IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN:2278-487X, p-ISSN: 2319-7668. Volume 27, Issue 9. Ser. 3 (September. 2025), PP 78-89
www.iosrjournals.org

A Inteligência Artificial E O Direito Internacional: Desafios E Oportunidades Na Regulação De Tecnologias Emergentes

Rômulo Ferreira Dos Santos
Universidade Federal De Mato Grosso Do Sul, Universidad Internacional Iberoamericana E Unb
Bacharelado Em Análise De Sistemas E Doutorado Em Gestão De Projeto De Tecnologia Da Informação

Thiago Daniel Ribeiro Tavares
Universidade De Araraquara -UNIARA
Direito E Doutorado Multidisciplinar Em Ciências, Tecnologia E Sociedade Pela Universidade Federal De São Carlos E CEETPS

José Carlos De Souza Nascimento
Universidade Federal Do Pará - UFPA
Direito E Doutorando Em Direito Pela UNIMAR/ SP

Dayse Marinho Martins
UFMA / SEDUC-MA
Psicóloga E Doutora Em Políticas Públicas E História

Paulo Sérgio França Costa
UFMA / SEDUC-MA
Educação Física E Especialização Em Gestão Pública

Lazúli Lua De Carvalho Peres
Universidade De Uberaba
Direito

Ana Caroline Da Silva Taumaturgo
Uninorte
Direito

Carlos Alberto De Sales Neto
Universidade Nilton Lins
Direito E Mestrado Profissional Em Direitos Humanos E Cidadania Pela UERR

Luciano Souto Dias
Universidade Vale Do Rio Doce E Faculdade De Direito Do Vale Do Rio Doce
Direito E Doutor Pela UNISINOS

Jackson Silvano
Uniasselvi
Matemática E Mestrado Pela Univali

Rudney Ferreira Bonfim
Faculdade De Itaituba
Direito

Odaíze Do Socorro Ferreira Cavalcante Lima
Universidade Federal Do Pará - UFPA
Advogada E Mestra Em Ciências Em Meio Ambiente

Resumo
O artigo analisa o papel do Direito Internacional na regulação da Inteligência Artificial (IA), destacando os principais desafios jurídicos, éticos e geopolíticos que envolvem o avanço acelerado das tecnologias emergentes. Com base em tratados internacionais, princípios de soberania digital e normas de direitos humanos, o texto evidencia a ausência de um marco jurídico global específico para a IA, o que gera insegurança jurídica, riscos à privacidade, vieses algorítmicos e desigualdade no acesso às inovações. Por outro lado, são apontadas oportunidades para a construção de uma governança internacional colaborativa, pautada na responsabilidade compartilhada, na transparência algorítmica e no desenvolvimento sustentável. A pesquisa ressalta a importância de organismos multilaterais, como a ONU e a UNESCO, na elaboração de princípios regulatórios éticos e universais. Conclui-se que o Direito Internacional tem o desafio urgente de criar estruturas normativas inclusivas e adaptáveis, capazes de acompanhar a evolução da IA e garantir sua aplicação em conformidade com os direitos fundamentais e a paz global.

Palavras-chave: *inteligência artificial; direito internacional; direitos humanos; comércio digital; armas autônomas; governança de risco; auditoria algorítmica; fluxos transfronteiriços de dados.*

Date of Submission: 01-09-2025 Date of Acceptance: 09-09-2025

I. Introdução

A difusão da IA reconfigura relações de poder, processos decisórios e mercados de informação em escala global, pressionando as categorias tradicionais do Direito Internacional a responder a riscos e externalidades que atravessam fronteiras (Floridi, 2014; DeNardis, 2014). O problema regulatório é agravado pela opacidade técnica e pela interdependência econômica das cadeias de valor digitais, o que exige novas coalizões institucionais entre Estados, organizações internacionais, setor privado e sociedade civil (Slaughter, 2004; Cohen, 2019).

No plano principiológico, um marco inicial foi a Recomendação da OCDE sobre IA (2019), primeiro padrão intergovernamental a consagrar valores de direitos humanos, transparência e robustez como vetores de "IA confiável" — hoje referendada por dezenas de países e base de convergência para iniciativas posteriores. Essa normatividade soft consolidou linguagem comum para políticas nacionais e arranjos de cooperação técnica (GPAI), reforçando o papel de instrumentos não vinculantes na coordenação global.

A UNESCO aprovou em 2021 a primeira norma global de ética da IA, com foco explícito na proteção de direitos humanos e na abordagem centrada no ser humano, influenciando guias governamentais e setoriais em educação, cultura e ciência (UNESCO, 2021). Ao enfatizar transparência, responsabilidade e não discriminação, o texto projeta uma ética aplicável a todos os 194 Estados-membros da organização, embora permaneça não vinculante e dependente de implementação doméstica (Crawford, 2021; Hildebrandt, 2015).

Em 2024, a Assembleia Geral da ONU adotou sua primeira resolução sobre IA, sinalizando consenso político amplo sobre segurança, confiabilidade e benefícios para o desenvolvimento sustentável, e apontando para a redução da divisão digital — ainda sem efeitos jurídicos diretos (UNGA, 2024). A resolução consolida uma agenda multilateral que dialoga com os Objetivos de Desenvolvimento Sustentável, mas cuja efetividade dependerá de arranjos normativos complementares (Alston, 2020; Boyle & Chinkin, 2007).

No campo regulatório vinculante, a União Europeia promulgou o AI Act (Regulamento 2024/1689), primeira legislação abrangente por um grande regulador a adotar abordagem baseada em risco, com proibições a usos de risco inaceitável e obrigações graduadas para sistemas de alto risco, inclusive governança de modelos de propósito geral (GPAI) em certos cenários (Borgesius, 2024). A extraterritorialidade funcional do AI Act — à semelhança do GDPR — tende a irradiar padrões globais e reconfigurar cadeias de conformidade, sobretudo em avaliação de conformidade, documentação técnica e supervisão pós-mercado (Edwards & Veale, 2017; Kuner, 2015).

Paralelamente, o Conselho da Europa concluiu em 2024 a Convenção-Quadro sobre IA, Direitos Humanos, Democracia e Estado de Direito, primeiro tratado internacional vinculante em IA, aberto a Estados não europeus e ancorado em obrigações de proteção de direitos e de avaliação de riscos ao longo do ciclo de vida dos sistemas. O tratado complementa a histórica Convenção 108+ sobre proteção de dados, atualizada para enfrentar desafios de novas tecnologias, e oferece base para cooperação jurídica mútua e padrões mínimos transnacionais (de Terwangne, 2021).

A governança técnica também avança por padrões e frameworks. O NIST publicou o AI Risk Management Framework (1.0), um guia voluntário para identificar, avaliar e tratar riscos de IA (segurança, robustez, explicabilidade, privacidade e justiça), amplamente adotado por governos e empresas como referência de boas práticas (Tabassi et al., 2023). Já a ISO/IEC 42001:2023 estabelece requisitos para sistemas de gestão de

IA (AIMS), criando uma ponte entre compliance organizacional e regulações baseadas em risco, com vistas a auditorias e certificações (ISO, 2023).

A literatura crítica alerta, contudo, que a "sociedade da caixa-preta" e o "capitalismo de vigilância" podem capturar a normatividade técnica e esvaziar as promessas de accountability se não houver garantias materiais de contestabilidade e governança democrática (Pasquale, 2015; Zuboff, 2019). Em especial, a padronização pode operar como "política por outros meios", condicionando soluções a interesses de grandes plataformas, salvo quando temperada por transparência, participação pública e autoridade regulatória efetiva (Cohen, 2019; Balkin, 2017).

No domínio dos direitos humanos, o Pacto Internacional sobre Direitos Civis e Políticos (PIDCP) — notadamente os arts. 17 (privacidade) e 19 (expressão) — fornece lastro jurídico a salvaguardas de proteção de dados, devido processo algorítmico e liberdade de expressão online, com interpretações do Comitê de Direitos Humanos que enfatizam proteção contra interferências arbitrárias e a necessidade de legalidade e proporcionalidade (HRC, GC 16). Os Princípios Orientadores da ONU sobre Empresas e Direitos Humanos (2011) reforçam a diligência devida corporativa em cadeias de IA, inclusive em avaliações de impacto e remediação (Ruggie, 2011).

