IX ENCONTRO VIRTUAL DO CONPEDI

DIREITO, GOVERNANÇA E NOVAS TECNOLOGIAS V

D598

Direito, governança e novas tecnologias V [Recurso eletrônico on-line] organização CONPEDI

Coordenadores: Vivianne Rigoldi; Jéssica Fachin; Rosane Teresinha Porto – Florianópolis: CONPEDI, 2026.

Inclui bibliografia
ISBN: 978-65-5274-589-7
Modo de acesso: www.conpedi.org.br em publicações
Tema: Constituição, Processo e Justiça Social: o papel da jurisdição constitucional na efetivação de direitos

2. Direito. 3. Governança e novas tecnologias. IX Encontro Virtual do CONPEDI (2: 2026: Florianópolis, Brasil).

CDU: 34

Conselho Nacional de Pesquisa
e Pós-Graduação em Direito Florianópolis
Santa Catarina – Brasil
www.conpedi.org.br

IX ENCONTRO VIRTUAL DO CONPEDI

DIREITO, GOVERNANÇA E NOVAS TECNOLOGIAS V

Apresentação

O Conselho Nacional de Pesquisa e Pós-graduação em Direito (CONPEDI), referência na difusão de pesquisas que abordam os novos fenômenos envolvendo o Direito, firma-se como espaço qualificado para apresentação de pesquisas produzidas na área das novas tecnologias em face do Direito.

O tema, recorrente nos artigos apresentados - o uso da inteligência artificial no âmbito jurídico, demonstra que a maior sociedade científica na área jurídica do Brasil se mantém atenta às inovações tecnológicas, ao avanço de diversas tecnologias, novas experiências no judiciário e nos escritórios de advocacia. Artigos, de alta relevância acadêmica e científica, destacam que o Direito não está imune a essas transformações e reiteram a necessidade de atenção aos desafios regulatórios do uso da inteligência artificial no mundo jurídico.

Diferentes pesquisas, apresentadas no Grupo de Trabalho 'Direito, Governança e Novas Tecnologias V', alertam para os efeitos deletérios da opacidade algorítmica,

referindo-se à falta de transparência sobre como sistemas de inteligência artificial e seus algoritmos produzem decisões, dentro e fora do judiciário. Nos textos, a governança

digital é debatida, com rigorosa visão científica, objetivando garantir o uso estratégico de tecnologias de informação e comunicação para gerir serviços, promover a segurança

de dados e integrar processos, de forma transparente, eficiente, com foco na participação ativa do cidadão na tomada de decisões.

algorítmico e tecnológico do sistema de justiça, englobando a investigação de cibercrimes, a digitalização de procedimentos policiais e judiciais, provas digitais e a

regulação de plataformas para combater conteúdos ilícitos.

Destarte, os artigos apresentados e os debates produzidos no Grupo de Trabalho ‘Direito, Governança e Novas Tecnologias V’ são essenciais para o

fortalecimento das diretrizes nacionais de regulação das novas tecnologias, pela formação de parâmetros normativos e procedimentais adequados à proteção dos

direitos humanos e fundamentais, à garantia da segurança da informação, à gestão de riscos e ao fortalecimento da transparência e do alinhamento legal no uso das novas

tecnologias no âmbito jurídico.

Por essas razões, as coordenadoras do GT convidam para a leitura do teor integral dos artigos, agradecendo a participação dos autores pesquisadores desta edição.

Vivianne Rigoldi/Jéssica Fachin/Rosane Porto

A ALEGORIA DA CAVERNA DE PLATÃO E OS DESAFIOS DA TRANSPARÊNCIA EM INTELIGÊNCIA ARTIFICIAL

PLATO'S ALLEGORY OF THE CAVE AND THE CHALLENGES OF TRANSPARENCY IN ARTIFICIAL INTELLIGENCE

Djônata Winter

Resumo

Partindo da Alegoria da Caverna de Platão como metáfora para a condição de desconhecimento em que se encontram os indivíduos diante de decisões algorítmicas cujos fundamentos e mecanismos internos lhes são inacessíveis, este artigo investiga a opacidade dos sistemas de Inteligência Artificial contemporâneos e a necessidade de transparência. A pesquisa utiliza metodologia qualitativa de caráter exploratório e descritivo, combinando revisão bibliográfica e análise documental de normas, recomendações e projetos de lei relacionados à transparência em IA. O estudo demonstra que a opacidade inerente aos sistemas de aprendizado de máquina não constitui mera limitação técnica, mas um problema de dimensões éticas, políticas e jurídicas, capaz de perpetuar vieses e discriminações algorítmicas e comprometer princípios democráticos fundamentais. Examina as dimensões da transparência em IA, os vieses algorítmicos e suas injustiças sistêmicas, as implicações democráticas do uso de IA na Administração Pública e as iniciativas regulatórias internacionais e nacionais, incluindo os princípios da OCDE, a Recomendação da UNESCO, o AI Act europeu e o Projeto de Lei n. 2.338/2023 no Brasil. Conclui que a superação da opacidade algorítmica exige não apenas marcos regulatórios adequados, mas também uma educação cívica abrangente e a participação ativa dos cidadãos na governança da IA, configurando a transparência como condição necessária para uma governança algorítmica legítima e democrática.

Palavras-chave: Inteligência artificial, Alegoria da caverna, Transparência, Opacidade algorítmica, Governança de ia

and discrimination and undermining fundamental democratic principles. It examines the dimensions of transparency in AI, algorithmic biases and their systemic injustices, the democratic implications of AI use in Public Administration, and international and national regulatory initiatives, including the OECD principles, the UNESCO Recommendation, the European AI Act, and Bill No. 2,338/2023 in Brazil. It concludes that overcoming algorithmic opacity requires not only adequate regulatory frameworks, but also comprehensive civic education and active citizen participation in AI governance, establishing transparency as a necessary condition for legitimate and democratic algorithmic governance.

Keywords/Palabras-claves/Mots-clés: Artificial intelligence, Allegory of the cave, Transparency, Algorithmic opacity, Ai governance

1. INTRODUÇÃO

A Inteligência Artificial (IA) tem se integrado progressivamente aos mais diversos setores da sociedade, incluindo a Administração Pública, onde sistemas algorítmicos são cada vez mais utilizados para otimizar processos decisórios, alocar recursos e avaliar políticas públicas.

Segundo o Instituto de Ética Empresarial de Londres (2018, p. 1), a IA pode ser caracterizada como a simulação de elementos do processo de inteligência humana por máquinas e sistemas computacionais, sendo fundamentalmente definida por três características principais:

> 1) Aprendizado – a capacidade de adquirir informações relevantes e as regras para usá-las;
>
> 2) Raciocínio – a capacidade de aplicar as regras adquiridas e usá-las para chegar a conclusões aproximadas ou definitivas;
>
> 3) Iterativo – a capacidade de modificar o processo com base em novas informações adquiridas .

