2

Responsabilidade na era da IA: transparência e Mercosul

Responsibility in the AI Age: Transparency and Mercosul

DOI: 10.29327/5798915.1-4

Daniela América da Silva[4]
Johnny Cardoso Marques[4]
Delmo Mattos da Silva[4]

Resumo: A delegação de decisões à Inteligência Artificial (IA) introduz um dilema de responsabilidade crítico para os seres humanos. Esse desafio é exacerbado pela velocidade e complexidade da cadeia de desenvolvimento e uso (o problema das "muitas mãos" e "muitas coisas") e, fundamentalmente, pela falta de conhecimento e explicabilidade (XAI) inerente às operações da IA. A Transparência é identificada como fator mitigador essencial. Esta pesquisa busca definir a transparência, avaliar seu papel na redução de vieses algorítmicos e analisar sua abordagem específica nos princípios emergentes de IA dos países membros do Mercosul, propondo, por fim, uma visão macro dos requisitos de transparência em toda a região.

Palavras-chave: Inteligência Artificial, Responsabilidade, Transparência, Mercosul, Vieses Algorítmicos, Princípios

Abstract: Delegating decisions to Artificial Intelligence (AI) introduces a critical responsibility dilemma for humans. This challenge is exacerbated by the speed and complexity of the development and usage chain (the "many hands" and "many things" problem), and fundamentally, by the lack of

[4] Instituto Tecnológico de Aeronáutica, São José dos Campos, SP, Brasil, damerica@ita.br, johnny@ita.br, delmo@ita.br.

knowledge and explainability (XAI) inherent in AI operations. Transparency is identified as the essential mitigating factor. This research seeks to define transparency, evaluate its role in reducing algorithmic bias, and analyze its specific approach within the emerging AI principles of Mercosul member countries, ultimately proposing a macro vision of the transparency requirements across the region.

Keywords: Artificial Intelligence, Accountability, Transparency, Mercosul, Algorithmic Biases, Principles

INTRODUÇÃO

A utilização da Inteligência Artificial (IA) na tomada de decisões, que possibilita a delegação de tarefas à máquina em uma extensão superior à habitual, suscita um problema central: a atribuição de responsabilidade (Coeckelbergh, 2024). Para os seres humanos, a causalidade de um efeito sobre o mundo e sobre terceiros implica a responsabilidade pelas consequências, o que, segundo Aristóteles, constitui a condição primordial para a responsabilidade moral. No entanto, ao delegar decisões e ações à IA, cujas consequências podem ser eticamente significativas, depara-se com um dilema. Uma vez que a IA carece de consciência e de raciocínio moral, ela não pode ser considerada responsável pelos seus atos (Coeckelbergh, 2024).

Máquinas podem ser consideradas agentes, mas não agentes morais, visto que lhes são inerentes a falta de consciência e de livre-arbítrio. Consequentemente, os seres humanos delegam a agência, mas retêm a responsabilidade. Não obstante, a velocidade de processamento da IA e a complexidade do seu histórico de desenvolvimento dificultam o processo de atribuição de responsabilidade (Coeckelbergh, 2024). A título de exemplo, uma IA pode ter sido concebida no âmbito de um projeto científico universitário e, subsequentemente, ser implementada em um setor de saúde ou em um contexto militar, levantando a questão: quem será o responsável? (Coeckelbergh, 2024).

Os sistemas de tecnologia implicam os conceitos de "muitas mãos" e "muitas coisas". O cenário de "muitas mãos" estabelece que a responsabilidade por falhas, a exemplo de um acidente envolvendo um veículo autônomo, é di-

fusa. Esta abrange programadores, a empresa desenvolvedora, o usuário, órgãos reguladores, entre outros agentes. Por sua vez, o conceito de "muitas coisas" refere-se à multiplicidade de componentes técnicos: algoritmos que interagem com sensores, dados, *hardware* e *software*. No contexto do aprendizado de máquina (*machine learning*), há diversas fases (coleta, tratamento e treinamento), cada uma delas com elementos técnicos e decisões humanas, o que complexifica a causalidade (Coeckelbergh, 2024).

A responsabilidade humana pelas ações da IA está intrinsecamente conectada ao problema do conhecimento: para ser considerado responsável, o ser humano necessita saber e ser capaz de explicar o funcionamento da IA. A responsabilidade exige, portanto, elucidação e explicabilidade (Coeckelbergh, 2024). A dificuldade reside no problema da transparência. Em certos sistemas, o processo de tomada de decisão da IA não é claro nem explicável, configurando o que se denomina caixa-preta. Quando o mecanismo pelo qual a IA atinge um resultado não pode ser rastreado ou justificado, torna-se um desafio responsabilizar o ser humano, uma vez que este se vê impedido de "responder pelas ações" do sistema (Coeckelbergh, 2024).

Diante do exposto, e com o objetivo de elucidar o papel da transparência e os problemas correlatos no contexto da IA, propõem-se as seguintes Questões de Pesquisa (QPs):

- QP1: O que é a transparência da IA?
- QP2: Como a transparência auxilia na mitigação de vieses?
- QP3: Como a transparência é abordada nos princípios para a IA nos países do Mercosul?

Em consonância com as questões de pesquisa estabelecidas, o trabalho será estruturado da seguinte forma: primeiramente, a seção *Contexto* abordará a compreensão do conceito de transparência e seus desafios. Em seguida, a seção *Discussão* apresentará a relevância da transparência para a mitigação de vieses e o desenvolvimento desse conceito nos países do Mercosul. Posteriormente, a seção *Modelo Proposto* delineará uma visão macro dos principais requisi-

tos para a transparência da IA aplicáveis aos países do Mercosul. Por fim, a seção *Conclusão* resumirá as principais contribuições e realizações do estudo conduzido.

Contexto

Considerando o jogo de xadrez, embora os programadores detenham o conhecimento sobre o funcionamento da IA, eles desconhecem o modo como esta concebeu uma jogada específica. Tal fato representa um problema de responsabilidade, pois inviabiliza que os seres humanos expliquem uma decisão particular do sistema (Coeckelbergh, 2024). Essa dificuldade inerente a alguns algoritmos de IA é denominada problema da caixa-preta, visto que, embora os programadores originais conheçam o código, os desenvolvedores subsequentes e os usuários não têm conhecimento preciso do que a IA está, de fato, executando (Coeckelbergh, 2024).

