Acesso em: 22 set. 2024.

IBM. **O que é Inteligência Artificial (IA)?** Disponível em: <https://www.ibm.com/br-pt/topics/artificial-intelligence>. Acesso em: ago. 2024.

IBM. **IBM | LinkedIn**. Disponível em: <https://www.linkedin.com/company/ibm/people/> . Acesso em: 19 out. 2024.

INTERNATIONAL TELECOMUNICATION UNION. **AI for Good**. Disponível em: <https://aiforgood.itu.int>. Acesso em: 20 set. 2024.

JOBIN, A.; IENCA, M.. VAYENA, E. The global landscape of AI ethics guidelines. **Nature Machine Intelligence**, v. 1, n. 9, p. 389-399, set. 2019.

KUMAR, W. **Ethical AI: Bridging Perspectives for Responsible Technology and Interdisciplinary Discourse.** Disponível em: <https://easychair.org/publications/preprint/j1gkT>. Acesso em: set. 2024.

KUTZ, J. et al. AI-based Services - Design Principles to Meet the Requirements of a Trustworthy AI. **AHFE international**, 1 jan. 2023.

KORDZADEH, N.; GHASEMACHAEI, M. Algorithmic bias: review, synthesis, and Future Research Directions. **European Journal of Information Systems**, v. 31,n. 3, p. 1-22, 6 jun. 2021.

LUNDGREN, B. In defense of ethical guidelines. AI and Ethics, 17 jan. 2023.

MICROSOFT. **Microsoft | LinkedIn**. Disponível em: <https://www.linkedin.com/company/microsoft/people/> . Acesso em: 19 out. 2024.

Microsoft Mais Brasil - Relatório de Impacto no Brasil. Disponível em: <https://www.microsoft.com/pt-br/maisbrasil> . Acesso em: 19 out. 2024.

MICROSOFT. **Princípios e Abordagem de IA Responsável**. Disponível em: <https://www.microsoft.com/pt-br/ai/principles-and-approach> . Acesso em: 02 out. 2024.

MOERTL, P.; EBINGER, N. The Development of Ethical and Trustworthy AI Systems Requires Appropriate Human-Systems Integration. **Studies in computational intelligence**, p.11-27, 1 jan. 2024.

NASSAR, A.; KAMAL, M. Ethical Dilemmas in AI-Powered Decision-Making: A Deep Dive into Big Data-Driven Ethical Considerations. **International Journal of Responsible Artificial Intelligence** , v. 11, n.8, p. 1-11, 6 ago. 2021.

NAZER, L. et al. Bias in artificial intelligence algorithms and recommendations for mitigation. **PLOS Digital Health**, v.2, n.6, p. e0000278-e0000278, 22 jun. 2019

OBERMEYER, Z. et al. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. **Science**, v. 366, n. 6464, p. 447-453, 25 out. 2019.

ONU. **High-Level Advisory Body on Artificial Intelligence / Office of the Secretary-General's Envoy on Technology**. Disponível em: <https:www.un.org.techenvoy/ai-advisory-body>. Acesso em: 10 out. 2024.

PEREZ, J. A. et al. Artificial Intelligence and Robotics. **arXiv:1803.10813 [cs]**, 28 mar. 2018.

PÖHLER, LUKAS D; DIEPOLD, K.; WALLACH, W. A Practical Multilevel Governance Framework for Autonomous and Intelligent Systems. **arXiv (Cornell University)**, 21 abr. 2024.

REQUIÃO, M.; COSTA, D. Discriminação algorítimica: ações afirmativas como estratégia de combate. **civilistica.com**, v. 11, n. 3, p. 1-24, 25 dez. 2022.

RIESKAMP, J. et al. Approaches to Improve Fairness when Deploying AI-based Algorithms in Hiring - Using a Systematic Literature Review to Guide Future Research. **Proceedings of the … Annual Hawaii International Conference on System Sciences/Proceedings of the Annual Hawaii International Conference on System Sciences**, 1 jan. 2023.

ROBLES, P.; MALLINSON, D. J. Catching up with AI: Pushing toward a cohesive governance framework. v. 51, n.3, p. 355-372, 15 maio 2023.

Salesforce. **Salesforce | LinkedIn**. Disponível em: <https://www.linkedin.com/company/salesforce/people/> . Acesso em: 19 out. 2024.

SALESFORCE. **Responsible AI and technology**. Disponível em: <https://www.salesforce.com/company/responsible-ai-and-technology/>. Acesso em: 22 set. 2024.

SIGFRIDS, A. et al. How Should Public Administrations Foster the Ethical Development and Use of Artificial intelligence? A Review of Proposals for Developing Governance of AI. **Frontiers in Human Dynamics**, v. 4, 19 maio 2022.

SIMÕES-GOMES, L; ROBERTO E.; MENDONÇA, J. Viés algorítmico - um balanço provisório. **Estudos de Sociologia**, v. 25, n. 48, 24 jul 2020.

SINGH, J. P. AI Ethics and Societal Perspectives: A Comparative Study of Ethical Principle Prioritization Among Diverse Demographic Clusters. **Journal of Advanced Analytics in Healthcare Management**, v. 5, n. 1, p. 1-8, 13 jan. 2021.

STAHL, B., B. Addressing Ethical Issues in AI. **SpringerBriefs in Research and Innovation Governance**, p. 55-79, 2021.

STAHL, B., B. Concepts of Ethics and Their Application to AI. **Artificial Intelligence for a Better Future**, p. 19-33, 18 mar. 2021.

TAEIHAGH, A. Governance of artificial intelligence. Policy and Society, v. 40, n. 2, p. 137-157, 3 abr. 2021.

THE WHITE HOUSE. **FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence**. Disponível em: <www.whitehouse.gov/briefing-room/statements-release/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-inteligence/>. Acesso em: 05 out. 2024.

UNIÃO EUROPEIA. **Hiroshima Process International Guiding Principles for Advanced AI system | Shaping Europe's digital future**. Disponível em: <https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system>. Acesso em: 05 out. 2024.

Catalogação na publicação
Seção de Catalogação e Classificação

P475p Pessoa, Vanessa Gabriele Lima.
Princípios éticos na IA: o impacto da governança ética na redução de vieses / Vanessa Gabriele Lima Pessoa. - João Pessoa, 2024.
23 f. : il.

Orientação: Valdecir Becker.
TCC (Graduação) - UFPB/Informática.

1. Princípios éticos na IA. 2. Governança ética. 3. Vieses. I. Becker, Valdecir. II. Título.

UFPB/CI CDU 004.8

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738