Entretanto, à medida que estas tecnologias se tornam mais sofisticadas e pervasivas, emerge a preocupação com sua natureza frequentemente opaca:

> No entanto, diferentemente dos sistemas de informação convencionais, os algoritmos incorporados em aplicações de IA podem ser "caixas-pretas". Anteriormente, aqueles que desenvolviam aplicações podiam explicar completamente como um algoritmo funcionava. Dado um input, eles podiam dizer qual seria o output e por quê, porque os sistemas aplicavam regras criadas por humanos. Isso não é mais verdade para aplicações baseadas em IA. A aplicação cria estruturas internas que determinam os outputs, mas estas são incompreensíveis para observadores externos, e mesmo os programadores não podem lhe dizer por que um output específico foi gerado. (ASATIANI et al, 2020, p. 259, nossa tradução)

Neste cenário, a Alegoria da Caverna de Platão oferece uma metáfora para compreender os limites e potenciais deturpações que os sistemas de IA podem perpetuar. Como afirma Karpouzis (2024, p. 1):

> Esta alegoria serve como uma poderosa metáfora para compreender as limitações e potenciais representações equivocadas que os sistemas de IA podem perpetuar, destacando a importância de examinar criticamente os dados que alimentam essas entidades digitais, bem como os algoritmos que impulsionam seu treinamento e implementação.

Sob esta perspectiva, este artigo se propõe a analisar a analogia entre a Alegoria da Caverna de Platão e os sistemas de Inteligência Artificial contemporâneos, focalizando a problemática da opacidade algorítmica e suas implicações para os ideais democráticos e para o exercício da cidadania no contexto digital. Busca-se investigar como a falta de transparência em sistemas de IA pode limitar a compreensão e a contestação de decisões algorítmicas que afetam direitos fundamentais, especialmente no âmbito da Administração Pública.

Os problemas centrais que orientam esta investigação são: Em que medida a opacidade dos sistemas de IA contemporâneos se assemelha à condição dos prisioneiros na alegoria da caverna platônica? Quais as dimensões e requisitos da transparência em sistemas algorítmicos? Como a transparência se relaciona com a legitimidade democrática e o controle social, particularmente no contexto da Administração Pública? Quais os desafios técnicos, jurídicos e sociais para implementação da transparência em sistemas de IA e que iniciativas regulatórias têm emergido neste campo?

O desenvolvimento deste artigo está estruturado em seis seções principais. A primeira seção (A Caverna Digital) trabalha a analogia entre a Alegoria da Caverna e os sistemas de IA contemporâneos. A segunda seção examina as dimensões da transparência em sistemas algorítmicos. A terceira seção discute os vieses algorítmicos e as injustiças sistêmicas que podem ser perpetuadas por sistemas opacos. A quarta seção analisa as implicações do uso de IA na Administração Pública. A quinta seção aborda as iniciativas regulatórias para promoção da transparência, com destaque para o Projeto de Lei n. 2.338/2023. A sexta seção enfatiza o papel da educação e da participação cívica na governança de sistemas de IA. Por fim, são apresentas as conclusões do estudo.

A pesquisa utiliza metodologia qualitativa de caráter exploratório e descritivo, combinando revisão bibliográfica e análise documental de normas, recomendações e projetos de lei relacionados à transparência em IA. O método interpretativo é empregado para relacionar a alegoria platônica ao contexto contemporâneo da inteligência artificial, estabelecendo paralelos conceituais que permitem uma reflexão crítica sobre as implicações éticas, sociais e políticas da opacidade algorítmica.

2. DESENVOLVIMENTO

2.1. A Caverna Digital

No Livro VII da República, Platão (2001, p. 315) apresenta sua famosa alegoria: prisioneiros acorrentados em uma caverna desde o nascimento, forçados a olhar apenas para a parede onde são projetadas sombras de objetos transportados por trás deles. Para esses prisioneiros, as sombras constituem a realidade. Quando um prisioneiro é libertado e vê o mundo exterior, inicialmente sofre com o deslumbramento, resistindo à nova realidade que contradiz suas antigas certezas.

Na obra, Sócrates descreve a situação de forma vívida:

> Suponhamos uns homens numa habitação subterrânea em forma de caverna, com uma entrada aberta para a luz, que se estende a todo o comprimento dessa gruta. Estão lá dentro desde a infância, algemados de pernas e pescoços, de tal maneira que só lhes é dado permanecer no mesmo lugar e olhar em frente; são incapazes de voltar a cabeça, por causa dos grilhões. (Platão, 2001, p. 315)

Atrás dos prisioneiros, há uma fogueira acesa em uma colina e, entre o fogo e os prisioneiros, uma estrada ascendente com um pequeno muro. Ao longo desse muro, homens transportam objetos de diversos tipos, alguns em silêncio, outros falando. Os prisioneiros, por sua vez, só podem ver as sombras desses objetos projetadas na parede à sua frente, tomando-as como a única realidade existente.

Esta alegoria milenar pode nos servir para analisar os sistemas contemporâneos de IA e sua relação com a sociedade. Conforme destacado por Karpouzis (2024, p. 1, tradução nossa):

> Esta alegoria serve como uma poderosa metáfora para compreender as limitações e potenciais representações errôneas que os sistemas de IA podem perpetuar, destacando a importância de examinar criticamente os dados que alimentam essas entidades digitais, bem como os algoritmos que impulsionam seu treinamento e implantação.

Como observa Murphy (2023, tradução nossa):

> Em muitos aspectos, a IA pode ser vista como uma extensão desta alegoria. Os algoritmos de IA são construídos a partir de grandes quantidades de dados, e seu 'conhecimento' do mundo é limitado àquilo em que foram treinados. Em certo sentido, eles estão acorrentados aos dados que lhes foram fornecidos, e só podem ver os padrões que existem dentro desses dados.

Essa limitação fundamental é ainda mais evidente quando consideramos os resultados produzidos por estes sistemas. Nas palavras do mesmo autor (Murphy, 2023, tradução nossa), "a saída dos algoritmos de IA pode ser pensada como sombras projetadas numa parede. Essas saídas são representações dos dados em que o algoritmo foi treinado, e nem sempre refletem a verdadeira realidade".

A analogia se estende também à epistemologia platônica, que distingue entre conhecimento verdadeiro ("episteme") e mera opinião ("doxa"). Karpouzis (2024, p. 8, tradução nossa) argumenta que:

> Embora modelos de IA como o ChatGPT possam produzir resultados que parecem convincentes e informativos, é importante reconhecer que esse conteúdo é, em última análise, derivado de padrões nos dados de treinamento e pode não constituir compreensão genuína ou sabedoria no sentido platônico.

Este questionamento é particularmente relevante quando consideramos a natureza da IA contemporânea. Como destaca Jungherr (2023, p. 6, tradução nossa):

> a IA não tem compromisso com a verdade de um argumento ou observação; ela está apenas imitando sua semelhança conforme encontrada em dados passados. A IA de hoje está comprometida apenas com a representação do mundo, um objeto ou um argumento disponível para ela, não com o mundo, objeto ou argumento como tal.

Ademais, assim como os prisioneiros da caverna eram incapazes de perceber o mecanismo que gerava as sombras, os usuários e até mesmo os desenvolvedores de sistemas de IA frequentemente não compreendem completamente os processos pelos quais estes sistemas chegam a determinados resultados.