A transparência da IA está intrinsecamente relacionada aos conceitos de explicabilidade e interpretabilidade, uma vez que estes fornecem *insights* que auxiliam na resolução do problema da caixa-preta. Tais conceitos também contribuem para a resposta da Questão de Pesquisa (QP1), que indaga: "O que é a transparência da IA?" (Jonker *et al*., 2025). É importante notar que estes conceitos apresentam definições e casos de uso distintos:

- Explicabilidade da IA: Busca responder: "Como o modelo chegou a estes resultados?" (Jonker *et al*., 2025). É o conjunto de métodos que permite aos usuários humanos compreender e confiar nas saídas geradas. A explicabilidade examina o caminho percorrido para um resultado específico, complementando a interpretabilidade ao analisar o processo real da decisão.
- Interpretabilidade da IA: Visa responder: "Como o modelo toma decisões?" (Jonker *et al.*, 2025). Torna todo o processo de funcionamento da IA compreensível, medindo a taxa de sucesso com que humanos podem prever a saída do sistema, focando na lógica subjacente, relevância e consequências esperadas.

- Transparência da IA: É o conceito mais abrangente, buscando responder: "Como o modelo foi criado, quais dados o treinaram e como ele toma decisões?" (Jonker *et al*., 2025). Ela engloba todos os fatores do desenvolvimento e implementação, incluindo dados de treinamento e políticas de acesso.

Soma-se a isso um problema de conhecimento relacionado à IA, dado que muitos usuários desconhecem o funcionamento do sistema, seus potenciais efeitos e, em diversas ocasiões, sequer estão cientes de sua utilização (Coeckelbergh, 2024). Tais problemas podem, adicionalmente, ser avaliados sob a perspectiva da confiança, visto que a ausência de transparência culmina na diminuição da credibilidade tanto na tecnologia quanto nos indivíduos que a empregam (Coeckelbergh, 2024).

A transparência é crucial para a reflexão social, pois o risco não é apenas a dominação por uma elite tecnológica, mas uma sociedade tecnológica onde ninguém se responsabiliza pelos impactos gerados (Joker *et al*., 2025). A abertura da "caixa-preta" beneficia a tecnologia, permitindo a identificação e eliminação de correlações espúrias, aprimorando o sistema (Jonker *et al*., 2025). Contudo, a implementação da transparência é desafiadora. Instituições frequentemente evitam revelar seus algoritmos para proteger interesses comerciais e propriedade intelectual, o que pode comprometer a divulgação necessária (Jonker *et al*., 2025). O equilíbrio entre o benefício social e a proteção comercial é, portanto, central.

A explicabilidade da IA, um pilar da transparência, transcende a mera comunicação de decisões, levantando questões filosóficas e científicas sobre a natureza da explicação em si. Estudos em psicologia e ciência cognitiva (Miller, 2019) sugerem que, como nas interações humanas, as explicações da IA devem ser sociais, adaptadas às crenças do receptor, em vez de um encadeamento causal completo.

A transparência é vital para a IA Responsável, pois permite aos *stakeholders* avaliar a precisão preditiva, a imparcialidade e a mitigação de vieses nos modelos (Jonker *et al*., 2025; da Silva *et al.*, 2024; da Silva e Marques, 2025).

Essa abordagem visa alinhar o desenvolvimento, projeto e implementação da IA com padrões legais, éticos e os interesses das partes envolvidas.

Considerando que os sistemas de IA têm uso regional ou global, a transparência exige o conhecimento dos princípios de diversas tradições culturais. Compreender essas nuances é crucial para a aplicabilidade da explicabilidade da IA em diferentes contextos, fomentando uma explicação socialmente contextualizada da tecnologia (da Silva *et al*., 2025; Coeckelbergh, 2024).

Neste âmbito, o viés configura-se como um problema de natureza tanto ética quanto social. Quando uma IA toma ou recomenda decisões, estas podem se manifestar como injustas ou desleais para indivíduos ou grupos específicos. Embora o viés possa surgir em aplicações de IA clássica – por exemplo, em um sistema especialista ou um banco de dados –, ele está prevalentemente associado a aplicações de aprendizado de máquina (Coeckelbergh, 2024). A compreensão desse fenômeno propicia a resposta à Questão de Pesquisa (QP2): "Como a transparência auxilia na mitigação de vieses?"

O viés, via de regra, não é intencional, manifestando-se frequentemente devido à falha dos desenvolvedores da IA em prever efeitos discriminatórios contra grupos ou indivíduos específicos. Essa ocorrência se deve a múltiplos fatores, como a insuficiente compreensão do sistema de IA, a falta de ciência acerca do problema de enviesamento, o desconhecimento dos próprios preconceitos ou a incapacidade de antecipar as consequências não intencionais da tecnologia. Adicionalmente, observa-se a insuficiência de reflexão sobre tais consequências e a ausência de contato com as partes interessadas relevantes (*stakeholders*) (Coeckelbergh, 2024; da Silva *et al*., 2024).

Entretanto, os impactos decorrentes de decisões tendenciosas da IA podem gerar consequências graves para o acesso a recursos e para as liberdades individuais. Indivíduos podem ser impedidos de obter emprego ou crédito, ser submetidos à prisão ou sofrer violência. Ademais, os efeitos de tais decisões não se restringem aos indivíduos, podendo

afetar comunidades inteiras (Coeckelbergh, 2024; da Silva *et al.*, 2021; da Silva *et al*., 2023).

Um caso notório é o do sistema COMPAS (utilizado para predizer a reincidência de réus na Flórida). Uma análise conduzida pela ProPublica demonstrou que o sistema produziu uma taxa desproporcional de falsos positivos para réus negros (aqueles previstos para reincidir, mas que não o fizeram) e uma taxa desproporcional de falsos negativos para réus brancos (aqueles previstos para não reincidir, mas que reincidiram), o que contribuiu para a perpetuação de disparidades raciais (Coeckelbergh, 2024).