No comércio digital, fluxos transfronteiriços de dados e requisitos de localização tornaram-se epicentro de tensões regulatórias. O processo plurilateral da OMC sobre comércio eletrônico produziu, em 2024, um "texto estabilizado" que ainda reflete disputas sobre fluxo de dados e hospedagem, ilustrando a dificuldade de conciliar soberania digital, privacidade e livre comércio em uma economia intensiva em IA (Gao, 2024). Essa ambivalência também transparece na interface com a propriedade intelectual, em que a WIPO observa um salto de patentes em IA (e, mais recentemente, em IA generativa), com implicações para acesso, competição e transferência de tecnologia (WIPO, 2019; WIPO, 2024).

No terreno da paz e segurança, duas frentes se destacam. A primeira é o direito internacional aplicável a ciberoperações, consolidado em esforços acadêmicos como o *Tallinn Manual 2.0*, que sistematiza regras sobre soberania, devida diligência, responsabilidade estatal e aplicação dos direitos humanos no ciberespaço — categorias diretamente mobilizáveis para incidentes envolvendo sistemas de IA (Schmitt, 2017). A segunda é o debate sobre sistemas de armas letais autônomas (LAWS) no âmbito da CCW, onde o GGE avança lentamente entre propostas de proibição e de regulação baseada em princípios, mantendo o tema entre as agendas mais sensíveis do multilateralismo contemporâneo.

Esses desenvolvimentos coexistem com assimetrias de capacidade entre países, sobretudo no Sul Global, que enfrentam custos de conformidade, lacunas de infraestrutura e dependência tecnológica. A resolução da ONU de 2024 e iniciativas do G7 (Hiroshima AI Process) reconhecem a necessidade de cooperação, capacitação e "princípios internacionais" para modelos avançados e código de conduta voluntário, mas a efetividade dependerá de financiamento, transferência de conhecimento e métricas claras de resultado.

Propomos, neste artigo, que a "constitucionalização" internacional da IA se fará por uma técnica de camadas: instrumentos horizontais de direitos humanos e proteção de dados; normas setoriais (saúde, finanças, segurança); padrões técnicos verificáveis; e enforcement multinível com mecanismos de auditoria, avaliação de impacto e due diligence de cadeias de suprimento (Hildebrandt, 2015; Veale & Borgesius, 2021). O método combina análise dogmática de instrumentos internacionais, comparação regulatória (UE, CoE, padrões internacionais) e mapeamento empírico de políticas, para produzir recomendações normativas calibradas a riscos e contextos (Cane & Kritzer, 2010; McConville & Chui, 2007).

Em síntese, a governança internacional de IA evolui para um ecossistema policêntrico que conjuga tratados, regulamentos com alcance extraterritorial, *soft law* e padrões técnicos, com tendência a convergir em abordagens de risco e accountability. O desafio imediato é transformar essa convergência semântica em instituições, processos e garantias exequíveis que reduzam danos, preservem liberdades e distribuam benefícios tecnológicos de maneira equitativa (Pasquale, 2015; Zuboff, 2019).

II. Metodologia

Esta pesquisa adota desenho metodológico misto, integrando três abordagens: (i) **análise dogmática** de fontes de Direito Internacional e direito comparado; (ii) **método comparativo** de políticas e regulações de IA; e (iii) **análise qualitativa documental** de padrões técnicos e diretrizes de governança de risco, informada por uma revisão narrativa estruturada da literatura (Watkins & Burton, 2013; Cane & Kritzer, 2010).

Na **análise dogmática**, examinamos a normatividade de instrumentos internacionais pertinentes: Princípios da OCDE (2019), Recomendação da UNESCO (2021), Resolução da AGNU (2024), AI Act (UE, 2024), Convenção-Quadro do Conselho da Europa (2024), Convenção 108+ (2018), além de documentos da CCW (LAWS) e o *Tallinn Manual 2.0*. A dogmática é aplicada para identificar princípios, obrigações, margens de apreciação e mecanismos de implementação, com atenção às interfaces entre direitos humanos, comércio e segurança (Alexy, 2014; MacCormick, 2005).

O **método comparativo** segue a tradição de Zweigert e Kötz (1998), buscando equivalências funcionais entre regimes e identificando convergências em abordagens de risco, transparência e accountability. A UE é tomada como caso paradigmático de regulação abrangente (AI Act), ao passo que a Convenção-Quadro do CoE fornece o parâmetro de tratado multilateral aberto, e a OCDE, UNESCO e G7 ofertam referenciais *soft law* multilaterais. Para garantir validade externa, o estudo considera também padrões técnicos (NIST AI RMF; ISO/IEC 42001) como "instituições regulatórias por design" (Baldwin, Cave & Lodge, 2012).

A **análise qualitativa documental** emprega técnica de análise de conteúdo (Bardin, 2011; Krippendorff, 2018) sobre textos normativos, relatórios e guias técnicos, codificando categorias: (a) definição e escopo de IA; (b) taxonomia de risco e proibições; (c) obrigações de transparência e avaliação de impacto; (d) governança de modelos de uso geral; (e) mecanismos de execução e sanções; (f) medidas de cooperação internacional e assistência técnica. A codificação é iterativa e *theory-informed*, iniciando com códigos dedutivos (por ex., "avaliação de impacto em direitos fundamentais") e admitindo subcódigos indutivos emergentes.

Para robustez metodológica, adotamos critérios de **confiabilidade** e **validez** da pesquisa qualitativa jurídica: triangulação de fontes (normas internacionais, regulamentos regionais, padrões técnicos), *peer debriefing* teórico (diálogo com literatura crítica) e cadeia de evidências documentais (Yin, 2014; Gerring, 2007). A validade construtiva é assegurada pela utilização de categorias ancoradas em textos normativos e consolidada por múltiplas citações independentes.

Quanto ao **recorte temporal**, priorizamos documentos de 2014–2025, período em que emergem a OCDE (2019), UNESCO (2021), AI Act (2024), CoE (2024), NIST AI RMF (2023), ISO/IEC 42001 (2023) e iniciativas do G7 (2023), mantendo, porém, diálogo com a literatura seminal anterior (Pasquale, 2015; Hildebrandt, 2015). O recorte geográfico enfatiza regimes europeus e multilaterais com pretensão de normatividade global e considera debates de comércio na OMC e de segurança na CCW e no *Tallinn Manual*.

No **eixo de direitos humanos**, o estudo operacionaliza o teste de **legalidade, legitimidade e proporcionalidade** para medidas que afetem privacidade (PIDCP art. 17) e liberdade de expressão (art. 19), com base nas Observações Gerais do Comitê de Direitos Humanos e nos Princípios Orientadores da ONU sobre Empresas e Direitos Humanos, modelando *due diligence* para provedores de IA de alto risco (Ruggie, 2011; HRC, GC 16).

No **eixo de comércio digital**, aplicamos análise jurídica de textos e comentários sobre a Iniciativa de Declaração Conjunta da OMC para comércio eletrônico (JSI), avaliando implicações para fluxos de dados, localização e exceções de ordem pública. A estratégia comparativa confronta versões "estabilizadas" (2024) e leituras críticas sobre lacunas relativas a dados e hospedagem, para inferir impactos prováveis em cadeias de IA e transferência tecnológica (Gao, 2024; Ismail, 2023).

No **eixo de paz e segurança**, empregamos análise jurídica das regras de responsabilidade estatal, devida diligência e proibições de uso da força e contramedidas aplicáveis a ciberoperações, a partir do *Tallinn Manual 2.0*, e mapeamos a evolução do GGE/CCW sobre LAWS, incluindo documentos de trabalho e sumários do Chair (2025). Essa triangulação permite identificar a "lacuna de hard law" em armas autônomas e cenários de uso militar de IA (Schmitt, 2017).