Esta opacidade não é acidental, mas inerente à natureza dos complexos sistemas de aprendizado de máquina, que desenvolvem representações internas e padrões de decisão que podem escapar à compreensão humana. Como destacado pelo Instituto de Ética Empresarial de Londres (2018, p. 3, tradução nossa):

> À medida que os algoritmos de IA aumentam em complexidade, torna-se mais difícil compreender como eles funcionam. Em alguns casos, as aplicações de IA têm sido referidas como uma "caixa preta" onde nem mesmo os engenheiros conseguem decifrar porque a máquina tomou determinada decisão. Isso pode prejudicar significativamente sua eficácia e causar preocupação. O uso de algoritmos do tipo "caixa preta" dificulta não apenas identificar quando algo dá errado, mas também determinar quem é responsável no caso de qualquer dano ou lapso ético.

Não sabemos realmente o que estes sistemas "aprenderam" ou quais critérios estão utilizando para suas decisões, criando um abismo na cadeia de responsabilidade e compreensão. Conforme apontado no levantamento de Jungherr e Schroeder (2023, p. 170, tradução nossa):

> [...] o desafio da avaliabilidade vai ainda mais fundo. Em seu estado atual, muita IA—especialmente a que depende de aprendizado profundo—vem com uma opacidade inerente, e frequentemente não está claro como ela alcança seu sucesso. Sabemos que a IA aprende a conexão entre sinais disponíveis e saídas predefinidas nos dados disponíveis, mas quais sinais ela usa para fazer isso? Erros potenciais abundam. Por um lado, a IA pode captar sinais espúrios que estavam correlacionados com resultados de interesse nos dados de treinamento, mas não conectados causalmente, levando a IA a falhar uma vez implementada ou a cair vítima de ataques direcionados (Szegedy *et al.*, 2014). Olhando de fora, permanece pouco claro o que a IA aprendeu para prever resultados, então também permanece pouco claro se um resultado buscado pela IA realmente se alinha com os objetivos dos atores que a implementam (CHRISTIAN, 2020).

Esta condição de opacidade é agravada pelo crescente papel que algoritmos desempenham na mediação de experiências sociais e políticas fundamentais. Villani (2018, p. 114, tradução nossa) observa:

> Uma grande proporção das considerações éticas levantadas pela IA tem a ver com a natureza obscura desta tecnologia. Apesar de seu alto desempenho em muitos domínios, da tradução às finanças, bem como à indústria automotiva, muitas vezes se prova extremamente difícil explicar as decisões que ela toma de uma maneira que a pessoa média possa entender. Este é o notório 'problema da caixa-preta': é possível observar dados de entrada (*input*) e dados de saída (*output*) em sistemas algorítmicos, mas suas operações internas não são muito bem compreendidas.

A opacidade algorítmica pode criar um cenário onde, como na caverna platônica, as pessoas tomam por realidade o que são apenas representações limitadas geradas por mecanismos obscuros. O paradoxo está justamente no fato de que, diferentemente de sistemas

computacionais anteriores, estamos diante de ferramentas cujas decisões internas podem ser inacessíveis mesmo para aqueles que as projetaram, estabelecendo uma relação de confiança necessária, porém potencialmente problemática.

2.2. Dimensões da Transparência em Sistemas de Inteligência Artificial

A transparência em sistemas de IA é um conceito multifacetado que envolve diferentes níveis e aspectos. Burrell (2016, p. 1-2, tradução nossa) identifica três formas distintas de opacidade em algoritmos:

> (1) opacidade como autoproteção e ocultação institucional ou corporativa intencional e, junto com ela, a possibilidade de enganação consciente; (2) opacidade decorrente do estado atual das coisas onde escrever (e ler) código é uma habilidade especializada; e (3) uma opacidade que decorre da incompatibilidade entre a otimização matemática em alta dimensionalidade característica do aprendizado de máquina e as demandas de raciocínio em escala humana e estilos de interpretação semântica.

Esta classificação ilustra como a transparência vai além da mera abertura do código-fonte. Como observa Rahwan (2017, p. 7, tradução nossa):

> No contexto dos algoritmos, isso não significa ter acesso ao código-fonte do computador, por mais intuitiva que essa noção possa parecer. Ler o código-fonte de um algoritmo moderno de aprendizado de máquina nos diz pouco sobre seu comportamento, porque é frequentemente através da interação entre algoritmos e dados que coisas como discriminação emergem.

Millar, Etzioni e Hibbard (2018, p. 49-50, tradução nossa) destacam que a transparência exige clareza sobre "quais dados são coletados e para qual finalidade", “qual é o objetivo dos algoritmos que apoiam e tomam decisões” e também sobre "a possibilidade de rastrear a pureza e integridade dos dados que sustentam a tomada de decisões baseada em inteligência artificial".

Sob outro aspecto, os conceitos de "black box" e "white box" (ou sistemas opacos e transparentes) são centrais nesta discussão. Hassija *et al.* (2024, p. 47, tradução nossa) explicam que:

> Um modelo black-box em XAI refere-se a um modelo de aprendizado de máquina que opera como um sistema opaco onde o funcionamento interno do modelo não é facilmente acessível ou interpretável. Esses modelos fazem previsões com base em dados de entrada, mas o processo de tomada de decisão e o raciocínio por trás das previsões não são transparentes para o usuário.

Em contraste, modelos "white box" ou transparentes permitem que o funcionamento interno e o raciocínio por trás das previsões sejam acessíveis e interpretáveis.

A iniciativa "Explainable AI" (XAI), lançada pela Agência de Projetos de Pesquisa Avançada de Defesa Norte Americana (DARPA) busca desenvolver sistemas de IA que:

> 1) Produzam modelos mais explicáveis, mantendo um alto nível de desempenho de aprendizado (precisão de previsão); 2) Permitam que usuários humanos compreendam, confiem apropriadamente e gerenciem efetivamente a geração emergente de parceiros artificialmente inteligentes (Institute of Business Ethics, 2018, p. 3, tradução nossa).

De acordo com a DARPA, os "novos sistemas de aprendizado de máquina terão a capacidade de explicar seu raciocínio, caracterizar seus pontos fortes e fracos, e transmitir um entendimento de como se comportarão no futuro" (Institute of Business Ethics, 2018, p. 3, tradução nossa).