Outra fonte de viés reside nas bases de dados, as quais podem carecer de representatividade da população global. A base ImageNet, a título de exemplo, apresenta uma forte representação dos Estados Unidos da América (EUA), enquanto países como China e Índia estão sub-representados. Este desequilíbrio é atribuído ao fato de as imagens terem sido coletadas por meio de buscas tendenciosas (Coeckelbergh, 2024; Yang *et al.*, 2020).

Vieses de gênero também se manifestam, sendo evidentes, por exemplo, em Modelos de Linguagem Natural (LLMs) aplicados à área da saúde. Estudos apontam que a linguagem gerada para descrever as necessidades clínicas femininas demonstrou maior propensão a minimizar as condições das pacientes, em contraste com a linguagem mais direta empregada para o público masculino. Tal discrepância pode resultar em disparidades no acesso a serviços de saúde, visto que estes são alocados com base na necessidade clínica (Coeckelbergh, 2024; Rickman, 2025).

Por fim, há o viés relacionado ao uso humano da IA, que envolve o risco de confiança excessiva nos algoritmos, levando os tomadores de decisão a desconsiderar seu próprio julgamento. Em modelos de linguagem, essa confiança exacerbada pode se manifestar como comportamentos de apego excessivo, nos quais a IA valida em demasia os sentimentos do usuário, o que pode prejudicar a saúde mental (Coeckelbergh, 2024; O'Donnell, 2025; Slattery *et al.*, 2024).

Contudo, um algoritmo totalmente imparcial é inviável, pois não há consenso sobre justiça ou equidade perfeita. Os dados da IA são abstrações e refletem escolhas humanas, não sendo neutros (Coeckelbergh, 2024). Assim, o tratamento do viés na IA é uma questão política e filosófica, ligada ao modelo de sociedade desejado. Se a decisão humana é injusta, a IA poderia ter o papel de revelar nossos preconceitos inerentes e instruir-nos sobre a condição humana (Coeckelbergh, 2024).

Discussão

O arcabouço regulatório em torno da IA encontra-se em constante evolução, sendo a Lei de IA da União Europeia (UE) o primeiro *framework* abrangente a nível global. Esta legislação adota uma abordagem baseada em risco, que implica a aplicação de regras distintas e, em certos casos, a proibição de usos específicos. Para a devida conformidade, a lei exige transparência, governança rigorosa e uma eficaz gestão de riscos (Jonker *et al*., 2025).

Especialistas preveem um "Efeito Bruxelas", análogo ao Regulamento Geral de Proteção de Dados (GDPR), onde a Lei de IA da UE atuará como catalisadora para o desenvolvimento de padrões globais de ética e governança. Embora a maioria dos países ainda careça de legislação abrangente, a existência de *frameworks* e de projetos de lei em blocos regionais como o Mercosul evidencia que o tema é incipiente, mas está em ascensão. A transparência dos processos é, portanto, crucial para atender às exigências de auditores e órgãos reguladores (Jonker *et al*., 2025). Exemplos relevantes no Mercosul são sumarizados no Quadro 1, que demonstra os projetos de lei que foram recentemente aprovados ou estão em tramitação nos países do bloco (Blanchet, 2025).

Existe, ainda, uma declaração conjunta dos países do bloco, publicada em 2023, intitulada "Mercosul– Declaração de Princípios de Direitos Humanos sobre Inteligência Artificial" (Mercosul, 2023). No que concerne à transparência, a declaração estabelece que:

> A transparência e a explicabilidade algorítmica, bem como o controle humano da tecnologia, são essenciais

> para garantir que as decisões tomadas pelos sistemas de inteligência artificial sejam compreensíveis para as partes envolvidas, afetadas e interessadas. Os sistemas de inteligência artificial devem ser projetados de maneira a permitir uma supervisão adequada, de acordo com as normas de cada país, e a explicação de seus processos e resultados, de forma universal por idôneos na matéria.

O documento enfatiza a importância da supervisão adequada em conformidade com a legislação interna de cada nação.

Quadro 1 - Principais projetos de lei no Mercosul e princípios considerados.

País	Situação da Lei de Inteligência Artificial
Argentina	Iniciativas legislativas em desenvolvimento, com ênfase na proteção de dados pessoais e na incorporação de princípios de transparência, equidade e proteção de dados, ainda sem marco regulatório específico para IA (Argentina, 2023).
Brasil	Marco consolidado de proteção de dados por meio da LGPD, complementado por estratégias nacionais de IA, e projeto de lei em tramitação no Congresso Nacional voltado à regulação do uso da inteligência artificial (Brasil, 2024; 2025).
Chile	Projeto de Lei de Inteligência Artificial e Robótica em tramitação, com enfoque na regulação de sistemas de IA a partir dos princípios de não discriminação, inclusão e transparência (Chile, 2025).
México	Propostas legislativas em tramitação que articulam a proteção de dados pessoais e os direitos humanos, com iniciativas setoriais voltadas à regulação da inteligência artificial (Mexico, 2024a; 2024b).
Peru	Marco normativo vigente que inclui a Estratégia Nacional de Inteligência Artificial e legislação específica, orientada por princípios de sustentabilidade, inclusão e direitos humanos (Peru,2023).
Colômbia	Legislação de proteção de dados pessoais vigente, acompanhada de propostas normativas em desenvolvimento para a construção de um regime regulatório específico de inteligência artificial (Colombia, 2023; 2025).

Equador	Projeto de Lei de Governança da Inteligência Artificial em tramitação, com foco em ética, segurança e proteção de dados no uso de sistemas de IA (Equador, 2025).
Uruguai	Estratégia nacional e iniciativas legislativas em tramitação que tratam da inteligência artificial e das neurotecnologias, fundamentadas nos princípios de IA responsável, neurodireitos e não discriminação (Uruguai, 2025).

Dada a relevância de se compreender as normativas internas de cada país, a seção subsequente apresentará a descrição da transparência nas principais legislações das nações do Mercosul. O objetivo é, assim, responder à Questão de Pesquisa (QP3): "Como a transparência é abordada nos princípios para a IA nos países do Mercosul?"