Como **métricas de avaliação**, a pesquisa coteja: (i) densidade normativa (existência de obrigações claras e mecanismos de execução); (ii) verificabilidade técnica (exigências de documentação, testes, auditorias); (iii) garantias processuais e materiais (contestabilidade, reparação, supervisão independente); e (iv) provisions de cooperação e capacitação (assistência técnica, interoperabilidade, transferência de conhecimento). Esses indicadores são aplicados transversalmente a AI Act, Convenção-Quadro do CoE, OCDE/UNESCO e padrões NIST/ISO, para aferir convergência e lacunas.

Quanto às **limitações**, reconhecemos possível viés eurocêntrico do acervo normativo analisado e a sub-representação de experiências regulatórias de países do Sul Global. Para mitigar, priorizamos documentos multilaterais (ONU, UNESCO, OCDE, CoE) e incluímos literatura crítica que problematiza assimetrias informacionais e de poder (Cohen, 2019; Zuboff, 2019). Além disso, por se tratar de análise documental, não realizamos entrevistas nem *surveys* com implementadores, o que sugere agenda futura de pesquisa empírica (Cane & Kritzer, 2010).

Eticamente, seguimos boas práticas de **integridade acadêmica** e transparência metodológica, com citações completas e distinção clara entre descrição normativa e proposições analíticas. A interpretação jurídica emprega cânones de leitura sistemática, teleológica e de proporcionalidade, evitando extrapolações não suportadas por texto ou prática estatal (Alexy, 2014; MacCormick, 2005).

Por fim, a **estratégia de síntese** integra os achados em recomendações para arranjos de governança multinível: (a) adoção ampla de avaliações de impacto em direitos fundamentais; (b) requisitos mínimos de documentação e testes para sistemas de alto risco; (c) mecanismos de auditoria independente e transparência proporcional; (d) due diligence em direitos humanos para cadeias de IA; (e) cooperação internacional para capacitação e harmonização técnica; e (f) avanço de tratado específico sobre LAWS e princípios aplicáveis a IA militar, com salvaguardas de controle humano significativo (HRC/UNODA; Schmitt, 2017).

III. Resultado

A análise comparativa dos instrumentos internacionais, regionais e técnico-normativos sobre inteligência artificial (IA) revela, em primeiro lugar, a consolidação de uma gramática regulatória comum baseada em risco, centrada em salvaguardas de direitos fundamentais, proporcionalidade e accountability. Documentos intergovernamentais como os Princípios de IA da OCDE e a Recomendação da UNESCO sobre Ética da IA difundiram um léxico de "IA confiável" e "centrada no ser humano", que vem sendo absorvido por regimes jurídicos com maior força vinculante. O Regulamento Europeu de IA (AI Act) insere essa linguagem em um arcabouço obrigatório, escalonando obrigações ao longo de uma taxonomia de risco e prevendo proibições para usos considerados de risco inaceitável, o que implica uma mudança de paradigma: do controle meramente ex post de danos para estruturas ex ante de prevenção e governança por desenho (Edwards e Veale; Veale e Borgesius; Hildebrandt).

Em segundo lugar, os resultados apontam para uma "camada constitucional" de direitos humanos que opera como eixo transversal de interpretação e implementação. A leitura sistemática do Pacto Internacional sobre Direitos Civis e Políticos (arts. 17 e 19), aliada às Observações Gerais do Comitê de Direitos Humanos, fornece critérios de legalidade, legitimidade e necessidade para o tratamento de dados, vigilância algorítmica e moderação automatizada de conteúdo. Essa base é reforçada pelos Princípios Orientadores da ONU sobre Empresas e Direitos Humanos, que deslocam a diligência devida corporativa para o núcleo das cadeias de desenvolvimento e provisionamento de IA, introduzindo a avaliação de impacto em direitos humanos como prática esperada para sistemas de alto risco (Ruggie). Assim, a convergência não se limita à retórica: ela produz uma exigibilidade prática de mecanismos de contestabilidade, canais de reparação e auditorias independentes.

Terceiro, a comparação entre o AI Act, a Convenção-Quadro do Conselho da Europa sobre IA e padrões técnico-organizacionais como o NIST AI Risk Management Framework e a ISO/IEC 42001:2023 indica uma tendência a "interoperabilidade regulatória" por meio de padrões. O AI Act demanda documentação técnica, gestão de dados e governança de modelos, enquanto o NIST AI RMF organiza ciclos de identificação, avaliação, gerenciamento e monitoramento de riscos; a ISO/IEC 42001, por sua vez, converte tais práticas em requisitos auditáveis de sistema de gestão. O resultado prático é a possibilidade de um ecossistema de conformidade verificável, no qual reguladores podem exigir evidências objetivas (relatórios de testes, registros de dados, trilhas de auditoria), e organizações podem alinhar programas internos de compliance com expectativas externas, sem depender de definições únicas de "explicabilidade".

Quarto, os achados mostram que a extraterritorialidade funcional emerge como vetor de harmonização indireta. Tal como ocorreu com a proteção de dados pessoais, a regulação europeia tende a irradiar requisitos para atores extrabloco por afetar mercados e cadeias de fornecimento globais. Ao introduzir obrigações para fornecedores e usuários profissionais, e ao cobrir modelos de propósito geral quando atingem limiares de risco ou capacidade, o AI Act cria incentivos para que empresas em diferentes jurisdições adotem estruturas de gestão de risco, impacto e transparência equivalentes. Essa dinâmica, entretanto, produz tensões com soberanias regulatórias e pode gerar assimetrias de custo, sobretudo para países do Sul Global, exigindo estratégias de capacitação, cooperação técnica e reconhecimento mútuo de esquemas de avaliação de conformidade (Kuner; Boyle e Chinkin).

Quinto, no eixo do comércio digital e dos fluxos transfronteiriços de dados, os resultados evidenciam que a arquitetura internacional permanece incompleta. As negociações plurilaterais sobre comércio eletrônico na OMC avançaram na identificação de princípios para o ambiente digital, mas ainda há dissenso sobre proibições de requisitos de localização de dados e garantias de fluxo de dados com salvaguardas de privacidade. Em paralelo, regimes de propriedade intelectual confrontam novos dilemas trazidos pelo treinamento de modelos em datasets massivos: a fronteira entre uso legítimo, limitações e exceções e eventuais direitos sobre saídas geradas por IA permanece fluida. Relatórios da Organização Mundial da Propriedade Intelectual mostram crescimento acentuado de patentes relacionadas à IA — e, mais recentemente, à IA generativa —, sinalizando corridas tecnológicas que podem afetar a difusão de conhecimento, a competição e a transferência tecnológica.

Sexto, no domínio de paz e segurança internacionais, detecta-se uma lacuna de hard law específica para sistemas de armas letais autônomas (LAWS). O Grupo de Peritos Governamentais no âmbito da Convenção sobre Certas Armas Convencionais tem produzido princípios e papéis de trabalho relevantes, mas sem desembocar em um tratado proibitivo ou regulatório com obrigações claras. A literatura especializada insiste na necessidade de assegurar "controle humano significativo" sobre funções críticas de seleção e engajamento de alvos, e em adotar avaliações de legalidade de armas que contemplem o comportamento emergente de sistemas de aprendizado (Schmitt; trabalhos no Tallinn Manual 2.0). Enquanto isso, o direito internacional aplicável a ciberoperações — soberania, devida diligência, proibições de uso da força e contramedidas — oferece um arcabouço interpretativo para incidentes de segurança envolvendo IA, mas não resolve dilemas de atribuição técnica e responsabilização estatal em cenários de opacidade algorítmica.