Por sua vez, o debate sobre transparência tem influenciado diversas linhas de pesquisa em Inteligência Artificial, fundamentais para sua evolução e melhor compreensão:

> a) Interpretabilidade do modelo: Há um foco crescente no desenvolvimento de modelos de IA que sejam interpretáveis, significando que seu processo de tomada de decisão pode ser compreendido e explicado aos usuários.
>
> b) Requisitos regulatórios: Em algumas indústrias e regiões, regulamentações estão sendo introduzidas para exigir que sistemas de IA sejam explicáveis, a fim de garantir responsabilidade e transparência.
>
> c) Confiança e adoção: A explicabilidade pode desempenhar um papel na construção de confiança em sistemas de IA, o que é crucial para sua ampla adoção e uso.
>
> d) Validação de modelo: A explicabilidade pode ajudar a validar as decisões tomadas por modelos de IA, garantindo que estejam livres de vieses e erros.
>
> e) Depuração e melhoria: A explicabilidade pode fornecer insights sobre como os modelos de IA estão tomando decisões, tornando mais fácil identificar e abordar fontes de erro ou viés. (Hassija *et al.*, 2024, p. 49, tradução nossa)

Contudo, é importante reconhecer que a transparência pode ter diferentes níveis de profundidade dependendo do público-alvo. Chaudhary (2024, p. 110-111, tradução nossa) propõe uma transparência qualificada que oferece graus variados de transparência baseados nas necessidades e níveis de compreensão das partes interessadas, identificando três categorias destas:

> Sujeitos de dados: Indivíduos diretamente afetados por decisões algorítmicas devem ter acesso a explicações claras do resultado, dos fatores que o influenciaram e do potencial de viés. Isso poderia envolver resumos dos dados utilizados, do processo de tomada de decisão e dos riscos e limitações associados.
>
> Reguladores e Auditores: Órgãos reguladores encarregados de supervisionar a equidade algorítmica e a conformidade requerem acesso mais profundo a detalhes técnicos, incluindo arquitetura de algoritmos, qualidade dos dados de treinamento e metodologias de teste. Isso lhes permite avaliar efetivamente os riscos potenciais e garantir a adesão às regulamentações.
>
> Partes Interessadas Internas: Desenvolvedores, engenheiros e gerentes responsáveis pelo design e manutenção do algoritmo precisam de acesso abrangente ao funcionamento interno do sistema. Isso permite que identifiquem e abordem problemas, melhorem o desempenho e garantam práticas de desenvolvimento responsáveis.

Esta abordagem reconhece que nem todos os aspectos técnicos de um sistema de IA precisam ser compreensíveis para todos os públicos, mas que diferentes níveis de transparência são necessários para diferentes contextos e finalidades.

2.3. Vieses Algorítmicos e Injustiças Sistêmicas

A opacidade dos sistemas de IA pode potencializar problemas específicos relacionados a vieses e discriminações algorítmicas. Como observa a UNESCO (2022, p. 23), os sistemas de IA levantam questões éticas que incluem:

> [...] seu impacto na tomada de decisões, emprego e trabalho, interação social, assistência médica, educação, meios de comunicação, acesso à informação, exclusão digital, dados pessoais e proteção ao consumidor, meio ambiente, democracia, Estado de direito, segurança e policiamento, dupla utilização, e direitos humanos e liberdades fundamentais, incluindo liberdade de expressão, privacidade e não discriminação. Além disso, novos desafios éticos são criados pela possibilidade de que os algoritmos de IA reproduzam e reforcem vieses existentes e, assim, agravem formas já existentes de discriminação, preconceitos e estereótipos.

O problema do viés algorítmico está intrinsecamente ligado aos dados utilizados para treinar sistemas de IA. Como explicam Asatiani *et al.* (2020, p. 264, tradução nossa):

> a forma como um sistema de IA atua é determinada pelas características dos dados usados para treiná-lo. Um conjunto de dados de treinamento enviesado leva a decisões enviesadas, mesmo que não haja nada de errado com a funcionalidade do algoritmo ensinado pelos dados.

Estes vieses podem surgir em diversos pontos do desenvolvimento e implementação de sistemas de IA. Floridi *et al.* (2020, p. 1786, tradução nossa) explicam:

> Os desenvolvedores de IA geralmente dependem de dados que podem conter vieses socialmente significativos. Esse viés pode ser transferido para a tomada de decisão algorítmica que sustenta muitos sistemas de IA, de maneiras que são injustas para os sujeitos do processo decisório (Caliskan, Bryson e Narayanan 2017) e, portanto, podem violar o princípio da justiça. Essas decisões podem ser baseadas em fatores de importância ética (por exemplo, motivos étnicos, de gênero ou religiosos) e irrelevantes para a tomada de decisão em questão, ou podem ser relevantes, mas legalmente protegidos como características não discriminatórias (Friedman e Nissenbaum 1996). Além disso, as decisões impulsionadas por IA podem ser amalgamadas a partir de fatores que não têm importância ética óbvia e, no entanto, constituem coletivamente uma tomada de decisão injustamente tendenciosa.

O problema torna-se ainda mais complexo devido à natureza auto-reforçadora de muitos sistemas algorítmicos. Os mesmos autores (FLORIDI *et al.*, 2020, p. 1787, tradução nossa) descrevem este ciclo vicioso que "começaria com um conjunto de dados enviesado informando uma primeira fase de tomada de decisão por IA, resultando em ações discriminatórias, levando à coleta e uso de dados enviesados por sua vez".

Um caso emblemático citado por Villani (2018, p. 123, tradução nossa) ilustra bem este problema:

> Em maio de 2016, jornalistas da Propublica (um jornal investigativo americano) revelaram que o algoritmo COMPAS usado na estimativa do risco de reincidência pelo sistema legal americano e desenvolvido pela empresa Northpointe, era racista e ineficiente. Uma análise das pontuações atribuídas aos prisioneiros revelou que este algoritmo sistematicamente superestimava o risco de reincidência de prisioneiros americanos negros em duas vezes o dos americanos brancos.

Este exemplo demonstra como vieses presentes nos dados de treinamento podem perpetuar e até amplificar injustiças sociais existentes, especialmente quando o funcionamento do algoritmo é opaco e suas decisões não podem ser facilmente auditadas ou contestadas.

Jungherr (2023, p. 7, tradução nossa) destaca outro aspecto importante dessa questão: "A visibilidade das pessoas para a IA depende de sua representação passada nos dados. A IA tem dificuldade em reconhecer aqueles que pertencem a grupos sub-representados nos dados usados para treiná-la". Esta observação indica que certos grupos sociais podem se tornar "invisíveis" para sistemas de IA, resultando em exclusão sistemática.

A opacidade dos sistemas de IA torna particularmente difícil identificar e corrigir tais vieses. Como argumenta Chaudhary (2024, p. 117, tradução nossa), "as chances de obter um resultado justo sem transparência algorítmica são pequenas em campos como emprego e justiça criminal. Decisões raciais e injustas podem ser devastadoras para indivíduos e toda a sociedade".

2.4. Inteligência Artificial na Administração Pública

O uso crescente de sistemas de IA na Administração Pública levanta questões significativas sobre transparência e legitimidade democrática. Jungherr (2023, p. 2, tradução nossa) destaca que "a presença crescente de sistemas baseados em IA na sociedade exige uma interrogação, em linhas semelhantes, de como a IA afeta a ideia e a prática da democracia".