Argentina

Embora não exista uma regulamentação específica para a IA, a norma 87/2023 estabelece diretrizes para a proteção de dados. O Artigo 6º define o Princípio de licitude, lealdade e transparência, nos seguintes termos:

- "Os dados pessoais devem ser tratados de maneira lícita, leal e transparente. O tratamento é considerado lícito se for realizado de acordo com o estabelecido na presente Lei e complementaridade normativa. Considera-se leal se o responsável pelo tratamento se abster de tratar os dados através de meios enganosos ou fraudulentos. É transparente se as informações que estiverem vinculadas ao tratamento dos dados forem facilmente acessíveis e utilizarem uma linguagem simples e clara" (Argentina, 2023).

Brasil

No que se refere à Lei Geral de Proteção de Dados (LGPD) do Brasil (Brasil, 2024), o princípio da transparência é estabelecido nos seguintes artigos:

- Art. 6º As atividades de tratamento de dados pessoais deverão observar a boa-fé e os seguintes princípios: VI – Transparência: Garantia, aos titulares, de informações claras, precisas e facilmente acessíveis sobre a realiza-

ção do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial.

- Art. 9º O titular tem direito ao acesso facilitado às informações sobre o tratamento de seus dados, que deverão ser disponibilizadas de forma clara, adequada e ostensiva acerca de, entre outras características previstas em regulamentação para o atendimento do princípio do livre acesso:
    - § 1º Na hipótese em que o consentimento é requerido, este será considerado nulo caso as informações fornecidas ao titular tenham conteúdo enganoso ou abusivo ou não tenham sido apresentadas previamente com transparência, de forma clara e inequívoca.
- Art. 10 O legítimo interesse do controlador somente poderá fundamentar tratamento de dados pessoais para finalidades legítimas, consideradas a partir de situações concretas, que incluem, mas não se limitam a:
    - § 2º O controlador deverá adotar medidas para garantir a transparência do tratamento de dados baseado em seu legítimo interesse.
- Art. 40 A autoridade nacional poderá dispor sobre padrões de interoperabilidade para fins de portabilidade, livre acesso aos dados e segurança, assim como sobre o tempo de guarda dos registros, tendo em vista especialmente a necessidade e a transparência.

Em relação ao Projeto de Lei (PL) nº 2.338/2023 (Brasil, 2025), destacam-se os seguintes pontos sobre a transparência:

- Art. 2º O desenvolvimento, a implementação e o uso de sistemas de inteligência artificial no Brasil têm como fundamentos:
    - VI – Transparência, explicabilidade, inteligibilidade e auditabilidade.
- Art. 18. Caberá à autoridade competente atualizar a lista dos sistemas de inteligência artificial de risco excessivo ou de alto risco, identificando novas hipóteses, com base em, pelo menos, um dos seguintes critérios:

    - VII – Baixo grau de transparência, explicabilidade e auditabilidade do sistema de inteligência artificial, que dificulta o seu controle ou supervisão.

- Art. 19. Os agentes de inteligência artificial estabelecerão estruturas de governança e processos internos aptos a garantir a segurança dos sistemas e o atendimento dos direitos de pessoas afetadas, nos termos previstos no Capítulo II desta Lei e da legislação pertinente, que incluirão, pelo menos:
    - I – Medidas de transparência quanto ao emprego de sistemas de inteligência artificial na interação com pessoas naturais, o que inclui o uso de interfaces ser-humano-máquina adequadas e suficientemente claras e informativas;
    - II – Transparência quanto às medidas de governança adotadas no desenvolvimento e emprego do sistema de inteligência artificial pela organização.
- Art. 24. A metodologia da avaliação de impacto conterá, ao menos, as seguintes etapas:
    - i) Medidas de transparência ao público, especialmente aos potenciais usuários do sistema, a respeito dos riscos residuais, principalmente quando envolver alto grau de nocividade ou periculosidade à saúde ou segurança dos usuários, nos termos dos artigos 9 e 10 da Lei nº 8.078, de 11 de setembro de 1990 (Código de Defesa do Consumidor).

Justificativa: Além de estabelecer direitos básicos e transversais para qualquer contexto de interação entre máquina e ser humano, como informação e transparência, tal obrigação é intensificada quando o sistema de IA produz efeitos jurídicos relevantes ou impacta os sujeitos de maneira significativa (por exemplo: direito de contestação e intervenção humana).

No que tange à governança dos sistemas, o projeto de lei elenca as medidas a serem adotadas para garantir a transparência e a mitigação de vieses; estabelece medidas adicionais para sistemas de alto risco e para sistemas gover-

namentais de inteligência artificial; e normatiza o procedimento para a avaliação de impacto algorítmico.

CHILE

O Boletim n.º 15.869-19 e 16.821-19 (Chile, 2025) constitui o projeto de lei que visa regulamentar os sistemas de IA, do qual se extraem as seguintes disposições:

- Transparência e Identificação
    - Os sistemas de IA serão desenvolvidos e utilizados de modo a facilitar uma rastreabilidade adequada, em conformidade com o ordenamento jurídico vigente. Adicionalmente, o sistema deverá identificar-se como agente artificial em cada oportunidade de interação com seres humanos, permitindo que estes tomem conhecimento de forma clara e precisa e estejam conscientes de que a comunicação ou interação ocorre com um sistema de IA.
- Explicabilidade
    - Os sistemas de IA serão criados, desenvolvidos, inovados, implementados e utilizados de maneira que seus resultados ou *outputs* sejam compreensíveis e inteligíveis para as pessoas por eles impactadas; e promoverão a transparência e a rastreabilidade em todas as suas operações.
- Art. 5º Obrigações de transparência em determinados sistemas de IA. Todo operador de sistemas de IA que gere conteúdo sintético de áudio, imagem, vídeo ou texto deverá zelar para que seus resultados ou *outputs* sejam identificáveis como gerados ou manipulados de maneira artificial.
- Riscos
    - Os sistemas de IA de risco limitado devem garantir condições de transparência e segurança proporcionais ao seu nível de risco, de modo que os indivíduos sejam informados de forma clara e precisa e possam reconhecer que estão interagindo com um sistema de IA.