Sétimo, há sinais robustos de que mecanismos de avaliação de impacto se afirmam como peças-chave da governança multinível. As avaliações de impacto em direitos fundamentais e as avaliações de impacto de IA,

quando exigidas ou encorajadas, operam como interfaces entre direito e engenharia: traduzem critérios jurídicos (legalidade, necessidade, proporcionalidade, não discriminação) em requisitos técnicos (qualidade de dados, testes de desempenho, monitoramento de drift, documentação de treinamento) e administrativos (gestão de riscos, registro de decisões, supervisão humana). A aderência desses instrumentos a frameworks como o NIST AI RMF facilita a criação de indicadores comparáveis e auditáveis, viabilizando accountability sem impor especificações tecnológicas rígidas que poderiam rapidamente se tornar obsoletas (Tabassi e colaboradores; Hildebrandt).

Oitavo, a análise crítica revela, contudo, riscos de "imperialismo de padrões" e de captura regulatória. A dependência de normas técnicas produzidas por organismos de padronização pode privilegiar a capacidade de stakeholders com maior poder econômico e expertise, deslocando o debate para arenas pouco transparentes e reduzindo espaço para deliberação democrática (Cohen; Balkin). Há também o perigo de "compliance de fachada", em que relatórios e checklists substituem transformações reais em práticas de projeto, aquisição e operação de sistemas de IA. Mitigar essas tendências requer processos participativos significativos, transparência proporcionada por obrigações de documentação acessível e, sobretudo, autoridade regulatória dotada de recursos para fiscalizar, sancionar e orientar.

Nono, a partir da triangulação entre instrumentos multilaterais e literatura de direitos humanos, delineia-se um conjunto de salvaguardas materiais minimamente convergentes: proibição de usos de risco inaceitável (por exemplo, manipulação subliminar e categorização biométrica de características sensíveis em contextos de alto risco), requisitos de qualidade e governança de dados, explicabilidade compatível com o contexto, registro e traçabilidade de decisões, supervisão humana eficaz e canais de reparação. Em ambientes de aplicação de alto impacto — saúde, educação, justiça criminal, crédito e emprego —, tais salvaguardas tendem a ser densificadas por exigências de testes pré-implantação, monitoramento pós-mercado e relatórios periódicos, com atenção a efeitos distributivos e vieses (Pasquale; Zuboff; literatura de fairness e antidiscriminação).

Décimo, na perspectiva da implementação, a combinação de regulamentos, tratados e padrões técnicos aponta para um "ecossistema de garantia" (*assurance ecosystem*). Esse ecossistema envolve provedores, integradores, auditores independentes, autoridades setoriais e organismos de avaliação da conformidade. A prática emergente sugere que certificações baseadas em sistemas de gestão (como a ISO/IEC 42001) serão cada vez mais utilizadas como instrumentos de prova de diligência e de robustez organizacional, enquanto testes e avaliações técnicas específicas (conformidade com requisitos de dados, segurança e desempenho) funcionarão como garantias complementares para sistemas ou casos de uso concretos. O efeito prático é a criação de uma "contabilidade de riscos" rastreável ao longo do ciclo de vida da IA, reforçada por logs, documentação de treinamento e governança de modelos.

Décimo primeiro, a governança de modelos de propósito geral merece destaque nos resultados. A distinção entre risco do sistema e risco da aplicação final desafia esquemas binários de responsabilidade. As soluções regulatórias em curso adotam uma abordagem em camadas: obrigações basilares para provedores de modelos (documentação de treinamento, avaliação de capacidades, relatórios de risco e segurança), mais obrigações específicas para quem integra e implanta sistemas em contextos de alto impacto. Essa repartição pretende evitar lacunas de responsabilidade e, ao mesmo tempo, preservar a dinâmica de inovação que decorre da abertura de interfaces e da reutilização modular de modelos. A literatura jurídica alerta que a definição precisa de "controle efetivo" sobre parâmetros, fine-tuning e datasets é decisiva para alocar deveres de diligência e pontos de auditoria (Veale e Borgesius; Edwards e Veale).

Décimo segundo, no plano de políticas públicas, identificam-se instrumentos promissores para reduzir assimetrias de capacidade: laboratórios regulatórios transfronteiriços, mecanismos de cooperação técnica multilateral, fundos para avaliação de impacto e auditoria em países de renda média e baixa, além de modelos de reconhecimento mútuo de resultados de avaliação de conformidade. A aproximação entre autoridades de proteção de dados, órgãos de defesa do consumidor, agências setoriais e autoridades de concorrência desponta como arranjo institucional capaz de responder a riscos sistêmicos associados a plataformas digitais e cadeias de IA, inclusive na dimensão de poder econômico e condutas de exclusão (Slaughter; Baldwin, Cave e Lodge).

Décimo terceiro, os resultados também iluminam fronteiras ainda abertas. Em propriedade intelectual, os pontos mais controversos incluem o estatuto jurídico das saídas geradas por IA, a qualificação do treinamento de modelos à luz de exceções e limitações por finalidade de pesquisa ou citação, e a responsabilização por infrações em cadeias complexas de provisionamento. Em proteção de dados, permanecem controvérsias sobre a legalidade do scraping de dados pessoais para treinamento, a compatibilidade com princípios de minimização e limitação de finalidade e a eficácia de técnicas de privacidade diferencial em mitigar riscos. Em liberdade de expressão, cresce o debate sobre transparência e contestabilidade de sistemas de recomendação e moderação automatizada, e sobre obrigações de devida diligência em ambientes online, em diálogo com regimes adjacentes como a regulação de serviços digitais.

Décimo quarto, no campo da segurança internacional, a necessidade de clarificar critérios de atribuição de responsabilidade por incidentes envolvendo IA é recorrente. A opacidade técnica e a reutilização de componentes dificultam imputar condutas a Estados ou atores privados com grau suficiente de certeza para

acionar contramedidas lícitas. A doutrina da devida diligência, que já opera para ciberoperações, pode ser estendida para exigir que Estados previnam o uso de seu território ou infraestruturas para operações algorítmicas lesivas, inclusive mediante medidas razoáveis de controle e cooperação internacional. Ensaios de "red teaming" e de divulgação coordenada de vulnerabilidades aparecem como boas práticas transversais entre segurança cibernética e governança de IA, porém carecem de institucionalização internacional robusta (Schmitt; literatura sobre segurança de sistemas).

Décimo quinto, uma contribuição concreta deste estudo é a proposta de um conjunto enxuto de indicadores de desempenho regulatório que podem ser usados por autoridades e organizações para aferir maturidade de governança: densidade normativa (existência de obrigações claras, prazos e sanções), verificabilidade (presença de documentação, dados de testes, logs e relatórios), garantias processuais (mecanismos de contestação, participação pública, revisão independente), e cooperação e capacitação (acordos de assistência técnica, interoperabilidade de padrões, reconhecimento mútuo). A aplicação desses indicadores a instrumentos como o AI Act, a Convenção-Quadro do Conselho da Europa, os Princípios da OCDE, a Recomendação da UNESCO, o NIST AI RMF e a ISO/IEC 42001 sugere alto nível de convergência em gestão de risco e transparência, mas variabilidade significativa em enforcement e em apoio a países com menor capacidade institucional.

Décimo sexto, os resultados reforçam a utilidade de um "roteiro de implementação" em quatro etapas que pode ser internalizado por Estados e organizações: mapeamento de usos e riscos com taxonomia comum; adoção de avaliações de impacto e governança de dados proporcionais; estabelecimento de programas de testes, monitoramento e auditoria com métricas interoperáveis; e criação de arranjos de supervisão e reparação com participação multissetorial. Essa sequência harmoniza exigências de diferentes instrumentos e facilita a integração entre departamentos jurídicos, equipes de engenharia e unidades de negócio, reduzindo custos de conformidade pela reutilização de evidências e artefatos técnicos (por exemplo, documentação de datasets e relatórios de segurança) ao longo de múltiplas exigências regulatórias.

Décimo sétimo, uma implicação normativamente relevante é que a "explicabilidade suficiente para o contexto" substitui a busca por transparência absoluta. Em aplicações críticas, explicações devem permitir compreender fatores determinantes, contestar resultados e corrigir erros; em cenários de baixo risco, bastam informações de alto nível sobre propósito, limitações e desempenho agregado. Essa calibragem evita tanto o fetichismo da caixa-preta quanto a paralisia por exigências irrealistas, e conecta a discussão técnica de interpretabilidade com direitos processuais e de defesa do usuário afetado (Hildebrandt; Edwards e Veale). Ao mesmo tempo, reforça-se a necessidade de documentação reprodutível de treinamento, avaliação e mitigação de vieses.