Uma das preocupações é que a opacidade algorítmica pode minar princípios democráticos fundamentais, como a transparência dos atos públicos e a prestação de contas aos cidadãos. Jungherr e Schroeder (2023, p. 171, tradução nossa) assinalam que:

> A combinação atual de incógnitas e incognoscíveis introduzidos na arena pública através da IA reduz sua avaliabilidade. Isso é especialmente preocupante no que diz respeito a determinar se a IA está em conformidade ou em conflito com o funcionamento das infraestruturas de mídia como subsistemas sociais distintos, e sua autorregulação junto a normas específicas... Pessoas, elites e até mesmo atores que administram ou contribuem para as estruturas da arena pública tornam-se incertos sobre seu funcionamento, e podem perder sua fé em sua justiça.

A preocupação com a transparência também foi objeto da Recomendação sobre Ética na Inteligência Artificial da UNESCO, a qual dimensiona o princípio da transparência da seguinte forma:

> A explicabilidade significa tornar inteligível e fornecer informações sobre o resultado dos sistemas de IA. A explicabilidade desses sistemas também se refere à compreensibilidade sobre a entrada, a saída e o funcionamento de cada bloco de construção dos algoritmos e como ele contribui para o resultado dos sistemas. Assim, a explicabilidade está estreitamente relacionada com a transparência, pois os resultados e os subprocessos que conduzem a resultados devem almejar serem compreensíveis e rastreáveis, conforme o contexto. Os atores de IA devem se comprometer a garantir que os algoritmos desenvolvidos sejam explicáveis. No caso de aplicativos de IA que afetam o usuário final de uma forma que não seja temporária, facilmente reversível ou de baixo risco, deve-se garantir que seja fornecida uma explicação significativa para qualquer decisão que resultou na ação tomada, para que o resultado seja considerado transparente (UNESCO, 2022, p. 23).

Porém, o mesmo documento reconhece a necessidade de sopesar o requisito da transparência com outros princípios relevantes no contexto da IA:

> Embora sejam necessários esforços para aumentar a transparência e a explicabilidade dos sistemas de IA, incluindo aqueles com impacto extraterritorial, ao longo de seu ciclo de vida para apoiar a governança democrática, o nível desses dois aspectos sempre deve ser adequado ao contexto e ao impacto, pois pode haver uma necessidade de equilíbrio entre transparência, explicabilidade e outros princípios, como privacidade, segurança e proteção (UNESCO, 2022, p. 22).

A transparência é também um componente fundamental para a confiança pública em sistemas de IA utilizados pelo governo. Floridi *et al.* (2018, p. 701, tradução nossa) afirmam que "é especialmente importante que a IA seja explicável, pois a explicabilidade é uma ferramenta crítica para construir a confiança e a compreensão públicas na tecnologia".

Há ainda o risco de que a crescente dependência de sistemas de IA possa transformar fundamentalmente a relação entre cidadãos e governantes. Conforme adverte Jungherr (2023, p. 6, tradução nossa):

> Este aparente aumento no poder dos especialistas para orientar as sociedades na resposta aos desafios pode reduzir o espaço de opções disponível para a tomada de decisão democrática, deslocando a questão de se as pessoas podem para se elas devem decidir por si mesmas. Nisto, a IA poderia induzir uma transição do autogoverno para o governo dos especialistas e, assim, enfraquecer a democracia.

Esta transferência de poder decisório para sistemas técnicos opacos representa um desafio significativo para os princípios democráticos de autogoverno e soberania popular, em especial o controle dos atos governamentais. Por outro lado, requisitos de transparência podem ajudar a superar esse desafio.

Como destacado por Jungherr (2023, p. 4, tradução nossa):

> Não parece que a modelagem de ambientes de informação digital conduzida por IA leve inevitavelmente à deterioração do acesso às informações necessárias para que as pessoas exerçam seu direito ao autogoverno. No entanto, há muita opacidade na forma como os ambientes de comunicação digital são moldados. Quanto maior o papel desses ambientes nas democracias, maior a necessidade de avaliação do papel da IA na sua formação.

A opacidade compromete a capacidade dos cidadãos de compreender e avaliar como decisões que os afetam são tomadas, enfraquecendo assim a legitimidade democrática dos processos governamentais e dificultando o exercício do direito de contestação das decisões adotadas pela administração pública.

Além dos desafios à legitimidade democrática, a natureza "black box" dos sistemas de IA pode perpetuar e amplificar vieses e discriminações existentes. Este problema é particularmente grave quando sistemas algorítmicos são utilizados em áreas sensíveis da administração pública, pois podem reforçar desigualdades históricas sem que haja qualquer possibilidade de identificação da origem destes vieses, tornando impossível para os cidadãos contestarem decisões injustas baseadas em preconceitos arraigados nos dados que alimentam estes sistemas.

A transparência e mesmo o direito à explicação surgem, neste contexto, como um elemento fundamental para a preservação da autonomia dos cidadãos frente às decisões algorítmicas. Conforme a UNESCO recomenda:

> As pessoas devem ser plenamente informadas quando uma decisão é fundamentada ou tomada com base em algoritmos de IA, inclusive quando ela afeta sua segurança ou seus direitos humanos; nessas circunstâncias, elas também devem ter a oportunidade de solicitar explicações do ator de IA ou das instituições do setor público pertinentes (UNESCO, 2022, p. 22).

Sem estas explicações, os cidadãos ficam impossibilitados de exercer seu direito ao contraditório e à ampla defesa quando afetados por decisões automatizadas, transformando sistemas de IA governamentais em uma espécie de "oráculo místico cujos pronunciamentos devem ser aceitos por fé" (Cerutti e Grassi, 2018, p. 1, tradução nossa).

2.5. Regulação e Implementação da Transparência

Diante dos desafios apresentados pela opacidade dos sistemas de IA, diversas iniciativas regulatórias têm emergido para promover maior transparência.

A Organização das Nações Unidas no documento Princípios para o Uso Ético da Inteligência Artificial no Sistema das Nações Unidas destaca a transparência como um dos princípios do uso ético da IA:

> As organizações do sistema das Nações Unidas devem garantir a transparência e a explicabilidade dos sistemas de IA que utilizam em todas as etapas de seu ciclo de vida e dos processos de tomada de decisão envolvendo sistemas de IA. A explicabilidade técnica requer que as decisões tomadas por um sistema de IA possam ser compreendidas e rastreadas por seres humanos. Os indivíduos devem ser significativamente informados quando uma decisão que pode ou irá impactar seus direitos, liberdades fundamentais, direitos, serviços ou benefícios, é baseada ou tomada com base em algoritmos de IA e ter acesso às razões para uma decisão e à

> lógica envolvida. As informações e razões para uma decisão devem ser apresentadas de maneira que seja compreensível para eles. (NAÇÕES UNIDAS, 2022, p. 5, tradução nossa)

Ainda no âmbito internacional, a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) estabeleceu os Princípios sobre Inteligência Artificial (IA) como o primeiro padrão intergovernamental nessa área. Esses princípios visam promover uma IA inovadora e confiável que respeite os direitos humanos e os valores democráticos.

Adotados em 2019 e atualizados em 2024, eles fornecem diretrizes práticas e flexíveis para formuladores de políticas e atores de IA.