MÉXICO

No que diz respeito à proteção de dados (México, 2024a), o princípio da transparência é contemplado nos seguintes artigos:

- Art. 52. Para o tratamento de dados pessoais em serviços, aplicações e infraestrutura na denominada computação em nuvem (*cloud computing*), em que o responsável adere aos mesmos mediante condições ou Cláusulas Gerais de Contratação, somente será permitido utilizar estes serviços nos fornecedores que: […] demonstrem transparência nas subcontratações que envolvam informações sobre quem presta o serviço.
- Art. 80. Os esquemas de autorregulação podem ser traduzidos em códigos deontológicos ou de boas práticas profissionais, selos de confiança, políticas de privacidade, regras de privacidade corporativas e outros mecanismos, os quais incluirão regras ou padrões específicos e atenderão aos seguintes objetivos primordiais:
    - VIII. Promover o compromisso dos responsáveis com a entrega de informações e a adoção de políticas internacionais consistentes com critérios externos, bem como auspiciar mecanismos para implementar políticas de privacidade, incluindo ferramentas, transparência, supervisão interna contínua, avaliações de risco, verificações externas e sistemas de remediação...

No que tange à Inteligência Artificial (IA) (México, 2024b), a legislação faz referência ao tema da seguinte forma:

- Em conformidade com a recomendação da UNESCO, busca-se garantir que as transformações digitais promovam os direitos humanos e contribuam para a conquista dos Objetivos de Desenvolvimento Sustentável, abordando questões relacionadas à transparência, à transmissão de informações e à privacidade. Para tal, são incluídos capítulos com políticas orientadas para a ação sobre governança de dados, educação, cultura, trabalho, atenção sanitária e economia.

- Art. 9º Todos os membros do Conselho são obrigados a observar as disposições contidas na Constituição Política dos Estados Unidos Mexicanos em matéria de serviço público, transparência, administração pública, direitos humanos, inclusão e perspectiva de gênero, bem como cumprir as demais leis que dela emanam.

Peru

Em relação à Inteligência Artificial (IA) no Peru (Peru, 2023), destacam-se os seguintes artigos:

- Art. 1º O objeto da presente Lei consiste em promover o uso da inteligência artificial no âmbito do processo nacional de transformação digital, privilegiando a pessoa e o respeito aos direitos humanos. O fim é fomentar o desenvolvimento econômico e social do país, em um ambiente seguro que garanta seu uso ético, sustentável, transparente, replicável e responsável.
- Art. 4º A Autoridade Nacional, no contexto da transformação digital, desenvolve e articula ações para promover e impulsionar:
    - (e) A adoção de diretrizes éticas para um uso sustentável, transparente e reaplicável da inteligência artificial.

Colômbia

No que se refere à proteção de dados na Colômbia (Colômbia, 2023), o princípio da transparência é apresentado no seguinte artigo:

- Art. 4º Princípios para o tratamento de dados pessoais.
    - Princípio da Transparência: O Tratamento deve garantir o direito do Titular de obter, junto ao Responsável pelo Tratamento ou ao Encarregado do Tratamento, a qualquer momento e sem restrições, informações sobre a existência de dados que lhe digam respeito.

No que concerne à Inteligência Artificial (IA) na Colômbia (Colômbia, 2025), a regulamentação estabelece o seguinte:

- Art. 3º Princípios diretores para o desenvolvimento e uso da Inteligência Artificial:
    - 3.4. Ética e sustentabilidade: Promover o desenvolvimento e o uso da IA com base em princípios éticos que respeitem e protejam os direitos humanos, garantindo a transparência, a equidade, a não discriminação e a proteção da privacidade. Além disso, a ética e a solidez almejadas fomentam o equilíbrio por meio do projeto de tecnologias que minimizem seu impacto ambiental, promovam o desenvolvimento social e econômico inclusivo e priorizem o bem-estar das gerações presentes e futuras.
    - 3.5. Transparência e explicação: Os sistemas de IA devem ser projetados e utilizados com princípios que garantam a clareza, a acessibilidade e a rastreabilidade dos processos e decisões inteligentes. Tais sistemas devem garantir que suas decisões sejam compreensíveis para os usuários e autoridades reguladoras, facilitando sua verificação, auditoria e capacidade de questionar seus resultados.
- Art. 4º Definições:
    - Sistemas de alto risco: Sistemas de IA que podem afetar o exercício dos direitos à privacidade, liberdade de expressão, transparência ou acesso à informação pública, a qualquer momento de sua implementação ou posterior.
    - Sistemas de Inteligência Artificial de risco limitado: São sistemas que, sem implicar uma ameaça significativa a direitos ou segurança […] exigem o cumprimento de obrigações de transparência e oferecem informações claras ao usuário sobre a natureza artificial do sistema e permitem sua desativação.
    - Sistemas de Inteligência Artificial de baixo risco: São sistemas com um risco mínimo para a segurança ou os direitos dos indivíduos e, por isso, estão sujeitos a uma regulamentação focada

na gestão do risco, princípios gerais de ética, transparência e boas práticas.

- Princípios e regras para a governança e o uso responsável pela Inteligência Artificial na administração pública:
    - As decisões tomadas com o apoio de sistemas de IA devem ser claras, compreensíveis e transparentes para os cidadãos, permitindo sua revisão e questionamento quando necessário. O uso destes sistemas não exime os servidores públicos nem as autoridades competentes de responderem por suas ações, nos termos estabelecidos pela lei.

EQUADOR

Em relação à Inteligência Artificial (IA) no Equador (Equador, 2025), a regulamentação apresenta as seguintes diretrizes:

- Exposição de Motivos e Princípios
    - Transparência, explicação e transmissão de informações: O funcionamento dos sistemas de IA deve ser compreensível, auditável e explicável para as pessoas afetadas. Serão promovidos mecanismos de publicação dos dados, modelos e especificações utilizados, e será exigida aos operadores de sistemas de alto risco a documentação detalhada de seus processos de projeto e funcionamento.
- Direito dos Titulares
    - Transparência e explicabilidade: Garante o direito a conhecer quando uma decisão é tomada por IA ou com a assistência de IA, a compreender seus fundamentos e a solicitar revisão humana.
- Setor Público e Acesso à Informação
    - Avaliação de impacto (sistemas de alto risco): Os responsáveis pelos sistemas de IA sujeitos a estas avaliações deverão publicar proativamente em seus sites ou canais de comunicação, em formatos abertos e acessíveis, um resumo executivo e outras informações relevantes sobre a avaliação.