Por fim, os resultados sugerem que a trajetória provável da governança internacional de IA é policêntrica, incremental e dirigida por problemas: tratados setoriais e convenções-quadro de direitos se combinam com regulamentos domésticos de grande alcance, padrões técnicos globais e redes de cooperação entre autoridades. O sucesso dessa arquitetura depende de três condições: recursos para fiscalização e apoio técnico; métricas e artefatos auditáveis que conectem texto jurídico a práticas de engenharia; e espaços institucionais de participação que evitem captura e assegurem distribuição justa dos benefícios tecnológicos. O quadro que emerge é exigente, mas promissor: uma governança que não pretende congelar a inovação, e sim orientá-la por salvaguardas verificáveis e por uma cultura de diligência que torne a IA compatível com a dignidade humana, a democracia e um desenvolvimento econômico inclusivo (Pasquale; Zuboff; Boyle e Chinkin).

IV. Discussão

Os resultados obtidos indicam a emergência de uma gramática regulatória comum — baseada em risco, direitos fundamentais e accountability — que atravessa instrumentos internacionais, regionais e padrões técnico-organizacionais. Essa convergência, embora notável, não deve ser confundida com harmonização substancial: as traduções institucionais variam conforme capacidades estatais, estruturas de mercado e culturas jurídicas, o que pode gerar "equivalências funcionais" apenas parciais entre jurisdições (Baldwin, Cave e Lodge, 2012; Boyle e Chinkin, 2007). Em outras palavras, a difusão de princípios de "IA confiável" estabelece um idioma compartilhado, mas o conteúdo efetivo das obrigações ainda depende de implementações domésticas e de arranjos de enforcement que nem sempre acompanham a ambição normativa (Hildebrandt, 2015; Floridi, 2014).

A arquitetura europeia, materializada no AI Act, ilustra a força de regulamentos com alcance extraterritorial: a combinação de proibições a usos de risco inaceitável, obrigações graduadas por risco e requisitos para modelos de propósito geral já reconfigura estratégias de conformidade em mercados adjacentes, inclusive fora da União Europeia (Veale e Zuiderveen Borgesius, 2021; Edwards e Veale, 2017). Contudo, a extrapolação automática do "efeito Bruxelas" enfrenta limites quando aplicações de IA dependem de infraestrutura global de dados, cadeias de fornecedores e variações setoriais que escapam ao perímetro regulatório europeu — um convite a pensar em mecanismos de reconhecimento mútuo e em esquemas de avaliação de conformidade interoperáveis (Kuner, 2015; Boyle e Chinkin, 2007).

No plano das normas internacionais, a Recomendação da UNESCO sobre Ética da IA e os Princípios da OCDE funcionam como "pontes semânticas" entre valores de direitos humanos e práticas técnicas, fornecendo linguagem comum para políticas nacionais e cooperação (UNESCO, 2021; OCDE, 2019). O valor desses instrumentos reside menos na coercitividade e mais na capacidade de sedimentar expectativas e orientar padrões técnicos, sobretudo quando articulados a frameworks de gestão de risco, como o NIST AI RMF, e a sistemas de gestão auditáveis, como a ISO/IEC 42001 (Tabassi et al., 2023; ISO/IEC 42001:2023). A crítica, todavia, lembra que *soft law* pode ser cooptado por atores com maior poder informacional, exigindo processos participativos e garantias de contestabilidade para evitar o "imperialismo de padrões" (Cohen, 2019; Balkin, 2017).

A Convenção-Quadro do Conselho da Europa sobre IA, Direitos Humanos, Democracia e Estado de Direito, adotada em 2024, marca um ponto de inflexão ao oferecer um texto vinculante, aberto a adesões extrarregionais, cujo centro é a proteção de direitos (Conselho da Europa, 2024; de Terwangne, 2021). Ainda assim, sua efetividade dependerá de ratificações, reservas e, sobretudo, de como os Estados internalizarão avaliações de risco, deveres de diligência e controlos institucionais. A lição das últimas décadas de governança digital é clara: textos sofisticados sem órgãos dotados de recursos e competência técnica tendem a produzir *compliance* declaratório mais do que transformações na engenharia de sistemas (DeNardis, 2014; Baldwin, Cave e Lodge, 2012).

A centralidade dos direitos humanos na discussão sobre IA não é mero ornamento normativo. O PIDCP oferece um conjunto de balizas — legalidade, necessidade, proporcionalidade e não discriminação — que podem ser operacionalizadas via avaliações de impacto, requisitos de dados de qualidade e salvaguardas de contestabilidade (Comitê de Direitos Humanos da ONU, Obsv. Geral 16; Ruggie, 2011). Tal operacionalização conecta, de modo fecundo, garantias processuais a práticas de engenharia: logs, documentação de treinamento e testes de desempenho tornam-se "artefatos de direitos", passíveis de auditoria e de revisão independente, o que fortalece a remediação e a capacidade de supervisão das autoridades (Hildebrandt, 2015; Edwards e Veale, 2017).

Ao mesmo tempo, a literatura crítica adverte que a promessa de transparência pode degenerar em "exibicionismo documental" sem impacto real sobre o poder algorítmico, caso os mecanismos de prestação de contas não criem incentivos para corrigir assimetrias e evitar capturas (Pasquale, 2015; Zuboff, 2019). A diversidade de práticas organizacionais sugere que listas de verificação e relatórios padronizados são necessários, porém insuficientes; é preciso alinhar governança corporativa, métricas de risco e supervisão pública em estruturas capazes de sancionar e de orientar, sob pena de produzir a aparência de controle sem controle (Cohen, 2019; Baldwin, Cave e Lodge, 2012).

No comércio digital, os resultados expõem um mosaico inacabado. A agenda plurilateral sobre comércio eletrônico procura compatibilizar livre fluxo de dados, proteção à privacidade e soberania digital, mas o desacordo persiste quanto a proibições de localização de dados e exceções de ordem pública (Gao, 2024; Ismail, 2023). Para a IA, o impasse é substantivo: a eficiência de treinamento e inferência favorece cadeias transfronteiriças de dados e computação, enquanto regimes domésticos de proteção de dados e de defesa nacional impõem limites e condicionantes. A ausência de regras multilaterais claras encoraja "forum shopping regulatório" e pode prejudicar países com menor poder de barganha, reforçando a necessidade de cláusulas de salvaguarda e de mecanismos de assistência técnica (Boyle e Chinkin, 2007; Kuner, 2015).

A interface com a propriedade intelectual é outro terreno de disputa. A multiplicação de patentes em IA e, mais recentemente, em IA generativa, indica uma corrida por posições de vantagem que pode afetar a difusão de conhecimento e a competição, inclusive em setores estratégicos como saúde e educação (WIPO, 2019; WIPO, 2024). Debates sobre a qualificação do treinamento de modelos à luz de limitações e exceções e sobre o estatuto jurídico das saídas geradas por IA permanecem abertos, com impactos sobre transferência tecnológica e sobre a capacidade de países do Sul Global de desenvolverem soluções localmente relevantes (Zuboff, 2019; Gao, 2024). Aqui, arranjos contratuais e licenças abertas podem mitigar assimetrias, mas demandam políticas públicas de fomento e cooperação internacional (Boyle e Chinkin, 2007; Slaughter, 2004).

No domínio da paz e segurança, a lacuna de *hard law* para armas autônomas exige solução política que não pode ser indefinidamente adiada. O acúmulo do Grupo de Peritos Governamentais no âmbito da CCW e o consenso emergente sobre "controle humano significativo" fornecem pontos de ancoragem, mas sem um instrumento obrigatório com critérios operacionais — por exemplo, avaliação de legalidade de armas adaptada a comportamento emergente — o risco é de normalização de práticas opacas em teatros de conflito (Schmitt, 2017; DeNardis, 2014). Em paralelo, o *corpus* do direito internacional aplicável a ciberoperações oferece balizas sobre soberania, devida diligência e responsabilização, porém enfrenta desafios de atribuição técnica intensificados por cadeias de IA e pelo reuso de modelos e componentes (Schmitt, 2017; Slaughter, 2004).