No referido documento, o Princípio 1.3 trata da transparência nos seguintes termos:

> 1.3 Transparência e explicabilidade. Os Atores de IA devem comprometer-se com a transparência e a divulgação responsável em relação aos sistemas de IA. Para este fim, eles devem fornecer informações significativas, apropriadas ao contexto e consistentes com o estado da arte: i. para promover uma compreensão geral dos sistemas de IA, incluindo suas capacidades e limitações, ii. para conscientizar as partes interessadas sobre suas interações com sistemas de IA, inclusive no local de trabalho, iii. quando viável e útil, fornecer informações claras e de fácil compreensão sobre as fontes de dados/entrada, fatores, processos e/ou lógica que levaram à previsão, conteúdo, recomendação ou decisão, para permitir que os afetados por um sistema de IA compreendam o resultado, e, iv. fornecer informações que permitam àqueles adversamente afetados por um sistema de IA contestar seu resultado. (OCDE, 2019, p. 8-9, tradução nossa)

Na União Europeia, o Regulamento Geral de Proteção de Dados (GDPR) introduziu o "direito à explicação". Tal qual observa Hassija *et al.* (2024, p. 51, tradução nossa), a capacidade de explicar como os algoritmos de IA chegam às suas decisões é um requisito legal na Europa. O Regulamento Geral de Proteção de Dados da União Europeia (GDPR) determina o direito de um indivíduo à explicação.

Ainda na União Europeia, o “Artificial Intelligence Act” também trouxe normas acerca da transparência, em especial em seus artigos 13 e 50. Especificamente quanto aos sistemas de alto risco, o art. 13 estabelece que:

> Os sistemas de IA de alto risco devem ser projetados e desenvolvidos de forma a garantir que sua operação seja suficientemente transparente para permitir que os implementadores interpretem a saída do sistema e a utilizem adequadamente. Um tipo e grau apropriados de transparência devem ser assegurados, visando o cumprimento das obrigações relevantes do fornecedor e do implementador estabelecidas na Seção 3. (UNIÃO EUROPEIA, 2024, tradução nossa)

Na França, este direito tem raízes ainda mais antigas. Villani (2018, p. 124, tradução nossa) lembra que:

> [...] em 1978, a Lei de Proteção de Dados francesa estabeleceu o princípio segundo o qual 'nenhuma decisão judicial ou outra envolvendo consequências legais para um indivíduo pode ser tomada exclusivamente com base no processamento automatizado de dados pessoais destinados a definir o perfil da pessoa em questão ou a avaliar certos

> aspectos de sua personalidade', acrescentando que 'um indivíduo tem o direito de conhecer e contestar essas informações e a lógica subjacente ao processamento automatizado quando esses resultados lhe são negados'.

No Brasil a transparência dos sistemas de IA é objeto de discussão legislativa. Em dezembro de 2024 foi aprovado pelo Senado Federal, o Projeto de Lei n. 2.338/2023 (Brasil, 2023), que "dispõe sobre o desenvolvimento, o fomento e o uso ético e responsável da inteligência artificial com base na centralidade da pessoa humana". Referido projeto de lei foi aprovado na forma de substitutivo, englobando diversos outros projetos de lei e emendas relativos ao tema, e encaminhado para a Câmara de Deputados, onde atualmente encontra-se em tramitação.

O Projeto de Lei n. 2.338/2023 traz várias disposições acerca da transparência. Logo no Artigo 3º, inciso VI, a proposta legislativa apresenta a transparência como princípio norteador, observando o equilíbrio necessário entre a abertura informacional e a proteção do segredo industrial, considerando a complexidade das cadeias de valor em IA. Este princípio se materializa no direito fundamental à informação, garantido pelo Artigo 5º, inciso I, que assegura às pessoas e grupos afetados por sistemas de IA o conhecimento sobre suas interações com tais sistemas, de maneira acessível, gratuita e compreensível, contemplando inclusive a ciência quanto ao caráter automatizado dessas interações.

Em relação aos sistemas classificados como de alto risco, o projeto aprofunda as garantias de transparência, avançando para um efetivo direito à explicação, conforme disposto no Artigo 6º, inciso I. Esta explicabilidade assume contornos práticos no Artigo 7º, que determina seu fornecimento mediante processo gratuito e em linguagem acessível, adequada à compreensão do resultado da decisão ou previsão automatizada. Ademais, o grau de transparência, explicabilidade e auditabilidade constitui critério relevante para a própria classificação de risco dos sistemas, sendo considerado de alto risco aquele cuja opacidade dificulte significativamente seu controle ou supervisão, conforme estabelece o Artigo 15, inciso VI, reforçando a centralidade desses elementos para a regulação proposta.

Na perspectiva da implementação técnica, o Projeto de Lei impõe aos desenvolvedores, por meio do Artigo 18, inciso II, alínea "d", a adoção de medidas que viabilizem tanto a aplicabilidade dos resultados dos sistemas quanto o fornecimento de informações adequadas para sua interpretação.

No âmbito da administração pública, o Artigo 24 atribui ao Poder Executivo federal a responsabilidade de estabelecer padrões mínimos de transparência para sistemas utilizados no setor público, bem como de fomentar esta prática em todos os órgãos e entidades públicas. Esta

orientação é reiterada no Artigo 69, inciso IV, que determina expressamente que os sistemas de IA utilizados por entes públicos devem perseguir a garantia de transparência, consolidando um compromisso institucional com a abertura informacional e a explicabilidade em todas as interações algorítmicas estatais.

A implementação efetiva da transparência em sistemas de IA enfrenta, contudo, desafios significativos, tanto técnicos quanto jurídicos. Um desses desafios é o aparente conflito entre transparência e eficiência. Como observa Villani (2018, p. 115, tradução nossa), "a responsabilidade de sistemas baseados em aprendizado de máquina constitui um verdadeiro desafio científico, que está criando tensão entre nossa necessidade de explicações e nossos interesses em eficiência". Quanto mais complexos os sistemas, maior a dificuldade de implementação de mecanismos de transparências e explicabilidade.

Outro desafio diz respeito aos segredos comerciais e à propriedade intelectual. O Center for Democracy & Technology (2025, tradução nossa) observa que:

> Muitos algoritmos comerciais são considerados proprietários, o que também complica a questão da explicação. Embora alguns tenham proposto a transparência algorítmica como uma solução para injustiça e explicabilidade, esta não é uma solução viável. Revelar detalhes técnicos sobre como a tecnologia funciona, mesmo a serviço de explicá-la ao público, poderia correr o risco de expor detalhes lucrativos aos concorrentes.

É importante notar, contudo, que o direito à explicação não implica necessariamente a divulgação de todos os aspectos técnicos de um sistema de IA. A mesma organização esclarece que:

> Embora explicações técnicas possam ser desafiadoras de fornecer, o conceito de explicação não se limita a uma descrição da mecânica. Muitas declarações de princípios que incluem esse valor estão focadas em um usuário entender a lógica por trás de uma decisão, em vez de ter acesso a detalhes técnicos. (Center for Democracy & Technology, 2025, tradução nossa)

A abordagem de "transparência qualificada", proposta por Chaudhary e acima já exposta, também oferece um caminho possível para equilibrar as necessidades de diferentes partes interessadas, desde indivíduos afetados por decisões algorítmicas até reguladores e desenvolvedores.