- Aplicação Geral
    - Esta Lei será aplicada de forma complementar e sem prejuízo das obrigações estabelecidas na normativa nacional vigente em matéria de proteção de dados, transparência e acesso à informação pública, defesa do consumidor, competência e direito administrativo.

URUGUAI

Em relação à Inteligência Artificial (IA) no Uruguai (Uruguai, 2025), a legislação estabelece o seguinte:

- A estratégia nacional para IA deverá ser fundamentada nos princípios de equidade, não discriminação, responsabilidade, transmissão de informações, transparência, auditoria e inovação segura, com o devido respeito à dignidade humana, ao sistema democrático e à forma republicana de governo. Os princípios de proteção de dados pessoais, contidos na Lei n.º 18.331, de 11 de agosto de 2008, farão parte integrante da estratégia mencionada.
- Como componente dessa estratégia nacional de inteligência artificial, é estabelecido um prazo de cento e oitenta dias para a apresentação ao Poder Legislativo de um relatório e de recomendações para sua regulamentação legal, visando seu desenvolvimento ético, a proteção dos direitos humanos e, simultaneamente, o fomento da inovação tecnológica.

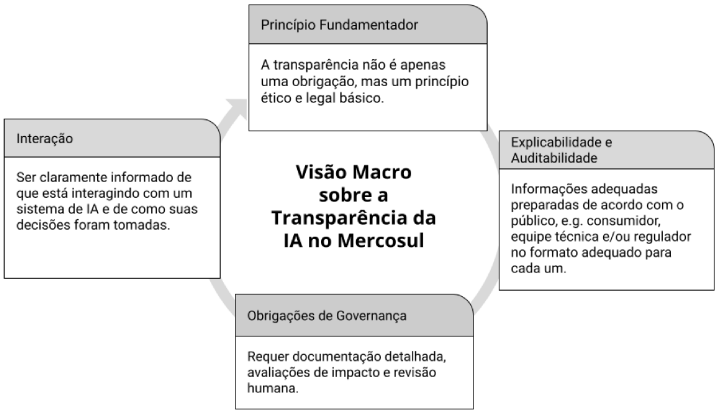
CONSIDERAÇÕES SOBRE O MERCOSUL

O denominador comum mais evidente para a governança da IA no Mercosul reside no foco na Transparência, desdobrado em quatro pilares principais, conforme detalhado a seguir e ilustrado na Figura 1, que apresenta a visão macro sobre a transparência da IA no bloco:

- Princípio Fundamentador: A transparência é estabelecida não apenas como uma obrigação, mas como um princípio ético e legal básico (Brasil, Peru, Colômbia, Equador, Uruguai).

- Explicabilidade e Auditabilidade: A transparência exige que os sistemas de IA sejam compreensíveis, inteligíveis e passíveis de auditoria (Brasil, Chile, Colômbia, Equador).
- Transparência na Interação: O usuário deve ser claramente informado de que está interagindo com um sistema de IA (identificação) e de como as decisões foram tomadas (Chile, Colômbia, Equador).
- Obrigações de Alto Risco/Governança: A transparência é intensificada para sistemas de maior risco, exigindo documentação detalhada, avaliações de impacto públicas e mecanismos de revisão humana (Brasil, Colômbia, Equador, Uruguai).

Figura 1 - Visão Macro sobre a Transparência da IA no Mercosul



MODELO PROPOSTO

Dada a relevância da transparência como um princípio central para os países do Mercosul, e considerando-se que o provimento de transparência da IA varia em função do caso de uso, da organização e do setor, existem estratégias que as empresas podem adotar ao conceber sistemas de IA (Jonker *et al*., 2025). Tais estratégias incluem o estabelecimento de princípios claros de confiança e transparência, a sua implementação prática e a sua incorporação em todo o ciclo de vida da IA (Jonker *et al*., 2025).

No caso de uso, o setor e o público-alvo do modelo deverão orientar as informações a serem divulgadas, sendo que sistemas de alto risco exigirão dados mais abrangentes do que sistemas de baixo risco (Jonker *et al*., 2025). Cada etapa do ciclo de vida de desenvolvimento do sistema de IA pode contribuir com informações, distribuindo a responsabilidade por todo o ecossistema, em vez de atribuí-la a um único indivíduo. Plataformas de *software* e ferramentas estão disponíveis para auxiliar na automação da coleta de informações e em outras atividades de governança da IA (Jonker *et al*., 2025). A seguir estão descritas as principais características para descrever sobre um modelo.

Formato e documentação

O formato da informação depende, adicionalmente, do público-alvo. Por exemplo, se a informação é destinada ao consumidor, ela deve ser facilmente compreensível; se for voltada para equipes técnicas (como cientistas de dados ou reguladores), exigirá um alto nível de detalhamento técnico. Os tipos de documentos a serem abrangidos podem incluir:

- Um documento vivo de conformidade.
- Páginas oficiais de políticas, detalhando como a organização implementa as iniciativas de transparência da IA.
- Recursos educacionais para auxiliar os usuários a compreenderem como a IA é empregada e como pode afetar a experiência do cliente.
- Atividades oficiais focadas na ética da IA na organização.
- Artigos de pesquisa e/ou outras comunicações para oferecer *insights* sobre o uso da IA na organização.

Itens de divulgação do modelo

A divulgação do modelo pode abranger a totalidade ou parte das seguintes informações (Jonker *et al*., 2025):

- Identificação: Nome do modelo, Propósito, Nível de risco, Política de modelo, Geração de modelos, Domínio pretendido, e Informações de contato.

- Desempenho e Qualidade: Dados de treinamento, Precisão de treinamento e teste, Viés, Métricas de robustez adversarial, Métricas de imparcialidade e Métricas de explicabilidade.

Esta lista é considerada abrangente e inclui itens de documentação frequentemente exigidos em regulamentações de Inteligência Artificial, sobretudo para sistemas de Alto Risco. Os itens serão detalhados a seguir, sendo mais explicitamente previstos na legislação, como o Projeto de Lei n.º 2.338/2023 (PL 2338/2023) do Brasil e em projetos similares alinhados ao modelo da União Europeia (como o do Equador), visto que estes exigem o estabelecimento de uma estrutura de governança e gestão de riscos baseada em documentação técnica e não técnica.