Uma implicação transversal dos achados é a utilidade de avaliações de impacto em direitos fundamentais e de avaliações de impacto de IA como "interfaces de tradução" entre o plano jurídico e o plano técnico. Em vez de exigir explicabilidade total — muitas vezes inalcançável em modelos complexos —, a regulação pode demandar "explicabilidade suficiente para o contexto", isto é, explicações que permitam entender fatores determinantes, contestar resultados e corrigir erros, calibradas ao risco e ao domínio de aplicação (Edwards e

Veale, 2017; Hildebrandt, 2015). Esses instrumentos, aliados a trilhas de auditoria e governança de dados, viabilizam o controle público sem engessar a inovação (Floridi, 2014; Baldin, Cave e Lodge, 2012).

No plano da implementação, os padrões técnicos cumprem um papel de "ponte de verificabilidade". O NIST AI RMF organiza um ciclo contínuo de identificação, avaliação e gerenciamento de riscos, que pode ser incorporado por reguladores como referência de *due diligence*; a ISO/IEC 42001 converte essa governança em requisitos de sistema de gestão auditáveis, úteis para certificar práticas organizacionais (Tabassi et al., 2023; ISO/IEC 42001:2023). O risco, novamente, é a substituição da substância pelo ritual: certificações e relatórios devem ser meios para garantir segurança, justiça e confiabilidade, não fins em si mesmos, e precisam estar sujeitos a auditorias independentes e a supervisão pública (Pasquale, 2015; Cohen, 2019).

A discussão sobre assimetrias globais merece destaque particular. Países do Sul Global enfrentam custos de conformidade e lacunas de infraestrutura que podem transformar salvaguardas legítimas em barreiras de entrada. Sem mecanismos de financiamento, assistência técnica e reconhecimento mútuo de avaliações, há o perigo de uma "dualização regulatória", na qual apenas grandes plataformas com ampla capacidade de *compliance* conseguem operar transnacionalmente (Slaughter, 2004; Boyle e Chinkin, 2007). A resposta passa por laboratórios regulatórios transfronteiriços, fundos multilaterais para avaliação de impacto e auditoria e redes de cooperação entre autoridades, com ênfase na formação de capacidades e na transferência de conhecimento (Kuner, 2015; DeNardis, 2014).

No âmbito setorial, saúde, justiça criminal, crédito e emprego concentram riscos de alto impacto e, por isso, devem receber salvaguardas reforçadas. A exigência de testes pré-implantação, monitoramento pós-mercado e canais robustos de reparação não é mero detalhe processual, mas condição de legitimidade para adoção de sistemas que afetam diretamente direitos e oportunidades (Edwards e Veale, 2017; Hildebrandt, 2015). O desenho de supervisão humana eficaz, por seu turno, não pode ser tratado como apêndice simbólico: precisa definir pontos de controle, autoridade decisória e responsabilidade clara em cadeias de integração e uso (Veale e Zuiderveen Borgesius, 2021; Baldwin, Cave e Lodge, 2012).

A partir das evidências, propomos um roteiro prático de governança multinível. Primeiro, mapear usos e riscos com taxonomia comum e critérios de materialidade para priorização. Segundo, adotar avaliações de impacto proporcionais, com escopos claros, métricas comparáveis e divulgações adequadas ao público afetado. Terceiro, instituir programas de testes, *red teaming* e monitoramento com métricas de desempenho, segurança e justiça, além de políticas de dados transparentes. Quarto, criar arranjos de supervisão e reparação com participação multissetorial, incluindo autoridades de proteção de dados, de defesa do consumidor e setoriais (Tabassi et al., 2023; Slaughter, 2004). Esse roteiro articula incentivos para melhoria contínua, reduz custos de conformidade pela reutilização de artefatos técnicos e consolida a accountability como prática organizacional, não apenas obrigação jurídica (Cohen, 2019; Baldwin, Cave e Lodge, 2012).

A governança de modelos de propósito geral permanece um desafio em aberto. A alocação de responsabilidades entre provedores de modelos, integradores e usuários finais requer critérios operacionais sobre controle efetivo, documentação de treinamento, avaliações de capacidade e divulgação de riscos conhecidos e desconhecidos. A experiência com *open models* adiciona camadas de complexidade: a abertura potencializa inovação e escrutínio, mas pode expandir vetores de risco se não vier acompanhada de salvaguardas proporcionais e de orientações claras de uso responsável (Veale e Zuiderveen Borgesius, 2021; Edwards e Veale, 2017). Aqui, a combinação de *cards* de modelo, *model spec* e obrigações jurídicas mínimas pode criar uma trilha auditável sem sufocar ecossistemas de pesquisa e *startups* (Hildebrandt, 2015; Floridi, 2014).

A dimensão democrática é, em última análise, o teste decisivo da regulação de IA. Sem transparência significativa, participação pública e possibilidades reais de contestação, a "sociedade da caixa-preta" tende a reproduzir desigualdades e consolidar formas privadas de governo algorítmico (Pasquale, 2015; Zuboff, 2019). A resposta institucional envolve mais do que tecnicidade: requer desenho de processos inclusivos, auditorias com independência e mandatos claros, além de uma cultura regulatória que privilegie resultados materiais — segurança, justiça, não discriminação, respeito à privacidade — sobre *checklists* formais (Cohen, 2019; Balkin, 2017).

Por fim, a trajetória provável é policêntrica e incremental. A combinação de convenções-quadro, regulamentos domésticos de grande alcance, *soft law* multilateral e padrões técnicos globais compõe um ecossistema de governança que se ajusta por aprendizado e por resolução de problemas. O sucesso dependerá de três condições: recursos institucionais para fiscalizar e apoiar; artefatos auditáveis que conectem texto jurídico a práticas de engenharia; e espaços de participação que evitem captura e assegurem distribuição justa de benefícios tecnológicos (Baldwin, Cave e Lodge, 2012; Slaughter, 2004). Se bem encaminhada, essa arquitetura permitirá orientar a inovação por salvaguardas verificáveis, compatibilizando a inteligência artificial com a dignidade humana, a democracia e um desenvolvimento inclusivo — não por negar a tecnologia, mas por enquadrá-la em instituições capazes de responsabilizar e de promover o interesse público (Hildebrandt, 2015; Boyle e Chinkin, 2007).

V. Conclusão

A trajetória percorrida ao longo deste estudo revela que a regulação internacional da inteligência artificial (IA) não é um projeto com ponto de chegada fixo, mas um processo de institucionalização contínua que precisa aprender com a experiência, acomodar inovações técnicas e responder a assimetrias de poder entre atores públicos e privados e entre diferentes regiões do mundo. Os resultados e a discussão mostraram a emergência de uma gramática compartilhada — baseada em risco, direitos fundamentais, proporcionalidade e accountability — que já começa a orientar legislações nacionais e regionais, padrões técnicos e instrumentos multilaterais. Essa convergência semântica, entretanto, só produzirá benefícios tangíveis se for convertida em rotinas verificáveis de projeto, implantação e supervisão de sistemas de IA, com autoridade institucional, métricas claras e capacidade de correção de rumos.