2.6. Educação e Participação Cívica

Para além da regulação, a educação e a participação cívica emergem como elementos cruciais para promover maior transparência e responsabilização em sistemas de IA. A UNESCO destaca que:

> A conscientização e a compreensão públicas das tecnologias de IA e do valor dos dados devem ser promovidas por meio de educação aberta e acessível, engajamento cívico, habilidades digitais e treinamento em ética da IA, alfabetização midiática e informacional, além de treinamento conduzido em conjunto por governos, organizações intergovernamentais, sociedade civil, universidades, meios de comunicação, líderes comunitários e setor privado, e considerando a diversidade linguística, social e cultural existente, para garantir a participação efetiva do público (UNESCO, 2022, p. 22).

Coeckelbergh (2024, p. 5-6, tradução nossa) enfatiza a necessidade de uma educação cívica que vá além do modelo humanista tradicional:

> Um papel mais ativo e republicano para os cidadãos, no entanto, não deve ser apoiado apenas por conhecimentos externos, mas também pressupõe conhecimentos e habilidades por parte dos cidadãos e, portanto, requer uma melhor educação para todos. Requer o que se pode chamar de educação "cívica": educação na e para a democracia, voltada para o florescimento da pólis e seus cidadãos – de fato, voltada para o bem comum. Isso implicaria o treinamento de habilidades críticas e deliberativas. Em contraste com a educação humanista clássica, no entanto, essa educação não apenas possibilitaria uma relação mais crítica com os textos, mas também desenvolveria um melhor conhecimento de (outras) tecnologias e uma melhor consciência dos aspectos éticos e políticos dessas tecnologias.

Essa educação cívica deve incluir não apenas a compreensão técnica de sistemas de IA, mas também uma consciência crítica de suas implicações éticas, sociais e políticas, inclusive para os desenvolvedores. Villani (2018, p. 124, tradução nossa) defende que:

> Para poder treinar especialistas para serem mais responsáveis, o ensino de ética — e das ciências sociais em geral — deve ser incluído em todos os currículos de cursos de engenharia e ciência da computação. Em última análise, o objetivo seria produzir graduados com a expertise técnica necessária para desenvolver sistemas eficientes e as habilidades em ciências sociais necessárias para entender o impacto de seus desenvolvimentos na sociedade e em seus cidadãos.

A participação multidisciplinar no desenvolvimento e governança de sistemas de IA é outro aspecto fundamental:

> À medida que os sistemas de IA se tornam cada vez mais sofisticados e integrados em vários domínios da atividade humana, é crucial considerar como essas tecnologias podem ser desenvolvidas e implantadas de maneiras que se alinhem com princípios éticos e promovam o florescimento humano. Isso requer colaboração interdisciplinar contínua entre pesquisadores, desenvolvedores, formuladores de políticas e eticistas para garantir que os sistemas de IA sejam projetados com transparência, *accountability* e respeito pelos valores humanos. (Karpouzis, 2024, p. 8, tradução nossa)

Finalmente, Coeckelbergh (2024, p. 5, tradução nossa) sugere que a participação ativa dos cidadãos na governança da IA deve ser vista não apenas como um direito, mas como uma virtude cívica:

> Vista dessa perspectiva, a questão então não é apenas: como os governos e corporações devem respeitar meus direitos e como eles podem ouvir minha voz e me permitir beneficiar do bem comum criado por eles, mas também como eu e nós (digamos, comunidades) podemos contribuir para o bem comum usando IA e outras tecnologias, e deliberando sobre seu uso. Fazer isso seria uma obrigação cívica e uma virtude cívica.

Esta visão de participação cívica ativa na governança da IA juntamente com a implementação de mecanismos de transparência oferecem um caminho para a "libertação" coletiva da caverna digital algorítmica.

3. CONSIDERAÇÕES FINAIS

A investigação conduzida neste artigo evidenciou que a analogia entre a Alegoria da Caverna de Platão e os sistemas de Inteligência Artificial contemporâneos transcende o exercício meramente retórico, constituindo uma ferramenta analítica relevante para a compreensão dos desafios impostos pela opacidade algorítmica. Assim como os prisioneiros da caverna tomavam as sombras projetadas na parede como a totalidade do real, usuários e mesmo desenvolvedores de sistemas de IA permanecem, em grande medida, limitados a uma percepção fragmentária dos processos decisórios que essas tecnologias executam.

Conforme demonstrado ao longo deste estudo, a opacidade inerente aos sistemas de aprendizado de máquina não é uma mera limitação técnica, mas um problema de dimensões éticas, políticas e jurídicas profundas. Os vieses algorítmicos, alimentados por dados historicamente enviesados, podem reproduzir e amplificar injustiças sistêmicas, especialmente quando operam sob o véu da opacidade que impossibilita sua identificação e contestação. No âmbito da Administração Pública, essa condição se revela particularmente grave, na medida em que decisões automatizadas passam a afetar direitos fundamentais dos cidadãos sem que haja mecanismos adequados de compreensão, fiscalização e controle social.

Nesse contexto, a transparência emerge não como uma solução isolada, mas como condição necessária para a construção de uma governança algorítmica legítima e democrática. Como sugere a metáfora de Floridi, referida pelo Instituto para Ética nos Negócios de Londres (2018, p. 3), diante de um ambiente desconhecido e obscuro, a resposta mais produtiva não reside no temor, mas na iluminação, isto é, na promoção de maior abertura e explicabilidade dos sistemas de IA, permitindo que a sociedade navegue com segurança pelo universo algorítmico.

As iniciativas regulatórias analisadas, tanto no plano internacional, como os princípios da OCDE, a Recomendação da UNESCO e o AI Act europeu, quanto no âmbito nacional, com o Projeto de Lei n. 2.338/2023, sinalizam um avanço significativo no reconhecimento da transparência como princípio estruturante da governança de IA. Todavia, a efetividade dessas normativas dependerá de sua capacidade de equilibrar o imperativo da explicabilidade com a

proteção de segredos comerciais, a viabilidade técnica e a adequação aos diferentes públicos e contextos.

Por fim, a "libertação" da caverna digital não se esgota na esfera regulatória. Tal como na alegoria platônica, em que o prisioneiro liberto retorna para despertar seus companheiros, a superação da opacidade algorítmica exige também uma educação cívica ampla, que capacite os cidadãos a compreender criticamente os sistemas que mediam suas experiências sociais e políticas, e uma participação ativa e multidisciplinar na governança dessas tecnologias. A transparência em IA é, em última análise, um imperativo democrático: condição para que os sistemas algorítmicos sirvam ao bem comum, e não se convertam em instrumentos de poder insuscetíveis ao escrutínio público.

4. REFERÊNCIAS

ASATIANI, Aleksandre *et al.* Challenges of Explaining the Behavior of Black-Box AI Systems. **MIS Quarterly Executive**, v. 19, n. 4, p. 259-278, dez. 2020. Disponível em: https://aisel.aisnet.org/misqe/vol19/iss4/7. Acesso em: 3 fev. 2025..