Por exemplo, a documentação de sistemas de inteligência artificial (Jonker *et al*., 2025), especialmente daqueles classificados como de alto risco, exige a identificação clara e formal do **nome do modelo**, elemento essencial para fins de rastreabilidade, auditoria e registro oficial, conforme implícito no sistema de registro previsto no Projeto de Lei nº 2338/2023. Associado a isso, o **propósito do modelo** deve ser explicitamente definido, abrangendo sua finalidade, o contexto de uso e as tarefas a que se destina, aspecto central para a determinação do nível de risco e para assegurar que o sistema seja empregado de acordo com o uso pretendido, nos termos do artigo 17 do referido projeto de lei (Brasil, 2023).

Outro elemento fundamental é a definição do **nível de risco** do sistema de IA (Jonker *et al*., 2025), classificado em categorias como risco inaceitável, alto, limitado ou mínimo, com base no potencial impacto sobre direitos fundamentais e a segurança. Essa classificação é determinante para a aplicação das obrigações regulatórias correspondentes, conforme os artigos 17 e 19 do PL 2338/2023 (Brasil, 2023). Complementarmente, a **descrição da geração do modelo** (Jonker *et al*., 2025) deve contemplar o processo de desenvolvimento, incluindo métodos de treinamento, arquitetura empregada e decisões tomadas ao longo do ciclo de vida do sistema, atendendo às exigências implícitas de documentação técnica e qualidade previstas para sistemas de alto risco.

A definição do **domínio pretendido** também é indispensável, especificando os setores, ambientes e contextos geográficos ou demográficos para os quais o modelo foi projetado e validado, de modo a prevenir usos indevidos, conforme a lógica de avaliação de risco baseada na finalidade e no contexto (Jonker *et al*., 2025). No que se refere aos **dados de treinamento** (Jonker *et al*., 2025), é necessária a apresentação de informações detalhadas sobre a origem, curadoria, representatividade, volume e metodologias empregadas no tratamento e mitigação de vieses, em consonância com o artigo 21, §1º, do PL 2338/2023 (Brasil, 2023).

A avaliação do desempenho do sistema deve incluir as **métricas de precisão de treinamento e teste** (Jonker *et al*., 2025), como *accuracy*, *recall* e *F1-score*, a fim de verificar se o modelo atende aos padrões de desempenho estabelecidos, conforme as obrigações de gestão de risco e qualidade previstas nos artigos 19 e 21 (Brasil, 2023). De igual modo, é imprescindível a **análise de vieses** (Jonker *et al*., 2025), com o registro dos potenciais vieses sistêmicos identificados nos dados ou no modelo, bem como das medidas adotadas para mitigá-los, em observância aos princípios de não discriminação, conforme o artigo 19, inciso III (Brasil, 2023).

No âmbito da segurança e confiabilidade, devem ser consideradas as **métricas de robustez adversarial** (Jonker *et al*., 2025), que avaliam a resistência do sistema a ataques destinados a manipular suas saídas, em alinhamento com as exigências de segurança cibernética e gestão de riscos. Adicionalmente, as **métricas de imparcialidade** (Jonker *et al*., 2025) permitem avaliar quantitativamente se o modelo produz resultados discriminatórios com base em atributos protegidos, complementando a análise de vieses e reforçando a proteção contra discriminação prevista no marco legal.

Por fim, a documentação deve contemplar **métricas de explicabilidade** (Jonker *et al*., 2025), indicando o grau de transparência e interpretabilidade do modelo, inclusive por meio do uso de técnicas de inteligência artificial explicável (XAI), de modo a assegurar o cumprimento dos princípios de transparência e explicabilidade, bem como a realização

de avaliações de impacto algorítmico, conforme os artigos 2º, inciso VI, e 24 do PL 2338/2023 (Brasil, 2023). A inclusão de **informações de contato** do responsável pelo sistema de IA, como o controlador ou operador e o Encarregado de Dados, é igualmente necessária para garantir a comunicação com autoridades reguladoras e usuários afetados, em consonância com as exigências de documentação técnica e transparência ao público (Jonker *et al*., 2025).

Conclusão

O presente estudo teve como objetivo central abordar a delegação de decisões a sistemas de Inteligência Artificial (IA) e a consequente discussão sobre a responsabilidade humana. Esta problemática é acentuada pela velocidade, pela complexidade da cadeia de desenvolvimento e uso (*multiple hands/things*), e, sobretudo, pela lacuna de conhecimento e explicabilidade inerente às operações da IA.

O desenvolvimento do trabalho incluiu, ademais, uma abordagem conceitual da transparência nas normativas dos países do Mercosul, identificando-a como um fator mitigador essencial para a redução de vieses. A análise da legislação sobre proteção de dados e IA na região culminou na proposição de uma visão macro dos requisitos de transparência regionais.

O resultado alcançado é um modelo de governança que delineia as principais tarefas requeridas para a IA Responsável no Mercosul. Este modelo considera os aspectos mais robustos das legislações nacionais e a sinergia regulatória entre os países do bloco. Em suma, o estudo oferece uma ferramenta valiosa para que desenvolvedores possam refletir, de forma contínua, sobre a importância da transparência na região e as melhores práticas para que modelos de IA criados em um país do Mercosul sejam empregados em outro, garantindo a mitigação de riscos éticos e a interoperabilidade regulatória.