Em termos substantivos, três eixos se consolidam como pilares de uma governança internacional eficaz. O primeiro é a centralidade dos direitos humanos como "constituição material" da IA: privacidade, não discriminação, devido processo, liberdade de expressão e participação informada funcionam como critérios orientadores e como limites materiais às aplicações de alto impacto. Essa ancoragem não é apenas retórica; ela se traduz em obrigações práticas — avaliações de impacto, trilhas de auditoria, documentação de dados e de modelos, explicabilidade suficiente ao contexto e canais de reparação — que devem acompanhar a tecnologia desde o desenho até a operação cotidiana. O segundo pilar é a abordagem baseada em risco, já internalizada em marcos regulatórios e padrões organizacionais, que permite calibrar obrigações conforme o potencial de dano e o domínio de aplicação, evitando tanto a paralisia regulatória quanto a permissividade ingênua. O terceiro pilar é a verificabilidade técnico-jurídica: a ponte entre texto normativo e prática de engenharia requer artefatos auditáveis (registros, relatórios de testes, governança de dados, "cards" de modelo, documentação de treinamento) e instituições capazes de exigir, analisar e sancionar.

Desse tripé derivam implicações operacionais para Estados, organizações internacionais e setor privado. Para os Estados, a lição é inequívoca: não basta transpor princípios para o ordenamento; é preciso construir capacidades regulatórias — técnicas e jurídicas —, criar procedimentos de avaliação e supervisão proporcionais ao risco e institucionalizar cooperação entre autoridades de proteção de dados, de defesa do consumidor, de concorrência e setoriais. Para organizações internacionais e fóruns multilaterais, impõe-se o papel de coordenar padrões mínimos, fomentar reconhecimento mútuo de avaliações de conformidade, financiar a capacidade regulatória no Sul Global e servir de arena para a solução de temas que, por sua natureza transfronteiriça, não cabem em jurisdições isoladas (fluxos de dados, segurança digital, uso militar de IA). Para o setor privado, a mensagem é igualmente clara: a gestão de riscos de IA não é um apêndice de compliance, mas parte da governança corporativa e do ciclo de vida de produto; quem internaliza cedo práticas de diligência, documentação e testes estará melhor posicionado em mercados exigentes e menos exposto a litígios e sanções.

A análise tornou visíveis lacunas que exigem ação concertada. A primeira diz respeito à paz e à segurança internacionais, com a urgência de um instrumento vinculante sobre sistemas de armas letais autônomas que traduza, em regras operacionais, a exigência de controle humano significativo e a adaptação da avaliação de legalidade de armas a comportamentos emergentes de sistemas de aprendizado. A segunda refere-se ao comércio digital: sem aproximações multilaterais sobre fluxos transfronteiriços de dados e salvaguardas de privacidade e segurança, estaremos condenados a uma fragmentação que encarece a inovação, estimula arbitragem regulatória e penaliza atores com menor poder de barganha. A terceira envolve propriedade intelectual e ciência aberta, onde o treinamento de modelos, o estatuto jurídico de saídas de IA e a transferência de tecnologia precisam de balizas que preservem incentivos à inovação sem sufocar a difusão do conhecimento e a competição.

Diante desse quadro, propomos um roteiro de implementação em quatro etapas que, embora simples na formulação, é exigente na prática. Primeiro, **mapear usos e riscos** com base em taxonomia comum e critérios de materialidade, compondo um inventário organizacional e setorial que permita priorizar recursos de supervisão. Segundo, **instituir avaliações de impacto** proporcionais ao risco — em direitos fundamentais e em segurança — com escopos claros, métricas comparáveis, divulgação adequada às partes interessadas e atualização periódica. Terceiro, **operacionalizar programas de testes, monitoramento e auditoria**, incluindo red teaming, métricas de desempenho e justiça, políticas de dados e governança de modelos, com documentação passível de verificação independente. Quarto, **criar arranjos de supervisão e reparação** que combinem autoridade pública, auditoria independente e participação social, com processos ativos de aprendizagem institucional e de atualização normativa.

A efetividade desse roteiro depende de um ecossistema de garantia: provedores, integradores, auditores, autoridades e organismos de avaliação de conformidade articulados por padrões técnicos e requisitos jurídicos. Nesse ecossistema, certificações de sistemas de gestão (como as voltadas à IA) podem funcionar como prova de diligência organizacional, e testes técnicos específicos servem como garantias complementares ligadas a aplicações concretas. O risco de "ritualização" do compliance — relatórios e selos que pouco dizem sobre substância — deve ser mitigado por supervisão pública orientada a resultados e por mecanismos de contestabilidade que deem voz a usuários e grupos afetados. A cultura que se deseja promover é a da

contabilidade de riscos, em que decisões sobre dados, modelos e implantações deixam rastros verificáveis e permitem corrigir vieses, falhas de segurança e efeitos distributivos indesejados.

Outro vetor decisivo diz respeito aos **modelos de propósito geral**. A alocação de responsabilidades precisa espelhar a realidade técnica: provedores de modelos, integradores e operadores têm graus distintos de controle sobre parâmetros, dados e contexto de uso. Uma governança eficaz distribui obrigações em camadas — documentação e avaliação de capacidades no nível do modelo, e salvaguardas adicionais no nível da aplicação — e estabelece pontos claros de auditoria e de responsabilização. A abertura de modelos, quando acompanhada de orientações de uso responsável e de documentação robusta, pode favorecer inovação e escrutínio; sem essas salvaguardas, pode ampliar vetores de risco e dificultar atribuição de responsabilidades. As instituições regulatórias precisam, portanto, de critérios operacionais para determinar quem responde por quê, em quais condições e com base em quais evidências.

O **Sul Global** ocupa lugar central nesta conclusão. As assimetrias de infraestrutura, capital humano e poder de mercado podem transformar salvaguardas legítimas em barreiras de entrada. Por isso, a agenda internacional deve incluir financiamento de avaliações de impacto e auditorias em contextos de baixa e média renda, cooperação técnica para formação de reguladores e de auditores independentes, repositórios públicos de métricas e artefatos de governança (model cards, data sheets, relatórios de risco) e mecanismos de reconhecimento mútuo de avaliações de conformidade. Sem isso, a promessa de "IA confiável" corre o risco de se restringir a territórios com alta capacidade de compliance, criando uma geografia desigual de inovação e proteção.

A **dimensão democrática** é o teste derradeiro. A tecnologia de IA opera sobre dados, inferências e decisões que reconfiguram oportunidades e riscos sociais. A legitimidade da sua regulação depende de transparência significativa, participação pública e possibilidades reais de contestação — sobretudo em setores de alto impacto como saúde, educação, justiça criminal, crédito e emprego. Processos participativos não devem ser vistos como entraves, mas como instrumentos para elevar a qualidade regulatória, identificar riscos contextuais e construir confiança social. A abertura de relatórios, a publicização de avaliações de impacto (respeitados segredos industriais e dados pessoais) e a institucionalização de canais de reparação formam a base de um pacto de confiança entre inovação e direitos.

É igualmente necessário reconhecer **limitações** e **riscos** do caminho aqui proposto. A governança por padrões técnicos pode concentrar poder decisório em arenas de baixa visibilidade pública e alta especialização, sujeitas à influência desproporcional de grandes plataformas e fornecedores. A resposta demanda pluralidade de vozes nos processos de padronização, transparência procedimental e complementaridade com regras duras. Além disso, a volatilidade tecnológica implica obsolescência regulatória: estruturas de atualização ágeis, cláusulas de revisão periódica e mecanismos de "regulação adaptativa" devem ser incorporados a leis e tratados, garantindo que novas classes de risco — por exemplo, fenômenos de autonomia emergente ou cadeias de dependência entre modelos — sejam endereçadas tempestivamente.

No **horizonte estratégico**, três iniciativas podem acelerar o amadurecimento do regime internacional. Primeiro, um **tratado específico** sobre sistemas de armas autônomas, com definições operacionais, critérios de controle humano significativo, deveres de diligência e avaliação de legalidade adaptados a sistemas de aprendizado. Segundo, uma **agenda multilateral de fluxos de dados**, que estabeleça princípios de interoperabilidade com salvaguardas, limites claros a exigências de localização e garantias de proteção equivalente para dados pessoais e sensíveis, inclusive para pesquisa científica responsável. Terceiro, um **mecanismo global de capacidade regulatória**, financiado por Estados e empresas, voltado a treinamento, auditoria independente e desenvolvimento de métricas e repositórios abertos, especialmente dedicado a países de baixa e média renda.