BRASIL. Projeto de Lei n. 2.338, de 2023. Dispõe sobre o uso da Inteligência Artificial. Senado Federal, Brasília, DF, 2023. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Acesso em: 24 fev. 2025.

BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. **Big Data & Society**, jan./jun. 2016. Disponível em: https://doi.org/10.1177/2053951715622512. Acesso em: 03 fev. 2025.

CENTER FOR DEMOCRACY & TECHNOLOGY. **AI & Machine Learning**. Disponível em: https://cdt.org/ai-machine-learning/. Acesso em: 13 fev. 2025.

CERUTTI, Federico; GRASSI, Alessia; VALLATI, Mauro. Unveiling the Oracle: Artificial Intelligence for the XXI Century. **AI Communications**, v. 31, n. 6, p. 1-9, 2018. Disponível em: https://content.iospress.com/articles/ai-communications/aic789. Acesso em: 3 fev. 2025.

CHAUDHARY, Gyandeep. Unveiling the Black Box: Bringing Algorithmic Transparency to AI. **Masaryk University Journal of Law and Technology**, v. 18, n. 1, p. 93-122, 2024. Disponível em: https://journals.muni.cz/mujlt/article/download/36881/32877 Acesso em: 10 fev. 2025.

COECKELBERGH, Mark. Artificial intelligence, the common good, and the democratic deficit in AI governance. **AI and Ethics**, 2024. Disponível em: https://doi.org/10.1007/s43681-024-00492-9. Acesso em: 3 fev. 2025.

EUROPEAN UNION. **Artificial Intelligence Act**. Disponível em: https://artificialintelligenceact.eu/. Acesso em: 04 fev. 2025.

FLORIDI, Luciano *et al.*. How to Design AI for Social Good: Seven Essential Factors. **Science and Engineering Ethics**, v. 26, n. 3, p. 1771-1796, 2020. Disponível em: https://doi.org/10.1007/s11948-020-00213-5. Acesso em: 15 fev 2025.

FLORIDI, Luciano; COWLS, Josh; BELTRAMETTI, Monica; CHATILA, Raja; CHAZERAND, Patrice; DIGNUM, Virginia; LUETGE, Christoph; MADELIN, Robert; PAGALLO, Ugo; ROSSI, Francesca; SCHAFER, Burkhard; VALCKE, Peggy; VAYENA, Effy. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. **Minds and Machines**, v. 28, p. 689-707, 2018. Disponível em: https://doi.org/10.1007/s11023-018-9482-5. Acesso em: 03 fev. 2025.

FLORIDI, Luciano. Should we be afraid of AI? **Aeon**, 2016. Disponível em: https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible. Acesso em: 03 fev. 2025.

HASSIJA, Vikas; CHAMOLA, Vinay; MAHAPATRA, Atmesh; SINGAL, Abhinandan; GOEL, Divyansh; HUANG, Kaizhu; SCARDAPANE, Simone; SPINELLI, Indro; MAHMUD, Mufti Sultan; HUSSAIN, Amir. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. **Cognitive Computation**, v. 16, n. 1, p. 45-74, 2024. Disponível em: https://doi.org/10.1007/s12559-023-10179-8. Acesso em: 3 fev. 2025.

INSTITUTE OF BUSINESS ETHICS. Business Ethics and Artificial Intelligence. **Business Ethics Briefing**, n. 58, Londres, jan. 2018. Disponível em: https://www.ibe.org.uk/static/5f167681-e05f-4fae-ae1bef7699625a0d/ibebriefing58businessethicsandartificialintelligence.pdf. Acesso em: 03 fev. 2025.

JUNGHERR, Andreas. Artificial Intelligence and Democracy: A Conceptual Framework. **Social Media + Society**, v. 9, n. 3, p. 1-14, 2023. Disponível em: https://doi.org/10.1177/20563051231186353. Acesso em: 03 fev. 2025.

JUNGHERR, Andreas e SCHROEDER, Ralph. Artificial intelligence and the public arena. **Communication Theory**, v. 33, n. 2-3, p. 164-173, 2023. Disponível em: https://doi.org/10.1093/ct/qtad006. Acesso em: 01 fev. 2025.

KARPOUZIS, Kostas. Plato's Shadows in the Digital Cave: Controlling Cultural Bias in Generative AI. **Electronics**, v. 13, n. 1457, p. 1-13, 2024. Disponível em: https://doi.org/10.3390/electronics13081457. Acesso em: 3 fev. 2025.

MILLAR, Janna; ETZIONI, Oren; HIBBARD, Robin. **Work in the age of artificial intelligence:** Four perspectives on the economy, employment, skills and ethics. Ministry of Economic Affairs and Employment of Finland. Helsinki, 2018. Disponível em: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160980/TEMjul_21_2018_Work_in_the_age.pdf. Acesso em: 14 fev. 2025.

MURPHY, Joseph. **Throwing Shade or Shadows:** AI as an Extension of The Allegory of the Cave. LinkedIn, 24 abr. 2023. Disponível em: https://www.linkedin.com/pulse/throwing-shade-shadows-ai-extension-allegory-cave-joseph-murphy/. Acesso em: 25 fev. 2025.

NAÇÕES UNIDAS. **Principles for the Ethical Use of Artificial Intelligence in the United Nations System**. Inter-Agency Working Group on Artificial Intelligence, High-Level Committee on Programmes (HLCP), Chief Executives Board for Coordination (CEB), 20 set. 2022. Disponível em: https://unsceb.org/sites/default/files/2022-09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf Acesso em: 7 fev. 25.

ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. **Recommendation of the Council on Artificial Intelligence**. Paris, 2019. Disponível em: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Acesso em: 20 jan. 2025.

PLATÃO. **A República**. Introdução, tradução e notas de Maria Helena da Rocha Pereira. 9. ed. Lisboa: Fundação Calouste Gulbenkian,2001.

RAHWAN, Iyad. Society-in-the-Loop: Programming the Algorithmic Social Contract. **Ethics of Information Technology**, 20 jul. 2017. Disponível em: https://arxiv.org/abs/1707.07232. Acesso em: 03 fev. 2025.

UNESCO. **Recomendação sobre a Ética da Inteligência Artificial**. Paris, 2022. Disponível em: https://on.unesco.org/EthicsAI. Acesso em: 03 fev. 2025.

UNIÃO EUROPEIA. **Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024, que estabelece disposições harmonizadas relativas à**

inteligência artificial (Artificial Intelligence Act - AIA) e que altera os regulamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, o Regulamento (UE) 2018/1139 e a Diretiva 2020/1828, e que revoga o Regulamento (UE) 2019/1020. Diário Oficial da União Europeia, L 248, 13 de junho de 2024. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX:32024R1689. Acesso em: 24 fev. 2025.

VILLANI, Cédric. **For a Meaningful Artificial Intelligence**: Towards a French and European Strategy. Paris, 2018. Disponível em: https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf. Acesso em: 03 fev. 2025.