Referências

ARGENTINA. Poder Executivo Nacional. *Resolução n. 87/2023*. Projeto de Lei de Proteção de Dados Pessoais. [Mensagem]. Buenos Aires, [2023]. Disponível em:

https://www.argentina.gob.ar/sites/default/files/mensajeyproyecto_leypdp2023.pdf. Acesso em: 30 out. 2025.
BLANCHET, Atahualpa. *Inteligência artificial e direitos humanos*. [Aula aberta online]. Disponivel em: https://www.instagram.com/p/DO9AbNmE67s/. [S. l.]: [s. n.]. Acesso em: 29 set. 2025.
BRASIL. *Lei n. 13.709, de 14 de agosto de 2018*. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Senado Federal, [2024]. Disponível em: https://www2.senado.leg.br/bdsf/bitstream/handle/id/658231/Lei_geral_protecao_dados_pessoais_1ed.pdf. Acesso em: 30 out. 2025.
BRASIL. Senado Federal. *Projeto de Lei n. 2.338, de 2023*: dispõe sobre o uso da inteligência artificial no Brasil. Brasília, DF: Senado Federal, [2025]. Disponível em: https://legis.senado.leg.br/sdleg-getter/documento?dm=9347622&ts=1742240889313&disposition=inline. Acesso em: 30 out. 2025.
CHILE. Senado Federal. *Boletim n. 15.869-19 e 16.821-19*: regula os sistemas de inteligência artificial. [Projeto de Lei]. [S. l.], [2025]. Disponível em: https://www.camara.cl/verDoc.aspx?prmID=35352&prmTIPO=OFICIOPLEY. Acesso em: 30 out. 2025.
COECKELBERGH, Mark. *Ética na inteligência artificial*. São Paulo: Ubu Editora, 2024.
COLÔMBIA. Governo Federal. *Lei Estatutária n. 1.581, de 17 de outubro de 2012*. Dispõe sobre a proteção de dados pessoais. Bogotá, [2023]. Disponível em: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981. Acesso em: 31 out. 2025.
COLÔMBIA. Ministério da Ciência, Tecnologia e Inovação. *Projeto de Lei*: por meio do qual se regula a inteligência artificial na Colômbia para garantir seu desenvolvimento ético e responsável e se ditam outras disposições. [2025]. Disponível em: https://minciencias.gov.co/sites/default/files/upload/noticias/pl_ia_finalizado.pdf. Acesso em: 31 out. 2025.
DA SILVA, Daniela A. *et al.* Principais características para o uso responsável da IA. In: CONFERÊNCIA LATINO-AMERICANA DE ÉTICA EM INTELIGÊNCIA ARTIFICIAL, 2024. São Paulo: SBC, 2024. p. 125–128.
DA SILVA, Daniela A.; MARQUES, J. Ethical considerations when using LLMs. In: AMERICAS CONFERENCE IN INFORMATION SYSTEMS, 2025.

DA SILVA, Daniela A.; MARQUES, Johnny C.; DA SILVA, Delmo M. Ética na inteligência artificial: Princípios para a tomada de decisão responsável. In: INTERNATIONAL WORKSHOP ETHICS AND AI — AIRES S2/E1 — RTAIM, 2025, Rio de Janeiro; Porto. Trabalho apresentado. Rio de Janeiro, Brasil: UFRJ; Porto, Portugal: UP, 2025.

DA SILVA, Daniela A.; MARQUES, Johnny C.; TASINAFFO, Paulo M. Mapping the asymmetries of graduate programs in brazil: modelling, visualization and reporting of estimates. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, v. 18, n. 3, 2023.

DA SILVA, Daniela A.; MARQUES, Johnny C.; TASINAFFO, Paulo M. Why machine learning bias is bad? In: WORKSHOP OF DATA SCIENCE– CEDS ITA, 2021. Trabalho apresentado. São José dos Campos, SP: ITA, 2021.

EQUADOR. Assembleia Nacional. Resolução n. [s.n.]. *Lei Orgânica de Regulação e Promoção da Inteligência Artificial no Equador*. [Projeto de Lei]. [2025]. Disponível em: https://www.asambleanacional.gob.ec/es/multimedios-legislativos/97303-proyecto-de-ley-organica-de-regulacion Acesso em: 31 out. 2025.

JONKER A. *et al*. *O que é a transparência de IA*. [S. l.]: IBM, 2025. Disponível em: https://www.ibm.com/br-pt/think/topics/ai-transparency. Acesso em: 29 out. 2025.

MERCOSUL. *Declaração de ministros e altas autoridades sobre direitos humanos dos Estados Partes do Mercosul sobre os princípios dos direitos humanos no âmbito da Inteligência Artificial*. [2023]. Disponível em: https://documentos.mercosur.int/simfiles/declaraciones/98118_ATTXNW0C.docx. Acesso em: 27 out. 2025.

MÉXICO. Senado Federal. *Lei Federal de Proteção de Dados Pessoais em Posse de Particulares* (LFPDPPP). Cidade do México, [2024a]. Disponível em: https://www.gob.mx/indesol/documentos/ley-federal-de-proteccion-de-datos-personales-en-posesion-de-los-particulares. Acesso em: 01 fev. 2026.

MÉXICO. Senado Federal. *Lei para a Regulação Ética da Inteligência Artificial para os Estados Unidos Mexicanos*. [Projeto de Lei]. Cidade do México, [2024b]. Disponível em: http://sil.gobernacion.gob.mx/Archivos/Documentos/2023/04/asun_4551867_20230420_1680209419.pdf. Acesso em: 31 out. 2025.

MILLER, Tim. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, v. 267, p. 1–38, 2019.

O'DONNELL, J. A IA deve nos lisonjear, nos corrigir ou apenas nos informar? *MIT Technology Review Brasil*, [2025]. Disponível em: https://mittechreview.com.br/companion-reinforcing-behaviors-ia/. Acesso em: 27 out. 2025.

PERU. Governo Federal. *Lei n. 31.814, de 24 de julho de 2023*. Lei que promove o uso da inteligência artificial em favor do desenvolvimento econômico e social do país. Lima, [2023]. Disponível em: https://www.gob.pe/institucion/congreso-de-la-republica/normas-legales/4565760-31814 Acesso em: 31 out. 2025.

RICKMAN, S. Evaluating gender bias in large language models in long-term care. *BMC Medical Informatics and Decision Making*, v. 25, n. 1, p. 274, 2025.

SLATTERY, Philip *et al. The AI risk repository*: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. [S. l.]: [s. n.], 2024. [Disponível em: [s. l.]. Acesso em: [s. d.].]

URUGUAI. Lei de Neurotecnologia e Proteção de Informação. *Lei n. 20.212*, art. 74. [Projeto de Lei sobre IA e Neurotecnologias]. [2025]. Disponível em: https://www.impo.com.uy/bases/leyes/20212-2023. Acesso em: 31 out. 2025.

YANG, Kai; QINAMI, Kenzo; FEI-FEI, Li; DENG, Jia; RUSSAKOVSKY, Olga. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2020. p. 547–558.