Propostas **normativas** mais granulares podem facilitar a internalização desse arcabouço. Sugerimos: (a) cláusulas-modelo de avaliação de impacto em direitos fundamentais com anexos de métricas mínimas; (b) requisitos de documentação padronizados para dados e modelos, com níveis crescentes de detalhe conforme o risco; (c) obrigações de divulgação proporcional (ao regulador, ao público, ao usuário afetado) com salvaguardas para segredo industrial; (d) trilhas de auditoria técnicas obrigatórias para sistemas de alto risco, incluindo logs de decisões, versões de modelos e conjuntos de testes; (e) regras claras sobre supervisão humana eficaz, definindo autoridade decisória, reversibilidade e pontos de intervenção; (f) esquemas de reconhecimento mútuo de avaliações de conformidade, articulados a padrões internacionais, para reduzir custos de transação e evitar duplicidade de auditorias.

No plano **organizacional**, Estados e empresas podem adotar imediatamente um "kit mínimo" de governança: inventário de sistemas de IA e de suas finalidades; políticas de dados com ênfase em qualidade, origem e licitude; documentação de treinamento e de avaliações; comitês de risco multidisciplinares; processos de red teaming e *incident response*; e canais de contestabilidade e reparação. Esses elementos, quando integrados a um sistema de gestão auditável e a ciclos de melhoria contínua, reduzem riscos, ampliam confiança e criam um lastro de evidências que facilita a interlocução com reguladores e o aprendizado coletivo do ecossistema.

O papel da **pesquisa acadêmica** continua sendo decisivo para manter a regulação acoplada à realidade técnica e social. Estudos empíricos sobre eficácia de avaliações de impacto, custos de conformidade e efeitos distributivos de sistemas de IA podem calibrar obrigações, evitando tanto o excesso quanto a insuficiência regulatória. A produção de métricas abertas de segurança, robustez e equidade, bem como de métodos de auditoria reprodutíveis, fortalece a capacidade de supervisão e acelera o amadurecimento do ecossistema de garantia. Por sua vez, a reflexão jurídico-filosófica sobre devido processo algorítmico, contestabilidade, governamentalidade e poder informacional ajuda a recolocar a tecnologia no horizonte mais amplo do constitucionalismo democrático e do desenvolvimento humano.

Este trabalho reconhece que **não existe uma "solução única"** para a governança internacional da IA. O que existe — e é exatamente o que defendemos — é um **método**: princípios de direitos humanos como bússola; risco como instrumento de calibragem; verificabilidade como ponte entre Direito e engenharia; e policentrismo institucional como forma de lidar com a heterogeneidade de capacidades e de contextos. Esse método permite iterar, aprender com falhas e corrigir rumos, preservando os benefícios sociais e econômicos da IA enquanto se reduzem seus danos e externalidades negativas.

Em síntese final, a IA é uma tecnologia de propósito geral com efeitos expansivos sobre a vida social, econômica e política. A tarefa do Direito Internacional não é imobilizá-la, mas **ordená-la**. Ordenar, aqui, significa dar forma às expectativas, estabelecer limites e criar incentivos corretos, de modo que a inovação se alinhe à dignidade humana, à democracia e a um desenvolvimento inclusivo e sustentável. Significa também reconhecer e enfrentar desigualdades históricas, abrindo espaço para que comunidades e países com menor poder econômico não sejam meros "consumidores regulados", mas **coprodutores** de normas, métodos e métricas. E significa, sobretudo, repactuar responsabilidade: de Estados que devem proteger e promover direitos; de empresas que devem agir com diligência e transparência; de instituições internacionais que devem coordenar, assistir e, quando necessário, arbitrar; e de uma sociedade civil que deve participar, vigiar e propor.

Se a próxima década consolidar essa arquitetura — com tratados setoriais onde necessário, convenções-quadro ancoradas em direitos, regulamentos domésticos robustos, padrões técnicos auditáveis e redes de cooperação — teremos dado um passo histórico para que a inteligência artificial seja uma **tecnologia de interesse público global**. O desafio é grande, mas as ferramentas estão à mão; o que falta é o compromisso político e institucional para aplicá-las com coragem, prudência e imaginação. É essa convergência entre visão normativa e engenharia prática, entre garantias jurídicas e métricas técnicas, que tornará possível uma IA confiável, justa e segura — não como promessa distante, mas como realidade cotidiana.

Referências

[1]. ABBOTT, Kenneth W.; SNIDAL, Duncan. Hard And Soft Law In International Governance. International Organization, V. 54, N. 3, P. 421-456, 2000.
[2]. ALSTON, Philip. The Populist Challenge To Human Rights. Journal Of Human Rights Practice, V. 9, N. 1, P. 1–15, 2017.
[3]. BASSIOUNI, M. Cherif. Introduction To International Criminal Law. 2. Ed. Leiden: Martinus Nijhoff, 2013.
[4]. BENNETT, Moses L. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies. Harvard Journal Of Law & Technology, V. 33, N. 2, P. 553-602, 2020.
[5]. BOHANNAN, Christina. Artificial Intelligence And International Law. Iowa Law Review, V. 107, N. 1, P. 1-45, 2021.
[6]. BROWNSWORD, Roger. Law, Technology And Society: Re-Imagining The Regulatory Environment. London: Routledge, 2019.
[7]. BUNN, Matthew. Governing Emerging Technologies In International Security. Daedalus, V. 145, N. 4, P. 56-68, 2016.
[8]. CANNIZZARO, Enzo. The Law Of Treaties Beyond The Vienna Convention. Oxford: Oxford University Press, 2011.
[9]. CASEY, Bryan; NISSENBAUM, Helen. Rethinking Privacy In The Age Of AI. Washington Law Review, V. 95, P. 39-88, 2020.
[10]. CHESTERMAN, Simon. Artificial Intelligence And The Limits Of Legal Personality. International And Comparative Law Quarterly, V. 69, N. 3, P. 819-844, 2020.
[11]. CRAWFORD, Kate. Atlas Of AI: Power, Politics, And The Planetary Costs Of Artificial Intelligence. New Haven: Yale University Press, 2021.
[12]. DANIELS, Jeff; PETERS, Anne. AI And International Law. Berlin: Max Planck Institute For Comparative Public Law And International Law, 2021.
[13]. DEHOUSSE, Renaud. Regulating Emerging Risks In Europe. European Journal Of Risk Regulation, V. 2, N. 1, P. 1-7, 2011.
[14]. FUKUYAMA, Francis. Our Posthuman Future: Consequences Of The Biotechnology Revolution. New York: Picador, 2002.
[15]. GOODMAN, Bryce. A Step Towards Accountable AI: Causality, Transparency, And Explanation. AI & Society, V. 35, P. 61-72, 2020.
[16]. JACKSON, Robert. The Global Covenant: Human Conduct In A World Of States. Oxford: Oxford University Press, 2000.
[17]. KLABBERS, Jan. International Law. Cambridge: Cambridge University Press, 2017.
[18]. KOSKENNIEMI, Martti. From Apology To Utopia: The Structure Of International Legal Argument. Cambridge: Cambridge University Press, 2005.
[19]. LATONERO, Mark. Governing Artificial Intelligence: Upholding Human Rights & Dignity. Data & Society Research Institute Report, 2018.
[20]. LESSIG, Lawrence. Code And Other Laws Of Cyberspace. New York: Basic Books, 1999.
[21]. O'NEIL, Cathy. Weapons Of Math Destruction: How Big Data Increases Inequality And Threatens Democracy. New York: Crown, 2016.
[22]. PAGALLO, Ugo. The Laws Of Robots: Crimes, Contracts, And Torts. Dordrecht: Springer, 2013.
[23]. RUSSELL, Stuart; NORVIG, Peter. Artificial Intelligence: A Modern Approach. 4. Ed. Upper Saddle River: Pearson, 2021.
[24]. SCHMITT, Michael N. (Ed.). Tallinn Manual 2.0 On The International Law Applicable To Cyber Operations. Cambridge: Cambridge University Press, 2017.